Title
Data-driven Approaches to Inventory Management

Author
Cao, Ying

Publication Date
2019

Peer reviewed|Thesis/dissertation

Data-driven Approaches to Inventory Management

by

Ying Cao

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 Philip M. Kaminsky
Assistant Professor Anil Aswani
Assistant Professor Scott Moura

Spring 2019

Data-driven Approaches to Inventory Management

Abstract

Data-driven Approaches to Inventory Management

by

Ying Cao

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun Max Shen, Chair

With the advances in technologies and the growing popularity of e-commerce, huge datasets and massive computational power have never been more accessible. Moreover, the rising machine learning and distributionally robust optimization techniques bring new opportunity for more effective inventory decision making in the data-rich environments. The goal of this dissertation is, thus, to explore data-driven approaches to inventory management problems that are efficient and practical.

We address the challenges in this field from three different aspects: firstly, we aim at proposing a flexible model for capturing real-world demand process accurately with as few assumptions as possible; then, we take into account additional features in the demand model and derive robust inventory policies with out-of-sample performance guarantees under milder assumption than current literature; and finally, we explore the usage of a group of decomposition algorithms to tackle the increasing computational difficulty as the data size grows. Chapter 2, Chapter 3 and Chapter 4 each delves into one of these three directions respectively.

In Chapter 2, we leverage the universal approximating capability of neural network structures to approximate an arbitrarily complex autoregressive demand process without any parametric assumptions. By adopting a quantile loss in training, we allow our neural network to output directly an estimation of the critical quantile, which is indeed the inventory policy for classical newsvendor problem. In addition, in contrast to the prevalent feedforward neural networks which are asymptotically stationary, the special structure we choose is capable of handling nonstationary time series. To the best of our knowledge, this is the first approach which deals with nonstationary time series without any parametric assumption or preprocessing to capture the components like trend or seasonality. Though theoretical guarantees are sacrificed due to a lack of assumption on the underlying real process, empirical studies validate the performance of our approach on real-world nonstationary demand process. Moreover, we establish the optimality of the myopic policy to the multi-period newsvendor problem where unmet demand and excess inventory can be carried over to next period, and argue that our approach is also a data-driven solution.

The second project in Chapter 3 addresses the data-driven newsvendor problem from a different angle with the goal to achieve robust policies as well as theoretical support. We start with a simple linear demand model to incorporate information from other features related to demand, such as price, materials and etc. And to hedge against uncertainty of the demand distribution, the idea of distributionally robust optimization (DRO) is applied. We contribute to the current literature of DRO applications in supervised learning by adopting a fixed design interpretation of the features. Thus, similar to the neural network approach, we are also able to relax the assumption of identical and independent sample points, which is more applicable in real-world scenarios. Then, by leveraging results from fixed design linear regression, we propose a two-step framework to obtain a newsvendor solution. Moreover, Wasserstein metric is chosen for constructing the ambiguity set of all candidate distributions, and based on which our data-driven policy can be obtained efficiently in polynomial time and attains both finite-sample and asymptotic performance guarantees.

Finally, in Chapter 4, we put some effort in dealing with practical issues of implementing such data-driven approaches when massive datasets are available. Specifically, we consider a group of decomposition algorithms which are suitable for large-scale multi-block convex optimization problems with linear constraints. This problem setting covers a lot of applications in machine learning and data-driven problems. We focus on a special case of such algorithms which can be guaranteed to converge under mild conditions with linear rate, and also enjoys the convenience of parallel implementable subproblems. We modify an adaptive parameter tuning scheme to achieve faster convergence in practice. And at the end, we further show that when parameters are chosen appropriately, global convergence can be established even if the primal subproblems are only solved approximately.

We reckon that our results are just some primary attempts at achieving the goal of efficient decision making in data-driven environment, and hope that this dissertation can serve as a catalyst for other research in this field. Thus, we list a number of directions for future research in the last chapter after the concluding remarks.

To my parents and grandmother

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest appreciation to my advisor, Professor Zuo-Jun Max Shen, for his persistent help and encouragement through the many different phases of my PhD program, internship and job search. Throughout this journey, he provided tremendous insightful discussions into the research field and always made himself available whenever I needed guidance. Most importantly, he has taught me to be open-minded and to embrace new ideas and techniques. Without his advice and support this dissertation would not have been possible.

I would like to thank my dissertation committee members: Professor Phil Kaminsky, Professor Anil Aswani, and Professor Terry Taylor for their great support and invaluable suggestions that has helped to shape the work detailed in this dissertation. It is their enthusiasm and valuable insights that made my qualifying exam and dissertation workshop thoughtful and rewarding.

I would also like to extend my deepest gratitude to my collaborators. I am very grateful to Professor Gemma Berenguer and Professor Borja Peleato from Purdue University for financially supporting me during my summer research there and introducing the idea of Alternating Direction Method of Multipliers. I learned many good research habits from them. I must also thank my coauthor and good friend, Meng Qi, who brought me the idea of distributionally robust optimization using Wasserstein metric. Together, we came up with the idea and results in Chapter 3 of this dissertation.

In addition, many thanks to my friends in Berkeley who have made it an enjoyable journey. These include but are not limited to Sheng Liu, Junyu Cao, Mengxin Wang, Chao Mao, Cris Pais Martínez, Renyuan Xu, Haoyang Cao, Xu Rao, Nan Yang, Quico Spaen, Shiman Ding, Dan Bu, Yanqiao Wang, just to name a few.

Finally, my deepest gratitude to my family - my parents, who have always been unconditionally supporting and encouraging me to pursuit my dream since I was little; my grandmother, a wise old lady who never had a chance to go to school when she was young due to the war and old traditions, yet has cultivated a desire for knowledge in my mind. This dissertation would not have been possible without their warm love, continued patience, and endless support.

Chapter 1

Introduction and Background

In this competitive market, especially with the success of e-commerce and online retail, the importance of inventory management cannot be stressed enough. Classical inventory management models use stochastic programming technique to determine optimal policies or decisions, thus require the full knowledge of random phenomena such as distribution of demand. Indeed, with these models, researchers have gained a great deal of insight into the inventory control problems and provided elegant solutions such as (r, Q) policies, base-stock policies [101] and etc. In reality, however, managers need to make decisions without knowing the demand distribution. Often, specific probabilistic assumptions are made and parameters are then estimated via observed data points. Consequently, the resulting policies are sensitive to the parametric assumption and cannot deal with the case when independently and identically distributed (i.i.d.) samples are not available.

The ever-evolving technologies have not only made us faster at producing the things we need, but also helped us bring down error and uncertainty in ways that make us better equipped than ever. Especially with the advances in technologies which result in increasing speed of data generation, processing and analysis, we are able to make more accurate inventory control decisions with less assumptions. The goal of this dissertation is to explore inventory problems under the data-driven environment, and develop more practical yet robust solutions for real-world applications.

The following three chapters of this dissertation line up to work towards this goal using three different strategies. Firstly, we aim at modeling complicated nonstationary demand process without parametric assumptions; then, we introduce the technique of distributionally robust optimization and derive robust newsvendor solutions under mild conditions; finally, we investigate a group of decomposition algorithms which can be helpful in practice when the problem size is large. The outline of this thesis is as follows:

In Chapter 2, we study the newsvendor problem, one of the most fundamental inventory control models, under a framework of time series. Rather than using the prevalent two-step approach which first estimates some parameters from data and then solves an optimization problem with the estimation plugged in, we propose an integrated approach that incorporates the inventory-optimization into the machine learning training step. With the

universal approximating theorem, our neural-network-based approach is nonparametric in spirit. More importantly, by adding shortcuts in classical feedforward neural networks, the special structure we consider is capable of capturing nonstationary components such as trend and seasonality in the time series. With numerical studies on real-world time series, our results show that this approach gives better newsvendor decisions than the popular two-stage approach or other parametric methods in the literature. To the best of our knowledge, this is first data-driven approach to newsvendor that can deal with nonstationary demand process without signal decomposition. Additionally, we show the generalizability of our method to multi-period newsvendor problem.

Then, in Chapter 3, we look at the same problem from another perspective and take into consideration information from other features related to demand in addition to the time series. Instead of trying to give a nonparametric flexible model for the demand process, we start with a very simple and restrictive linear model, but put our focus on developing a robust solution with theoretical performance guarantees. In contrast to current literature where the feature vector is interpreted as random with i.i.d. observations available, we proceed with the fixed design interpretation where the features are regarded as deterministic (such as from controlled experiments). As a result, we also assume all features are observable before the decision making step, so that we do not have to account for the randomness of covariates in the objective function. By leveraging the properties of ordinary least squares estimators and the technique of distributionally robust optimization with Wasserstein metric, we propose a two-step approach for generating a robust newsvendor solution. We show that not only this solution is asymptotic optimal, but its out-of-sample performance is also bounded with high probability in a finite-sample scenario. Ultimately, we demonstrate that our distributionally robust optimization can be reformulated to a simple tractable equivalent problem. And the distributionally robust solution can be obtained in polynomial time with a single iteration of linear regression and then sorting.

The goal of Chapter 4 focuses on tackling data-driven problems when the data size is huge so that special care should be taken to improve the practicality of such approaches. Specifically, we consider a decomposition algorithm based on the application of proximal point method. Compared to classical alternating direction method of multipliers which updates each block of variables sequentially, the subproblems of our algorithm can be implemented in parallel. Thus, it can better take advantage of the arising popularity of distributed computing infrastructures. Moreover, this algorithm is guaranteed to be globally convergent under mild assumptions with linear rate. We modify an existing adaptive parameter tuning scheme to achieve faster convergence in practice. In addition, for a special case of such algorithm, we prove that global convergence still holds even if all subproblems are only solved approximately.

Finally, in Chapter 5, we conclude the dissertation and suggests some thoughts on future directions for research.

Chapter 2

Data-driven Approach for Newsvendor under Nonstationary Demand

2.1 Introduction

In varies fields of production/ inventory management, economic, engineering etc., predicting quantiles of a random process provides essential information for decision making which is ignored by traditional point estimation of the conditional expectation. Moreover, many of these applications emphasize short-term forecasting where time series-based models, which take into account the internal structure of a process involving over time, are often preferable to explanatory approaches [114, 105]. Therefore, a time series model is considered in this chapter. And we start with its application in the newsvendor problem, one of the most fundamental stochastic inventory models, as it is well-known that the optimal stock level is the so-called critical quantile. Moreover, we consider the extension to multiperiod inventory control problem, where excess inventory and backlogged demand can be carried over to the next period, and prove the optimality of the myopic policy. Meanwhile, our results can be easily extended to predicting any quantile of a random process in other fields.

In practice, when the distribution of demand is unknown, managers need to decide the inventory level based on historical demand observations. In the current analysis, we assume that the demand observations are uncensored, that is, the sales reveal real demand. This problem essentially boils down to predicting the critical quantile of the future demand.

To tackle the unknown demand distribution, a standard treatment is to first estimate the distribution, and then use the estimation in the decision making step (see [41, 67, 122] for a review). Often, the parametric form of the distribution is specified in advance, and the parameters are then estimated from a random sample. Consequently, policies derived in such case are very sensitive to the parametric assumption of the demand distribution. Moreover, many a time, the available information is not sufficient to postulate an accurate

model, or the real distribution is too complicated to be represented by any commonly used models like Gaussian, exponential and etc.

To handle this problem, nonparametric and data-driven approaches have been developed, which generally combine historical data and optimization techniques [11, 6, 100]. However, most of the above mentioned methods, either parametric or nonparametric, assume that demand process is i.i.d. in consecutive sales periods. Such assumptions. This assumption though facilitates the establishment of asymptotic optimality of the resulting strategies, suffers from a major practical limitation that demand in real life changes over time and is in general time-correlated. For instance, the consumption of ice cream is obviously much higher in summer than the rest of a year, thus its demand exhibits yearly seasonality; demand for fast-fashion products often shows a product life cycle and low demand in past periods is an indicator for customers' lack of interest in one product in the future. Thus, intuitively, managers should take into account these correlation and patterns, and always adjust their inventory decision once new related information is available.

Indeed, for the reasons mentioned above, some authors have studied inventory models with time-correlated demand (see [20] for a detailed review). Most of these papers either assume perfect knowledge of the demand evolution or focus only on bounds of the objective function via robust optimization. When the real demand evolution is unknown, mean square error (MSE) criteria is used to estimate the parameters of a predefined demand model. For example, linear demand models, especially the simplest $AR(1)$ model is very popular ([45, 31]). However, simple models such as $AR(1)$, which considers linear relationship only, is in practice unrealistic. Moreover, as in the parametric distribution-fitting case, the choice of evolution model may generate drastic errors in the inventory policy. Thus, data-driven approaches under the time-series framework is in need.

Despite the existence of some nonlinear parametric autoregressive (NLAR) models (see [47] for a short review), one of the goals of this chapter is therefore to provide a single-step nonparametric solution for quantile forecasting of potentially time-correlated or even nonstationary [1] time series. Then, we argue that it serves as a data-driven approach for making inventory decisions in the newsvendor problem setting and its multiperiod extension. To the best of our knowledge, this is the first work in the data-driven inventory management field that deals with a general autoregressive demand process of unknown form, which also works with nonstationary demand. In addition, we show that the myopic policy is still optimal in the multiperiod newsvendor problem setting under autoregressive demand, without requiring the demand process to be statistically increasing as did in the previous literature. Moreover, comparing with the existing neural network based methods for quantile prediction by [103, 18] and [117], our method enjoys the advantage of being able to handle nonstationary time series; and not requiring previous quantile values as input, which are not observable in practice. In fact, this is also the first time that a nonstationary time series can be dealt with

[1] When talking about ***stationarity***, there are two interpretations. Besides the formal definition of stationary process in mathematics and statistics, we often also indicate the case where the demand distributions do not change over time (i.i.d.) in inventory literature. In this study, we use its mathematical definition.

without prespecified components of seasonality or trend.

The remainder of this chapter is organized as follows. In Section 2.2 we give a comprehensive literature review of related topics and techniques which inspired our research. In Section 2.3, we introduce a special neural network structure known as the double parallel feedforward neural networks (DPFNN). Then, we describe a quantile autoregression technique based on this structure and denote it by DPFNN-QAR. In Section 2.4, we discuss the application of quantile autoregression for solving data-driven newsvendor problem and illustrate its efficiency using numerical studies; then, in Section 2.5, we consider the extension to multiperiod inventory control problems, where excess inventory and backlogged demand can be carried over to the next period. Finally, Section 2.6 contains the concluding remarks.

2.2 Literature Review

As discussed above, in this chapter, we aim at proposing a quantile forecasting method to deal with nonstationary time series in a single step and argue that it's a data-driven approach for newsvendor problem using only historical sales data. Thus, We review three streams of literature related to our initial work. We first go through some popular existing data-driven approaches for inventory control with time series data, highlight their assumptions and drawbacks, and then identify the improvement that we aim to achieve. Next, we introduce the widely used econometric models for representing a time series, and the theory of quantile regression for estimating any quantile of a random process. In next section, we propose a model consolidating these ideas to achieve our goal.

2.2.1 Data-driven Inventory Models

Though parametric methods provide mathematical convenience and lead to some useful theoretical insights into the problem, different choices of the distribution family can yield to different solutions. To address this limitation, methods based directly on available demand observations are developed. Liyanage et al. proposed to use the empirical distribution instead of assuming a prior distribution family [83]. Levi et al. further analyzed this sample average approximation (SAA) approach, and obtained an analytical bound on the probability that the relative regret of the SAA solution exceeds a threshold [75]. Huh et al. presented how to use Kaplan-Meier Estimator to deal with censored demand [65]. And Bisi, using an online convex optimization procedure, proposed a nonparametric adaptive algorithm for both perishable and non-perishable inventory system [13]. In addition, other data-driven policies such as the Concave, Adaptive Value Estimation (CAVE) [43], which estimate the value function instead of demand distribution using sample gradient is proposed.

The research detailed above is restricted to solving single-period newsvendor problem with exclusive usage of uncensored i.i.d. historical demand data, while another few papers have explored more complicated scenarios. Burnetas et al. developed a framework for jointly determining price and ordering quantity [15], which is still restricted to the i.i.d. demand

setting. [64] considered inventory decision for multiple products with a warehouse-capacity constraint and relax the identical assumption. Likewise, Levi et al. allowed independent but non-identical demand and proposed a near-optimal sampling-based policy [76]. Both methods, however, assumed that multiple independent sample paths can be generated, such information, however, is not available in real life.

Moreover, demand can be correlated, or exhibit some trend or seasonality as time evolves, and learning these patterns will be the key to making informative decisions with only past demand data. [9] assumed that demand is Markovian with given transition matrix, which is still a strong assumption. Levina et al. applied weak aggregating algorithm (WAA) as an online approach, to adaptively select ordering quantity from a pool of fixed expert advice, when the demands in subsequent periods are i.i.d. [77]. Zhang et al. extended this algorithm, by allowing each expert randomly switching his advice, to cope with a slightly non-stationary environment [122]. However, the last two methods require predetermined finite countable decisions and possible cost at each period and suffer from curse of dimensionanity. Beutel and Minner introduced exogenous variables such as price and temperature, and estimated the demand as linear combination of them [12]. Later, they take into consideration the case of censored demand [96] by estimating the uncensored demand using a heuristic as a first step. Similar to the aforementioned literature, we focus on solving inventory problems with the demand data exclusively in this chapter, but our method can be easily extended to include external signals.

2.2.2 Time Series Analysis

When a sequence of historical demand realizations is taken at successive equally spaced points in time, our goal is to learn the behaviour of this process so as to predict the future. This falls in the field of time series analysis:

> Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.[29]

Let $\{Y_t\}$ be the process of interest and $\{y_t\}$ be the observed realizations. Among the rich and rapidly growing techniques for analyzing time series, we review two groups of models that are most closely related to our research, autoregressive models and the Holt-Winters methods:

Autoregressive model is widely used to describe time-varying processes in nature, economics, etc. It specifies that the output variable depends linearly on its own previous (lagged) values and on a i.i.d. stochastic term. For example, an autoregressive model of order p, denoted by $AR(p)$ is defined as

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t, \tag{2.1}$$

where ε_t is i.i.d. or at least white noise. To avoid this restrictive linearity assumption, a natural extension is the nonlinear autoregressive (NLAR) model (see [47] for a short review). Again, the choice of a prespecified parametric form is crucial. More general autoregressive models based on neural network have been proposed, most of them use a multilayer perceptron with a single layer of hidden neurons (i.e. a three-layer feedforward neural network). The capability of such models to forecast nonstationary time series has been a topic of dispute, limited empirical studies yield mixed results. While selected research [121, 91] find that prior detrend and deseasonality are essential for artificial neural networks (ANN) models to work, others [69] conclude the opposite. We will have a more detailed discussion on this issue in Section 2.3.

Different extensions of AR models have also been developed to handle different nonstationary series. For example, if there is a trend with stochastic mean in data, integrated AR model can be used; and if the data shows seasonality, seasonal AR models are useful. However, the order of differencing and frequency also need to be determined in advance, and the resulting seasonal AR models generally perform poorly in long-term prediction. Another group of techniques, **Holt-Winters'** methods, work better in modeling data with trend or seasonality by decomposing the data into level, trend and seasonality components. Depending on the form of seasonality assumed, there are two versions of Holt-Winters' formulation, i.e. Additive Holt-Winters' (HWA) and Multiplicative Holt-Winters' (HWM). The addictive version of Holt-Winters (HWA) assumes a linear trend and additive seasonal components:

$$\textbf{\textit{Level :}} \quad L_t = \alpha(y_t - S_{t-k}) + (1-\alpha)(L_{t-1} + T_{t-1}), \tag{2.2}$$

$$\textbf{\textit{Trend :}} \quad T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}, \tag{2.3}$$

$$\textbf{\textit{Seasonality :}} \quad S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-k}, \tag{2.4}$$

where k is the length of cyclical period, and should be specified as input. Then the one-step-ahead forecast of the time series is

$$\hat{Y}_t = L_{t-1} + T_{t-1} + S_{t-k}. \tag{2.5}$$

For the multiplicative seasonal version (HWM), while Equation (2.3) remains unchanged, the other equations are modified as

$$\textbf{\textit{Level :}} \quad L_t = \alpha(y_t/S_{t-k}) + (1-\alpha)(L_{t-1} + T_{t-1}), \tag{2.6}$$

$$\textbf{\textit{Seasonality :}} \quad S_t = \gamma(y_t/L_t) + (1-\gamma)S_{t-k} \tag{2.7}$$

and

$$\hat{Y}_t = (L_{t-1} + T_{t-1})S_{t-k}. \tag{2.8}$$

One of the common practices in time series analysis is to find the parameters α, β, γ such that the sum of the squares of the forecast errors (SSE) is minimized, that is,

$$\begin{aligned} &\min_{\alpha,\beta,\gamma} \sum_{t=2}^{N} (y_t - \hat{Y}_t)^2 \\ &\text{s.t.} \quad 0 \leq \alpha, \beta, \gamma < 1. \end{aligned} \tag{2.9}$$

Then, a quantile is obtained by assuming the time series has i.i.d. normal innovations with the unknown variance estimated via observed residuals. Both versions are used as benchmark algorithms in the numerical studies in this chapter. We refer the readers to [88] for more implementation details of these two algorithms.

2.2.3 Quantile Regression

To define a time series by any parametric models including AR and Holt-Winters' methods reviewed above, the major challenge is to decide the parameter values. Given observations from the process $\{y_1, y_2, ..., y_N\}$, one common practice is to estimate the parameters by ordinary least squares (OLS) method with the goal of minimizing the sum of the squares of forecast errors [90]. Take the $AR(p)$ model as an example, the forecasting error is defined as

$$e_t = y_t - (\beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i}). \tag{2.10}$$

The OLS estimator, equivalent to the maximum likelihood estimator (MLE), is found by minimizing

$$\tilde{\beta} = \min_{\beta} \sum_{t=p+1}^{N} e_t^2. \tag{2.11}$$

Under certain regularity conditions, the obtained one-step-ahead forecast,

$$\tilde{Y}_t = \tilde{\beta}_0 + \sum_{i=1}^{p} \tilde{\beta}_i y_{t-i}, \tag{2.12}$$

converges asymptotically to the conditional mean of y_t given its lagged values. Nevertheless, in many practical situations, such a point estimator is not informative enough. Especially when the cost of overestimation and underestimation is asymmetrical, it is more desirable to estimate some quantile of the distribution instead of its mean. In 1978, Koenker proposed the method to estimate the quantiles in a linear regression model [72]. Later, Koenker and Xiao extended it to the autoregressive case [73]. Sample quantile loss instead of the squared loss in equation (2.11) is minimized in order to determine the parameter values. For any $\tau \in (0, 1)$, parameters are estimated via

$$\hat{\beta} = \min_{\beta} \sum_{t=p+1}^{N} \rho_\tau (y_t - (\beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i})), \tag{2.13}$$

where $\rho_\tau(\cdot)$ is the usual check function, given as

$$\rho_\tau(u) = \begin{cases} \tau u & \text{if } u > 0 \\ (\tau - 1)u & \text{if } u \leq 0 \end{cases}. \tag{2.14}$$

Then, an estimate of the τth quantile of y_t conditional on its previous values is given by

$$\hat{Y}_t = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i y_{t-i}. \tag{2.15}$$

Similar to the NLAR models, researchers also came up with different nonlinear quantile autoregressive models [28, 82, 3], most of which work for stationary time series. Likewise, quantile versions of the Holt-Winters' methods are also developed in [2]. Denoted by QHWA and QHWM respectively, they are also used as benchmark algorithms to compare with our own method. We reckon that quantile regression is closely related to inventory control since we are also considering asymmetric costs of understocking and overstocking, and thus can be used as a data-driven approach for newsvendor-like problems.

2.2.4 Motivation

Considering the limitation of parametric time series models and the lack of data-driven approaches under time-correlated demand, we aim to extend the linear AR models to a more general form without having to specify its parametric form in advance. Meanwhile, inspired by quantile regression and the fact that the optimal solution to newsvendor problem is the famous critical quantile, we propose a data-driven algorithm for making inventory decisions under such circumstances.

2.3 Quantile Forecasting with Neural Networks

The popular $AR(p)$ models allow only linear terms, which oversimplifies most real-world processes. Thus, we consider a more general autoregressive model to define a process with potentially complicated nonlinear structure:

$$Y_t = g(Y_{t-1}, ..., Y_{t-p}) + \varepsilon_t, \tag{2.16}$$

where $g(\cdot)$ can be any continuous function of unknown form, random innovations $\{\varepsilon_t\}$ are i.i.d. with mean 0 and unknown variance σ^2, but are not necessarily normally distributed. Note that this is essentially a stronger assumption compared with the white noise innovation assumed in traditional linear regression and autoregression models. The identical ε_t is crucial in the quantile case, which allows us to consider the entire conditional distribution of Y_t, not merely its conditional mean.

Let Z_τ denote the τth quantile of the common cumulative density function of $\{\varepsilon_t\}$, that is $Pr(\varepsilon_t \leq Z_\tau) = \tau$, it follows that the τth conditional quantile of Y_t can be written as

$$Q_{Y_t}(\tau|y_{t-1}, ..., y_{t-p}) = g(y_{t-1}, ..., y_{t-p}) + Z_\tau. \tag{2.17}$$

However, neither $g(\cdot)$ nor Z_τ is known in practice, and are traditionally estimated separately in two consecutive steps. Given a predefined parametric form of $g(\cdot)$, OLS can be used to set its parameters. Next, Z_τ can be calculated using the forecast residuals by assuming normality of ε_t. With quantile regression technique, however, we can avoid making the normality assumption of ε_t. In addition, by taking advantage of the universal approximation capability of neural networks ([63] and [62]), the following approach we propose is nonparametric in spirit and can be used to deal with any continuous function of $g(\cdot)$.

2.3.1 Structure of DPFNN

In this project, we propose to use neural networks as an nonlinear extension of linear quantile autoregression, which is an universal approximator. In particular, we use a standard three layer feedforward network (FNN) with shortcuts directly from the input nodes to the output node, a.k.a. a double parallel feedforward network (DPFNN). This structure is better at capturing linear mapping compared with normal FNN configuration, while remains sensitive in nonlinear relationships ([109]). Moreover, the analysis of [74] and [104] suggested that while FNN based autoregressive models are asymptotically stationary, adding shortcuts between inputs and output allows them to model integrated time series. While the current neural network based quantile estimation literature deal with nonstationary time series either by preprocessing (e.g. [121]) or combining other models (e.g. [120]), we aim at capturing the nonstationarity directly within the network structure. Thus, we reckon that DPFNN is suitable for modeling the potentially nonstationary demand process. As far as we know, we are the first to use this structure in quantile autoregression studies and treat stationary and nonstationary time series using exactly the same procedure.

The utilization of neural network as the baseline model arises from the famous Universal Approximation Theorem stated as follows:

Theorem 2.1 (Universal Approximation Theorem by [62])**.** *For any continuous function $g(\cdot)$ on a compact set K, there exists a feedforward neural network (FNN), having only a single hidden layer, which uniformly approximates $g(\cdot)$ to within an arbitrary $\epsilon > 0$ on K.*

$$\begin{gathered}\textit{i.e. given any } \epsilon > 0, \textit{ there exits } m, \textit{ and parameters } v_i, b_i, w_i \textit{ and} \\ H(x) = \textstyle\sum_{i=1}^{m} v_i f(w_t^T x + b_i) \textit{ such that} \\ |H(x) - g(x)| < \epsilon \quad \forall x \in K\end{gathered}$$

where $f(\cdot)$ is a nonconstant, bounded, and monotonically-increasing continuous function (e.g. sigmoid function $f(x) = \frac{1}{1+e^{-x}}$).

This theorem for standard FNN by [63] and [62] can be easily extended to the case of DPFNN, since every FNN is a special case of a corresponding DPFNN when all weights on shortcuts from inputs to output are set to zero. Then for any continuous $g(\cdot)$, the right-hand-side of equation (2.17) can be arbitrarily closely approximated by some DPFNN structure with sufficiently large number of hidden neurons and appropriately chosen parameters. We denote such a DPFNN configuration by $H(\cdot;\theta_0)$ with θ_0 containing all parameters of the network. That is, the following equation holds with arbitrarily small gap (for simplicity, we assume it holds with equality).

$$\begin{aligned} Q_{Y_t}(\tau|y_{t-1},...,y_{t-p}) &= H(y_{t-1},...,y_{t-p};\theta_0) \\ &= H(x_t;\theta_0), \end{aligned} \tag{2.18}$$

where $X_t = (Y_{t-1},...,Y_{t-p})$ and x_t being the observed vector. Figure 2.1 shows a general DPFNN structure considered in this project. It is a three-layer network with p input nodes, m hidden nodes and a single output. As in most neural network training, the constant term will be captured by assigning biases, and thus no input node is needed for it.

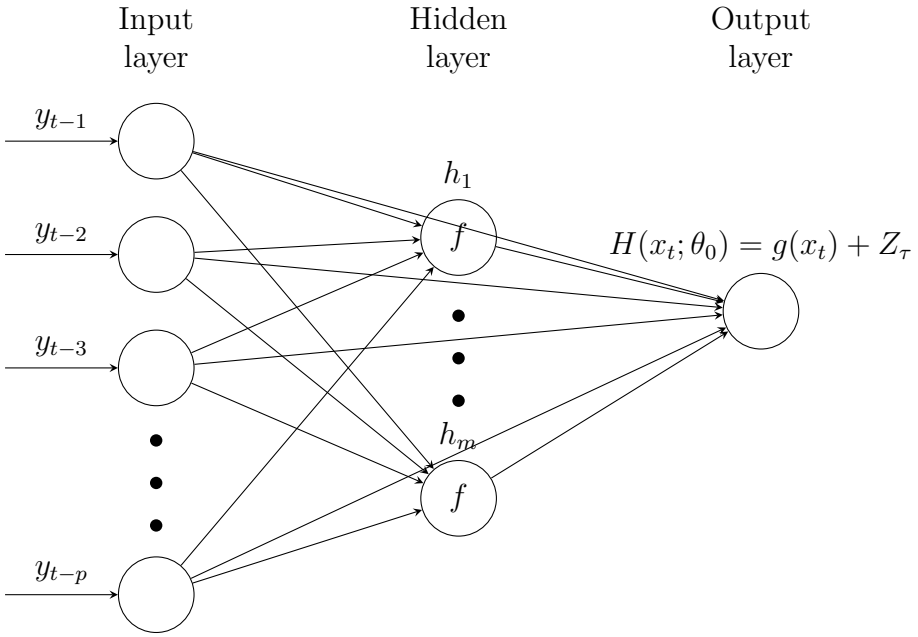


Figure 2.1: A General DPFNN Model.

We denote weight matrix connecting the input layer and the hidden layer by $W_{p\times m}$ where W_{ij} is the linear weight on the link from input node i to hidden node j. Similarly, we can denote the weight vector from the hidden layer to output node by $u = (u_1, u_2, ..., u_m)^T$ and that from input layer to output node by $v = (v_1, v_2, ..., v_p)^T$. Use $b^h = (b^{h_1}, b^{h_2}, ..., b^{h_m})^T$ and b^o as the biases at the hidden layer and output node respectively. And for representation simplicity, we stack all these parameters together as θ. In addiction, an activation $f(\cdot)$,

sigmoid function in this project, is used at all hidden nodes. So the output of this network is given as

$$H(x_t;\theta) = \sum_{i=1}^{p} y_{t-i}v_i + \sum_{j=1}^{m} f(\sum_{i=1}^{p} y_{t-p}w_{ij} + b^{h_j})u_j + b^o. \tag{2.19}$$

Observing the structure of DPFNN, we see that many widely used time series models are special cases of DPFNN. For example, an $AR(p)$ model is just a DPFNN with p input nodes and 0 hidden nodes; moreover, a DPFNN with only nonzero weight on the shortcut link from node y_{t-s} $(s < p)$ and node y_{t-1} to the output node is a seasonal $AR(1)_s$ model. In general, a DPFNN model can be regarded as the combination of a simple stationary FNN and a seasonal ARIMA (Autoregressive Integrated Moving Average) model, where the former captures time-correlation and the latter adjusts for seasonality. With properly chosen parameters, this DPFNN structure is able to capture both stationary time series and non-stationary ones with trends and seasonality, thus it is selected in our research.

2.3.2 DPFNN-based Quantile Autoregression

Once the structure of a DPFNN is given (p and m can be selected by cross-validation as suggested in many other neural network literature), our goal is to determine the value of θ_0 based on the historical realizations of the time series. And the quantile regression technique plays an important role. Inspired by Taylor's quantile regression neural network model ([103]), which captures nonlinear relationships between the process of interest and multiple exogenous features, we consider a network structure in the context of time series data instead. The model we use also departs from more recent work of [117] in the sense that we do not require historical quantiles, which are generally not observable, as inputs. Furthermore, a DPFNN structure instead of classical FNN, as in the previous mentioned literature, is used. To the best of our knowledge, it is the first time this structure has been used for quantile autoregression, with which we provide a single-step framework to deal with nonstationary time series and relax the assumption on trend and seasonality.

By the property of quantile, as θ_0 corresponds to the real quantile value function, we have that

$$\theta_0 \in \arg\min_{\theta} E_{Y_t|x_t}[\rho_\tau(Y_t - H(x_t;\theta))] \quad \forall x_t. \tag{2.20}$$

So that an estimator $\hat{\theta}_N$ can be chosen such that the following empirical analogue of expected loss in equation (2.20) is minimized

$$TC(\theta) = \frac{1}{N-p}\sum_{t=p+1}^{N} \rho_\tau(y_t - H(x_t;\theta)). \tag{2.21}$$

Meanwhile, we can rewrite equation (2.16) as

$$Y_t = H(X_t;\theta_0) + u_t, \tag{2.22}$$

where $\{u_t = \varepsilon_t - Z_\tau\}$ are i.i.d. with $Pr(u_t \leq 0) = \tau$. Note that this is indeed time series forecasting under asymmetrical error loss, as the minima of each summand in equation (2.21) is zero when the estimator $H(x_t; \theta)$ equals the real value y_t.

As in all neural network training cases, the loss function (2.21) is neither convex nor concave. Stochastic gradient-based optimization methods such as Adam (A Method for Stochastic Optimization), though do not guarantee the convergence to global optimum, have been empirically found to outperform other methods in such cases ([70]). However, an obstacle to apply these most widely used gradient-based methods to optimize (2.21) is that it is not differentiable everywhere. So we follow the treatment in [18] and approximate the error function using the finite smoothing method from [26]. Thus, the following cost function is used in our experiments instead of (2.21):

$$\hat{TC}(\theta) = \sum_{t=p+1}^{N} \hat{\rho}_\tau(y_t - H(x_t; \theta)), \tag{2.23}$$

where the Huber function, as defined in (2.24), is used to approximate the check function (2.14) and smooths the turning points by quadratic functions.

$$h(u) = \begin{cases} u^2/(2\epsilon) & \text{if } 0 \leq |u| \leq \epsilon \\ |u| - \epsilon/2 & \text{if } |u| > \epsilon \end{cases} \tag{2.24}$$

for some small constant ϵ. And

$$\hat{\rho}_\tau(u) = \begin{cases} \tau h(u) & \text{if } u > 0 \\ (1-\tau)h(u) & \text{if } u \leq 0 \end{cases}. \tag{2.25}$$

Training cycles repeated with decreasing values of ϵ. And [26] showed that as ϵ goes to zero, the algorithm converges to the minimum of the original error function. As this is essentially minimizing quantile autoregression costs with a DPFNN model, we denote the method by DPFNN-QAR.

2.3.3 Simulation

Now we want to numerically verify the efficiency of our method. We start with weakly stationary time series, which are commonly used in most current time series analysis literature. In next section, we will further establish some theoretical guarantees under this scenario. However, as in almost all cases real-world time series are not perfectly stationary, we generate data from the following nonlinear autoregressive model:

$$Y_t = 30 + 0.5 \times Y_{t-1} + 0.2 \times \frac{Y_{t-1} \times Y_{t-3}}{Y_{t-2}} + \varepsilon_t, \tag{2.26}$$

with ε_t i.i.d $N(0, 7^2)$. With these complete information, the real quantiles can be computed easily.

We initialize the generation with $y_1 = 100+\varepsilon_1$, $y_2 = 30+0.7\times y_1+\varepsilon_2$, $y_3 = 30+0.5\times y_2+ 0.2\times\frac{100\times y_2}{y_1}+\varepsilon_3$ and the remaining from the above formula. We discard data from a warm-up period of 500 points and keep the following 500 points. The first 400 points were used for model selection and training. Given different initial seeds, three random sample paths are generated. And it is verified that these series are stationary by Augmented Dickey-Fuller Test with p-values of $1.7e^{-12}$, $4.8e^{-10}$ and $4.2e^{-11}$ respectively.

With the remaining testing set of 100 data points, we compare DPFNN-QAR forecasts and real quantiles. The DPFNN structure is selected via Monte Carlo cross-validation (MCCV) in the following manner:

1. For each value of p, m, clean the data into $500 - p$ records in the form of (x_t, y_t) for $t = p + 1, ..., 500$, reserve the last 100 records as testing set and the other for training and validation;

2. Randomly select 80% records from the training set and train the DPFNN for parameters with $\tau = 0.5$;

3. Then, use the x_t from the remaining 20% validation set as the input to the trained network, and calculate the total quantile loss on the validation set;

4. Repeat Step 2) to 3) for 10 times and calculate the average cost for each (p, m) combination.

We present the cross-validation average costs in Table 2.1:

Table 2.1: Average Monte Carlo Cross-validation costs of Simulated Data (Sample 1).

	m=0	m=1	m=2	m=3	m=4
p=1	244.27	237.04	237.81	235.51	235.82
p=2	251.53	246.62	248.31	245.94	245.44
p=3	246.06	245.83	245.87	246.15	246.34
p=4	243.21	242.78	242.94	243.33	243.46
p=5	225.24	224.96	225.16	224.94	225.78
p=10	231.54	230.99	230.84	231.98	231.41

As shown in Table 2.1, the affect of m is minimal compared with that of p. In fact, many neural network literature has shown that it is sufficient to use about 3 to 5 hidden neurons to approximate an arbitrary continuous function, and our experiment consolidate the argument. Other complexity-based penalty criterion, such as AIC, can also be used for selecting p. The discussion is omitted here since its not the focus of our study.

Once a model with $p = 5$ and $m = 3$ is selected and trained, we predict different quantiles of several sample paths and compare them with the real values, and in Figure 2.2 we show the results for 3 samples. The curves for our quantile prediction almost overlap the real quantiles, indicating good performance of our DPFNN-based quantile prediction method. In

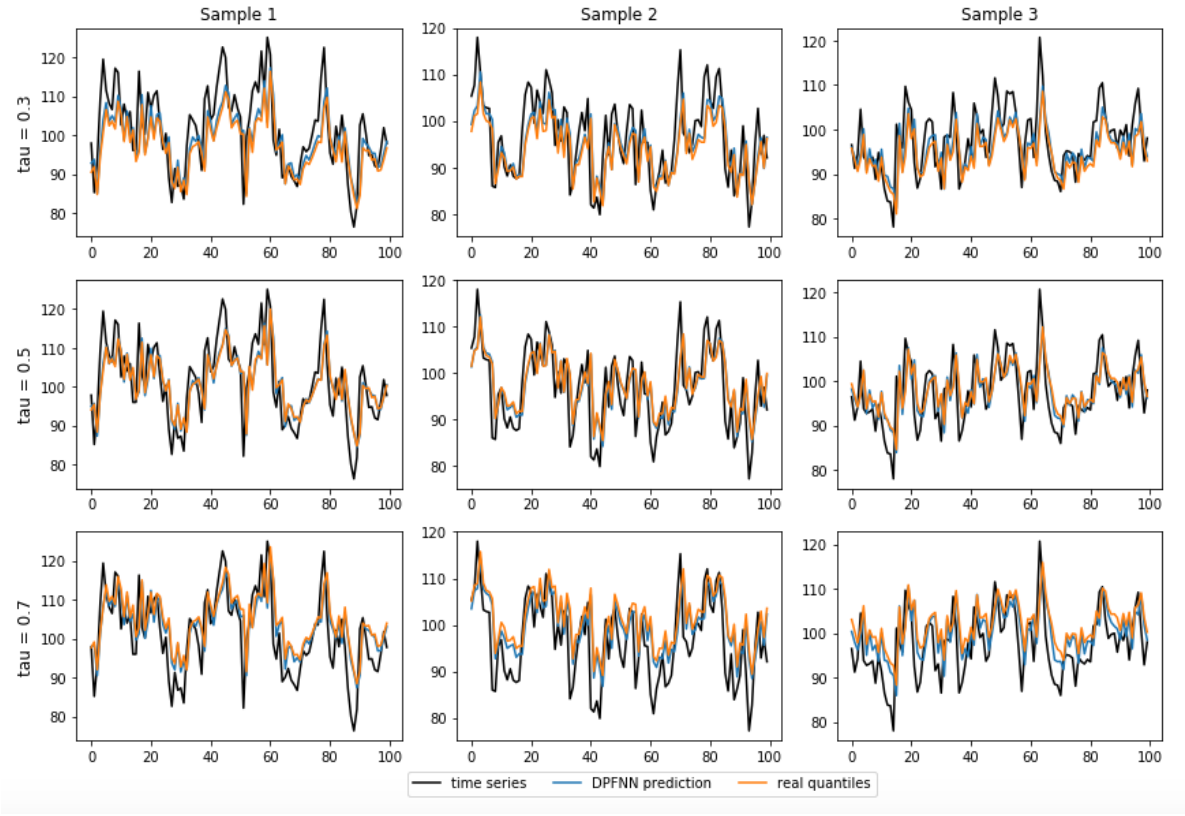


Figure 2.2: Comparison between real quantiles and DPFNN-based predictions on simulated data.

Table 2.2: Average Quantile Costs Differences between DPFNN-QAR Predictions and Real Quantiles.

τ	Sample 1	Sample 2	Sample 3
0.3	3.4%	1.4%	-1.4%
0.5	0.6%	-0.2%	0.0%
0.7	1.6%	3.1%	4.4%

Table 2.2, we summarize the percentage cost gaps between the DPFNN-QAR estimators and the real quantiles on test set. The results tell the same story where the gaps are negligible in many cases, and the average quantile costs of predictions are even lower in a few cases. We further demonstrate its application in predicting quantiles of nonstationary real time series in the next two sections.

2.4 Data-driven Newsvendor Problem

Now we consider the application of DPFNN-based quantile forecasting (DPFNN-QAR) in the field of inventory management, and start with the newsvendor problem, one of the most fundamental stochastic inventory models.

2.4.1 Problem Statement

The key element of this model is that the decision maker has a single opportunity to place an order - before the random demand is observed, no excess inventory can be carried over to the next period and all unmet demands are lost ([101]). It has important applications in stocking level management for a variety of perishable products, including newspapers, fresh produce, hotel and airline overbooking, and fashion goods etc. Specifically, we consider the newsvendor problem which is solved repeatedly in successive periods. At each iteration, the manager has to set the inventory level based on previous sales (we assume sales are uncensored demand). The following elements are taken into consideration:

Decisions:

S_t: order-up-to inventory level at period t, assuming immediate delivery (zero lead time)

Variables:

D_t: nonnegative random demand occurred at period t, and d_t is the realized value

Parameters:

H_t: history of the process up to the beginning of period t, based on which a manager makes the decision of S_t, e.g. $H_t = (d_1, d_2, ..., d_{t-1}, S_1, ..., S_{t-1})$

c: constant ordering cost per unit

h: holding cost per unit paid for excess inventory at hand at the end of each period after demand has been met

b: per unit understock cost (e.g. lost sales + penalty for unmet demand)

That is, at the beginning of time period t, the decision maker has to decide the order-up-to inventory level S_t for this period, in order to trade off the purchasing costs, overstock costs for excess inventory and understock costs for unmet demand. In the newsvendor setting, since no excess inventory can be carried over to the later periods, S_t is essentially the quantity that the decision maker has to order. To simplify the situation, we further assume that all orders arrive immediately with zero lead time. To achieve the minimal expected total cost,

it is well known that the optimal order-up-to level is given by the critical number ([101] Section 4.4.2):

$$C(S) = cS + \mathbb{E}_{D_t|H_t}\{h(S - D_t)^+ + b(D_t - S)^+\}. \tag{2.27}$$

And it follows that

$$\begin{aligned} S_t^* &= \arg\min_S C(S) \\ &= F^{-1}_{D_t|H_t}(\frac{b-c}{h+b}), \end{aligned} \tag{2.28}$$

where $F_{D_t|H_t}(\cdot)$ is the cumulative density function (cdf) of D_t given H_t.

While most literature on newsvendor assume that this cdf is unchanged and independent over time (i.i.d.), we propose to further explore the internal structure of the demand process by assuming it follows the autoregressive model (2.16). i.e.

$$D_t = g(D_{t-1}, D_{t-2}, ..., D_{t-p}) + \varepsilon_t, \tag{2.29}$$

where ε_t are i.i.d., following an unknown common distribution. Then, the optimal ordering quantity for any period t given all historical demand is

$$S_t^* = g(d_{t-1}, d_{t-2}, ..., d_{t-p}) + Z_\tau, \tag{2.30}$$

where Z_τ is the $\frac{b-c}{h+b}$th quantile of ε_t.

A practical limitation to use this result in real-life is that neither $g(\cdot)$ nor Z_τ is known. Instead, we need to determine the ordering quantities based on observable historical demand. Observing S_t^* follows the same structure as the conditional quantile as in (2.17), the problem of determining the ordering quantities boils down to finding this quantile (2.17) of the demand process with $\tau = \frac{b-c}{h+b}$. And it is thus naturally to use the DPFNN-QAR estimators to make the inventory decisions. Note that since the network is selected and trained with historical demand observations and its output is used directly as decisions, this is essentially a data-driven approach for solving the newsvendor problem.

2.4.2 A Data-driven Approach

Existing parametric methods first assume that $g(\cdot)$ has a simple linear structure such as $AR(1)$ or $AR(p)$ with unknown coefficients ([31] and [20]) and ε_t are normally distributed with mean 0 and unknown variance σ^2. Then, OLS estimator is used to replace the coefficients of $g(\cdot)$ and forecast errors are regarded as a sample of ε_t to estimate σ. As in all parametric approaches, the choice of the structure of $g(\cdot)$ impacts the results significantly. Moreover, the simple linear $g(\cdot)$ and normal ε_t assumptions are too restrictive in most real-life scenarios.

Inspired by the universal approximating capability of DPFNN and quantile regression, we propose to use DPFNN-QAR to estimate $S_t^* = H(x_t; \theta_0)$ in a single-step framework. And we can rewrite the demand process (2.22) as $D_t = H(x_t; \theta_0) + u_t$ where $u_t = \varepsilon_t - Z_\tau$ also i.i.d.. As discussed in Section 2.3, the structure of $g(\cdot)$ no longer needs to be defined in advance.

Instead, $g(\cdot)$ is allowed to be nonlinear, arbitrarily complicated, depending on data and the normality constraint of ε_t is also relaxed. Consequently, a natural practical policy is to use the ordering quantities $\hat{S}_t = H(x_t; \hat{\theta})$, where $H(x_t; \hat{\theta})$ is selected and trained following the DPFNN-based quantile autoregression (DPFNN-QAR) procedure described in Section 2.3.

From a second point of view, this DPFNN-QAR approach can also be interpreted as an integrated data-driven solution to newsvendor problem. To justify this argument, we refer readers to the data-driven linear programming proposed by [12], where they assumed a linear relationship between demand and some explanatory variables. This assumption leads to the observation that the required inventory level is also a linear combination of the same factors, whose coefficients can be determined by solving a data-driven cost model formulated as the following LP problem:

$$\begin{aligned} \min_{\beta} & \sum_{i=1}^{N} (hy_i + b(d_i - s_i) + c\beta^T x_i) \\ s.t. & y_i \geq \beta^T x_i - d_i \quad i = p+1, ..., n \\ & s_i \leq d_i \quad i = p+1, ..., n \\ & s_i \leq \beta^T x_i \quad i = p+1, ..., n \\ & s_i, y_i \geq 0 \quad i = p+1, ..., n. \end{aligned} \tag{2.31}$$

By assuming that the demand is some complicated nonlinear function of the previous p observations $x_t = (d_{t-1}, d_{t-2}, ..., d_{t-p})$, and then approximating it by a DPFNN structure $H(\cdot; \theta_0)$, we obtain the following integrated approach for determine the stock inventory level with historical demand $\{d_1, ...d_N\}$. The decision variables, in our case, are the weights θ and indirectly the excess inventory levels y_i and satisfied demands s_i. And the goal is likewise to minimize the total in-sample costs:

$$\begin{aligned} \min_{\theta} & \sum_{i=p+1}^{N} (hy_i + b(d_i - s_i) + cH(x_i; \theta)) \\ s.t. & y_i \geq H(x_i; \theta) - d_i \quad i = p+1, ..., n \\ & s_i \leq d_i \quad i = p+1, ..., n \\ & s_i \leq H(x_i; \theta) \quad i = p+1, ..., n \\ & s_i, y_i \geq 0 \quad i = p+1, ..., n. \end{aligned} \tag{2.32}$$

However, model (2.32) can no longer be solved as an LP as the (2.31) proposed in [12], and the fact that $H(\cdot; \theta)$ being neither convex nor concave makes it even more challenging. To characterize the optimal solution, we observe that regardless of the value of θ, the objective function and constraints always force $y_i = \max(H(x_i; \theta) - d_i, 0)$ and $s_i = \min(H(x_i; \theta), d_i)$. Thus, (2.32) is equivalent to

$$\begin{aligned}
&\min_{\theta} \sum_{i=p+1}^{n} \left(h \max(H(x_i;\theta) - d_i, 0) + b \max(d_i - H(x_i;\theta), 0) + cH(x_i;\theta)\right) \\
\Leftrightarrow &\min_{\theta} \sum_{i=p+1}^{n} \left((h+c) \max(H(x_i;\theta) - d_i, 0) + (b-c) \max(d_i - H(x_i;\theta), 0) + cd_i\right) \qquad (2.33) \\
\Leftrightarrow &\min_{\theta} \sum_{i=p+1}^{n} \left(\frac{h+c}{h+b} \max(H(x_i;\theta) - d_i, 0) + \frac{b-c}{h+b} \max(d_i - H(x_i;\theta), 0)\right).
\end{aligned}$$

The objective function is essentially the loss function used in DPFNN-QAR when choosing $\tau = \frac{b-c}{h+b}$. Thus solving the data-driven newsvendor problem (2.32) is equivalent to training a DPFNN-QAR model.

2.4.3 Asymptotic Optimality

When the demand process D_t is covariance stationary and some general regularity conditions hold, it's easy to verify the uniformly convergence of $H(\cdot, \hat{\theta}_N)$ to $H(\cdot, \theta_0)$ by following a similar argument as presented in [28] and Theorem 2.2 from [115], i.e. $\hat{S}_t$ converges to the real optimal solution S_t^* as the number of previous demand points goes to infinity. And we can theoretically support the performance of DPFNN-QAR demonstrated in Subsection 2.3.3.

Theorem 2.2. *Let $\{D_t\}$ be an ergodic stationary process. Suppose that (i) The parameter space Θ is compact; (ii) $E[\sup_{\theta\in\Theta} |H(X_t;\theta)|] < \infty$; (iii) $F_u(\cdot)$, the common cdf of $\{u_t\}$, is differentiable and has mass around 0 (i.e. if we denote its derivative by $f_u(\cdot)$, there exits $\epsilon > 0$ such that $f_u(s) > 0 \quad \forall s \in [-\epsilon, \epsilon]$) . Then, $H(\cdot, \hat{\theta}_N)$ converges to $H(\cdot, \theta_0)$ uniformly.*

Proof. See Appendix for the proof. □

Thus, the policy $\hat{S}_t$ is asymptotically optimal under the conditions of Theorem 2.2.

Corollary 2.1. *If the demand process satisfies conditions of Theorem 2.2, then $\hat{S}_t$ converges in probability to the optimal policy S_t^* as the number of historical records N goes to infinity, i.e. $\hat{S}_t \xrightarrow{p} S_t^*$ Moreover, $C(\hat{S}_t) \xrightarrow{p} C(S_t^*)$.*

Proof. It follows from the consistency of $\hat{S}_t$ and the continuous mapping theorem. □

Nevertheless, through numerical examples in the next subsection, we show that this method also works well on nonstationary time series.

2.4.4 Case Study

Up till now, we have already shown theoretically and numerically that DPFNN-QAR is efficient when deal with weakly stationary demand series. Many demand process in real world, however, exhibits certain patterns. For example, the sales of many foods are obviously seasonal. We apply the proposed method on such two real time series, and compare it to the Holt-Winters' methods with normal innovations and the quantile versions of Holt-Winters. Moreover, we also consider a two-step procedure by stationarizing the time series first and then training a model on the transformed data.

1) Datasets

Two time series are used for case study. The first one comes from the University of Wisconsin Dairy Marketing and Risk Management Program maintained by Prof. Brian W. Gould of the Dept. of Agricultural and Applied Economics. The time series contains the monthly regular ice cream production (measured in thousand gallons) in US. We selected data from January 1983 to January 2017, which contains 409 observations, where the first 360 were used for model selection and training and the remaining for testing.

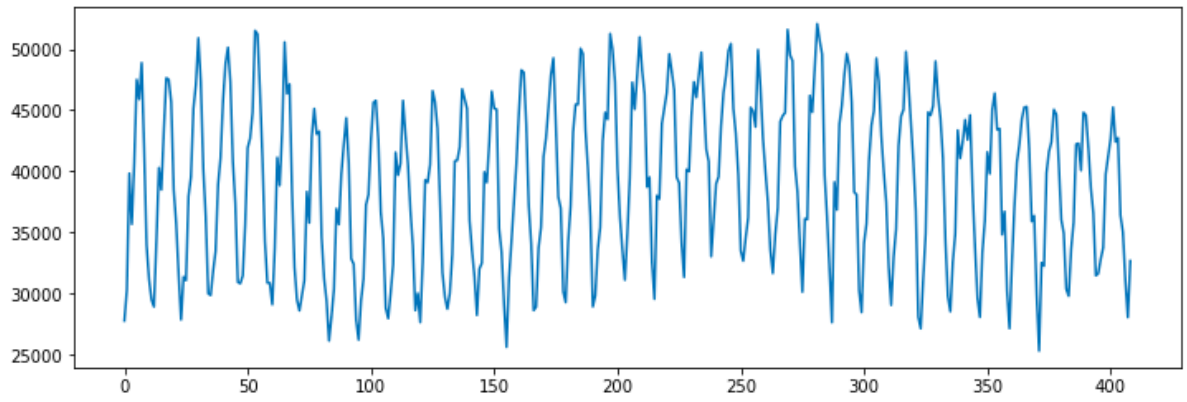


Figure 2.3: Ice-cream demand time series.

The other dataset contains monthly gasoline demand (measured in million gallons) in Ontario, Canada from January 1960 to December 1975. We got these 192 fact values from Datamarket [2]. While the first 143 points are used for training and validation, we tested the selected model on the remaining 49 months of data.

As demonstrated in Figure 2.3 and Figure 2.4, both time series show annual seasonality. Augmented Dickey-Fuller, with p-values of 0.40 and 0.99 further respectively, provides evidence for nonstationarity. Furthermore, there is an obvious increasing trend in gasoline

[2]https://datamarket.com/data/set/22of/monthly-gasoline-demand-ontario-gallon-millions-1960-1975#!ds=22of&display=line

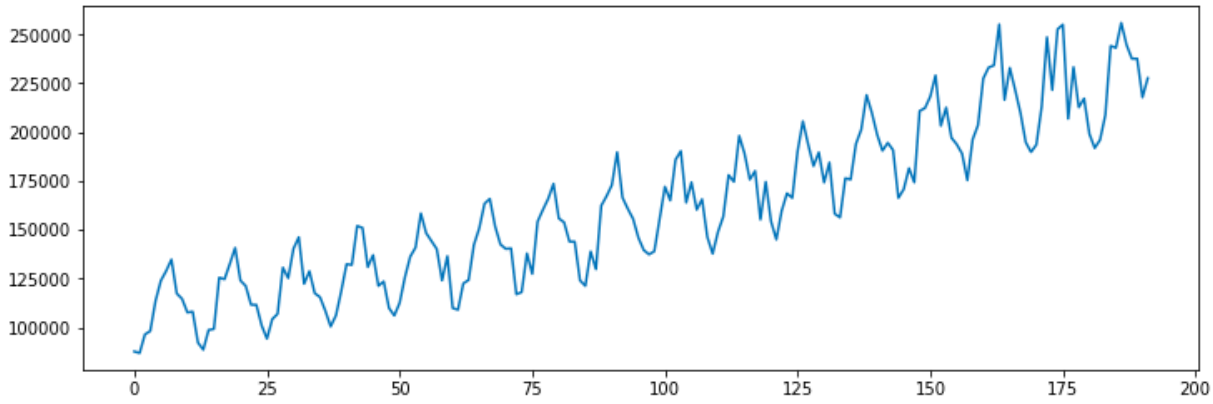


Figure 2.4: Gasoline demand time series.

demand, while there is no steady trend in the time series which imposes difficulties in data preprocessing. Though we have to admit that newsvendor decisions are generally made on store or warehouse levels, these nationwide time series should also be representative for some demand patterns and bring some insights.

2) Benchmark Methods

Due to the cyclical pattern of these two time series, we select the widely used Holt-Winters' triple exponential smoothing method, which is suitable for forecasting time series that exhibit both trend and seasonality ([101] and [88]), to illustrate the prediction power of our model. Depending on the form of seasonality assumed, there are two versions of Holt-Winters' formulation, i.e. Additive Holt-Winters' (HWA) and Multiplicative Holt-Winters' (HWM). Both versions decompose the data into level, trend and seasonality components, where parameters are estimated by minimizing mean squared errors of the training data. Then, a quantile is obtained by assuming the time series has White Noise innovations (Gaussian) and then estimating its variance via observed residuals. We refer the readers to [88] for details and initialization of these two algorithms. Later, [2] proposed quantile versions of Holt-Winters methods, denoted by QHWA and QHWM respectively, which replaces the MSE criterion by the quantile loss just as we did in our method. We also implemented and compared these methods with DPFNN-QAR. Please refer to Section 2.2 for the detailed formulation of these methods.

Moreover, it is under these nonstationary real-world time series, we can observe the major difference between our method and the quantile regression neural networks (QRNN) proposed by [103] and [18]. Without data preprocessing, the original QRNN fails and produces estimations that are almost constant over time at the sample quantile of the training data. We omit the details of QRNN results as they are far from being accurate. Instead, we conduct comparison with a closely related and intuitive method, that is to first remove trend and seasonality by differencing, train a QRNN model on the stationarized time series and

finally convert back to the original scale. We denote this method by QRNN with differencing (QRNN-D). Both order of simple and seasonal differences are selected to be 1 in our case, the the stationarity of the two transformed times series is validated by Augmented Dickey-Fuller test with p-values of $3.5e^{-7}$ and $2.2e^{-5}$. The structure of the QRNN is chosen via the same cross-validation procedure where $p = 18$ and $m = 3$ are chosen for both datasets following the same Monte Carlo cross-validation procedure as described in Subsection 2.3.3.

3) Experimental Design

The program is implemented in Python 3.5 using tensorflow 1.0 for modeling the neural network structure. AdamOptimizer ([70]) is used with learning rate= $(N - p) \times 10^{-4}$ for tuning the weights. And the algorithm terminates when a maximum number of 20000 is reached or the relative change of loss function is less than 10^{-9}. ϵ in (2.24) was initially set to 2^{-5} and was halved every 500 training epochs. As for the benchmark algorithms, they were also implemented in Python3.5 and the objective functions were optimized by L-BFGS-B.

4) Results and Discussion

Based on the results from cross-validation, a DPFNN with $p = 24$ and $m = 4$ were used for modeling the ice cream time series. Similarly $p = 23$ and $m = 1$ were chosen for the gasoline series. Then, we trained the selected models and tested them on the reserved testing set. Here, we demonstrate the predictions of the 0.8th quantile of the ice cream demand and the 0.4th quantile of the gasoline case as examples.

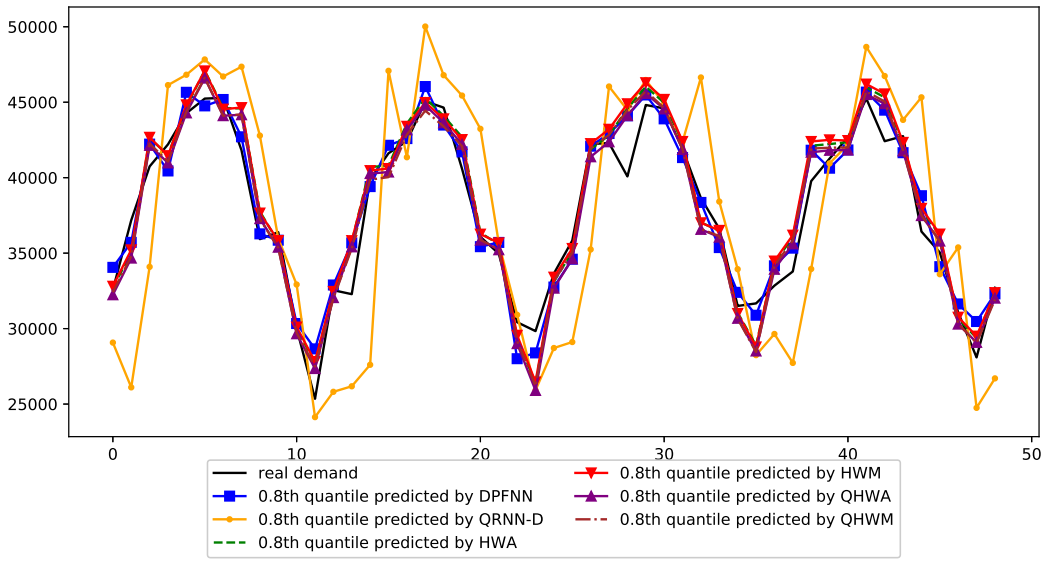


Figure 2.5: Ice-cream Demand and Predictions.

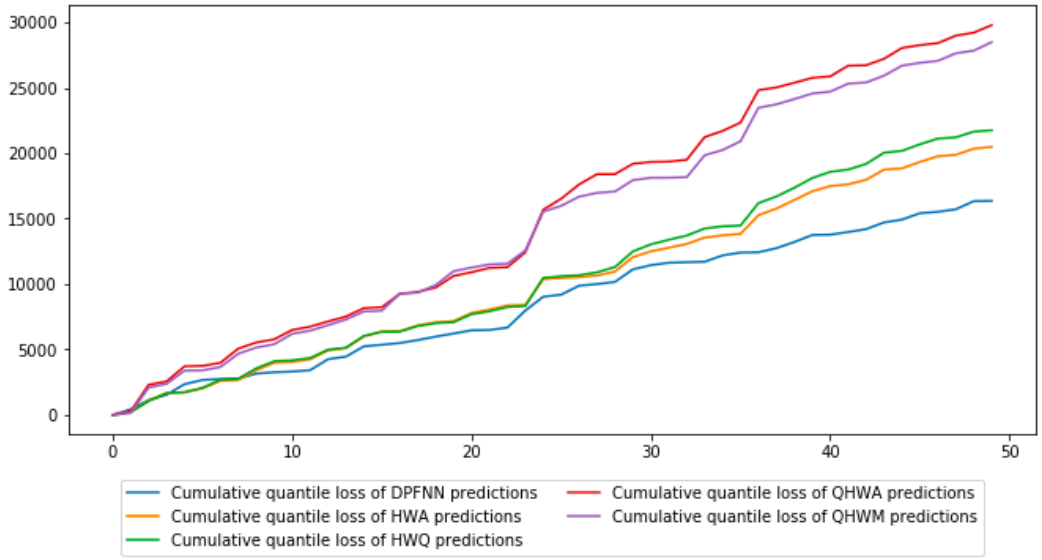


Figure 2.6: Cumulative quantile loss of different predictions of Ice Cream Demand.

Table 2.3: Relative Changes of Quantile Loss for Ice Cream Dataset.

Benchmark Algorithm	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.8$
HWA	-5.53%	-13.20%	-12.57%	-11.60%	-19.37%
HWM	-4.62%	-12.34%	-13.01%	-12.90%	-24.06%
QHWA	-32.07%	-15.38%	-13.83%	-17.57%	-44.58%
QHWM	-34.55%	-15.20%	-12.67%	-15.41%	-41.97%
QRNN-D	-78.72%	-73.47%	-71.97%	-73.53%	-83.18%

The testing set of US regular ice cream demand consists of data from 49 months. The real demand together with predictions by different methods are shown in Figure 2.5. The performance of QRNN-D is significantly worse than all other methods as shown in Figure 2.5 and Table 2.3. Two reasons may have contributed to this failure - as shown in 2.3, there is no steady linear trend in the time series and the order one differencing transformation may have overfitted to the training data; as argued earlier, quantile estimation requires i.i.d. innovations, such an assumption is rejected by the runs test as shown in Table 2.5. All the other five types of predictions lie above of the real demand most of the time as desired, since now we penalize underestimation more than overestimation. There are no significant difference between the four groups of different versions of Holt Winters predictions, while the curve corresponding to DPFNN-QAR is closer to the real demand indicating a better prediction. The efficacy of our method can also be seen by observing the cumulative quantile loss in Figure 2.6 (the result from QRNN-D is neglected as it's significantly higher than the other methods). In fact, the average quantile loss of DPFNN-QAR predictions in 49 month is significantly

lower than all benchmark algorithms. To further illustrate the efficiency of our method over the others, we ran experiments to predict different quantiles, and the relative changes of average out-of-sample total costs, calculated as $\frac{\text{average loss of DPFNN}-\text{average loss of benchmark}}{\text{average loss of benchmark}}$, are summarized in the Table 2.3. The results show that our DPFNN-QAR method significantly beat the five benchmark methods.

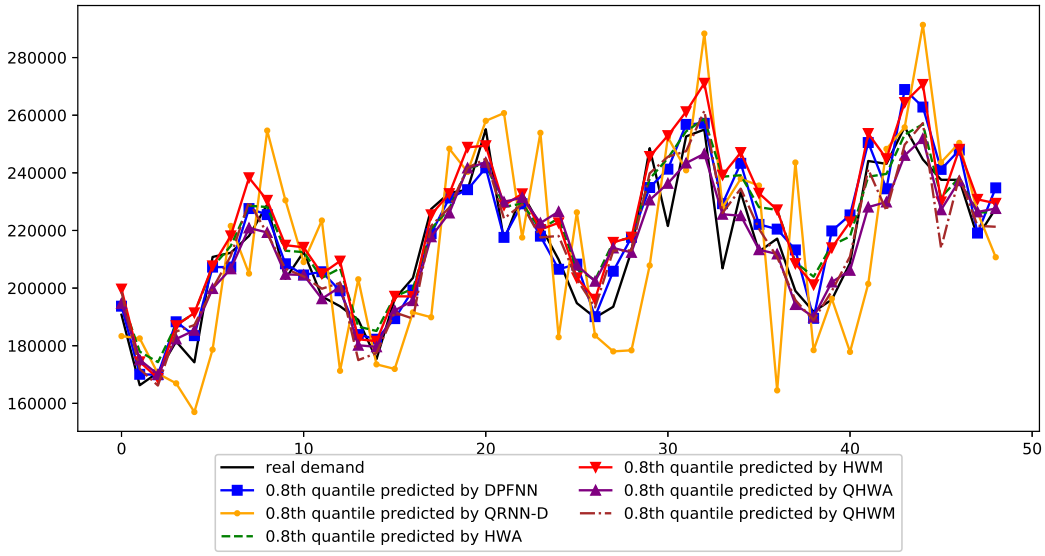


Figure 2.7: Gasoline Demand and Predictions.

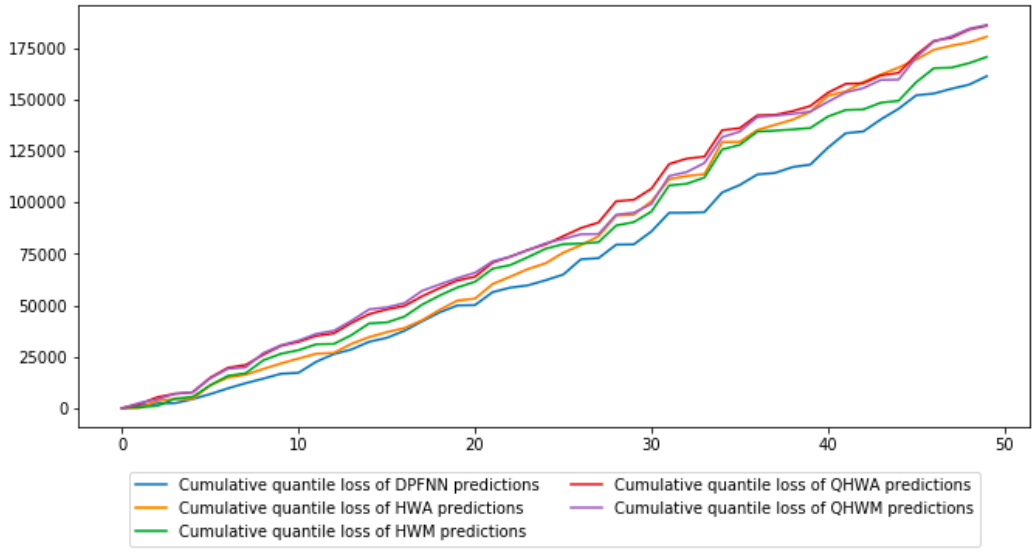


Figure 2.8: Cumulative quantile loss of different predictions of Gasoline Demand.

Table 2.4: Relative Changes of Quantile Loss for Gasoline Dataset.

Benchmark Algorithm	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.8$
HWA	-2.23%	-8.60%	-9.26%	-6.27%	4.16%
HWM	7.78%	-3.32%	-6.65%	-6.66%	7.69%
QHWA	-10.85%	-9.87%	-9.06%	-10.06%	-32.56%
QHWM	-13.22%	-11.09%	-10.71%	-11.04%	-28.77%
QRNN-D	-73.18%	-66.99%	-66.76%	-67.58%	-73.20%

Similar experiment is conducted to predict the 0.4th quantile of the Gasoline demand for 49 months. And the results are shown in Figure 2.7 and Figure 2.8 respectively. Again, although the curve for QRNN-D looks much better compared with the first dataset, its performance is still not promising in cost evaluation. One possible reason is that differencing transformation is not flexible enough to capture the real complicated nonstationarity as efficient as the other models. Once the orders of differencing are determined, this transformation assumes fixed linear trend and seasonality, while such elements are updated in the Holt-Winters' methods with each new data point and even more flexible structure is allowed in DPFNN. Even though there can be other methods which stationarize these datasets better, the best choice of such a method is unknown in practice. And the choice of transformation affects the final prediction dramatically, which is indeed the major shortcoming of such two-step methods. Meanwhile, it is seen that although now we penalize overestimation more, all Holt-Winters predictions tend to overestimate the process more than our method. Again, we can see that the out-of-sample average period cost of DPFNN-QAR is much lower in most cases than that of the Holt-Winters methods as shown in Figure 2.8 and Table 2.4.

One possible explanation for DPFNN-QAR performing less better in this dataset is that we have much fewer observations. It reveals one shortcoming of this method that a large number of records are needed for accurate estimation of those parameters.

To validate whether the i.i.d assumption of the random innovations is satisfied, we first train all models using MSE criterion. Then, we use the residuals as a proxy for the sample path of $\{\varepsilon_t\}$ and perform the turning point test and the runs test using R. Both methods are frequently used to test the null hypothesis that the remaining residuals are i.i.d. As shown, in Table 2.5, we fail to reject the null hypothesis under DPFNN and Holt-Winters models, hence the i.i.d. assumption is tenable.

2.5 Multiperiod Safety Stock

2.5.1 Problem Statement

Now we consider the extension to multiperiod newsvendor scenario, where excess inventory will be carried over to the next period and unmet demand is backlogged. We introduce the

Table 2.5: p-values of statistical tests for i.i.d. residuals.

	p-value	DPFNN	QRNN-D	HWA	HWM
Ice-cream	turning point test	0.818	0.818	0.818	0.645
	runs test	0.183	**0.002**	0.311	0.663
Gasoline	turning point test	0.565	0.206	0.908	0.908
	runs test	0.311	0.311	0.826	0.936

following notation:

Decisions:

S_t: order-up-to inventory level at period t, assuming immediate delivery (zero lead time)

x_t: initial inventory at the beginning of period t, negative x_t means backlogged demand

Variables:

D_t: nonnegative random demand occurred at period t, and d_t is the realized value

Parameters:

H_t: history of the process up to the beginning of period t, based on which a manager makes the decision of S_t, e.g. $H_t = (d_1, d_2, ..., d_{t-1}, S_1, ..., S_{t-1})$

c: constant ordering cost per unit

h: holding cost per unit paid for excess inventory at hand at the end of each period after demand has been met

b: per unit understock cost (e.g. lost sales + penalty for unmet demand)

T: number of periods in the planning time horizon

γ: discounting factor for calculating the present value of future costs

Furthermore, $x_1 = 0$ is given and it is required that $S_t \geq x_t$. To facilitate the derivation of closed-form ordering policy, we need further assume that all remaining inventory at the end of the planning horizon can be returned to supplier at the original price c, and all backlogged demand will also be satisfied at the same cost.

Then, our goal is to find a sequence of ordering quantities $\bar{S} = (S_1, ...S_T)$, such that the expected discounted total cost defined below is minimized.

$$f_T(\bar{S}) = \mathbb{E}\{\sum_{t=1}^{T} \gamma^{t-1}[c(S_t - x_t) + g(S_t, D_t)] - \gamma^T c x_{T+1}\} \tag{2.34}$$

where $x_1 = 0$ is known and

$$x_{t+1} = S_t - D_t \quad t = 1, 2, ..., T \tag{2.35}$$

$$g(S_t, D_t) = h \max(S_t - D_t, 0) + b \max(D_t - S_t, 0) \tag{2.36}$$

2.5.2 Proposed Method

A base-stock is proven optimal in [101]. However, the optimality of the myopic policy no longer holds since our demand process may not be stochastically increasing any more. Instead, we need to reestablish the derivation of the optimal ordering policy for our autoregressive demand process based on the results of [106].

Theorem 2.3. *If the cost at the end of the planning horizon is $-cx_{T+1}$, then the myopic base-stock quantity given by*

$$\begin{aligned} S_t^* &= F_{D_t|H_t}^{-1}(\frac{b-(1-\gamma)c}{h+b}) \\ &= g(d_{t-1}, ..., d_{t-p}) + Z_\tau \\ &= H(x_t; \theta_0) \end{aligned} \tag{2.37}$$

is optimal in every period.

Proof. See proof in the Appendix section. □

Thus, we propose to use the DPFNN-QAR prediction of the $\frac{b-(1-\gamma)c}{h+b}$th quantile, $\hat{S}_t = H(x_t; \hat{\theta})$, as the ordering quantity. Note that if $H(\cdot; \hat{\theta}) - H(\cdot; \theta_0) \leq \delta$ and demand is strictly bounded above from zero, say $D_t \geq 2\delta$, then it is always possible to order-up-to $H(x_t; \hat{\theta})$ in all periods since

$$\begin{aligned} & H(x_t; \hat{\theta}) - D_t \leq H(x_{t+1}; \hat{\theta}) \\ \Leftrightarrow & H(x_t; \hat{\theta}) - (H(x_t; \theta) - Z_\tau + \varepsilon_t) \leq H(x_{t+1}; \hat{\theta}) - H(x_{t+1}; \theta) + H(x_{t+1}; \theta) \\ \Leftrightarrow & (H(x_t; \hat{\theta}) - H(x_t; \theta)) - (H(x_{t+1}; \hat{\theta}) - H(x_{t+1}; \theta)) \leq g(x_{t+1}) + \varepsilon_t \end{aligned} \tag{2.38}$$

The LHS$\leq 2\delta$.

Again, following the same argument as in the classical newsvendor case, this DPFNN-QAR policy is asymptotically optimal.

2.6 Conclusion

Although there is an extensive literature on stochastic inventory control, most studies assume that future demand distribution is known in advance or the demand process evolves according to a given model. However, in reality, the only observable information is the past demands (or more precisely, the past sales). The majority of papers on data-driven approaches focus

on identically and independently distributed demands. While it is of importance to learn about and make use of the internal structure of a demand process when making inventory decisions, very limited work have done to solve for inventory policies with time-correlated demand observations.

Thus, we extend the $AR(p)$ models to a more general autoregressive evolution for defining the demand process so that the demand correlation can be captured and no parametric form need to be determined in advance. And inspired by the universal approximation capability of neural networks and the idea of quantile regression, we develop a neural network based framework for predicting quantiles of this process. A DPFNN structure is chosen considering its advantage of modeling nonstationarity over standard FNN structure, which is used in most previous work on neural network-based time series analysis. Implementation details of this algorithm is discussed, and its efficacy is shown by both theoretical analysis and computational experiments. Subsequently, we propose to use the DPFNN-QAR predictions as the ordering quantity in newsvendor setting and its multiperiod extension.

The contribution of this project is twofold. Aforementioned, we first extend the demand of the newsvendor model to be an nonparametric autoregressive process and propose a data-driven method for finding optimal ordering decisions. Second, our algorithm also uses a new structure for quantile autoregression, which works for some nonstationary process. However, due to the lack of assumptions on the real underlying demand process, we sacrifice the availability of theoretical guarantees of our data-driven approach. Although our numerical results support its efficiency, we cannot bound the out-of-sample costs, which is desirable in practice especially when the manager needs to be risk-averse. Moreover, for many products such as fashion goods, there may not be enough time series data to train such a neural network model. Thus, in Chapter 3, we switch to a causal demand model which takes into account external features. And based on the technique of distributionally robust optimization, we aim at deriving a solution that performs well even under worst-case scenario.

Chapter 3

Distributionally Robust Newsvendor under Causal Demand

3.1 Introduction

In Chapter 2, we have introduced a neural network based method for solving newsvendor problem under nonstationary demand, with the focus of providing more flexible and practical decision making support with exclusively demand time series data. However, even though the theoretical and numerical analysis have demonstrated the efficiency of such a method in dealing with both stationary and nonstationary real-world time series, we have to admit that there are still practical limitations with the approach. First, time series models, especially neural network models, require a lot of historical demand data for training. However, for new products and products with short life cycle, such as fashion goods, such information is not available. This is where external features can play an important role to help estimating demand based on side information and learning from similar products. For instance, in almost cases, price is negatively related to demand and products with same material, style may share similar demand pattern. With the growing availability of such data, feature-based methods are introduced to take them into account [12, 7].

A more problematic shortcoming is that while the DPFNN-QAR performs well on real-world nonstationary time series, we unfortunately cannot provide a theoretical guarantee that it will perform well in all general cases. Even with weakly stationary and ergodic time series, we can only establish asymptotic convergence, but cannot characterize the finite-sample performance. Moreover, from simple linear models to flexible neural networks, most aforementioned methods seek for estimators which minimize some certain loss function evaluated on the training data. Thus, in most cases, out-of-sample performance guarantees do not exist with small or moderate sample size. It is possible that small changes in the data or in sample size yield large changes in the resulting solution. We suggest [10] for a more detailed discussion regarding this phenomenon.

In this chapter, we seek to address these two issues simultaneously and propose a robust

feature-based newsvendor solution with finite-sample performance guarantees. We start with a simple model by assuming that demand is a linear combination of multiple features with unknown coefficients plus a random noise. Although we put our analysis under the umbrella of data-driven newsvendor solution, due to the fact that the so-called critical quantile serves as the optimal solution, our approach naturally solves any feature-based quantile prediction and falls in the field of supervised learning.

When talking about supervised learning, especially linear regression, depending on the nature of the design (feature) points, there are two versions. The fixed design version assumes that the features observations in the data points are deterministic as given; while the random design models correspond to the statistical learning setup where features are realized i.i.d. from some underlying random vector. Current applications of distributionally robust optimization in supervised learning fall in the second stream assuming a random covariate vector and that i.i.d. sample of the features with their corresponding dependent variable values are available for model training. As a result, they aim at minimizing the worst-case expected cost with respect to randomness from both the covariate vector and some noise. In this chapter, however, we consider a fixed design matrix instead. That is, we assume that our data points are given in a non-i.i.d. manner such as a series of controlled experiments and the features are observable or even designed before the decision making procedure. We reckon that this assumption makes more sense in many real-world applications. For instance, when predicting the demand of clothes, certain features such as colors and materials exhibit different popularity patterns due to the fashion trends over time, thus, our observations will be far from being i.i.d. This phenomena is also closely related to the study of covariate shift [102]. Moreover, since in many cases features such as price and style are predetermined before inventory decisions, we only need to cope with the conditional randomness of demand in the objective function.

The chapter is organized as follows. We next provide a brief literature review on robust optimization and feature-based and robust inventory management techniques in Section 3.2. Then in Section 3.3, we describe a distributionally robust feature-based newsvendor model setup and propose a two-step framework for solving it. Section 3.4 establishes both asymptotic and finite-sample performance guarantees. Section 3.5 rewrites it into a tractable convex optimization formulation. Finally, in Section 3.7, we conclude the insights with a discussion of the practical takeaways as well as limitations of the current setup. All proofs for supporting our analysis can be found in Appendix B.

3.2 Literature Review

As discussed, our work contributes to the following two areas of investigation in inventory management and robust optimization. On one hand, we extend the distributional robust optimization techniques to incorporate side information as observable features decision making without assuming i.i.d. data points; on the other hand, we give a robust data-driven solution to newsvendor problem with both finite-sample and asymptotic performance guarantees.

Thus, we review literature from these two fields in order to identify our contributions.

Compared with the numerous literature studying newsvendor problem, the work which incorporates feature information is, though not entirely new, relatively limited. [7] provided a detailed review of a few related work in comparison to their algorithms. [98] modeled demand as a linear function of random features and solved an approximation of a robust optimization for decision rules. Their theoretical results depend on the assumption that a few important statistics such as mean and support etc. of the demand are known to the decision maker. [50] took into account the information of a state feature and proposed to solve a weighted empirical stochastic optimization problem. However, this approach requires a discrete state variable and suffers from the curse of dimensionality when high-dimensional feature data is available. Moreover, there is no guarantee on its out-of-sample performance.

From the perspective of data-driven decision making, as described in Section 2.4.2, [12] proposed linear programming models to deal with the case when demand is a linear combination of some exogenous variables and a random shock. Their cost model is essentially the well-known linear quantile regression. Then, [96] extended it to the case where demand is censored. However, in both papers, the goal was to minimize in-sample costs following an empirical risk minimization (ERM) idea, also known as sample average approximation (SAA), which essentially assumes that the real demand follows the empirical distribution of the sample. This treatment in general suffers from the overfitting effects. Later, [7] incorporated regularization for dimension reduction and proposed another nonparametric approach based on kernel regression. With assumption of linear demand process, compact support of random shock density function and i.i.d. sample, the authors derived a high probability bound for finite-sample optimality gap. Nevertheless, their performance bound is loose when the feature space is of high dimension. And even with asymptotically infinite sample points, the true out-of-sample cost cannot be bounded tightly.

In addition, above listed literature all assume each feature and demand data point is drawn i.i.d. from an unknown joint distribution. Such assumption does not hold in many real-world applications. For example, certain clothes materials and colors are more popular in some time periods as fashion changes, and will be far from being i.i.d. Moreover, these feature data can be observed in advance before decision making. However, current literature did not take the feature information into account.

A another group of techniques which provide out-of-sample performance guarantees come from the concept of robust optimization. Its usage in solving inventory problems dates back to 1958 when Scarf proposed a min-max solution to newsvendor [97]. Considering all possible values of the demand, a worst-case solution is used as the decision. Studies fall in this stream also include [78, 92] etc. This treatment though increases robustness, is however too conservative in most cases. For example, optimality may be sacrificed to cope with some extreme values of demand that seldom occur.

In an effort to address these issues, distributionally robust optimization 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 reformulated with respect to the worst case expected cost over the choice of a distribution in this set. In this field, different methods

have been proposed for constructing the ambiguity sets, leading towards various theoretical properties. Many of them have been applied for assisting inventory decision making in the newsvendor problem. [89] studied the newsvendor problem with partial information about the demand distribution (e.g., mean, variance, symmetry, unimodality), and derived the order quantities that minimize the worst-case costs. With certain special cases, closed-form solutions were derived and their relationship to entropy maximization was established. Similarly, [124] focused on the case when only mean of the demand and one of its variance or support is known. Instead of worst-case newsvendor cost, they attempted to minimize the regret with respect to the expected cost based on complete information. [92] incorporated CVaR-based profit maximization under the assumption of ellipsoid or box discrete distribution; and [119] demonstrated that optimal ordering decision with discrete demand is very different from that with continuous demand. The similar idea has been extended to solve more complicated inventory management problems beyond newsvendor, e.g. [108, 49, 5].

To take advantage of the historical demand points, a common practice is to firstly guess a nominal distribution based on observed sample or expert advice (e.g. empirical distribution). Then an ambiguity set can be constructed around this nominal distribution consisting of those distributions that are not far from it, with different choices of metrics for measuring this distance. For example, [113] constructs an ambiguity set such that the observed data achieves an lower -bounded empirical likelihood; [10] uses the confidence region of a goodness-of-fit test; [93] limits the variance distance between the candidate distributions and the nominal; and [40] measures the distribution distance using the Wasserstein distance instead. Other choices of distance metrics include but are not limited to other forms of f-divergence, e.g. Kullback-Leibler divergence, Prokhorov metric etc. Among these options, Wasserstein distance stands out being able to measure the distance between discrete and continuous distributions and incorporates a notion of how close two points are to each other. Thus, Wasserstein distance has become more and more popular in recent years and is chosen in our study.

Nevertheless, aforementioned robust-optimization-based approaches all assume a stationary unchanging demand environment, so that when historical data is taken into consideration, the demand points can be regarded as i.i.d. sample. Thus, they do not pertain to the time-correlated and feature-dependent demand process we explore in this chapter. When it comes to data-driven robust optimization with non-i.i.d. data, very limited work is currently available. [116] allowed the sequence of future demands to evolve as a martingale, however, still restricted to given constant mean and support. [20] explored the worst-case demand realization under an autoregressive demand evolution, which can be too conservative in real-world applications. Moreover, there is no out-of-sample performance guarantees for these approaches. And again, for new or fast-fashion goods, there may not be sufficient demand points available, and feature-based data is still not accounted for.

Motivation and Contributions

Motivated by the lack of feature-based robust optimization techniques, we start with a linear demand model and propose a data-driven robust solution to the newsvendor problem. Our contributions can be summarized as follows:

- *Distributionally robust optimization formulation without i.i.d. sample points.* In contrast to the existing literature in robust or distributionally robust optimization, which assume that i.i.d. sample points of the randomness are available so that an ambiguity set can be constructed accordingly (e.g. [1]), we interpret the feature observations as deterministic. The randomness only comes from the i.i.d. noises which are not observable directly. Moreover, when defining the objective function, we assume that the features are already observed before the decision making process. To decouple the two sources of obscurity from the unknown linear coefficients and the distribution of random noise, we propose a two-step framework. Ordinary least squares (OLS) estimators are used as a substitute for the coefficients, based on which a distributionally robust estimator for quantile is derived.

- *Finite-sample and asymptotic performance guarantees under milder conditions.* Built upon the famous closed-form expression of OLS estimators and its well-studied properties, we are able to extend the finite-sample and asymptotic performance guarantees for classical distributionally robust optimization problems from [40] to the feature-based case we explore. Sufficient conditions for such performance guarantees are developed explicitly to guide the choice of parameters in our approach and to achieve high-probability optimality gap bounds. As a further matter, our theoretical analysis holds under much milder conditions than assuming the availability of i.i.d. samples. Moreover, assumption imposed on the random noise term is also much weaker. For instance, [7] requires a continuous density function of the random noise defined on known bounded domain, while we only require that the random noise has a light-tailed distribution.

- *A polynomial-time solvable reformulation.* Besides providing an ambiguity set of candidate distributions that are more reasonable than those resulting from other popular choices, the adaption of Wasserstein distance plays another important role, which facilitates us to set up a tractable reformulation of our distributionally robust optimization problem via duality. In fact, due to the special structure of quantile loss (piece-wise linear), our reformulation can be solved easily with a single iteration of linear regression and then sorting. Moreover, although our original formulation depends on the realization of features, the solution is in fact feature-independent. Thus, our proposed method can better enjoy good practicality in many real-world applications, and is efficient even when the problem is of large scale.

3.3 Formulation and Preliminaries

3.3.1 Problem Formulation

We consider the demand follows a linear model:

$$d = \beta_0^T x + \varepsilon, \tag{3.1}$$

where d is the random demand of interest, $x \in \mathcal{R}^p$ is a vector of explanatory variables (containing features such as price, material, color and etc.) and ε is random innovation. As in Chapter 2, ε is assumed to be independently and identically distributed (i.i.d.) among realizations with mean 0 and unknown variance σ^2, but the common distribution $\mathbb{P}_\varepsilon$ is not necessarily normal. We further assume that ε is independent of x. We denote the observed feature, demand tuples by $\{(x_i, d_i)\}_{i=1}^N$ where i is the index for data points.

As commonly seen in the linear regression, there are two interpretations regarding the nature of the feature vector X. One assumes that x is a random effect and we can observe an i.i.d. sample path of it via data, the other assumes that we will have a fixed design matrix, e.g. values of x_i are deterministic as given. When the first assumption is applied, it follows that the realizations of d are also i.i.d. and all former analysis from [40, 1, 34] follows directly. In this study, we approach with the second assumption, and the goal is to find the conditional τth quantile, $\tau \in (0, 1)$, of D given observed $x = c$ and historical observations $\{(x_i, d_i)\}_{i=1}^N$. Following the linear model of demand process, the conditional quantile of demand is also a linear combination of these features, denoted by $Q_\tau(d|x = c)$:

$$Q_\tau(d|x = c) = \beta_0^T c + s_\tau, \tag{3.2}$$

where s_τ is the τth quantile of ε. However, both values of β_0 and s_τ are unknown in practice. Considering the classical newsvendor problem as setup described in Section 2.4, it follows that the optimal ordering quantity is exactly the conditional quantile of demand with $\tau = \frac{b-c}{h+b}$.

When there is no intercept term in x, equation (3.2) is identifiable, it follows from the definition of quantile that β_0 and s_τ are the unique solution to the stochastic optimization problem:

$$\begin{aligned} \beta_0, s_\tau &= \arg\min_{\beta, s} \mathbb{E}_{d|x=c}[\rho_\tau(d - \beta^T c - s)] \\ &= \arg\min_{\beta, s} \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\beta_0^T c + \varepsilon - \beta^T c - s)], \end{aligned} \tag{3.3}$$

and

$$s_\tau = \arg\min_{s} \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s)], \tag{3.4}$$

where $\rho_\tau(\cdot)$ is the same check function defined in (2.14), and it imposes asymmetric costs to overestimation and underestimation. And when demand indeed follows the (3.1), equations

(3.3) and (3.4) hold for any value of c. i.e.

$$\begin{aligned}\beta_0, s_\tau &= \arg\min_{\beta,s} \mathbb{E}_{d|x}[\rho_\tau(d-\beta^T x - s)], \forall x \in \mathcal{R}^p \\ &= \arg\min_{\beta,s} \mathbb{E}[\rho_\tau(\beta_0^T x + \varepsilon - \beta^T x - s)], \forall x \in \mathcal{R}^p,\end{aligned} \tag{3.5}$$

However, conditional probability of $d|x = c$ is not available in practice, and the uncertainty comes from two resources, i.e. the undiscovered value of β_0 and the unknown distribution of ε. A common treatment is to replace the objective function in (3.5) by the total quantile loss evaluated on training data as (3.6), which is recognized as quantile regression [72].

$$\hat{\beta}_{SAA}, \hat{s}_{SAA} = \arg\min_{\beta,s} \sum_{i=1}^{i=N} \rho_\tau(d_i - \beta^T x_i - s). \tag{3.6}$$

Under the random x setting, SAA is essentially equivalent to replacing the real joint distribution of (x, d) by the empirical probability, which converges to the real as sample size grows large. However, in the deterministic design matrix setting we try to explore, this converging empirical distribution property no longer holds as our data points are not i.i.d. Moreover, regardless of the popularity and practical success of SAA, it suffers from the criticism of overfitting. To obtain a robust solution with out-of-sample performance guarantees, and to incorporate the feature information into consideration, we consider the following robust optimization formulation:

$$\hat{\beta}_{RO}, \hat{s}_{RO} = \arg\min_{\beta,s} \sup_{d \in \Xi(c)} \rho_\tau(d - \beta^T c - s). \tag{3.7}$$

Since demand d is a function of c, it follows that when an ambiguity set of all possible demand values is constructed, not only should we gather the information from historical data $\{(x_i, d_i)\}_{i=1}^N$, but also account for the new value of feature c in decision making. Thus, we have $\Xi(c)$ depend on c, which is also random considering the randomness of $\{y_i\}_{i=1}^N$. Meanwhile, as argued above, such robust formulations are many a time too conservative and may overemphasize some rare extreme cases. Thus, distributionally robust optimization (DRO) alternatives, which hedge against a chosen set of candidate distributions, are becoming more and more popular:

$$\hat{\beta}_{DRO}, \hat{s}_{DRO} = \arg\min_{\beta,s} \sup_{\mathbb{Q}_d \in \mathcal{P}_d} \mathbb{E}_{\mathbb{Q}_d}[\rho_\tau(d - \beta^T c - s)], \tag{3.8}$$

where $\mathbb{Q}_d$ is a candidate conditional distribution of d given $x = c$ and $\mathcal{P}_d$ is an ambiguity set constructed via observing data $\{(x_i, d_i)\}_{i=1}^N$. Similarly, $\mathcal{P}_d$ also depends on the feature value of interest, i.e. $x = c$. Equivalently, we can rewrite the formulation to expectation with respect to ε as we assume that it is an i.i.d. process and is independent of x.

$$\hat{\beta}_{DRO}, \hat{s}_{DRO} = \arg\min_{\beta,s} \sup_{\mathbb{Q}_\varepsilon \in \mathcal{P}_\varepsilon} \mathbb{E}_{\mathbb{Q}_\varepsilon}[\rho_\tau(\beta_0^T c + \varepsilon - \beta^T c - s)], \tag{3.9}$$

where $\mathbb{Q}_\varepsilon$ and $\mathcal{P}_\varepsilon$ are the candidate distribution and ambiguity set of ε respectively. As reviewed in Section 3.2, there are various criteria which can be used to construct $\mathcal{P}_\varepsilon$. In this research, we begin with the case where Wasserstein distance is chosen.

Compared with traditional random interpretation of x, there arises a major difficulty in our deterministic design matrix formulation. Our i.i.d. process ε is not explicitly observable, instead we can only obtain realizations of $\{d_i\}_{i=1}^N$ where $d_i = \beta_0^T x_i + \epsilon_i$ and the value of β_0 is also unknown. To decouple the two sources of unknownness, we introduce a new notation, $\varepsilon(\beta) = \varepsilon + (\beta_0 - \beta)^T c$. Its distribution is the same as ε with mean shifted by $(\beta_0 - \beta)^T c$. Then, the distributionally robust optimization formulation (3.9) can be rewritten as

$$\hat{\beta}_{DRO}, \hat{s}_{DRO} = \min_{\beta,s} \sup_{\mathbb{Q}_{\varepsilon(\beta)} \in \mathcal{P}(\beta)} \mathbb{E}_{\mathbb{Q}_{\varepsilon(\beta)}}[\rho_\tau(\varepsilon(\beta) - s)], \tag{3.10}$$

where $\mathbb{Q}_{\varepsilon(\beta)}$ and $\mathcal{P}(\beta)$ are the corresponding candidate distribution and ambiguity set of $\varepsilon(\beta)$ respectively. And we detect that our ambiguity set should not only rely on the data points and value of c, but also the value of β, which in turn need to be determined via the optimization. However, with the intention to employ Wasserstein distance for constructing an ambiguity set and achieve convergence, the center of the ball depends on the value of β. Consequently, as we will discuss more in next section, the choice of a good radius r_N also depends on β. Hence, it's very difficult to determine the optimal value of β simultaneously.

To simplify the problem described in (3.10) and obtain a tractable solution, we decide to proceed as a two-step framework by providing a good guess of β_0 first and then solving an approximated problem. Then, for any guess of β, we can construct a sample path $\{(\epsilon_i(\beta) = d_i - \beta^T x_i)\}_{i=1}^N$, and we consider the case where the ambiguity set is constructed as Wasserstein ball around the empirical distribution, i.e. $\mathcal{P}(\beta) := \mathbb{B}_{r_N}\big(\hat{\mathbb{P}}_N(\beta)\big)$ with $\hat{\mathbb{P}}_N(\beta) = \frac{1}{N}\sum_{i=1}^N \delta(\epsilon_i(\beta))$, where $\delta(\epsilon(\beta))$ denotes the unit mass on $\epsilon(\beta)$. In this chapter, we start with using the least squared estimator $\hat{\beta}_{OLS}^N$ from classical linear regression as the value for estimating β_0. It is well-known that $\hat{\beta}_{OLS}^N$ is a consistent estimator of β_0, and has a nice closed-form solution with well-studied properties. It is well-known that

$$\hat{\beta}_{OLS}^N = (X^T X)^{-1} X^T Y, \tag{3.11}$$

where $X \in \mathcal{R}^{N \times p}$ is the design matrix with x_i being its ith row and $Y \in \mathcal{R}^N$ is a column vector storing d_is.

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, its corresponding approximated sample path is $\{\epsilon_i^{OLS}\}_{i=1}^N$ with $\epsilon_i^{OLS} = d_i - (\hat{\beta}_{OLS}^N)^T x_i$ with an approximated empirical distribution $\hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N) = \frac{1}{N}\sum_{i=1}^N \delta(\epsilon_i^{OLS})$, resulting in a Wasserstein ball of $\mathcal{P}(\hat{\beta}_{\mathcal{OLS}}^{\mathcal{N}}) := \mathbb{B}_{r_N}\big(\hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N)\big)$. Note that ϵ_i^{OLS} is the linear regression residual of the ith data point, but not a real realization of the random variable ε'. Thus, compared with the current work of [40, 1], the ambiguity set of our formulation is no longer centered at an empirical distribution of the random variable. Finally, we aim at solving the following distributional robust optimization problem:

$$\min_{s} \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)], \tag{3.12}$$

with $\mathcal{P} := \mathbb{B}_{r_N}\big(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})\big)$, and let $\hat{s}_N$ denotes the optimal solution of (3.12).

To understand how a Wasserstein ball looks like, let us introduce the following definition from [1] with mildly modification to suit for our problem setting:

Definition 3.1. *Let $\mathcal{M}(\Xi^2)$ denote the set of probability distributions on $\Xi \times \Xi$. The Wasserstein distance between two distributions $\mathbb{P}$ and $\mathbb{Q}$ supported on Ξ is defined as*

$$d_W(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi \in \mathcal{M}(\Xi^2)} \big\{ \int_{\Xi^2} d(\xi, \xi') \Pi(\mathrm{d}\xi, \mathrm{d}\xi') : \Pi(\mathrm{d}\xi, \Xi) = \mathbb{Q}(\mathrm{d}\xi), \Pi(\Xi, \mathrm{d}\xi') = \mathbb{P}(\mathrm{d}\xi') \big\}, \tag{3.13}$$

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

We remark that there is a generalized p-Wasserstein metric defined on the kth moment for some distance measure for $k \geq 1$. In this chapter, we exclusively focus on the 1-Wasserstein distance as given in Definition 3.1 and pick $d(\xi, \xi')$ to be the $l2$ (Euclidean) norm in our analysis. In fact, since our analysis focus on a scalar random variable ε', and any choice of $d(\xi, \xi')$ will eventually become the absolute difference and does not affect the results.

Equivalently, the Wasserstein metric can also be defined in a dual representation.

Lemma 3.1. *([34], Theorem 3.2) For any distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\Xi)$ we have*

$$d_W(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{L}} \big\{ \int_{\Xi} f(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} f(\xi) \mathbb{Q}(\mathrm{d}\xi) \big\}, \tag{3.14}$$

where $\mathcal{L}$ denotes the space of all Lipschitz functions with $|f(\xi) - f(\xi')| \leq \|\xi - \xi'\|$ for all $\xi, \xi' \in \Xi$.

3.3.2 Preliminaries

Since our DRO solution relies on the OLS estimators which has an elegant closed form representation (3.11), its performance also depends on how good $\hat{\beta}^N_{OLS}$ is as an proxy for β_0. Below we review the important assumptions for guaranteeing the performance of $\hat{\beta}^N_{OLS}$, together with some theoretical results. And throughout our analysis for DRO, we also make the same assumptions.

Assumption 3.1. *The design matrix X is of full rank. This ensures the invertibility of matrix $X^T X$, so that the closed-form solution (3.11) is well-defined.*

Although we are proceeding with a fixed design model where x_i's should not be regarded as i.i.d. samples, we still need to make some assumptions to permit a law of large numbers for the purpose of establishing convergent results.

Assumption 3.2. *([17]) As sample size N goes to infinity, the $p * p$ matrix*

$$M_{xx} = \lim N^{-1} X^T X = \lim N^{-1} \sum_{i=1}^{N} x_i x_i^T \tag{3.15}$$

exists and is finite nonsingular.

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

$$\begin{aligned} \lim c^T (X^T X)^{-1} c &= \lim \frac{c^T (N^{-1} X^T X)^{-1} c}{N} \\ &= \frac{c^T (\lim N^{-1} X^T X)^{-1} c}{N} \\ &= \frac{c^T M_{xx}^{-1} c}{N} \\ &\to 0. \end{aligned} \tag{3.16}$$

Moreover, with the above two assumptions holding, by applying the Markov Law of Large Numbers (Theorem A.9 from [17]), it can be proved that $\hat{\beta}_{OLS}^N$ is a (strongly) consistent estimator.

Lemma 3.2. *Under Assumptions 3.1 and 3.2, 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", aka the projection matrix

$$H = X(X^T X)^{-1} X^T. \tag{3.17}$$

Its diagonal elements h_{ii}'s are widely known as the leverage scores of data points, measuring how far away the features of each observation are from those of the other observations. And nice properties of these measures as known in the following lemma.

Lemma 3.3. *([19]) Let h_{ii} denoting the ith diagonal element of the projection matrix H defined above, where $X \in R^{n \times p}$ is the design matrix. Then, the following results hold:*

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

The trace can be interpreted as the amount of information extracted from the observations or degrees of freedom for signal [107]. Thus, the sum of leverage scores of all observations equals to the number of features. Built upon the theoretical results for OLS estimators, and with techniques from robust and distributionally robust optimization, we are able to derive the performance guarantees of our DRO solution to (3.12) in next section.

3.4 Performance Guarantees

3.4.1 Notation and Assumptions

This section establishes the theoretical performance guarantees for our distributionally robust solution. With finite sample points, we provide sufficient conditions for choosing the Wasserstein ball radius such that the out-of-sample newsvendor loss can be bounded from above with high probability. Then, we prove its asymptotic optimality by demonstrating asymptotic consistency of the resulting estimators when sample size goes to infinity. To clarify our analysis, let us firstly introduce the following notation:

- J^*: the theoretical minimal expected out-of-sample cost which we target at, i.e.

$$J^* = \min_{\beta,s} \mathbb{E}_{d|c}[\rho_\tau(d - \beta^T c - s)] = \min_{\beta,s} \mathbb{E}_{\mathbb{P}_{\varepsilon(\beta)}}[\rho_\tau(\varepsilon(\beta) - s)]. \tag{3.18}$$

- $\hat{J}_N$: the objective value achieved by our distributionally robust optimization formulation (3.12), i.e.

$$\hat{J}_N = \min_{s} \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)],\ \varepsilon' = \varepsilon + (\beta_0 - \hat{\beta}^N_{OLS})^T c. \tag{3.19}$$

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

$$J_{oos} = \mathbb{E}_{\mathbb{P}_{\varepsilon'}}[\rho_\tau(\varepsilon' - \hat{s}_N)]. \tag{3.20}$$

The feasibility of $\hat{s}_N, \hat{\beta}^N_{OLS}$ to the original problem (3.9) implies that $J^* \leq J_{oos}$. However, this lower bound is of very limited use in practice as J^* is unknown and our primary concern should be to obtain an upper bound of our costs. Considering the fact that DRO evaluates the worst-case costs, $\hat{J}_N$, which is random due to the randomness of the sample data, bounds our costs from above if and only if the real distribution $\mathbb{P}_{\varepsilon'}$ actually falls into the ambiguity set we used. Thus, we will examine the ambiguity set

$$\begin{aligned} \mathcal{P} &:= \mathbb{B}_{r_N}\big(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})\big) \\ &:= \big\{\mathbb{Q} \in \mathcal{M}(\Xi) : d_W(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}), \mathbb{Q}) \leq r_N\big\}, \end{aligned} \tag{3.21}$$

which consists of all distributions within a Wasserstein ball of radius r_N and center at $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$. Note that $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$ is not constructed with an i.i.d. sample path, and the results from [40] and [34] no longer holds. To reestablish the desired performance guarantees, we decompose the estimation error into several sources and apply the attractive properties of $\hat{\beta}^N_{OLS}$. Throughout this chapter, we make the following additional assumptions.

Assumption 3.3. *(Light-tailed assumption [[34], Assumption 3.3]) There exists an exponent* $a > 1$ *such that*

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

This light tail assumption requires the tail of the distribution of ε to decay at an exponential rate, and it implies that the variance of ε is finite. With $a > 1$ and $\|\varepsilon\| \geq 0$, we have

$$\begin{aligned} \sigma^2 - \mathbb{E}_{\mathbb{P}_\varepsilon}[exp(\|\varepsilon\|^a)] &= \mathbb{E}_{\mathbb{P}_\varepsilon}[\varepsilon^2] - \mathbb{E}_{\mathbb{P}_\varepsilon}[exp(\|\varepsilon\|^a)] \\ &\leq \mathbb{E}_{\mathbb{P}_\varepsilon}[|\varepsilon|^2] - \mathbb{E}_{\mathbb{P}_\varepsilon}[exp(|\varepsilon|)] \\ &= \mathbb{E}_{\mathbb{P}_\varepsilon}[|\varepsilon|^2 - exp(|\varepsilon|)] \\ &< 0 \end{aligned} \tag{3.22}$$

provided that $x^2 - exp(x) < 0,\ \forall x \geq 0$. Hence, we have σ^2 is bounded by A . Thus, we can assume that there is a known constant M such that $\sigma^2 \leq M$.

3.4.2 Finite-sample Performance

Theorem 3.4. *(Measure Concentration) Suppose Assumptions 3.1, 3.2 and 3.3 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.

Corollary 3.1. *(Finite-sample Performance Guarantee) It follows immediately from Theorem 3.4 that $\hat{J}_N$ upper bounds the out-of-sample cost with high probability. i.e.*

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

The proof of Theorem 3.4 will rely on the following technical lemmas, where we decompose the Wasserstein distance of the real distribution $\mathbb{P}_{\varepsilon'}$ and the approximated empirical distribution $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$ into three components,

$$d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq d_W(\mathbb{P}_\varepsilon, \hat{\mathbb{P}}_N(\beta_0)) + d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) + d_W(\mathbb{P}_{\varepsilon'}, \mathbb{P}_\varepsilon), \tag{3.23}$$

by triangular inequality. And we propose sufficient conditions to bound each of these component separately.

When β takes it real value β_0, $\hat{\mathbb{P}}_N(\beta_0)$ boil down to an empirical distribution of ε based on i.i.d. samples. Thus, the result from [34] holds.

Lemma 3.5. *([34]) If Assumption 3.3 holds, then for any $\eta' \in (0,1)$, we can set r^1_N such that*

$$\mathbb{P}^N_\varepsilon\{d_W(\mathbb{P}_\varepsilon, \hat{\mathbb{P}}_N(\beta_0)) \geq r^1_N\} \leq \eta' \tag{3.24}$$

for all $N \geq 1$, where

$$r^1_N(\eta') := \begin{cases} (\frac{\log(c_1\eta'^{-1})}{c_2 N})^{\frac{1}{2}} & \text{if } N \geq \frac{log(c_1\eta'^{-1})}{c_2} \\ (\frac{\log(c_1\eta'^{-1})}{c_2 N})^{\frac{1}{a}} & \text{if } N < \frac{log(c_1\eta'^{-1})}{c_2} \end{cases} \tag{3.25}$$

where c_1, c_2 are positive constant that only depend on a and A (see [37] Theorem 2 for details regarding choice of c_1 and c_2). And note that r^1_N chosen according to (3.25) converges to 0 as $N \to 0$.

We further investigate the distance between a real empirical distribution $\hat{\mathbb{P}}_N(\beta_0)$ and the approximation with linear regression residuals $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$. By using the closed form solution, we can also bound it with certain probability.

Lemma 3.6. *If Assumption 3.1, 3.2 and 3.3 holds, then we can set* $r_N^2 = \sqrt{\frac{pM}{N\eta'}}$*, such that*

$$\mathbb{P}_\epsilon\{d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \geq r_N^2\} \leq \eta' \tag{3.26}$$

for all $\eta' \in (0,1)$*, and that* $r_N^2 \to 0$ *as* $N \to \infty$*.*

The proof of Lemma 3.6 is built upon the property of the $\hat{\beta}^N_{OLS}$ and can be found in Appendix B. Finally, it remains to bound the distance between distributions of ε and $\varepsilon' = \varepsilon + (\beta_0 - \hat{\beta}^N_{OLS})^T c$, where the latter one is just the first shifted by a constant. As a result, we have

Lemma 3.7. *If Assumption 3.3 holds, then for any* $\eta' \in (0,1)$*, we can set*

$$r_N^3 = \sqrt{\frac{Mc^T(X^TX)^{-1}c}{\eta'}}$$

so that

$$\mathbb{P}_\epsilon\{d_W(\hat{\mathbb{P}}_\varepsilon, \mathbb{P}_{\varepsilon'}) \geq r_N^3\} \leq \eta'. \tag{3.27}$$

And $r_N^3 \to 0$ as $N \to \infty$.

Eventually, it therefore follows from Lemma 3.5, 3.6 and 3.7 that, for any $\eta \in (0,1)$, we can pick $\eta' = \frac{\eta}{3}$ and select r_N^1, r_N^2 and r_N^3 according to the above derived conditions. Then, by setting $r_N = r_N^1 + r_N^2 + r_N^3$, and using the decomposition in (3.23), we will have

$$\begin{aligned}
\mathbb{P}^N(\mathbb{P}_{\varepsilon'} \in \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}))) =& \mathbb{P}^N(d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq r_N) \\
\geq& 1 - \mathbb{P}(d_W(\mathbb{P}_\epsilon, \hat{\mathbb{P}}_N(\beta_0)) > r_N^1) - \mathbb{P}(d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) > r_N^2) \\
& - \mathbb{P}(d_W(\mathbb{P}_{\epsilon'}, \mathbb{P}_\epsilon) > r_N^3) \\
\geq& 1 - 3\eta' \\
=& 1 - \eta.
\end{aligned} \tag{3.28}$$

In this manner, we complete the proof of the claims of Theorem 3.4 and Corollary 3.1.

3.4.3 Asymptotic Performance

In addition to bound the out-of-sample performance with finite sample points, it is also of interest to know how our method behaves as the sample size increases. Moreover, although we demonstrated that $\hat{J}_N$ can upper bound the best expected out-of-sample cost J^* with any desired probability, we have not been able to explicitly describe the gap between the two values. From this point onward, we study the asymptotic behavior or the DRO solution and show that the gap between $\hat{J}_N$ and J^* asymptotically converges to zero.

Theorem 3.8. *(Asymptotic Consistency) Suppose that Assumption 3.3 holds and that a sequence* $\eta_N \in (0,1)$ *satisfies* $\sum_{N=1}^{\infty} \eta_N < \infty$ *and* $\lim_{N\to\infty} r_N(\eta_N) = 0$.[1] *Then*

a) $\hat{J}_N \downarrow J^*$ *as* $N \to \infty$ *where* J^* *is the optimal value of (3.18);*

b) Any accumulation point of $\{(\hat{s}_N, \hat{\beta}^N_{OLS})\}$ *is* $\mathbb{P}^\infty$ *-almost surely one of the optimal solution for (3.18).*

The proof of Theorem 3.8 can be found in Appendix B. It relies on the following lemma where we first show the convergence of any sequence of distributions that comes from the corresponding sequence of Wasserstein balls. This result is essentially the same as Lemma 3.7 in [34].

Lemma 3.9. *(Convergence of distribution, Lemma 3.7 of [34]) Assume 3.3 holds and a sequence* $\eta_N \in (0,1)$ *that satisfies* $\sum_{N=1}^{\infty} \eta_N < \infty$ *and* $\lim_{N\to\infty} r_N(\eta_N) = 0$. *Then for any sequence* $\mathbb{Q}_N \in \mathbb{B}_{r_N(\eta_N)}(\mathbb{P}_N(\hat{\beta}^N_{OLS})), N \in \mathbb{N}$, *it converges under Wasserstein metric to* $\mathbb{P}_{\varepsilon'}$ *almost surely with respect to* $\mathbb{P}^\infty_\varepsilon$, *that is*

$$\mathbb{P}^\infty\{\lim_{N\to\infty} d_W(\mathbb{P}_{\varepsilon'}, \mathbb{Q}_N) = 0\} = 1.$$

3.5 Tractable Reformulation

In the previous sections of this chapter, we have proposed a DRO approach for solving the newsvendor problem with covariate information, which attains both finite-sample and asymptotic performance guarantees. However, formulation (3.12) consists of a seemly intractable optimization problem with infinite-dimensions due to the candidate probability distributions. To assure this approach's practicality, we further demonstrate that it can be re-represented as a tractable finite-dimensional convex program by following similar argument from [34].

Theorem 3.10. *For any* $r_N \geq 0$, *(3.12) can be reformulated as*

$$\min_{s,\lambda\geq\max\{\tau,1-\tau\}} \lambda r_N + \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(\epsilon_i^{OLS} - s) \tag{3.29}$$

The above theorem provides a convex program reformulation of (3.12) with finite number of summands, thus is tractable. The proof of this claim can be found in Appendix B. Surprisingly, the two components in the reformulation are actually decoupled, which leads to a simple closed-form solution for $\lambda^* = \max\{\tau, 1-\tau\}$ and $\hat{s}_N = \arg\min_s \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(\epsilon_i^{OLS} - s)$. Thus, $\hat{s}_N$ is essentially the sample τth quantile of the sequence $\{\epsilon_i^{OLS}\}_{i=1}^N$, and can be obtained

[1] A possible choice is $\eta_N = \exp\{-\sqrt{N}\}$.

in polynomial time by simply sorting the linear regression residuals. And our DRO solution $(\hat{\beta}_N^{OLS}, \hat{s}_N)$ coincides with one of the currently popular procedures for quantile estimation in linear models. And it follows that the asymptotic consistency of our DRO solution can be also shown by proving the convergence of the sample quantile of the linear regression residuals.

This observation is rather counterintuitive at first glance as the value of distributionally robust decision $\hat{s}_N$ is irrelevant to the radius r_N of the ambiguity set. This is because in the quantile regression case, our loss function $\rho_\tau(\cdot)$ is of simple piece-wise linear form and can be regarded as the maximum of two linear function, and our distance metric defined on the space of ε' also reduces to the absolute difference as it is a scalar variable. Thus, the worst-case scenario is always attained at the tuning points as seen in the dual representation (see Appendix B for more discussion). On the other hand, this is not to suggest that r_N and the previous DRO discussion is useless. As argued earlier, in practice we not only want an easily accessible solution but also want it to be robust with measurable performance. With finite data points, the conditions we derived for r_N allow us to obtain an upper bound for the out-of-sample cost with any desired probability.

3.6 Numerical Experiments

We validate the theoretical results of this chapter in the context of bike sharing. Specifically, our goal is to forecast the aggregated bike demand during a certain hour based on external information such as temperature, wind speed and etc. Accurate prediction of quantiles of the random demand process helps us to capture its distribution and thus, will facilitate better inventory management of bikes in the system.

Data used in our experiments is simulated based on a real dataset from [35]. By selecting all observations that are recorded on weekdays from their training set, we obtain a dataset with $7,412$ instances. We further preprocess the dataset by extracting features including hour of the day, day of the week and etc. Then, we create dummy variables for the categorical features, and remove some features by hypothesis testing to get rid of the multicollinearity issue. Eventually, we obtain a design matrix with 30 useful features. To check the linear relationship in the nature of this problem, we conducted a linear regression on 70% of the dataset and test its performance on the remaining 30%. This model achieves R^2 and OSR^2 of 0.8389 and 0.8340 respectively, indicating good fitness of the linear model.

In order to generate the dataset which satisfies all assumptions required by our DRO method for the numerical experiments, the coefficients of the aforementioned linear model are recorded as β_0, the coefficients for the underlying true linear relationship. Finally, values of the dependent variable, demand, are generated by adding simulated noise, i.e.,

$$d_i = \beta_0^T x_i + \varepsilon_i. \tag{3.30}$$

We considered two types of noises - Gaussian and uniformly distributed, both of which satisfies our light-tailed assumption. Different values of variance $\sigma = 0.2, 2, 20$ and different

quantiles $\tau = 0.3, 0.5, 0.7$ are tested respectively. Since our data point are generated from known ε_i distributions, we are able to evaluate the real optimal solutions. Furthermore, we also use the SAA method from [96, 7] as a benchmark algorithm.

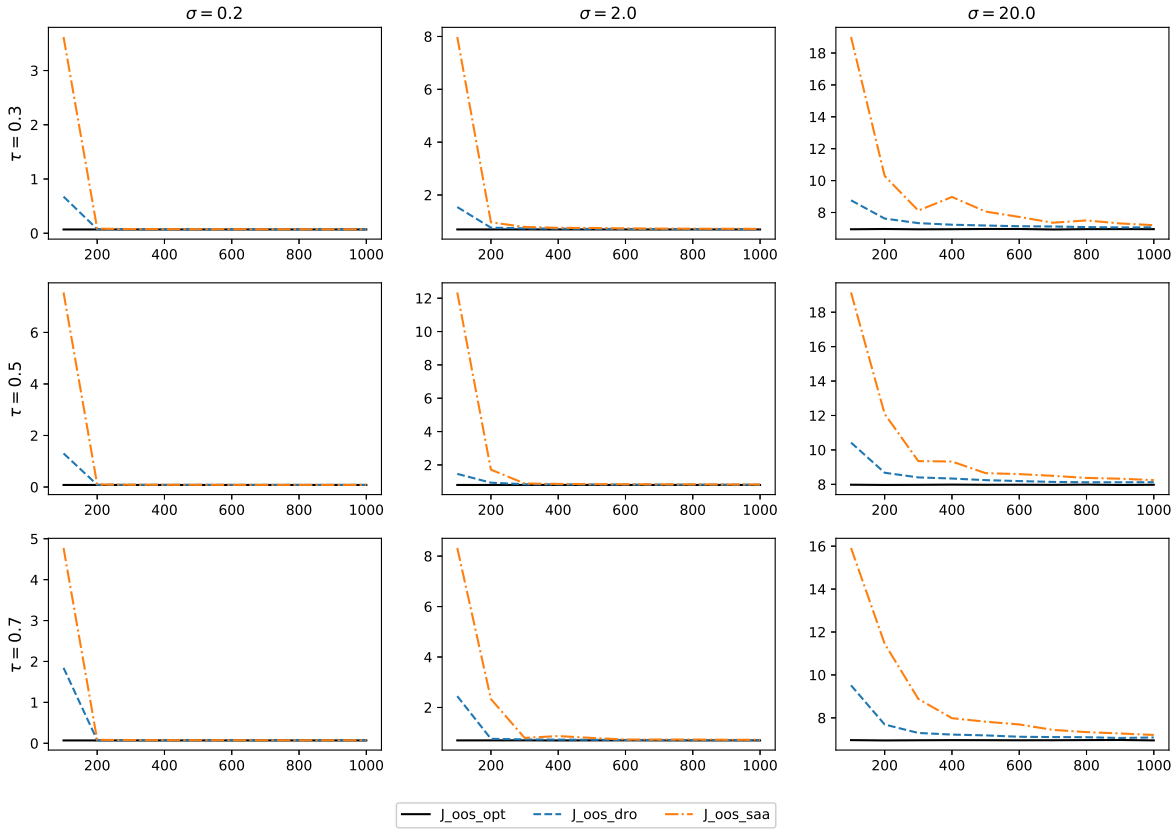


Figure 3.1: Comparison of optimal, SAA and DRO approaches with Gaussian noises.

Among the $7,412$ observations of covariates provided by the biking sharing dataset in total, 112 of them are randomly selected to serve as the covariates for testing purpose. For each configuration of noise and τ values, our experiment comprises 50 simulation runs. In each run we randomly select N, ranging from 100 to $7,200$, feature vectors from the training covariate set. Demand values for these N training points and the reserved 112 test feature vectors are generated according to (3.30). We calibrate the SAA and our DRO solutions to the N training points and evaluate the average performance of each method on the 112 test instances. Moreover, we calculate the average out-of-sample costs of each method across all simulation runs, and denote them by J_oos_opt, J_oos_saa and J_oos_dro respectively.

Figure 3.1 and Figure 3.2 visualizes how the out-of-sample costs of different strategies decreases as more data becomes available under various σ and τ values for Gaussian and uniform noises respectively. All subplots indicate that our DRO approach achieves lower out-of-sample cost compared with the SAA method under all scenarios, and the advantage is more significant especially when the training sample size is smaller. Moreover, as the

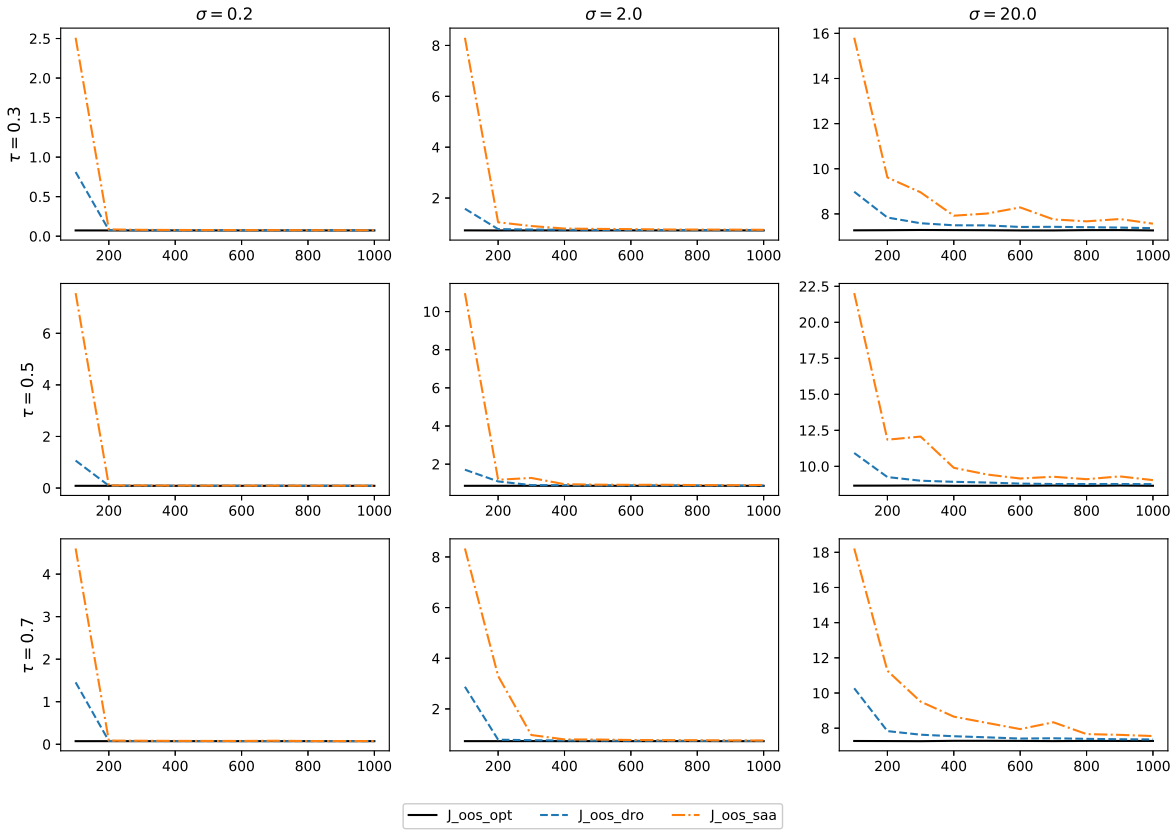


Figure 3.2: Comparison of optimal, SAA and DRO approaches with Uniform noises.

variance becomes larger, our DRO strategy dominates the SAA approach significantly even with moderate sample size. This observation is consistent with the results demonstrated in Table 3.1, where we quantify the average out-of-sample costs for all experiment settings. The percentage values within the parentheses correspond to the improvement of our DRO strategy compared with the SAA approach, calculated as $\frac{J\_oos\_dro - J\_oos\_saa}{J\_oos\_saa}$. With all of these values being significantly negative, it indicates that the DRO policy achieves better quantile prediction.

Moreover, even when the sample size become larger than 200 and the curves for SAA and DRO almost overlap with each other under small variance scenarios in Figure 3.1 and Figure 3.2, it can be shown that our DRO solution still dominates. For example, in Figure 3.3 and Figure 3.4, we plot the tails of cost curves for the $\tau = 0.3$ and $\sigma = 0.2$ case with sample sizes ranging from 400 to $7,200$. It is clearly seen that the DRO solution consistently achieves lower out-of-sample quantile loss. In fact, our empirical studies also provide evidence that the DRO performance among the 50 simulation runs achieves smaller variances compared with the performance of the SAA policy across all runs. Although the gap between the two policies seems to be small in terms of the average quantile loss when sample size is large enough, the real difference can be significant when we scale it back to the original newsvendor cost,

Table 3.1: Comparison of average out-of-sample costs.

Noise	σ	N	*J_oos_opt*			*J_oos_saa*			*J_oos_dro*					
			$\tau = 0.3$	0.5	0.7	0.3	0.5	0.7	0.3		0.5		0.7	
Gaussian	0.2	100	0.070	0.080	0.070	3.618	7.551	4.775	0.675	(−81.3%)	1.305	(−82.7%)	1.844	(−61.4%)
		500	0.070	0.080	0.070	0.074	0.084	0.074	0.072	(−2.8%)	0.082	(−2.5%)	0.072	(−2.8%)
		1000	0.070	0.080	0.070	0.072	0.082	0.072	0.071	(−1.2%)	0.081	(−1.1%)	0.071	(−1.3%)
	2.0	100	0.696	0.798	0.695	7.981	12.344	8.323	1.542	(−80.7%)	1.462	(−88.2%)	2.448	(−70.6%)
		500	0.696	0.797	0.696	0.751	0.846	0.792	0.717	(−4.5%)	0.824	(−2.6%)	0.720	(−9.1%)
		1000	0.696	0.798	0.694	0.717	0.826	0.720	0.705	(−1.8%)	0.810	(−1.9%)	0.708	(−1.6%)
	20.0	100	6.947	7.979	6.963	19.005	19.146	15.917	8.764	(−53.9%)	10.426	(−45.5%)	9.515	(−40.2%)
		500	6.965	7.975	6.959	8.058	8.596	7.821	7.183	(−10.9%)	8.245	(−4.7%)	7.184	(−8.1%)
		1000	6.956	7.976	6.950	7.204	8.247	7.203	7.068	(−1.9%)	8.125	(−1.5%)	7.086	(−1.6%)
Uniform	0.2	100	0.073	0.087	0.072	2.510	7.558	4.603	0.811	(−67.7%)	1.065	(−85.9%)	1.456	(−68.4%)
		500	0.073	0.087	0.073	0.077	0.093	0.078	0.075	(−3.4%)	0.089	(−5.3%)	0.075	(−4.3%)
		1000	0.073	0.087	0.073	0.075	0.089	0.075	0.074	(−1.8%)	0.087	(−1.9%)	0.074	(−2.1%)
	2.0	100	0.728	0.865	0.728	8.303	10.967	8.335	1.580	(−81.0%)	1.710	(−84.4%)	2.875	(−65.5%)
		500	0.728	0.867	0.729	0.787	0.924	0.791	0.746	(−5.1%)	0.886	(−4.2%)	0.748	(−5.5%)
		1000	0.727	0.865	0.727	0.752	0.910	0.753	0.737	(−2.0%)	0.874	(−4.0%)	0.737	(−2.1%)
	20.0	100	7.272	8.658	7.272	15.806	22.028	18.218	8.982	(−43.2%)	10.925	(−50.4%)	10.270	(−43.6%)
		500	7.277	8.648	7.281	8.016	9.433	8.300	7.491	(−6.5%)	8.876	(−5.9%)	7.481	(−9.9%)
		1000	7.268	8.650	7.275	7.566	9.042	7.554	7.368	(−2.6%)	8.764	(−3.1%)	7.360	(−2.6%)

which is the real metric that we care about. In addition, the newsvendor cost differences can accumulate along different sales periods. Thus, we believe that an inventory policy following our DRO approach can potentially reduce the newsvendor costs significantly, compared with the current data-driven method based on SAA, in real-world applications.

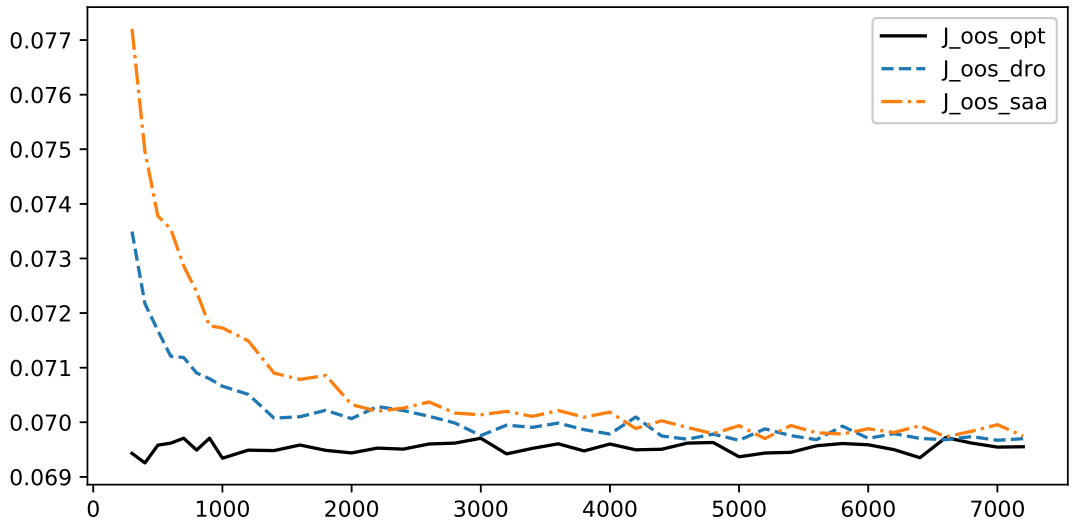


Figure 3.3: Comparison of optimal, SAA and DRO approaches with Gaussian noises, $\tau = 0.3$ and $\sigma = 0.2$.

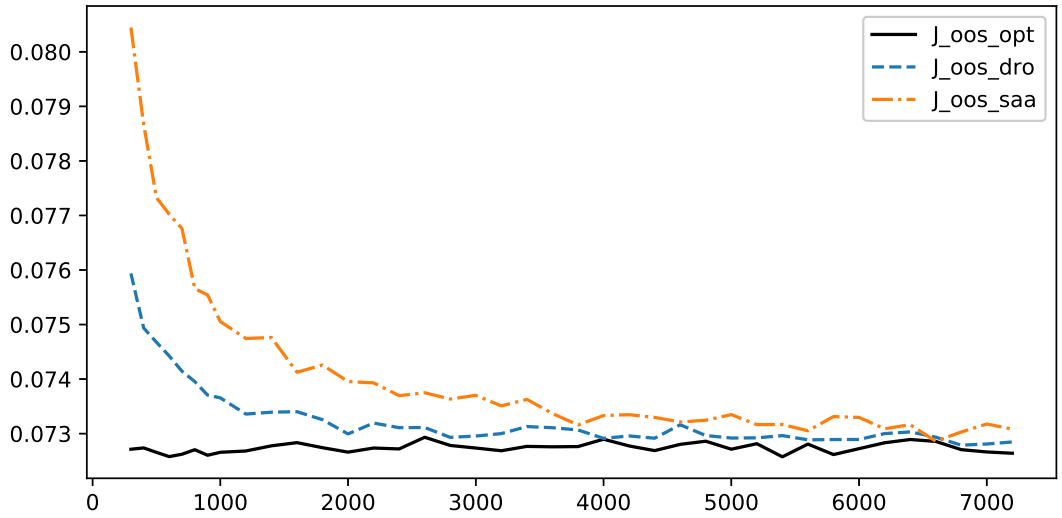


Figure 3.4: Comparison of optimal, SAA and DRO approaches with Uniform noises, $\tau = 0.3$ and $\sigma = 0.2$.

3.7 Conclusion

In this chapter, we considered the classical newsvendor problem where the demand is a linear function of some covariate variables plus i.i.d random noises. At the beginning of each sale period, the decision maker observes the values of the covariates first and has to decide the stocking level in order to minimize expected cost for overstocking or loss of sales. Provided that the optimal solution to newsvendor problem conforms to the well-known closed-form formula as a critical quantile, and that the expected costs are proportional to the corresponding expected quantile loss, this problem essentially boils down to quantile prediction with observable covariates information.

We explored the problem under a data-driven environment with the goal of developing a robust solution. In contrast to current literature which admits a random design of the feature matrix, and tries to minimize the expected cost with respect to random realization of the covariates, we proceed with a fixed design model. We reckon that this fixed feature interpretation better fits the real-life applications such as deciding inventory level of fashion goods, as features such as style and material follows trend over time and should not be regarded as i.i.d. samples and they are known before inventory decision has to be made and thus are not random in nature. Then, by leveraging the OLS estimators for linear coefficients and the recently developed distributionally robust optimization tools with Wasserstein metric, we proposed a two-step distributionally robust approach to the problem of interest.

Built upon the properties of the OLS estimators and the structure of Wasserstein distance, we bounded the expected our-of-sample cost with any desired probability and also demonstrated that the solution is asymptotic optimal. Moreover, our performance guarantees hold

under milder conditions compared with current literature [7]. Finally, by applying duality theories, we represented the original infinite-dimensional DRO formulation in a tractable equivalent reformulation. It is worth noting that, due to the special structure of quantile loss, this reformulation is polynomial solvable with a single iteration of linear regression followed by sorting.

Meanwhile, we will extend our discussion to algorithms that are suitable for more large-scale data-driven problems in Chapter 4 and more potential future research directions in Chapter 5.

Chapter 4

An Algorithm for Large-Scale Convex Optimization Problems with Linear Constraints

4.1 Introduction

In the previous two chapter, we delved into data-driven approaches for inventory management problems from two different perspectives. One attempts to provide a more flexible way for modeling time series demand process with less assumptions, and the other seeks to develop robust solution with performance guarantees. Both result in minimizing certain objective functions, (2.33) and (3.29) respectively, which are evaluated on historical data points. However, the first approach sacrifices theoretical performance guarantee and the latter requires a restrictive linear demand model. Thus, one naturally potential extension to incorporate these two ideas is to consider distributionally robust decisions under a demand process that is more flexible than linear models. One the other hand, given another demand model, the resulting reformulation of the DRO problem will not as easy to solve as (3.29). For instance, let us first consider the easy case where the cost function is still convex and that a random design with i.i.d. samples is assumed. Then, all analysis for DRO with Wasserstein metric in previous literature holds and leads to a convex finite-dimensional reformulation as shown in Theorem 4.2 of [34]. The number of variables and constraints both grow linearly as the number of data points increases, and soon falls in the field of big-data machine learning and large-scale data-driven problems. To enhance the application of such models with the ever-increasing data size, it is desired that distributed algorithms can be designed to take advantages of the fast growing computation infrastructures.

One group of formulations that are particularly popular in this stream of work belong to large-scale convex optimization problems with linear constraints, where the decision variables satisfy a multi-block structure. And our goal of this chapter is to explore efficient distributed algorithms for such a problem.

A general form can be expressed as the following problem (P):

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{m} \theta_i(x_i) \\ \text{s.t.} \quad & \sum_{i=1}^{m} A_i x_i = b \\ & x_i \in \chi_i,\ i = 1, \ldots, m, \end{aligned} \tag{4.1}$$

where $\chi_i \subseteq R^{n_i},\ i = 1, \ldots, m\ (m \geq 2)$ are closed convex sets, $A_i \in R^{l*n_i}$, $b \in R^l$ and $\theta_i(x_i) : R^{n_i} \to R, i = 1, \ldots, m$ are closed proper convex functions (not necessarily smooth). And we further denote $n = \sum_{i=1}^{m} n_i$, $A = [A_1\ A_2\ \ldots\ A_m]$, $x = [x_1^T\ x_2^T\ \ldots x_m^T]^T$and $\phi(x) := \sum_{i=1}^{m} \theta_i(x_i)$. The DRO reformulation can be rewritten in this multi-block form by creating duplicated variables as did in [85, 44].

Then, we obtain the corresponding Lagrangian function as

$$L(x, \lambda) = \phi(x) - \lambda^T(Ax - b), \tag{4.2}$$

where $\lambda \in R^l$ is the Lagrangian multiplier and let $\Omega := \chi_1 \times \chi_2 \times \ldots \times \chi_m \times R^l$. Note that the Lagrangian relaxation problem,

$$\min_{x} L(x, \lambda) = \min_{x_1, x_2, \ldots, x_m} L(x_1, x_2, \ldots, x_m, \lambda), \tag{4.3}$$

is decomposable with respect to x_i's. And throughout this chapter, we make the following assumptions.

Assumption 4.1. *The solution set of (P), Ω^*, is nonempty.*

Assumption 4.2. *There exists a saddle point $\omega^* = (x^*, \lambda^*)^T \in \Omega^*$ to the problem (P). That is, there exists ω^* such that*

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*),\ \forall x \in R^n, \forall \lambda \in R^l, \tag{4.4}$$

and

$$-A_i^T \lambda^* \in \partial\theta_i(x_i^*),\ \textit{for } i = 1, 2, \ldots, m, \tag{4.5}$$

$$Ax^* = \sum_{i=1}^{m} A_i x_i^* = b, \tag{4.6}$$

where $\partial\theta_i(x)$ denotes the subdifferentials of θ_i at x.

With the arising popularity of machine learning and other data-driven large-scale convex optimization problems, this formation has found wide applications in different fields (see, e.g., [86, 8, 4, 81, 14, 44, 85] etc.). Thus, it is of great interest and value that a decomposition scheme is available where the properties of θ_i's can be exploited individually. Among the

numerous splitting methods designed for (4.1), dual ascent is a classical idea that leads to decentralized algorithm, and is often referred to as dual decomposition. At each iteration, a Lagrangian relaxation problem (4.3), which is decomposable across x_i's and can be solved in parallel, is solved for updating the primal variables; then, a dual ascent update for the Lagrangian multiplier is performed (see Chapter 2 of [14] for a review). To bring robustness to the dual decomposition method and to ensure convergence without assumptions on the strict convexity or finiteness of the objective function, augmented Lagrangian methods were developed with an additional quadratic penalty term add to the Lagrangian function, i.e.

$$L_\rho(x, \lambda) = \phi(x) - \lambda^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|^2. \tag{4.7}$$

Application of dual ascent to this modified problem is known as the method of multipliers. Originally presented about four decades ago in [39, 94, 42], this method has been revived in recent years due to its usefulness in solving large-scale problems arising in image processing and statistical learning; see e.g., [14], [33] and references therein.

Unlike the original Lagrangian function, quadratic term in (4.7) destroys the separability. In order to decompose the large-scale problem and explore the property of each θ_i independently, the alternating direction method of multipliers (ADMM) is developed, which splits the problem by minimizing the augmented Lagrangian function with respect to each block of x_i alternatively in a Gauss-Seidel manner, and followed by an update for λ. Its application for solving two-block structured convex problems (i.e. $m = 2$) has been well studied in the literature. Global convergence is guaranteed with proper choice of ρ ([32, 58]) and convergence rate properties have been established (see, e.g., [14, 58, 57]). However, when it comes to the multi-block case (i.e. $m \geq 3$), its convergence has remained unclear for a long time. In [111], the authors proposed a strategy that first transforms a multi-block problem into an equivalent two-block problem and then solves it using the standard two-block ADMM. Although convergence is established, their approach is not as efficient as standard multi-block ADMM in practice (though the later lacks theoretical convergence guarantees). Recently, [48] showed that ADMM is globally convergent when θ_i's are further assumed to be strongly convex. Since then, this condition is relaxed to allow only parts of θ_i's to be strongly convex but may require some rank conditions on A_i's (see, e.g., [24] and [80]). Without imposing any strong convexity assumption, [25] gave a sufficient condition that ensure the convergence for the three-block problem (i.e. $m = 3$) which requires A_i's to be orthogonal; [79] demonstrated the convergence of standard multi-block ADMM when applied to a certain problem under some further conditions on the augmented Lagrangian function. Meanwhile, a counterexample was constructed in [25] showing that the standard ADMM is not necessarily convergent when the aforementioned conditions are violated. It is therefore of great interest to design algorithms which are convergent under more general conditions.

This has inspired researchers to develop extensions of ADMM with provable convergence on multi-block minimization problems under different conditions. Most of the modifications involve correcting the output of ADMM [56] or employing a proximal term to solve the

primal updates approximately, where the augmented function is replaced by

$$\tilde{L}^1_\gamma(x;\lambda) = \phi(x) - \lambda^T(Ax-b) + \frac{\rho}{2}\|Ax-b\|^2 + \|x-x^k\|^2_P, \tag{4.8}$$

where $||w||^2_P := w^T P w$.

This kind of approach has been given various names in different literature, sometimes it is referred to as the proximal alternating direction method of multipliers, or is also known as PPA-like method etc. The later name comes from the fact that it can be regarded as the application of the proximal point algorithm (PPA) to both the primal and dual problems (see [94]). Works fall in this category include [60, 99] with various recent modifications including the work of [55, 38, 54, 79, 22] and etc. Despite the different conditions required for convergence, i.e. strong convexity of all or parts of the objective functions and rank conditions on matrices A_i's, these methods are also compelled to a Gauss-Seidel implementation due to the coupling quadratic terms. Thus, it is clear that this type of sequential approaches will not be very efficient, especially when m is large. Moreover, the coefficient matrix of the quadratic terms requires additional computational effort for solving the subproblems. Both of these issues can be resolved if P is chosen carefully so the cross terms between x_i's can be cancelled out, thereby leading to a special case with easier subproblems (4.9):

$$\tilde{L}^2_\gamma(x_1,\ldots,x_m;\lambda) = \phi(x) - \lambda^T(Ax-b) + \frac{\gamma}{2}\sum_{i=1}^m \|x_i - q_i^k\|^2_{P_i}, \tag{4.9}$$

where q_i^k is some linear transformation of x_i^k.

Note that separability is preserved in (4.9) and in many cases the resulting subproblems are much easier to solve, even closed form solutions can be derived. Inspired by the fact that the $\sum_{i=1}^m ||x_i - q_i^k||^2_{P_i}$ term is similar to the proximal term in the classical PPA method, these techniques are sometimes known as customized PPA regularization or linearization, and have been extensively studied. Methods following this idea date back to more than two decades ago, when [27] presented a proximal-based decomposition method, known as predictor corrector proximal multiplier method (PCPM), for a special two-block minimization problem where $A_2 = -I$ and $b = \mathbf{0}$. The algorithm optimizes two subproblems in the form of (4.9) with $q_i^k = x_i^k$ simultaneously, and updates λ twice in a single iteration. However, when directly extended to the case with general linear constraints, known as the Primal-Dual Hybrid Gradient Algorithm (PDHG) by [123], the method became divergent. A variant by [21] ensures the convergence by a modified proximal operator (CP-method). Later, [61] and [59] accelerated the CP-method by adding a simple correction step, and revisited the proof of its convergence from a variational inequality point of view. We refer the readers to [52] for a review of PDHG and its extensions, where another convergent variant of PDHG is proposed without having to calculate q_i^k but requiring a more complicated correction on the primal variables and taking extra effort for matrix inversion and multiplication.

Heretofore, the aforementioned PPA-like methods consider either a general convex optimization or a very special two-block problem scenario. While in the multi-block case (4.1),

special attention to design decomposition schemes is deserved as we have observed the failures of many direct extensions of decomposition methods. In succession to their earlier relaxed CP-method, Bingsheng He et al. have proposed similar customized PPA algorithms to deal with the two-block and multi-block cases [16, 46]. The methods were originally proposed in the form of (4.7), and then part of the subproblems were converted into (4.9) by a linearization procedure under certain circumstance. Another group of PPA-based decomposition schemes take into account the idea of gradient descent. Replacing a_i^k by the gradient of the quadratic penalty in the original augmented Lagrangian function, [84] proposed the so-called alternating proximal gradient method (APGM) for a general two-block problem. It is then naturally extended to the multi-block case, however, with convergence only provable under the strong convexity of θ_i's [23].

Motivation. The algorithm considered in this chapter falls in the form of (4.9), and is the direct extension of PCPM to solve a general problem (P). It should be pointed out that, this method can be implemented Jacobian manner, and enjoys the advantages of parallel computing for the arising large-scale problems. In the Era of Big Data, millions of data records are gathered on daily basis, and even simple algorithms can become intractable when the problem size becomes extremely large and it is even impossible to store the whole algorithm structure with data in a single machine. Thus, the separability of an algorithm is the key to win a success in taking advantage of the big data and scaling the problem size in machine learning and data-driven decision making. We then realize that this is essentially a special case of the so-called Jacobi-Proximal ADMM algorithm proposed recently in [30]. Then, by leveraging the PPA interpretation of this algorithm and its equivalent variational inequality (VI) reformulation, we reestablish its globally convergence and linear convergence rate without assuming strong convexity of any objective function θ_i. Nor do we need to impose any column conditions on any matrix A_i. And our results is consistent with the analysis in [30].

However, despite of the theoretical linear convergence rate of this algorithm, its real-life application is very limited as the derived conditions for convergence guarantees are too conservative, leading to slow convergence in practice. To improve its performance, we provide a parameter tuning heuristic that is more flexible than the one proposed in [30]. Moreover, we show that a special case of such algorithm is still convergent even if we cannot, or it is too expensive to, solve the subproblems exactly.

The remainder of this chapter is organized as follows. In next section, we give a brief review of a few lemmas which form the foundation for many algorithm analysis in this field. Then, in Section 4.3, we introduce a multi-block decomposition algorithm with linear convergence guarantee, discuss the related study of such an algorithm in current literature and propose a new heuristic for parameter tuning. Afterward, in Section 4.4, we modify the algorithm so that the primal updates can be solved approximately and reestablish the convergence of the inexact version. Finally, we make some conclusion in Section 4.5.

4.2 Preliminaries

The reformulation of a convex optimization as variational inequalities (VI) has become an increasingly popular approach in algorithm design, as it significantly simplifies the convergence analysis process. In this section, we review two lemmas from the literature which illustrate this connection and state another lemma which is useful in our analysis In particular, the first two lemmas are taken from [54].

Lemma 4.1. *([54]) Let $\chi \subseteq R^n$ be a nonempty closed convex set, both $\phi(x) : R^n \to R$ and $g(x) : R^n \to R$ are closed convex functions. Further assume that $g(x)$ is smooth and differentiable. Then we have*

$$x^* \in \arg\min\{\phi(x) + g(x) | x \in \chi\} \tag{4.10}$$

if and only if

$$x^* \in \chi, \phi(x) - \phi(x^*) + (x - x^*)^T \nabla g(x^*) \geq 0 \quad \forall x \in \chi. \tag{4.11}$$

Lemma 4.2. *([54]) Let $\chi \subseteq R^n$ be a nonempty closed convex set, $\phi(x) : R^n \to R$ be a closed convex function. Consider the following constrained convex optimization problem*

$$x^* \in \arg\min\{\phi(x) | Ax = b, x \in \chi\}, \tag{4.12}$$

where $A \in R^{l*n}$, $b \in R^l$. Assume that the feasible set of (4.12) in nonempty, define $\lambda \in R^l$ as the dual variables corresponding to the m linear constraints and λ^* as the optimal dual solution. Then we can rewrite (4.12) as the variational inequalities:

$$\omega^* \in \Omega, \phi(x) - \phi(x^*) + (\omega - \omega^*)^T F(\omega^*) \geq 0 \quad \forall \omega \in \Omega \tag{4.13}$$

where

$$\omega = \begin{pmatrix} x \\ \lambda \end{pmatrix}, \quad \omega^* = \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix}, \quad F(\omega) = \begin{pmatrix} -A^T\lambda \\ Ax - b \end{pmatrix} \tag{4.14}$$

and

$$\Omega = \chi \times R^l.$$

Moreover,

$$(\omega - \bar{\omega})^T (F(\omega) - F(\bar{\omega})) = 0, \quad \forall \omega, \bar{\omega} \in \Omega, \tag{4.15}$$

which implies

$$\omega^* \in \Omega, \phi(x) - \phi(x^*) + (\omega - \omega^*)^T F(\omega) \geq 0 \quad \forall \omega \in \Omega. \tag{4.16}$$

These two lemmas imply that problem (P) can be rewritten as a set of variational inequalities. Thus, to show the convergence of this algorithm, it is equivalent to demonstrate that the accumulating point of this algorithm satisfies these inequalities. Moreover, the following lemma further helps with establishing its little o convergence rate, which is slightly stronger than the classical linear rate.

Lemma 4.3. *([30]) If a sequence $\{a_k\} \subseteq R$ obeys: (1) $a_k \geq 0$; (2) $\sum_{k=1}^{\infty} a_k \leq +\infty$; (3) a_k is monotonically non-increasing, then we have $a_k = o(1/k)$.*

4.3 The Exact Predictor Corrector Proximal Multiplier Method

4.3.1 Algorithm Description

Based on the properties of PPA and its primal-dual application, the predictor corrector proximal multiplier method (PCPM) was developed by Chen and Teboulle in 1994 [27]. It works on a special case of (P) where $m = 2$, $n_1 = n_2$, $A_2 = -I$ and $b = \mathbf{0}$. Further all, while their approach belongs to the case of (4.9), they consider a very special case where all P_is are diagonal matrices. The authors derived sufficient conditions to guarantee its global linear convergence under some extra conditions and allowed the primal updates to be solved approximately. The algorithm we consider applies the exact same logic to the general problem (P) but allows more general P_is, and starts with the case when all subproblems are solved exactly. Furthermore, we allow for a linear correction on the second dual update as this technique has been empirically shown to accelerate the convergence in practice, any all analysis still holds without this correction, i.e. $\gamma = 1$. Thus, we denote it the exact predictor corrector proximal multiplier method (EPCPM), as described in Algorithm 1.

Algorithm 1: EPCPM.

Input: ω^0, ρ, γ, P_i for $i = 1, 2, \ldots, m$.

for $k = 1, 2, \ldots$ **do**

1 **Step1.** Compute $\lambda_p^{k+1} = \lambda^k - \rho(Ax^k - b)$;

2 **Step2.** Solve

$$x_i^{k+1} = \underset{x_i \in \chi_i}{\arg\min} \left\{ L(x_1^k, \ldots, x_i, \ldots, x_m^k, \lambda_p^{k+1}) + \frac{1}{2}||x_i - x_i^k||_{P_i}^2 \right\}, \forall i = 1, 2, \ldots, m;$$

3 **Step3.** Update

$$\lambda^{k+1} = \lambda^k - \gamma\rho(Ax^{k+1} - b).$$

Same as PCPM, each iteration of Algorithm 1 consists of two steps for updating the dual variables, and a single primal variable update which is decomposable across each block of x. They can also be regarded as a proximal steps operated on the dual and primal problems of (P) respectively. And we will show in the following discussion that the additional dual update is essential for guaranteeing the convergence. Since a general instance of (P) may violate the structure required in the original PCPM paper, the convergence proof in [27] no longer hold. However, we later realize that Algorithm 1 is essentially a special case of the Jacobi-Proximal ADMM algorithm (Prox-JADMM) proposed in [30] with slightly differently defined P_i's. Thus, its convergence results follow directly from the general case. In order to clarify our notation and to facilitate our discussion on the inexact version of the algorithm,

we reestablish the result for Algorithm 1 explicitly from the perspective of PPA and VI properties.

4.3.2 Convergence Guarantees

The condition to ensure that Algorithm 1 is globally convergent is stated in the following theorem. The result is essentially the same as Theorem 2.1 from [30] with the prox-linear design. The difference in representation is mainly because that we use a different definition of the parameters.

To simplify the notation, let us first define that

$$R = \begin{pmatrix} P_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & P_m & 0 \\ 0 & \cdots & 0 & \frac{1}{\gamma\rho} I_l \end{pmatrix}, \tag{4.17}$$

$$G = \begin{pmatrix} \rho A_1^T A_1 & \cdots & \rho A_1^T A_m & \frac{1-\gamma}{\gamma} A_1^T \\ \vdots & \ddots & \vdots & \vdots \\ \rho A_m^T A_1 & \cdots & \rho A_m^T A_m & \frac{1-\gamma}{\gamma} A_m^T \\ 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{4.18}$$

and

$$Q = \begin{pmatrix} P_1 - \frac{\rho}{\delta_1} A_1^T A_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & P_m - \frac{\rho}{\delta_m} A_m^T A_m & 0 \\ 0 & \cdots & 0 & \frac{2-\gamma-\sum_{i=1}^m \delta_i}{\rho\gamma^2} I_l \end{pmatrix}, \tag{4.19}$$

where I_l is an identity matrix of size $l \times l$, $\mathbf{0}$ are matrices with all elements being zero and match the dimensions, and $\delta_i > 0,\ i = 1, 2, \ldots, m$ are a positive scalars that help us to derive sufficient conditions for achieving convergence.

Theorem 4.4. *Suppose that the solution set of the convex optimization problem (P) is nonempty, and the sequence $\{\omega^k\}$ is generated by Algorithm 1. Specifically, if one chooses $\gamma \in (0, 2)$ and some $\delta_i \in (0, 1]$ such that the other parameters ρ and P_i satisfy the following conditions:*

$$\begin{cases} P_i \succ \frac{\rho}{\delta_i} A_i^T A_i,\ i = 1, 2, \ldots, m, \\ \sum_{i=1}^m \delta_i < 2 - \gamma, \end{cases} \tag{4.20}$$

then, the sequence $\{\omega^k\}$ is convergent.

The proof is based upon the properties of PPA and VI conditions, which can be found in Appendix C. Furthermore, considering a special case where $P_i = \tau_i I_{n_i}$, and by letting each

$\delta_i < \frac{2-\gamma}{m}$, the condition (4.20) can be simplified to

$$\tau_i > \frac{m\rho}{2-\gamma}\|A_i\|^2 = \frac{m\rho}{2-\gamma}\lambda_{\max}(A_i^T A_i),\ i = 1, 2, \ldots, m, \tag{4.21}$$

where $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a matrix. Besides, it can also be shown that this algorithm achieves a $o(1/k)$ convergence rate.

Lemma 4.5. *When the conditions in Theorem 4.4 are satisfied, then*

i) $\|\omega^k - \omega^{k+1}\|_W^2 \leq \|\omega^{k-1} - \omega^k\|_W^2$,

ii) $\|\omega^k - \omega^{k+1}\|_W^2 = o(1/k)$,

where

$$W = \begin{pmatrix} P_1 - \rho A_1^T A_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & P_m - \rho A_m^T A_m & 0 \\ 0 & \cdots & 0 & \frac{1}{\rho\gamma} I_l \end{pmatrix}. \tag{4.22}$$

Now, let $\{\omega^{k_j}\}$ be a convergent subsequent of $\{\omega^k\}$, converging to ω^∞. Then, with assertion *ii)* and take $k_j \to \infty$, we can easily verify, from the standard analysis for contraction methods (see, e.g., [53]), that ω^∞ satisfies the VI corresponding to (P) and it is an optimal solution to (P).

4.3.3 A Special Case

In EPCPM algorithm, the choice of parameters P_is are crucial factors that decide the algorithm's performance. When P_is are too small, it takes risk of not converging. On the other hand, if P_is are chosen too large, the primal updates converges slowly as we penalize too much to change the primal solutions by imposing large values of $\|x_i - x_i^k\|_{P_i}^2$. Although Theorem 4.4 provides a sufficient condition for guaranteeing the linear convergence of EPCPM, which, however, is actually quite conservative, leading to slow convergence, because a few inequalities in the derivation are rather loose. In the following, we use a special case of EPCPM as an example and prove that conditions derived in Theorem 4.4 can be unnecessarily too conservative.

The setup of our example takes $\gamma = 1$, $P_i = \tau T_{n_i}$, $\forall i$ and $\rho = \frac{1}{\tau}$. We denote it as SPCPM (Special case of PCPM), which can be described as the above framework. Then, the sufficient conditions for global convergence from Theorem 4.4 is immediately $\tau \geq \max_i[\sqrt{m\lambda_{\max}(A_i^T A_i)}]$. On the other hand, we can actually achieve faster convergence by choosing less conservative parameters according to the following theorem.

Theorem 4.6. *Suppose that the solution set of the convex optimization problem (P) is nonempty, and the sequence $\{\omega^k\}$ is generated by Algorithm 2. If one chooses $\tau \geq \sqrt{2}\|A\|$, then, the sequence $\{\omega^k\}$ is convergent.*

Algorithm 2: SPCPM.

Input: ω^0 and τ.

for $k = 1, 2, \ldots$ **do**

Step1. Compute $\lambda_p^{k+1} = \lambda^k - \frac{1}{\tau}(Ax^k - b)$;

Step2. Solve

$$x_i^{k+1} = \arg\min_{x_i \in \chi_i} \left\{ L(x_1^k, \ldots, x_i, \ldots, x_m^k, \lambda_p^{k+1}) + \frac{\tau}{2}||x_i - x_i^k||^2 \right\}, \forall i = 1, 2, \ldots, m;$$

Step3. Update

$$\lambda^{k+1} = \lambda^k - \frac{1}{\tau}(Ax^{k+1} - b).$$

Thus, a sufficient condition for guaranteeing the convergence of Algorithm 2 is to choose $\tau \geq \min\{\max_i[\sqrt{m\lambda_{\max}(A_i^T A_i)}], \sqrt{2}\|A\|\}$. The condition given in Theorem 4.4 can be conservative.

4.3.4 Adaptive Parameter Tuning

Unfortunately, the approach for obtaining tighter convergence conditions does not apply to general case. Consequently, approaches like EPCPM are less frequently used in practice compared with ADMM, even though the latter lacks of a guarantee for convergence. In fact, however, we can see that in Appendix C to have a converging solution, it suffices to have the term $\|\omega^{k+1} - \omega^k\|_Q^2 > 0$. Thus, having Q being a positive definite matrix is a sufficient but not necessary condition. Based on this observation, [30] has proposed an Adaptive Parameter Tuning strategy (Strategy 1).

Strategy 1: Adaptive Parameter Tuning from [30].

Input: small $P_i^0 \succeq 0 (i = 1, 2, \ldots, m)$ and small $\eta > 0$.

for $k = 1, 2, \ldots$ **do**

if $\|\omega^{k-1} - \omega^k\|_Q^2 > \eta \cdot \|\omega^{k-1} - \omega^k\|^2$ **then**

$P_i^{k+1} \leftarrow P_i^k, \ \forall i$;

else

Increase P_i:

$$P_i^{k+1} \longleftarrow \alpha_i P_i^k + \beta_i Q_i \ (\alpha_i > 1, \ \beta_i > 1, \ Q_i \succ 0), \ \forall i;$$

Restart with $\omega^k \leftarrow \omega^{k-1}$.

Strategy 1 starts with small proximal parameters P_i and gradually increases them until

the contraction property holds. And the authors also showed that if P_i are initialized sufficiently small and then adaptively adjusted following this scheme, the algorithm converges to a solution to problem (P). They have also empirically demonstrated that it typically takes only a few iterations to get constant satisfactory parameters, which are usually much smaller than those proposed in Theorem 4.4. However, we realize that this strategy can be further improved as it suffers from the following three limitations:

1. The initial values of P_is are crucial to the performance of this strategy. If they are chosen too small, it will take more diverging iterations to find appropriate parameters; however, if they are initialized too large, the strategy cannot reversely tune the parameters to accelerate convergence.

2. Fixed step-sizes are used in this strategy, and we face almost the same difficulty in choosing adequate values for α_i and β_i for all $i = 1, 2, \ldots, m$. If these parameters are chosen too small, it takes more iterations to adjust P_is so that the algorithm converges; and if they are too large, the resulting stable P_is will be large leading to slow convergence of the algorithm.

To take account of the tradeoff, we propose the following modified Strategy 2, which allows two-directional tuning with changing step-sizes. And the step-sizes depend on the movement of the previous iteration. Thus, we name this new strategy as Adaptive Parameter Tuning with Feedback.

Strategy 2: Adaptive Parameter Tuning with Feedback .

Input: small $P_i^0 \succeq 0 (i = 1, 2, \ldots, m)$, $t > 0$ and small $\eta_1 > 0$, $\eta_2 > 0$ and $\eta_1 < \eta_2$.

for $k = 1, 2, \ldots$ **do**

Compute:

$$\Delta_1 = \|\omega^{k+1} - \omega^k\|_Q^2 - \eta_1 \|\omega^{k+1} - \omega^k\|^2;$$

$$\Delta_2 = \|\omega^{k+1} - \omega^k\|_Q^2 - \eta_2 \|\omega^{k+1} - \omega^k\|^2.$$

if $\Delta_1 < 0$ **then**

$P_i^{k+1} \leftarrow P_i^k - t\Delta_1 Q_i$; $\omega^k \leftarrow \omega^{k-1}$;

if $\Delta_2 > 0$ **then**

$P_i^{k+1} \leftarrow P_i^k - t\Delta_2 Q_i$; $\omega^k \leftarrow \omega^{k-1}$

else

$P_i^{k+1} \leftarrow P_i^k$, $\forall i$.

Intuition behind Strategy 2 is to bound the value of $\|\omega^{k+1} - \omega^k\|_Q^2$ within an interval of $[\eta_1 \|\omega^{k+1} - \omega^k\|^2, \eta_2 \|\omega^{k+1} - \omega^k\|^2]$ so that we can use small P_is which also guarantee convergence. Meanwhile, we also set the step-sizes to be proportional to the gap between $\|\omega^{k+1} - \omega^k\|_Q^2$ the value of and the boundaries. Thus, by setting small η_2, it is clear that the resulting parameters P_i are usually smaller than those obtained from Theorem 4.4 and

Strategy 1, thereby leading to faster algorithm convergence in practice. Moreover, compared with Strategy 1, it can be easily seen that the computation effort required at each iteration remains unchanged, while the new Strategy requires much less input parameters (a single value of t v.s. values of $\alpha_i, \beta_i, \ \forall i$).

4.3.5 Numerical Experiments

In [30], the authors compared the Jacobi-Proximal ADMM algorithm using parameter tuning strategy 1 with a couple of popular parallel splitting algorithms on two problems, i.e. Exchange Problem and l_1-minimization. The benchmark algorithms include:

- **Prox-JADMM**: Jacobi-Proximal ADMM (Algorithm 4 of [30] tuned according to Strategy 1)
- **VSADMM**: Variable Splitting ADMM (Algorithm 1 of [30])
- **Corr-JADMM**: Jacobi ADMM with correction steps

Moreover, the authors have kindly shared all codes and instances they used on their website[1]. We adopt the same experimental setting and all benchmark algorithms' parameters as given in the literature. To further compare with our approach, we initialize the EPCPM algorithm such that is is equivalent to the initialization of Prox-JADMM but tune the parameters according to Strategy 2 instead. All experiments are run in MATLAB (R2018b) on a MacBook with an Intel Core i7 CPU (2.2 GHz) and 8 GB of RAM.

Exchange Problem

The exchange problem aims at minimizing the total costs among N agents that exchange commodities subject to an equilibrium constraint:

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{N} f_i(x_i) \\ \text{s.t.} \quad & \sum_{i=1}^{N} x_i = 0. \end{aligned} \tag{4.23}$$

A special quadratic objective function in the form of $f_i(x_i) = \frac{1}{2}\|C_i x_i - d_i\|^2$ is considered, where $N = 100$, C_i are random Gaussian, $d_i = C_i x_i^*$ with random generated solution $x^* \in R^{90}$ according to [30], so are parameters for all benchmark algorithms. For EPCPM, we set γ to be 1, $\rho = 0.01$ as in Prox-JADMM and P_i initialized as $0.1(N-1)\rho I_{n_i} + \rho A_i^T A_i$ and then adaptively tuned according to Strategy 2.

[1] https://github.com/ZhiminPeng/Jacobi-ADMM

l_1-minimization

The l_1 minimization problem, also known as the basis pursuit problem, is another commonly used example in literature for demonstrating the efficiency of algorithms for large-scale convex optimization. It's mathematical formulation is

$$\begin{aligned} \min_{x} \quad & ||x||_1 \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{4.24}$$

where $A \in R^{l \times n}$ and $b \in R^l$. Similarly, all problem instances and benchmark algorithms are generated and set up as described in [30]: A is randomly generated from standard Gaussian distribution with size 300×1000; the optimal solution x^* is randomly generated with $k = 60$ non-zeros drawn from standard Gaussian distribution; and then b is set as $b = Ax^* + n$, where n $N(0, \sigma^2 I)$ is Gaussian noise with zero mean. Two cases are tested, i.e. the noise-free case ($\sigma = 0$) and the noisy case ($\sigma = 10^{-3}$). The problem is partitioned equally into 1000 blocks for Corr-JADMM, and 100 blocks for the other algorithms. Again, the initial parameters for EPCPM are set to be the same as Prox-JADMM.

Results

For both examples, 100 random instances are generated according to [30] and all algorithms are set to ran for at most 100 and 500 iterations respectively. While the benchmark algorithms performed similarly as in the literature, the green curves are those corresponding to Strategy 2. Below we compare the average performance over the 100 random trials.

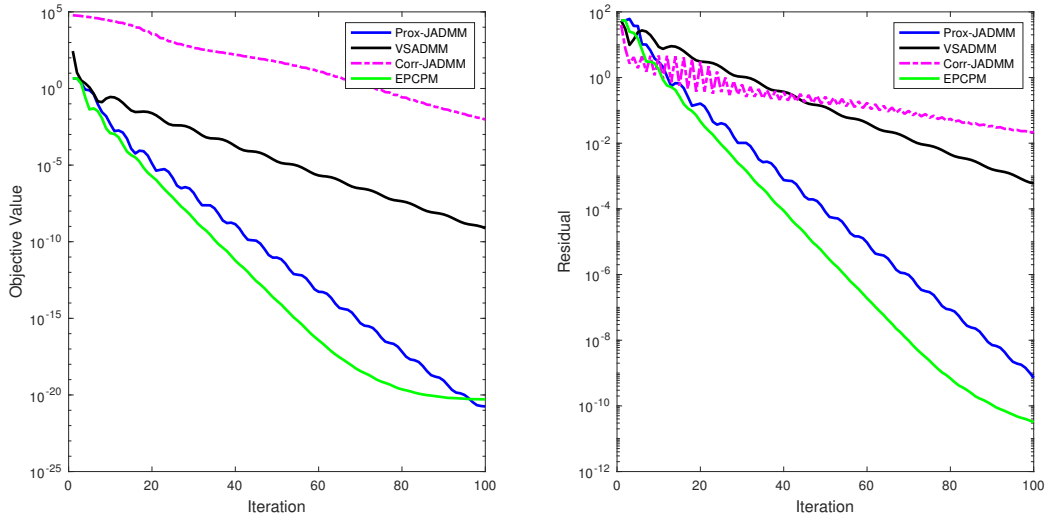
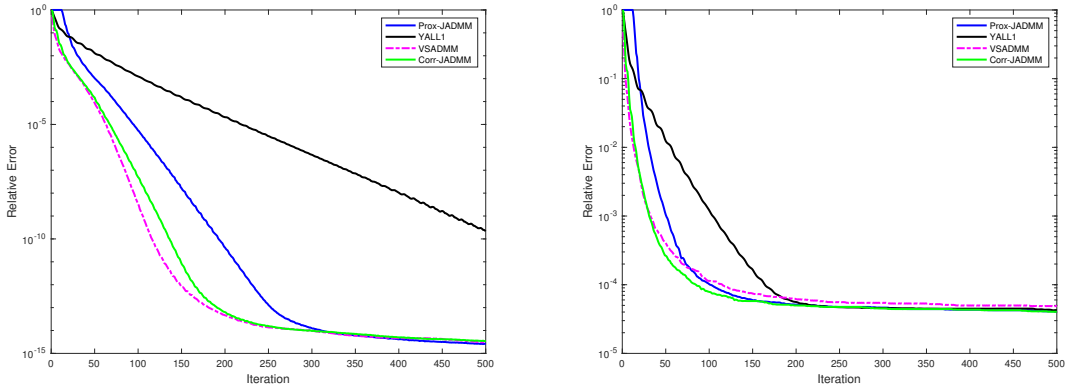


Figure 4.1: Exchange problem (N = 100, p = 80).



(a) No noise in data generation. (b) Noise added in data generation.

Figure 4.2: l_1 minimization problem.

Figure 4.1 shows the comparison of objective value and primal residual trajectories among all Algorithms. Same as demonstrated in [30], the Prox-JADMM is clearly the fastest one among all benchmark algorithms, while the green curves indicate that we can further accelerate its convergence significantly by using Strategy 2 to tune the parameters.

As for the l_1-minimization problem demonstrated in Figure 4.2, in both cases with or without noise, despite the slowly convergent VSADMM, all other three algorithms eventually became stable at similar accuracy level. Again, the utilization Strategy 2 helps us to achieve faster convergence.

Empirical evidence further shows that Strategy 2 in general requires less iterations to achieve constant parameters than the original Strategy 1. Moreover, the step to reduce the parameters is seldom performed. This is because that we use a nonconstant step-size which incorporates the size of $\|\omega^{k+1} - \omega^k\|$. At the beginning of the algorithm, P_is are small and the proximal terms penalize the updates of variables little, thus resulting in large step-size. On the other hand, as P_is grow larger and the current solution approaches the optimal solution, with Lemma C.1 indicating that $\|\omega^{k+1} - \omega^k\|$ will decrease, the step-size will decrease consequently. Hence, our new strategy is less likely to reach unnecessarily large parameters.

4.4 The Inexact Predictor Corrector Proximal Multiplier Method

In Section 4.3, we considered a decomposition algorithm and assume that all subproblems can be solved exactly. Many a time, it may be impossible or too time-consuming to find the exact solutions to all subproblems at each iteration. Thus, we further consider an extension of EPCPM to the case when oracles are available to approximately solve the subproblems,

and denote it by inexact predictor corrector proximal multiplier method (IPCPM). We adopt the notation for approximate minimization from [27].

Definition 4.1. *(Inexact Solution [27])*

$$\epsilon\text{-}\arg\min F(u) = \{v : F(v) \leq \inf F + \epsilon\}, \tag{4.25}$$

where $F(\cdot)$ is a given objective function and $\epsilon \geq 0$.

4.4.1 Algorithm Description

When modify the algorithm to the inexact version, we have to consider a special case when all P_is are diagonal matrices. In fact, the subproblems for primal updates in IPCPM are the same as the original PCPM algorithm from [27]. The scheme of IPCPM is described below.

Algorithm 3: IPCPM.

Input: ω^0, ρ, γ, P_i for $i = 1, 2, \ldots, m$.

for $k = 1, 2, \ldots$ **do**

Step1. Compute $\hat{\lambda}_p^{k+1} = \hat{\lambda}^k - \rho(A\hat{x}^k - b)$;

Step2. Solve

$$\hat{x}_i^{k+1} = \epsilon_i^k\text{-}\arg\min_{x_i \in \chi_i} \left\{ L(\hat{x}_1^k, \ldots, x_i, \ldots, \hat{x}_m^k, \hat{\lambda}_p^{k+1}) + \frac{\tau_i}{2}||x_i - \hat{x}_i^k||^2 \right\}, \forall i = 1, 2, \ldots, m;$$

Step3. Update

$$\hat{\lambda}^{k+1} = \hat{\lambda}^k - \gamma\rho(A\hat{x}^{k+1} - b).$$

Here, we require $\tau_i > 0$, $\forall i = 1, 2, \ldots, m$, and $\{\epsilon_i^k\}$ $(i = 1, 2, \ldots, m)$ are sequences which satisfy:

(1) $\epsilon_i^k \geq 0$;

(2) $\lim_{k\to\infty} \epsilon_i^k = 0$;

(3) $\{\epsilon_i^k\}$ are monotonically non-increasing.

Note that at any iteration, if we set $\epsilon_i^k = 0$, $i = 1, 2, \ldots, m$, then this IPCPM iteration is equivalent to an EPCPM iteration with exact minimization. And the contraction analysis for this iteration from EPCPM holds.

4.4.2 Convergence Guarantees

To establish the convergence of IPCPM, we leverage the results from our previous analysis of EPCPM. Let us consider the kth iteration of the algorithm with the sequence $\{\hat{\omega}^k\}$ generated by running IPCPM for the previous $k-1$ iterations. Now, suppose we can solve this iteration exactly and define:

$$\begin{cases} \hat{\lambda}_p^{k+1} = \hat{\lambda}^k - \rho(A\hat{x}^k - b) \\ \tilde{x}_i^{k+1} = \arg\min_{x_i \in \chi_i} \left\{ L(\hat{x}_1^k, \ldots, x_i, \ldots, \hat{x}_m^k, \hat{\lambda}_p^{k+1}) + \frac{\tau_i}{2}||x_i - \hat{x}_i^k||^2 \right\}, \forall i = 1, 2, \ldots, m; \\ \tilde{\lambda}^{k+1} = \hat{\lambda}^k - \gamma\rho(A\tilde{x}^{k+1} - b). \end{cases} \tag{4.26}$$

This is essentially performing EPCPM for a single iteration with an initial solution of $\hat{\omega}^k$, and the original analysis on EPCPM holds on this iteration. And by leveraging the analysis from [27], the convergence of IPCPM can be established by constructing a fundamental estimate between the exact and the inexact iterates from an optimal solution.

Theorem 4.7. *Suppose that the solution set of the convex optimization problem (P) is nonempty, and the sequence $\{\hat{\omega}^k\}$ is generated by Algorithm 3. Specifically, one can choose $\gamma \in (0, 2)$ and some $\rho > 0$ such that the other parameters τ_i and epsilon$_i^k$ satisfy the following conditions:*

i) $\epsilon_i^k \geq 0,\ \forall i,\ \forall k,$

ii) $\lim_{k\to\infty} \epsilon_i^k = 0,$

iii) $\tau_i I_{n_i} \succ \frac{m\rho}{2-\gamma} A_i^T A_i,$

Then, the sequence $\lim_{k\to\infty} \|\hat{\omega}^k\| = \lim_{k\to\infty} \|\tilde{\omega}^k\| = \omega^*$, *where ω^* is an optimal solution to (P).*

Again, same as the EPCPM case, the conditions listed in Theorem 4.7 are sufficient but not necessary, τ_i can be adaptively tuned following same strategies in practice and conditions for ϵ_i^k can be relaxed as long as it still holds that $\lim_{k\to\infty} \|\hat{\omega}^k - \tilde{\omega}^k\| = 0$.

4.5 Conclusion

When applying machine learning and distributionally robust optimization techniques to assist inventory management decision making in a data-driven environment, we eventually need to solve a minimization problem whose value is evaluated at all data points. The growing availability of data, on the one hand improves the accuracy in forecasting and results in decisions closer to optimal, on the other hand increases the difficulty in solving the optimization problems. Thus, due to the dramatically increasing need for dealing with big data, distributed algorithms which can be implemented in parallel and take advantage of the arising high performance computing infrastructures are of great interest.

In this chapter, we consider a group of convex optimization problems which can cover a lot of machine learning and data-driven optimization applications. The problem we explore has a special structure where the objective function is decomposable and the different blocks of decision variables are coupled together by only linear constraints. Inspired by ADMM, an algorithm based on the augmented Lagrangian function and designed specifically for the problem setup, we consider another framework with more flexible quadratic terms. This framework also enjoys the convenience of strongly convex subproblems, while having advantages over ADMM in the sense that each block of primal variables can be updated in parallel and its global convergence can be guaranteed in multi-block case. While we got our inspiration from the PCPM algorithm proposed in [27], the method turns out to be a special case of a Jacobi-ADMM algorithm introduced in [30]. We modified an adaptive parameter tuning strategy to improve the performance of this algorithm in practice, and we further allow the subproblems to be solved only approximately in some special cases while still ensure the convergence.

Chapter 5

Concluding Remarks and Future Work

Motivated by the rising popularity of big data in assisting decision making, we explore how to make optimal inventory management decisions using data-driven approaches from different perspectives in this thesis:

We first introduce a special neural network structure for modelling demand time series. Compared with traditional parametric time series analysis models, neural network works as a universal approximator and thus, is much more flexible and can capture a wide range of continuous demand functions. Specifically, the linear shortcuts in our model enables it to treat nonstationary time series in the same way as the stationary ones, while current methods in literature require additional parametric assumptions to capture components like trend and seasonality. Thus, we contribute to the current literature by proposing a model that deals with nonstationary time series without signal decomposition. Furthermore, by adopting quantile loss function, we allow the network to output the desired inventory level directly without generating an explicit demand forecast. With both theoretical and numerical studies, we demonstrate that our method can serve as data-driven solutions to the classical newsvendor problem and its multi-period extension.

In Chapter 3, we address the data-driven newsvendor problem from a different angle with the focus on getting robust data-driven solutions with theoretical performance guarantees. Moreover, we incorporate information from covariates in addition to time series data. However, in order to obtain the desired tractable formulation and theoretical results, we limit our analysis to consider a linear demand model. In contrast to the existing literature which consider the covariates as a random vector, we deploy a deterministic interpretation, reckoning that this fixed design better fits the real-world applications. A two-step framework is proposed by leveraging the techniques of ordinary least squares estimator and distributionally robust optimization. Specifically, the Wasserstein metric is chosen for constructing an ambiguity set in order to achieve nice out-of-sample performance guarantees and a tractable reformulation. As a matter of fact, we demonstrate that our data-driven solution can be obtained in polynomial time with linear regression and then sorting. Furthermore, with any

desired high probability we can bound the out-of-sample expected cost with finite sample points, and our robust solution converges to the real optimal solution when sample size grows to infinity. Moreover, while we conducted all analysis for these two projects in the newsvendor setting, our methods can be used naturally for solving a more general problem of quantile forecasting. To the best of our knowledge, this is the first work in the field of data-driven distributionally robust optimization that does not require an i.i.d. sample of the random component.

Finally, in Chapter 4, we investigate a class of decomposition algorithms to solve multi-block convex optimization problems with linear constraints, which can be used to tackle a wide range of problems in the field of machine learning and data-driven optimizations [14]. The algorithm we consider is a special case of the so-called proximal ADMM algorithms. It enjoys the convenience of global linear convergence under multi-block case without strongly convex objective functions, and all its primal subproblems can be solved in parallel such that modern distributed computing system can be used. However, parameters chosen according to the derived sufficient conditions to guarantee convergence are rather conservative in practice. Inspired by the adaptive parameter tuning scheme proposed in [30], we provide a modification with more flexible step-sizes and two-directional adjustments. Furthermore, we prove that, for a special case of such algorithms, convergence can also be established even if the subproblems are only solved approximately.

All together, the aforementioned three chapters work towards achieving the same goal, that is to provide more efficient and practical data-driven approaches for inventory management in the Big Data Era. With the purpose of extending the current research, possible future work includes:

- **Dealing with censored data.** Up till now, we have assumed that historical demand or covariates are fully observable. In reality, not only lost sales can not be tracked, there will also be some missing values. To cope with censored demand, the pattern-fitting method proposed by [96], which assumes demand occurs following the same pattern during each time period, can also be useful under our setting. For example, if decision is made on daily basis, Sachs et al. assume that the demand occurred during any hour is a fixed ratio of the total daily demand. Thus, if a product is sold out during the first few hours in a day, the demand in the remaining hours can be estimated by the previous hourly demands and ratios learned from other days. If, again, the demand is a time-correlated process, we can take account the autocorrelation to achieve more accurate estimation. For example, we may regard the hourly demand observations as a more frequently sampled time series. Based on which, we can come up with another way to estimate the real daily demand. Similarly, when missing value occurred, we may use techniques such as imputation and account for patterns over time to estimate the missing values. Especially, instead of using a two-step framework where we first estimate the real demand and missing values and then solve an optimization problem with the estimations, we might think about dealing with the censored data directly in the inventory management decision making models.

- **Integrating time series with covariates information.** While time series data help us to capture the internal structure of a random process over time, adding external signal as features can be especially useful when plan inventory for fast-fashion products. Retailers like Zara frequently introduces new products; e-commercial websites like Rue La La and VIP.com offer time-limited discounts. Under both circumstances, they need to make inventory decision without historical observations. It is crucial to make use of the information of similar products to make informative data-driven decisions. As demonstrated in [36], a study into this problem will be more data-oriented. While Chapter 2 and Chapter 3 each considers time series and covariates data respectively, we may consider models that account for both. Our goal is to propose a more theoretically and intuitively comprehensible framework to assist the decision-making procedure, rather than just try and compare different machine learning techniques.

- **Trading off the flexibility and robustness.** The two approaches we proposed in Chapter 2 and Chapter 3 address the newsvendor problem with two very different goals, and both have their own strengths and weaknesses. On the one hand, the DPFNN-QAR method works empirically well with little assumptions but lacks a theoretical guarantee for supporting the robustness of such approach; on the other hand, the distributionally robust optimization approach is easily solvable and attains elegant finite-sample and asymptotic performance guarantees, however, requires a very restrictive linear demand model. In practice, what is most desirable is probably something in between, a model that is more flexible than linear but still with some practical assumptions that enable theoretical results to be established. If we follow a similar distributionally robust optimization scheme as described in Chapter 3, the resulting reformulation will not be as simple as the one we have for linear model. However, it is likely that we will still need to solve a finite-dimensional problem with problem size growing as more data points are considered [34]. This is when decomposition algorithms, such as the one describe in Chapter 4, can be helpful.

- **Decomposition algorithms for more general problems.** Currently, popular decomposition algorithms based on augmented Lagrangian relaxation, such as ADMM and its variations as we discussed in Chapter 4, only work with convex optimization with linear constraints. In practice, however, there are may large-scale problems violating these assumptions. For example, the neural network based methods are not convex, and many other applications involve nonlinear constraints. Thus, it is desirable to extend them to solve more general formulations. Some empirical studies and special cases have been explored in this direction (e.g., [112, 110, 118] and etc.), but there still lacks more general and theoretical results. Hence, this leads to a potential future research direction that worth exploring.

- **More complicated inventory models.** In both Chapter 2 and Chapter 3, we have restricted our analysis in the environment of the newsvendor problem and its multi-period extension. These problem settings enjoy the convenience that the so-

called critical quantile is available as a closed-from optimal solution. Consequently, the inventory management problem eventually boils down to the forecasting of this quantile, and requires us to solve an unconstrained convex optimization formulation. Many a time, more complex problem setting is required in real-world applications. For instance, the limited resources is available and thus, our inventory control models should be constrained. It is of interest if we can propose new data-driven approaches and take into account of these additional constraints based on machine learning or distributionally robust optimization techniques.

– Besides aforementioned, there are other interesting areas to explore in the future, say inventory management for multiple products with substitutions, with non-zero fixed ordering costs, with non-zero or even random lead times, etc. In general, our goal for future research is to solve more complicated inventory problems and gain theoretical insights under more realistic assumptions, with the help of abundant data resources, powerful computing infrastructures and ever-evolving techniques in the field of machine learning and robust optimization, etc.

Appendix A

Supporting Results for Chapter 2

A.1 Proof of Theorem 2.2

The proof is inspired by [28] and uses results of [87] and [51]. We quote the useful lemma and theorem in the following:

Theorem A.1 (Theorem 2.1 of [87]). *If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at θ_0; (ii) Θ is compact; (iii) $Q_0(\theta)$ is continuous; (iv) $Q_N(\theta)$ converges uniformly in probability to $Q_0(\theta)$ ($\sup_{\theta\in\Theta} |Q_N(\theta) - Q_0(\theta)| \xrightarrow{p} 0$), then $\hat{\theta}_N \xrightarrow{p} \theta_0$ where $\hat{\theta}_N$ maximizes $Q_N(\theta)$*

Lemma A.2 (Lemma 7.2 of [51]). *Uniform law of large numbers: Let $\{w_t\}$ be an ergodic stationary process. Suppose that (i) the set Θ is compact; (ii) $m(w_t;\theta)$ is continuous in θ for all w_t; (iii) $m(w_t;\theta)$ is measurable in w_t for all $\theta \in \Theta$; and (iv) $E[\sup_{\theta\in\Theta} |m(w_t;\theta)|] < \infty$. Then $\frac{1}{n}\sum_{t=1}^{n} m(w_t;\cdot)$ converges uniformly in probability to $E[m(w_t;\cdot)]$ over Θ. Moreover, $E[m(w_t;\theta)]$ is a continuous function of θ.*

With the above theoretical support and assumptions, we are able to establish the consistency. Denote

$$h_\theta(x_t) = H(x_t;\theta) - H(x_t;\theta_0) \tag{A.1}$$

and define

$$\begin{aligned} L_N(\theta) &= \frac{1}{N-p}\sum_{t=p+1}^{N} [\rho_\tau(u_t - h_\theta(x_t)) - \rho_\tau(u_t)] \\ &= \frac{1}{N-p}\sum_{t=p+1}^{N} q_\tau(d_t, x_t, \theta), \end{aligned} \tag{A.2}$$

where

$$q_\tau(d_t, x_t, \theta) = \rho_\tau(u_t - h_\theta(x_t)) - \rho_\tau(u_t). \tag{A.3}$$

With (2.21) and (2.22), it is easy to see that

$$\hat{\theta}_N = \arg\min_{\theta\in\Theta} L_N(\theta). \tag{A.4}$$

Under Assumption (ii) and since $\tau \in [0,1]$,

$$\begin{aligned} E[\sup_{\theta\in\Theta} |q_\tau(D_t, X_t, \theta)|] &= E[\sup_{\theta\in\Theta} |\rho_\tau(D_t - H(X_t;\theta)) - \rho_\tau(D_t - H(X_t;\theta_0))|] \\ &\le E[\sup_{\theta\in\Theta} |H(X_t;\theta) - H(X_t;\theta_0)|] \\ &< \infty. \end{aligned} \tag{A.5}$$

Since Assumption (i) directly assumes the compactness of Θ, $q_\tau(D_t, X_t, \theta)$ is continuous in both (D_t, X_t) and θ by construction, thus measurable in (D_t, X_t) for all $\theta \in \Theta$. All conditions in Lemma A.2 are satisfied. Thus, by applying the uniform law of large numbers for stationary ergodic processes, we verify that $L_N(\theta)$ converges uniformly in probability to $E[L_N(\theta)]$, and the later is continuous in θ. Finally, it remains to verify that $H(\cdot, \theta_0)$ is the unique minimizer of $E[L_N(\theta)]$. With Knight's Identity from [71] we have

$$\rho_\tau(u - v) - \rho_\tau(u) = -v(\tau - \mathbb{1}(u < 0)) + \int_0^v \{\mathbb{1}(u \le s) - \mathbb{1}(u \le 0)\} ds. \tag{A.6}$$

Then, by the property of conditional expectation, and recall that $F_u(0) = \tau$, we have

$$\begin{aligned} E[q_\tau(D_t, X_t, \theta)] &= E[E[q_\tau(D_t, X_t, \theta)|X_t]] \\ &= E[E[\rho_\tau(u_t - h_\theta(X_t)) - \rho_\tau(u_t)|X_t]] \\ &= E[E[-h_\theta(X_t)(\tau - \mathbb{1}(u_t < 0))|X_t]] \\ &\quad + E[E[\int_0^{h_\theta(X_t)} \{\mathbb{1}(u_t \le s) - \mathbb{1}(u_t \le 0)\} ds|X_t]] \\ &= E[-h_\theta(X_t)(\tau - E[\mathbb{1}(u_t < 0)|X_t])] \\ &\quad + E[\int_0^{h_\theta(X_t)} E[\{\mathbb{1}(u_t \le s) - \mathbb{1}(u_t \le 0)\} ds|X_t]] \\ &= E[-h_\theta(X_t)(\tau - F_u(0))] + E[\int_0^{h_\theta(X_t)} (F_u(s) - F_u(0)) ds] \\ &= E[\int_0^{h_\theta(X_t)} (F_u(s) - F_u(0)) ds]. \end{aligned} \tag{A.7}$$

Under Assumption (iii), we know that $F_u(u)$ is a strictly increasing function on $[-\epsilon, \epsilon]$. If $h_\theta(x_t) > 0$,

$$\int_0^{h_\theta(x_t)} (F_u(s) - F_u(0)) ds \ge \int_0^{\min(\epsilon, h_\theta(x_t))} (F_u(s) - F_u(0)) ds > 0. \tag{A.8}$$

Similarly, in the case of $h_\theta(x_t) < 0$,

$$\int_0^{h_\theta(x_t)} (F_u(s) - F_u(0))ds \geq \int_{\max(-\epsilon, h_\theta(x_t))}^{0} (F_u(0) - F_u(s))ds > 0. \tag{A.9}$$

Thus, for any x_t, $E[q_\tau(Y_t, X_t, \theta)] = 0$ only if $h_\theta(X_t) = 0$. With

$$E[L_N(\theta)] = \frac{1}{N-p} \sum_{t=p+1}^{N} E[q_\tau(D_t, X_t, \theta)], \tag{A.10}$$

the minimum of $E[L_N(\theta)]$ is achieved only if $H(X_t; \theta) = H(X_t; \theta_0)$. However, the neural network model is unidentifiable, as there can be multiple θ_0 leading towards the same output. And we can not achieve the consistency of the parameters.

A.2 Proof of Theorem 2.3

Let

$$W(S_t, D_t) = cS_t + g(S_t, D_t) - \gamma c(S_t - D_t). \tag{A.11}$$

Then

$$\begin{aligned} f_T(\bar{S}) &= \mathbb{E}\{\sum_{t=1}^{T} \gamma^{t-1} W(S_t, D_t)\} \\ &= \sum_{t=1}^{T} \gamma^{t-1} \mathbb{E}\{\mathbb{E}_{D_t|H_t}[W(S_t, D_t)]\}. \end{aligned} \tag{A.12}$$

And our goal is to find $\bar{S}^* = \arg\min f_T(\bar{S})$.

Now, we can establish the optimality of a myopic policy by adapting the Theorem 6.1 from [106]:

Theorem A.3 (Theorem 6.1 from [106])**.** *if*

a) For any fixed H_t, $\tilde{S}_t = \arg\min \mathbb{E}_{D_t|H_t}[W(S_t, D_t)]$;

b) Under this policy, it is always true that $x_t \leq \tilde{S}_t \quad t = 1, 2, ..., T-1$;

then, $\bar{S}^* = (\tilde{S}_1, \tilde{S}_2, ..., \tilde{S}_T)$.

We first solve for $\tilde{S}_t$ as below

$$\begin{aligned} &\mathbb{E}_{D_t|H_t}[W(S_t, D_t)] = \mathbb{E}_{D_t|H_t}[cS_t + g(S_t, D_t) - \gamma c(S_t - D_t)] \\ &= \mathbb{E}_{D_t|H_t}[(h + (1-\gamma)c)\max(S_t - D_t, 0) + (b - (1-\gamma)c)\max(D_t - S_t, 0)] \\ &\quad + \mathbb{E}_{D_t|H_t}[cD_t] \\ &= \mathbb{E}_{D_t|H_t}[(h + (1-\gamma)c)\max(S_t - D_t, 0) + (b - (1-\gamma)c)\max(D_t - S_t, 0)] \\ &\quad + g(d_{t-1}, ..., d_{t-p}), \end{aligned} \tag{A.13}$$

which is similar to the single period model with a constant term. The minimum is

$$\tilde{S}_t = F_t^{-1}(\frac{b-(1-\gamma)c}{h+b}|l_t) = g(d_{t-1},...,d_{t-p}) + Z_\tau, \tag{A.14}$$

with $\tau = \frac{b-(1-\gamma)c}{h+b}$. Finally, it remains to verify condition (b) to show that this myopic policy is optimal. And since demand is always nonnegative, it is obviously true when $t = 1$, and for $t = 2, 3, ..., T-1$

$$\begin{aligned} x_t &= \tilde{S}_{t-1} - D_{t-1} \\ &= g(d_{t-2},...,d_{t-p-1}) + Z_\tau - g(d_{t-2},...,d_{t-p-1}) - \varepsilon_{t-1} \\ &= Z_\tau - \varepsilon_{t-1} \\ &\leq Z_\tau + g(d_{t-1},...,d_{t-p}) \\ &= \tilde{S}_t \end{aligned} \tag{A.15}$$

The inequality holds since demand is always nonnegative and ε_t are i.i.d.. This theorem also works when $T = \infty$. Thus, the myopic policy is optimal in the multiperiod scenario.

Appendix B

Supporting Results for Chapter 3

B.1 Proof of Lemma 3.6

In order to prove Lemma 3.6, let us first introduce the following intermediate results:

Lemma B.1. *For any sequence of random variables V_i, $i = 1, 2, \ldots, N$, we have*

$$Var(\frac{1}{N}\sum_{i=1}^{N}|V_i|) \leq \frac{1}{N}\sum_{i=1}^{N}Var(|V_i|) \leq \frac{1}{N}\sum_{i=1}^{N}Var(V_i)$$

Proof. Consider $U_i = |V_i| - \mathbb{E}|V_i|$, we have

$$Var(\sum_{i=1}^{N}U_i) = \mathbb{E}[(\sum_{i=1}^{N}U_i)^2].$$

Note that

$$\begin{aligned}(\sum_{i=1}^{N}U_i)^2 &= \sum_{i=1}^{N}U_i^2 + \sum_{i\neq j, i,j=1,\ldots N}U_iU_j \\ &\leq \sum_{i=1}^{N}U_i^2 + \sum_{i\neq j, i,j=1,\ldots N}\frac{U_i^2+U_j^2}{2} \\ &= \sum_{i=1}^{N}NU_i^2.\end{aligned} \tag{B.1}$$

Hence, $Var(\sum_{i=1}^{N}U_i) \leq N\sum_{i=1}^{N}Var(U_i)$, and the desired result follows directly. □

Consequently, to prove Lemma 3.6, let us consider

$$\begin{aligned} & d_W(\mathbb{P}_N(\beta_0), \mathbb{P}_N(\hat{\beta}^N_{OLS})) \\ =& \sup_{f\in\mathcal{L}}\{\frac{1}{N}\sum_{i=1}^{N} f(\varepsilon_i) - f(\varepsilon_i^{OLS})\} \\ \leq& \frac{1}{N}\sum_{i=1}^{N}|\varepsilon_i - \varepsilon_i^{OLS}| \\ =& \frac{1}{N}\sum_{i=1}^{N}|(\hat{\beta}_N^{OLS} - \beta_0)^T x_i| \\ =& \frac{1}{N}\sum_{i=1}^{N}|x_i^T(X^TX)^{-1}X^T\boldsymbol{\varepsilon}|, \end{aligned} \tag{B.2}$$

where ε_i^{OLS} denotes the linear regression residual of the ith data point.

The first equality follows from the dual representation of the Wasserstein metric [66], and the following inequality is justified by f being Lipschitz functions. Note that this distance and the residuals are random due to the randomness of the noise term in each of the data point, and we stack the random noises in all data points in vector $\boldsymbol{\varepsilon}$.

To simplify the notation and apply Lemma B.1, let us set $V_i = x_i^T(X^TX)^{-1}X^T\boldsymbol{\varepsilon}$. 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

$$\begin{aligned} Var(V_i) &= x_i^T(X^TX)^{-1}X^TVar(\boldsymbol{\varepsilon})X(X^TX)^{-1}x_i \\ &= x_i^T(X^TX)^{-1}X^T(\sigma^2)I_NX(X^TX)^{-1}x_i \\ &= \sigma^2 x_i^T(X^TX)^{-1}x_i \\ &\leq Mh_{ii}, \end{aligned}$$

where $h_{ii} = x_i^T(X^TX)^{-1}x_i$ is the leverage of the the ith data point in linear regression as defined in Section 3.3.2. It follows that the left-hand-side of inequality (B.2) is bounded from below with probability

$$\begin{aligned} \mathbb{P}(\frac{1}{N}\sum_{i=1}^{N}|V_i| > r_N^2) &\leq \frac{Var(\frac{1}{N}\sum_{i=1}^{N}|V_i|)}{(r_N^2)^2} \\ &\leq \frac{1}{N}\frac{\sum_{i=1}^{N}Var(V_i)}{(r_N)^2} \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\frac{Mh_{ii}}{(r_N^2)^2} \\ &= \frac{1}{N}\frac{pM}{(r_N^2)^2} \end{aligned} \tag{B.3}$$

The first inequality follows from the Chernoff's inequality, while the second inequality from Lemma B.1. And finally applying Lemma 3.3 as given in Section 3.3.2, the last equality holds. With (B.2), we have

$$\begin{aligned}\mathbb{P}_\epsilon\{d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \geq r_N^2\} &\leq \mathbb{P}(\frac{1}{N}\sum_{i=1}^{N}|V_i| > r_N^2) \\ &\leq \frac{1}{N}\frac{pM}{(r_N^2)^2}.\end{aligned} \tag{B.4}$$

Hence, with $r_N^2 = \sqrt{\frac{pM}{N\eta'}}$, we can bound the distance with any desired probability $(1-\eta') \in (0,1)$. And for any given value of η' since both p and M are finite constant, we have $r_N^2 \to 0$ as $N \to \infty$.

B.2 Proof of Lemma 3.7

Noting the fact that ε' is just the random variable ε shifted by a constant, we show that the following result holds for general case.

Lemma B.2. *Let μ and ν denote two probability measures supported on R. Suppose $d\nu(y) = d\mu(x)$ for all $x \in R$, $y = x + t$, where t is a constant. Then it holds that*

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

Proof. Due to the symmetry of this problem, without loss of generality, we may assume $t \geq 0$. Suppose the joint density function is $f(x,y)$, since $\mu_Y(x+t) = \mu_X(x)$,

$$\int f(y, x+t)dy = \int f(x,y)dy := g(x).$$

Hence,

$$\begin{aligned}\int |x-y|f(x,y)dxdy &\geq \int (y-x)f(x,y)dxdy \\ &= \int yf(x,y)dxdy - \int xf(x,y)dxdy \\ &= \int yg(y-t)dy - \int xg(x)dx \\ &= \int (y+t)g(y)dy - \int yg(y)dy \\ &= t.\end{aligned}$$

□

Thus, by applying Lemma B.2 to the left-hand-side of (3.27), we get the Wasserstein distance between $\mathbb{P}_\varepsilon$ and $\mathbb{P}_{\varepsilon'}$ is

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

Then, by Chernoff's Inequality, we have

$$\begin{aligned}
\mathbb{P}^N(|(\beta_0 - \hat{\beta}_{OLS}^N)^T c| > r_N^3) &\leq \frac{Var(|(\beta_0 - \hat{\beta}_{OLS}^N)^T c)|)}{(r_N^3)^2} \\
&\leq \frac{Var((\beta_0 - \hat{\beta}_{OLS}^N)^T c))}{(r_N^3)^2} \\
&= \frac{Var(((X^T X)^{-1} X^T \boldsymbol{\varepsilon})^T c)}{(r_N^3)^2} \\
&= \frac{\sigma^2 c^T (X^T X)^{-1} c}{(r_N^3)^2} \\
&\leq \frac{M}{(r_N^3)^2} (c^T (X^T X)^{-1} c).
\end{aligned} \tag{B.5}$$

Thus, it suffices to select $r_N^3 = \sqrt{\frac{Mc^T(X^TX)^{-1}c}{\eta'}}$ so that the above probability if bounded by η', and with Assumption 3.3 it converges to 0 as N goes to ∞.

B.3 Proof of Theorem 3.8

With DRO solution being feasible to the original stochastic program (3.18), we always have $J^* \leq J_{OOS}$. Then, for any sequence of $\eta_N \in (0, 1)$, we can choose $r_N(\eta_N)$ based on the conditions derived in Lemma 3.5-3.7, such that

$$\mathbb{P}^N\{J^* \leq \hat{J}_N\} \geq \mathbb{P}^N\{J_{OOS} \leq \hat{J}_N\} \geq 1 - \eta_N, \ \forall N \in \mathbb{N} \tag{B.6}$$

Then, by applying the Borel-Cantelli Lemma ([68], Theorem 2.18), we have

$$\mathbb{P}^\infty\{J^* \leq \hat{J}_N, \ \forall \text{ sufficiently large } N\} = 1. \tag{B.7}$$

By similar arguments, it also holds that

$$\mathbb{P}^\infty\{J_{OOS} \leq \hat{J}_N, \ \forall \text{ sufficiently large } N\} = 1. \tag{B.8}$$

Then, it remains to show that

$$\mathbb{P}^\infty\{\limsup_{N\to\infty} \hat{J}_N \leq J^*\} = 1. \tag{B.9}$$

Note that $\rho_\tau(\varepsilon - s)$ is a 1-Lipschitz function with respect to ε. Hence, for any $\delta > 0$, we may choose a δ-optimal solution s_δ of $\min_s \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s)]$, that is,

$$\mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s_\delta)] \leq J^* + \delta. \tag{B.10}$$

Consequently, we can always choose a sequence $\{\mathbb{Q}_N\}$ where $\mathbb{Q}_N \in \mathcal{P}, \forall N \in \mathbb{N}$ which satisfies

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}[\rho_\tau(\varepsilon - s_\delta)] \leq \mathbb{E}_{\mathbb{Q}_N}[\rho_\tau(\varepsilon - s_\delta)] + \delta. \tag{B.11}$$

Note that $\varepsilon' = \varepsilon + (\beta_0 - \hat{\beta}^N_{OLS})^T c \to \varepsilon$ almost surely due to the strongly consistency of the OLD estimator. Therefore,

$$\begin{aligned}
\limsup_{N\to\infty} \hat{J}_N \leq & \limsup_{N\to\infty} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}[\rho_\tau(\varepsilon - s_\delta)] \\
\leq & \limsup_{N\to\infty} \mathbb{E}_{\mathbb{Q}_N}[\rho_\tau(\varepsilon - s_\delta)] + \delta \\
\leq & \limsup_{N\to\infty} \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s_\delta)] + d_w(\mathbb{P}_{\varepsilon'}, \mathbb{Q}_N) + d_w(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) + \delta \\
= & \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s_\delta)] + \delta, \ \mathbb{P}^\infty - \text{almost surely} \\
\leq & J^* + 2\delta
\end{aligned} \tag{B.12}$$

where the first inequality follows from the optimality of $\hat{J}_N$ to (3.19); the second is based on inequality (B.11); followed by applying the dual representation of Wasserstein distance as defined in Lemma 3.1 and the fact that ρ_τ is 1-Lipschitz and finally with Lemma 3.9 to reach the second last equation.

Observing the equivalent reformulation of DRO in Section 3.5, we come up with an alternative proof of $\limsup \hat{J}_N \leq J^*$: Noting the fact that

$$\begin{aligned}
\limsup_{N\to\infty} \hat{J}_N = & \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \min_s \mathbb{E}_{\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})}(\rho_\tau(\varepsilon - s)) \\
\leq & \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \mathbb{E}_{\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})}(\rho_\tau(\varepsilon - s_\delta)) \\
\leq & \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \mathbb{E}_{\mathbb{P}_\varepsilon}(\rho_\tau(\varepsilon - s_\delta)) + d_w(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) + d_w(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \\
= & \mathbb{E}_{\mathbb{P}_\varepsilon}(\rho_\tau(\varepsilon - s_\delta)) \quad \mathbb{P}^\infty\text{-almost surely} \\
= & J^* + \delta,
\end{aligned} \tag{B.13}$$

where the first inequality follows from the feasibility of s_δ, and the second inequality holds due to the dual representation of Wasserstein distance as defined in Lemma 3.1 and the fact that ρ_τ is 1-Lipschitz. Further notice the fact that, $d_w(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) \to 0$ $\mathbb{P}^\infty-$ almost surely due to the almost surely convergence of $\hat{\beta}^N_{OLS}$ and $d_w(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}))$. This implies the second last equality together with the fact that $r_N \to 0$ and the fact that $\mathbb{P}_{\varepsilon'} \to \mathbb{P}_\varepsilon$, $\mathbb{P}^\infty-$ almost surely by Lemma 3.9.

Finally, since the above result holds for any arbitrarily small $\delta > 0$, thus the claim $\limsup_{N\to\infty} \hat{J}_N \leq J^*$ follows and leads to the conclusion that $\hat{J}_N \downarrow J^*$.

Now we aim at showing 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 (3.18) and that it is unique when the quantile function is identifiable. Due to the strongly consistency of $\hat{\beta}^N_{OLS}$ from Lemma 3.2, i.e. $\hat{\beta}^N_{OLS} \to \beta_0$ $-\mathbb{P}^\infty$ almost surely, it remains to show that $\bar{s}$, any limit point of $\hat{s}_N$, if exists, is s_τ.

Starting from J^* being the optimal object value and combining the previous results (B.8) and that $J^* = \lim_{N\to\infty} \hat{J}_N$, we have

$$\begin{aligned} J^* &\leq \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - \bar{s})] \\ &= \mathbb{E}_{\mathbb{P}_\varepsilon}[\lim_{N\to\infty} \rho_\tau(\varepsilon - \hat{s}_N)] \\ &\leq \lim \inf_{N\to\infty} \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - \hat{s}_N)] \\ &= \lim \inf_{N\to\infty} \mathbb{E}_{\mathbb{P}_{\varepsilon'}}[\rho_\tau(\varepsilon' - \hat{s}_N)] \\ &\leq \lim_{N\to\infty} \hat{J}_N \\ &= J^*. \end{aligned} \tag{B.14}$$

The second inequality follows by applying Fatou's Lemma. Hence, we complete the proof of Theorem 3.8.

B.4 Proof of Theorem 3.10

$$\begin{aligned} &\min_s \sup_{\mathbb{Q}_{\varepsilon'} \in \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}))} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)] \\ &= \begin{cases} \min_s \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{M}(E)} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)] \\ s.t. \quad d_w(\mathbb{Q}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq r_N \end{cases} \\ &= \begin{cases} \min_s \sup_{\Pi, \mathbb{Q}_{\varepsilon'}} \int_E \rho_\tau(\varepsilon' - s) \mathbb{Q}_{\varepsilon'}(d\varepsilon') \\ s.t. \int_{E\times E} |\varepsilon' - \xi| \Pi(d\varepsilon', d\xi) \leq r_N \\ \text{where } \varepsilon' \sim \mathbb{Q}_{\varepsilon'}, \xi \sim \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}), \text{and } \varepsilon', \xi \text{ has joint distribution } \Pi \end{cases} \end{aligned} \tag{B.15}$$

The first and second inequality holds due to the the definition of Wasserstein distance and our DRO formulation. Then let $\mathbb{Q}^i$ denote the conditional distributions of ε' given $\xi = \epsilon^{OLS}_i$, $\forall i = 1, \ldots, N$, i.e. $\Pi = \frac{1}{N} \sum_{i=1}^N \delta_{\epsilon^{OLS}_i} \otimes \mathbb{Q}^i$. Hence (B.15) can be further reformulated as:

$$(B.15) = \begin{cases} \min_s \sup_{\mathbb{Q}^i \in \mathcal{M}(E)} \frac{1}{N} \sum_{i=1}^N \int_E \rho_\tau(\varepsilon' - s) \mathbb{Q}^i(d\varepsilon') \\ s.t. \quad \frac{1}{N} \sum_{i=1}^N \int_E \|\varepsilon', \epsilon_i^{OLS}\| \mathbb{Q}^i(d\varepsilon') \leq r_N \end{cases} \tag{B.16}$$

$$= \min_s \sup_{\mathbb{Q}^i \in \mathcal{M}(E)} \inf_\lambda \frac{1}{N} \sum_{i=1}^N \int_E \rho_\tau(\varepsilon' - s) \mathbb{Q}^i(d\varepsilon') + \lambda(r_N - \sum_{i=1}^N \int_E \|\varepsilon' - \epsilon_i^{OLS}\| \mathbb{Q}^i(d\varepsilon')) \tag{B.17}$$

$$\leq \inf_{s,\lambda \geq 0} \sup_{\mathbb{Q}^i} \lambda r_N + \frac{1}{N} \sum_{i=1}^N \int_E (\rho_\tau(\varepsilon' - s) - \lambda |\varepsilon' - \epsilon_i^{OLS}|) \mathbb{Q}^i(d\varepsilon') \tag{B.18}$$

$$= \min_{s,\lambda \geq 0} \lambda r_N + \frac{1}{N} \sum_{i=1}^N \begin{cases} \rho_\tau(\epsilon_i^{OLS} - s) & \text{if } \lambda \geq \tau \text{ and } \lambda \geq 1 - \tau \\ \infty & \text{o.w.} \end{cases} \tag{B.19}$$

$$= \min_{s,\lambda \geq \max\{\tau, 1-\tau\}} \lambda r_N + \frac{1}{N} \sum_{i=1}^N \rho_\tau(\epsilon_i^{OLS} - s) \tag{B.20}$$

(B.17) follows from (B.16) by adding a dual multiplier λ. And the inequality (B.18) reduces to equality by strong duality which is guaranteed by Theorem 1 of [40].

Moreover, the $\mathbb{Q}^i$ that solves

$$\sup_{\mathbb{Q}^i} \frac{1}{N} \sum_{i=1}^N \int_E (\rho_\tau(\varepsilon' - s) - \lambda |\varepsilon' - \epsilon_i^{OLS}|) \mathbb{Q}^i(d\varepsilon') \tag{B.21}$$

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

$$\sup_{\varepsilon' \in R} \rho_\tau(\varepsilon' - s) - \lambda |\varepsilon' - \epsilon_i^{OLS}|. \tag{B.22}$$

As aforementioned, both $\rho_\tau(\cdot)$ and $\|\cdot\|$ are piece-wise linear functions, only their slops, i.e. τ, $\tau - 1$ and λ, play any role in determining the supremum of the function value. It can be easily shown that $\epsilon_i^* = \epsilon_i^{OLS}$ if $\lambda \geq 1 - \tau$ and $\lambda \geq \tau$. Otherwise, $\sup_\varepsilon \rho_\tau(\varepsilon' - s) - \lambda |\varepsilon' - \epsilon_i^{OLS}| = \infty$. Meanwhile, since this is an minimization problem and we can easily identify a feasible value (e.g. $\lambda = 1$) so that the objective value will be finite, the ∞ will never be achieved at optimal value.

Appendix C

Supporting Results for Chapter 4

C.1 Proof of Theorem 4.4

The justification of Theorem 4.4 depends on the following technical lemma.

Lemma C.1. *For any $k \geq 1$, and $\{\delta_i > 0\}$ such that $\sum_{i=1}^m \delta_i < 2 - \gamma$, we have*

$$\|\omega^k - \omega^*\|_R^2 \geq \|\omega^{k+1} - \omega^*\|_R^2 + \|\omega^{k+1} - \omega^k\|_Q^2. \tag{C.1}$$

Proof. By substituting the explicit expression of λ_p^{k+1} into Step 2 and considering the PPA interpretation of the dual update, the kth iteration of Algorithm 1 is equivalent to

$$\begin{cases} x^{k+1} = \arg\min_x \big(\phi(x) - \big\langle Ax, \lambda^k - \rho(Ax^k - b)\big\rangle + \frac{1}{2}\sum_{i=1}^m \|x_i - x_i^k\|_{P_i}^2\big), \\ \lambda^{k+1} = \arg\min_\lambda \big(\big\langle Ax^{k+1} - b, \lambda\big\rangle + \frac{1}{2\gamma\rho}\|\lambda - \lambda^k\|^2\big). \end{cases} \tag{C.2}$$

Then, according to Lemma 4.2, we can transform the optimization problems into the following variational inequality system

$$\begin{cases} \phi(x) - \phi(x^{k+1}) + \sum_{i=1}^m (x_i - x_i^{k+1})^T\big[- A_i^T\lambda^k + \rho A_i^T(Ax^k - b) + P_i(x_i^{k+1} - x_i^k)\big] \geq 0, \\ (\lambda - \lambda^{k+1})^T\big[(Ax^{k+1} - b) + \frac{1}{\gamma\rho}(\lambda^{k+1} - \lambda^k)\big] \geq 0. \end{cases} \tag{C.3}$$

Then, by substituting $\lambda^k = \lambda^{k+1} + \gamma\rho(Ax^{k+1} - b)$ into the first inequality results in

$$\begin{cases} \phi(x) - \phi(x^{k+1}) + \sum_{i=1}^m (x_i - x_i^{k+1})^T\big[- A_i^T\lambda^{k+1} - \rho A_i^T A(x^{k+1} - x^k) \\ \qquad\qquad - \frac{1-\gamma}{\gamma}A_i^T(\lambda^{k+1} - \lambda^k) + P_i(x_i^{k+1} - x_i^k)\big] \geq 0, \\ (\lambda - \lambda^{k+1})^T\big[(Ax^{k+1} - b) + \frac{1}{\gamma\rho}(\lambda^{k+1} - \lambda^k)\big] \geq 0. \end{cases} \tag{C.4}$$

By summing up these inequalities and rewrite in matrix form, it implies that

$$\phi(x) - \phi(x^{k+1}) + \big\langle \omega - \omega^{k+1}, F(\omega^{k+1}) + (R - G)(\omega^{k+1} - \omega^k)\big\rangle \geq 0, \tag{C.5}$$

which holds for all $\omega \in \Omega$. Thus, it is also true when we take $\omega = \omega^*$, i.e.

$$\phi(x^*) - \phi(x^{k+1}) + \left\langle \omega^* - \omega^{k+1}, F(\omega^{k+1}) + (R-G)(\omega^{k+1} - \omega^k) \right\rangle \geq 0. \tag{C.6}$$

On the other hand, let us reconsider (4.16) from Lemma 4.2 and evaluate it at $\omega = \omega^{k+1}$. It gives that

$$\phi(x^{k+1}) - \phi(x^*) + (\omega^{k+1} - \omega^*)^T F(\omega) \geq 0. \tag{C.7}$$

Adding the above two inequalities, we obtain

$$\left\langle \omega^* - \omega^{k+1}, R(\omega^{k+1} - \omega^k) \right\rangle \geq \left\langle \omega^{k+1} - \omega^*, G(\omega^k - \omega^{k+1}) \right\rangle. \tag{C.8}$$

Since, x^* is feasible to (P), it is always true that $Ax^* = b$ and hence, $\lambda^{k+1} = \lambda^k - \gamma\rho A(x^{k+1} - x^*)$. Thus, plug in the definition of G into the right-hand-side of the above inequality, we have

$$\begin{aligned}
\left\langle \omega^* - \omega^{k+1}, R(\omega^{k+1} - \omega^k) \right\rangle &\geq \rho(x^{k+1} - x^*)^T A^T A(x^k - x^{k+1}) + \frac{1-\gamma}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2 \\
&= \sum_{i=1}^m \big[\frac{1}{\gamma}(\lambda^{k+1} - \lambda^k)A_i(x_i^{k+1} - x_i^k)\big] + \frac{1-\gamma}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2 \\
&\geq -\frac{1}{2}\sum_{i=1}^m \big[\frac{\delta_i}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2 + \frac{\rho}{\delta_i}\|A_i(x_i^{k+1} - x_i^k)\|^2\big] \\
&\quad + \frac{1-\gamma}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2
\end{aligned} \tag{C.9}$$

where the second inequality follows from the triangle inequality. And hence, the claim in Lemma C.1 holds as

$$\begin{aligned}
\|\omega^k - \omega^*\|_R^2 &= \|\omega^{k+1} - \omega^k\|_R^2 + \|\omega^{k+1} - \omega^*\|_R^2 + 2\langle \omega^{k+1} - \omega^*, R(\omega^k - \omega^{k+1})\rangle \\
&\geq \|\omega^{k+1} - \omega^k\|_R^2 + \|\omega^{k+1} - \omega^*\|_R^2 + \frac{2(1-\gamma) - \sum_{i=1}^m \delta_i}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2 \\
&\quad - \rho\sum_{i=1}^m (\frac{1}{\delta_i}\|x_i^{k+1} - x_i^k\|_{A_i^T A_i}^2) \\
&= \|\omega^{k+1} - \omega^*\|_R^2 + \sum_{i=1}^m \|x_i^{k+1} - x_i^k\|_{(P_i - \frac{\rho}{\delta_i}A_i^T A_i)}^2 + \frac{2 - \gamma - \sum_{i=1}^m \delta_i}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2 \\
&= \|\omega^{k+1} - \omega^*\|_R^2 + \|\omega^{k+1} - \omega^k\|_Q^2 + \frac{2 - \gamma - \sum_{i=1}^m \delta_i}{\gamma^2\rho}\|\lambda^{k+1} - \lambda^k\|^2,
\end{aligned} \tag{C.10}$$

where last term is positive when $\sum_{i=1}^m \delta_i < 2 - \gamma$. □

Finally, when the condition of Theorem 4.4 holds, i.e. $P_i \succ \frac{\rho}{\delta_i}A_i^T A_i$, it is easy to verify that $Q \succ 0$ and $R \succ 0$. Thus, Lemma C.1 implies that

$$\|\omega^{k+1} - \omega^*\|_R^2 \leq \|\omega^k - \omega^*\|_R^2, \tag{C.11}$$

meaning that the sequence $\{\|\omega^k - \omega^*\|_R^2\}$ is monotonically non-increasing. Furthermore, $\{\omega^k\}$ is bounded and has at least one convergent subsequence (Bolzano-Weierstrass theorem). Hence, we complete the proof for Theorem 4.4.

C.2 Proof of Lemma 4.5

i) Similar to (C.3), let us consider the optimality condition for updating each block of x_i at the two successive iterations, the $k-1$th and kth iterations, we have

$$\theta_i(x_i) - \theta_i(x_i^k) + (x_i - x_i^k)^T\big[-A_i^T\lambda^{k-1} + \rho A_i^T(Ax^{k-1} - b) + P_i(x_i^k - x_i^{k-1})\big] \geq 0 \quad \text{(C.12)}$$

and

$$\theta_i(x_i) - \theta_i(x_i^{k+1}) + (x_i - x_i^{k+1})^T\big[-A_i^T\lambda^{k} + \rho A_i^T(Ax^{k} - b) + P_i(x_i^{k+1} - x_i^{k})\big] \geq 0, \quad \text{(C.13)}$$

for all $\omega \in \Omega$. By Evaluating these two inequalities as $\omega = \omega^{k+1}$ and $\omega = \omega^k$ respectively and summing them together, we derive that

$$(\Delta x_i^{k+1})^T[A_i^T\Delta\lambda^k - \rho A_i^T A\Delta x^k + P_i(\Delta x_i^k - \Delta x_i^{k+1})] \geq 0, \quad \text{(C.14)}$$

where $\Delta\omega^{k+1} := \omega^k - \omega^{k+1}$. Then, summing up over all i and rearranging the terms leads to

$$\langle A\Delta x^{k+1}, \Delta\lambda^k\rangle \geq \|\Delta x^{k+1}\|_P^2 - (\Delta x^k)^T(P - \rho A^T A)\Delta x^{k+1}, \quad \text{(C.15)}$$

where

$$P = \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_m \end{pmatrix}. \quad \text{(C.16)}$$

Since $Q \succ 0$, it follows that $P_x := P - \rho A^T A \succeq 0$, we have

$$2(\Delta x^k)^T P_x \Delta x^{k+1} \leq \|\Delta x^k\|_{P_x}^2 + \|\Delta x^{k+1}\|_{P_x}^2 \quad \text{(C.17)}$$

by triangular inequality and thus,

$$2\langle A\Delta x^{k+1}, \Delta\lambda^k\rangle \geq \|\Delta x^{k+1}\|_{P+\rho A^T A}^2 - \|\Delta x^k\|_{P_x}^2. \quad \text{(C.18)}$$

Note that $\Delta\lambda^{k+1} = \Delta\lambda^k - \gamma\rho A\Delta x^{k+1}$, it follows that

$$\begin{aligned} \frac{1}{\gamma\rho}(\|\Delta\lambda^k\|^2 - \|\Delta\lambda^{k+1}\|^2) &= 2\langle A\Delta x^{k+1}, \Delta\lambda^k\rangle - \gamma\rho\|A\Delta x^{k+1}\|^2 \\ &\geq \|\Delta x^{k+1}\|_{(P_x+(2-\gamma)\rho A^T A)}^2 - \|\Delta x^k\|_{P_x}^2. \end{aligned} \quad \text{(C.19)}$$

Hence,

$$(\|\Delta x^k\|_{P_x}^2 + \frac{1}{\gamma\rho}\|\Delta\lambda^k\|^2) - (\|\Delta x^{k+1}\|_{P_x}^2 + \frac{1}{\gamma\rho}(\|\Delta\lambda^{k+1}\|^2) \geq \|\Delta x^{k+1}\|_{(2-\gamma)\rho A^T A}^2 \geq 0. \quad \text{(C.20)}$$

The assertion i), thus, follows immediately by rearranging terms.

ii) By Lemma C.1, there exists some $\eta > 0$ such that

$$\|\omega^k - \omega^*\|_R^2 - \|\omega^{k+1} - \omega^*\|_R^2 \geq \|\omega^{k+1} - \omega^k\|_Q^2 \geq \eta\|\omega^{k+1} - \omega^k\|_W^2. \tag{C.21}$$

Summing over all iterations gives that $\sum_{k=1}^{\infty} \|\omega^{k+1} - \omega^k\|_W^2 < \infty$. On the other hand, the sequence $\{\|\omega^{k+1} - \omega^k\|_W^2\}$ is monotonically non-increasing. Thus, by Lemma 4.3, we have $\|\omega^{k+1} - \omega^k\|_W^2 = o(1/k)$ and conclude that EPCPM is convergent with rate $o(1/k)$ in a non-ergodic sense.

C.3 Proof of Theorem 4.6

The conditions proposed in Theorem 4.6 are based on the following two technical lemmas.

Lemma C.2. *(from [27]) Let $F(\cdot)$ be a closed proper convex function and*

$$u^{k+1} = \arg\min_u \{F(u) + \frac{\tau}{2}\|u - u^k\|^2\}. \tag{C.22}$$

Then, for all $k \geq 0$, we have

$$F(u^{k+1}) - F(u) \leq \frac{\tau}{2}\big(\|u^k - u\|^2 - \|u^{k+1} - u\|^2 - \|u^{k+1} - u^k\|^2\big),\ \forall u. \tag{C.23}$$

Lemma C.3. *If the sequences $\{\omega^k\}$ and $\{\lambda_p^{k+1}\}$ are generated from SPCPM, then for all $k \geq 0$, we have*

i)

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \|x^{k+1} - x^k\|^2 + \frac{2}{\tau}\langle\lambda_p^{k+1} - \lambda^*, Ax^{k+1} - b\rangle; \tag{C.24}$$

ii)

$$\begin{aligned}\|\lambda^{k+1} - \lambda^*\|^2 \leq & \|\lambda^k - \lambda^*\|^2 - \{\|\lambda_p^{k+1} - \lambda^{k+1}\|^2 + \|\lambda_p^{k+1} - \lambda^k\|^2\} \\ & + \frac{2}{\tau}\{\langle\lambda^{k+1} - \lambda_p^{k+1}, Ax^k - b\rangle + \langle\lambda^* - \lambda^{k+1}, Ax^{k+1} - b\rangle\}.\end{aligned} \tag{C.25}$$

Proof. Observing that Step 2 of SPCPM is essentially the application of PPA to Lagrangian function, $L(x, \lambda_p^{k+1})$, we can apply Lemma C.2 and evaluate it at $x = x^*$. Then, we obtain

$$L(x^{k+1}, \lambda_p^{k+1}) - L(x^*, \lambda_p^{k+1}) \leq \frac{2}{\tau}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 - \|x^{k+1} - x^k\|^2). \tag{C.26}$$

On the other hand, since ω^* is a stationary point for $L(x, \lambda)$, we also have

$$L(x^*, \lambda_p^{k+1}) - L(x^{k+1}, \lambda^*) \leq 0. \tag{C.27}$$

Assertion i) holds by summing up the above two inequalities and rearranging terms.

To prove assertion *ii*), we start with the PPA interpretation of the dual updates, i.e. Step 1 and Step 3 of SPCPM are equivalent to solving

$$\lambda_p^{k+1} = \arg\min_{\lambda}\{-L(x^k,\lambda) + \frac{\tau}{2}\|\lambda-\lambda^k\|^2\} \tag{C.28}$$

$$\lambda^{k+1} = \arg\min_{\lambda}\{-L(x^{k+1},\lambda) + \frac{\tau}{2}\|\lambda-\lambda^k\|^2\} \tag{C.29}$$

Then, if we apply Lemma C.2 with $F(\lambda) = -L(x^k,\lambda)$ and $F(\lambda) = -L(x^{k+1},\lambda)$, and further evaluate the inequalities at λ^{k+1} and λ^* respectively, it holds that

$$\frac{2}{\tau}[L(x^k,\lambda^{k+1}) - L(x^k,\lambda_p^{k+1})] \le \|\lambda^k-\lambda^{k+1}\|^2 - \|\lambda_p^{k+1}-\lambda^{k+1}\|^2 - \|\lambda_p^{k+1}-\lambda^k\|^2 \tag{C.30}$$

and

$$\frac{2}{\tau}[L(x^{k+1},\lambda^*) - L(x^{k+1},\lambda^{k+1})] \le \|\lambda^k-\lambda^*\|^2 - \|\lambda^{k+1}-\lambda^*\|^2 - \|\lambda^{k+1}-\lambda^k\|^2. \tag{C.31}$$

Summing the above two inequalities gives that

$$\begin{aligned}&\frac{2}{\tau}[L(x^k,\lambda^{k+1}) - L(x^k,\lambda_p^{k+1}) + L(x^{k+1},\lambda^*) - L(x^{k+1},\lambda^{k+1})]\\ \le&\|\lambda^k-\lambda^*\|^2 - \|\lambda^{k+1}-\lambda^*\|^2 - \|\lambda_p^{k+1}-\lambda^{k+1}\|^2 - \|\lambda_p^{k+1}-\lambda^k\|^2.\end{aligned} \tag{C.32}$$

At the same time, recall operations at Step 1 and Step 3, which leads to

$$\begin{aligned}&L(x^k,\lambda^{k+1}) - L(x^k,\lambda_p^{k+1}) + L(x^{k+1},\lambda^*) - L(x^{k+1},\lambda^{k+1})\\ =&\langle\lambda^{k+1}-\lambda_p^{k+1}, Ax^k-b\rangle + \langle\lambda^*-\lambda^{k+1}, Ax^{k+1}-b\rangle.\end{aligned} \tag{C.33}$$

The proof of assertion *ii*) holds by substituting (C.33) into (C.32) and rearranging terms. □

To complete the proof of Theorem 4.6, let us define

$$\begin{aligned}\Delta &= -\frac{2}{\tau}\langle\lambda^{k+1}-\lambda_p^{k+1}, A(x^{k+1}-x^k)\rangle\\ &= \frac{2}{\tau}\|A(x^{k+1}-x^k)\|^2\\ &\le 2\big(\frac{1}{\tau}\|A\|\big)\|x^{k+1}-x^k\|^2,\end{aligned} \tag{C.34}$$

where the last equation follows by Cauchy–Schwarz inequality. Finally, adding the two inequalities in Lemma C.3 results in

$$\begin{aligned}\|\omega^{k+1}-\omega^*\|^2 &\le \|\omega^k-\omega^*\|^2 - \|x^{k+1}-x^k\|^2 - \{\|\lambda_p^{k+1}-\lambda^{k+1}\|^2 + \|\lambda_p^{k+1}-\lambda^k\|^2\} + \Delta\\ &\le \|\omega^k-\omega^*\|^2 - \big[1-2(\frac{1}{\tau}\|A\|)^2\big]\|x^{k+1}-x^k\|^2\\ &\quad - \{\|\lambda_p^{k+1}-\lambda^{k+1}\|^2 + \|\lambda_p^{k+1}-\lambda^k\|^2\}.\end{aligned} \tag{C.35}$$

Hence, to ensure that SPCPM is convergent, it suffices to have the coefficient

$$\begin{aligned} & 1 - 2(\frac{1}{\tau}\|A\|)^2 \geq 0 \\ \Leftrightarrow \quad & \tau \geq \sqrt{2}\|A\|. \end{aligned} \tag{C.36}$$

And in many cases, we have $\sqrt{2}\|A\| < \max_i\{\sqrt{m}\|A_i\|\}$, leading to faster convergence than parameters chosen according to Theorem 4.4.

C.4 Proof of Theorem 4.7

We first show the following lemma as some intermediate result that can be used to prove this statement in Theorem 4.7.

Lemma C.4. *For any $k \geq 0$,*

$$\|\hat{\omega}^{k+1} - \tilde{\omega}^{k+1}\| \leq \sqrt{\beta \sum_{i=1}^{m} \frac{\epsilon_i^k}{\tau_i}}, \tag{C.37}$$

where $\beta = 2[1 + (\gamma\rho\|A\|)^2]$.

Proof. For any $k \geq 0$, we define the subproblems for primal updates as

$$\Psi_k(x_i) = \theta_i(x_i) - (\hat{\lambda}_p^{k+1})^T A_i x_i + \frac{\tau_i}{2}\|x_i - \hat{x}_i^k\|^2, \ \forall i. \tag{C.38}$$

Then, by the definition of $\tilde{x}_i^{k+1}$, we have $0 \in \partial\Psi_k(\tilde{x}_i^{k+1})$. While by the definition of ϵ-optimality, we have

$$0 \leq \Psi_k(\hat{x}_i^{k+1}) - \Psi_k(\tilde{x}_i^{k+1}) \leq \epsilon_i^k, \ \forall i. \tag{C.39}$$

Moreover, $\Psi_k(x_i)$ is strongly convex with modulus τ_i (see [95], Proposition 6). Hence,

$$\Psi_k(\hat{x}_i^{k+1}) - \Psi_k(\tilde{x}_i^{k+1}) \geq \frac{\tau_i}{2}\|\hat{x}_i^{k+1} - \tilde{x}_i^{k+1}\|^2. \tag{C.40}$$

Therefore,

$$\|\hat{x}_i^{k+1} - \tilde{x}_i^{k+1}\|^2 \leq \frac{2\epsilon_i^k}{\tau_i}. \tag{C.41}$$

Meanwhile, observing Step 3 of Algorithm 3, we have

$$\|\hat{\lambda}^{k+1} - \tilde{\lambda}^{k+1}\|^2 = \|\gamma\rho A(\hat{x}^{k+1} - \tilde{x}^{k+1})\|^2 \leq (\gamma\rho\|A\|)^2\|\hat{x}^{k+1} - \tilde{x}^{k+1}\|^2, \tag{C.42}$$

by Cauchy–Schwarz inequality. Finally,

$$\begin{aligned}\|\hat{\omega}^{k+1}-\tilde{\omega}^{k+1}\|^2 &= \|\hat{x}^{k+1}-\tilde{x}^{k+1}\|^2+\|\hat{\lambda}^{k+1}-\tilde{\lambda}^{k+1}\|^2\\ &\leq [1+(\gamma\rho\|A\|)^2]\sum_{i=1}^{m}\|\hat{x}_i^{k+1}-\tilde{x}_i^{k+1}\|^2\\ &\leq 2[1+(\gamma\rho\|A\|)^2]\sum_{i=1}^{m}\frac{\epsilon_i^k}{\tau_i}.\end{aligned} \tag{C.43}$$

□

To obtain the results in Theorem 4.7, we first apply the triangle inequality and get

$$\|\hat{\omega}^{k+1}-\omega^*\| \leq \|\hat{\omega}^{k+1}-\tilde{\omega}^{k+1}\|+\|\tilde{\omega}^{k+1}-\omega^*\|. \tag{C.44}$$

On the other hand, since $\tilde{\omega}^{k+1}$ can be treated as results from a single iteration of EPCPM with initial solution $\hat{\omega}^k$, it follows from Lemma C.1 that the contraction property holds:

$$\|\tilde{\omega}^{k+1}-\omega^*\| \leq \|\hat{\omega}^{k}-\omega^*\|. \tag{C.45}$$

On that account, and apply Lemma C.4, we have

$$\begin{aligned}\|\hat{\omega}^{k+1}-\omega^*\| &\leq \|\hat{\omega}^{k+1}-\tilde{\omega}^{k+1}\|+\|\hat{\omega}^{k}-\omega^*\|\\ &\leq \sqrt{\beta\sum_{i=1}^{m}\frac{\epsilon_i^k}{\tau_i}}+\|\hat{\omega}^{k}-\omega^*\|.\end{aligned} \tag{C.46}$$

When $\lim_{k\to\infty}\epsilon_i^k=0,\ i=1,2,\ldots,m$, it is a matter of course that

$$\lim_{k\to\infty}\sqrt{\beta\sum_{i=1}^{m}\frac{\epsilon_i^k}{\tau_i}}=0. \tag{C.47}$$

Henceforth, inequality (C.46) implies that sequence $\{\hat{\omega}^k\}$ is bounded, and the existence of

$$\lim_{k\to\infty}\|\hat{\omega}^{k+1}-\omega^*\|=\mu<\infty. \tag{C.48}$$

Furthermore, it is also implied by Lemma C.4 that

$$\lim_{k\to\infty}\|\hat{\omega}^{k+1}-\tilde{\omega}^{k+1}\|=0, \tag{C.49}$$

and thus,

$$\lim_{k\to\infty}\|\tilde{\omega}^{k+1}-\omega^*\|=\mu<\infty. \tag{C.50}$$

Finally, following exactly the same argument as [27] (we omit the duplication here), it can be seen that sequence $\{\tilde{\omega}^k\}$ has a unique limit point, which is a solution to problem (P).

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] A. Alexandre Trindade and Yanxun Xu. "Quantile versions of Holt-Winters forecasting algorithms". In: *Journal of Statistics: Advances in Theory and Applications* 5.1 (2011), pp. 15–35.

[3] Alexander Aue et al. "Piecewise quantile autoregressive modeling for nonstationary time series". In: *Bernoulli* 23.1 (2017), pp. 1–22.

[4] Francis Bach et al. "Structured sparsity through convex optimization". In: *Statistical Science* (2012), pp. 450–468.

[5] Qingguo Bai and Mingyuan Chen. "The distributionally robust newsvendor problem with dual sourcing under carbon tax and cap-and-trade regulations". In: *Computers & Industrial Engineering* 98 (2016), pp. 260–274.

[6] Gah-Yi Ban. "The Data-Driven (s, S) Policy: Why You Can Have Confidence in Censored Demand Data". In: (2015).

[7] Gah-Yi Ban and Cynthia Rudin. "The big data newsvendor: Practical insights from machine learning". In: *Operations Research* (2018).

[8] Johnathan M Bardsley, Sarah Knepper, and James Nagy. "Structured linear algebra problems in adaptive optics imaging". In: *Advances in Computational Mathematics* 35.2-4 (2011), p. 103.

[9] Alain Bensoussan, Metin Çakanyıldırım, and Suresh P Sethi. "A multiperiod newsvendor problem with partially observed demand". In: *Mathematics of Operations Research* 32.2 (2007), pp. 322–344.

[10] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. "Robust sample average approximation". In: *Mathematical Programming* (2017), pp. 1–66.

[11] Dimitris Bertsimas and Aurélie Thiele. *A data-driven approach to newsvendor problems*. Tech. rep. Technical report, Massechusetts Institute of Technology, Cambridge, MA, 2005.

[12] 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.

[13] Arnab Bisi, Karanjit Kalsi, and Golnaz Abdollahian. "A non-parametric adaptive algorithm for the censored newsvendor problem". In: *IIE Transactions* 47.1 (2015), pp. 15–34.

[14] Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine Learning* 3.1 (2011), pp. 1–122.

[15] Apostolos N Burnetas and Craig E Smith. "Adaptive ordering and pricing for perishable products". In: *Operations Research* 48.3 (2000), pp. 436–443.

[16] Xingju Cai et al. "A relaxed customized proximal point algorithm for separable convex programming". In: *Optimization Online* (2011).

[17] A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[18] Alex J Cannon. "Quantile regression neural networks: Implementation in R and application to precipitation downscaling". In: *Computers & Geosciences* 37.9 (2011), pp. 1277–1284.

[19] 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.

[20] Emilio Carrizosa, Alba V Olivares-Nadal, and Pepa Ramírez-Cobo. "Robust newsvendor problem with autoregressive demand". In: *Computers & Operations Research* 68 (2016), pp. 123–133.

[21] Antonin Chambolle and Thomas Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of mathematical imaging and vision* 40.1 (2011), pp. 120–145.

[22] Xiaokai Chang et al. "Convergent prediction–correction-based ADMM for multi-block separable convex programming". In: *Journal of Computational and Applied Mathematics* 335 (2018), pp. 270–288.

[23] Miantao Chao and Caozong Cheng. "A note on the convergence of alternating proximal gradient method". In: *Applied Mathematics and Computation* 228 (2014), pp. 258–263.

[24] Caihua Chen, Yuan Shen, and Yanfei You. "On the convergence analysis of the alternating direction method of multipliers with three blocks". In: *Abstract and Applied Analysis*. Vol. 2013. Hindawi. 2013.

[25] Caihua Chen et al. "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent". In: *Mathematical Programming* 155.1-2 (2016), pp. 57–79.

[26] Colin Chen. "A finite smoothing algorithm for quantile regression". In: *Journal of Computational and Graphical Statistics* 16.1 (2007), pp. 136–164.

[27] Gong Chen and Marc Teboulle. "A proximal-based decomposition method for convex minimization problems". In: *Mathematical Programming* 64.1-3 (1994), pp. 81–101.

[28] Xiaohong Chen, Roger Koenker, and Zhijie Xiao. "Copula-based nonlinear quantile autoregression". In: *The Econometrics Journal* 12.s1 (2009), S50–S67.

[29] Carroll Croarkin, Paul Tobias, and Chelli Zey. *Engineering statistics handbook*. NIST iTL, 2002.

[30] Wei Deng et al. "Parallel multi-block ADMM with O (1/k) convergence". In: *Journal of Scientific Computing* 71.2 (2017), pp. 712–736.

[31] Lingxiu Dong and Hau L Lee. "Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand". In: *Operations Research* 51.6 (2003), pp. 969–980.

[32] Jonathan Eckstein and Dimitri P Bertsekas. "On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators". In: *Mathematical Programming* 55.1-3 (1992), pp. 293–318.

[33] Jonathan Eckstein and W Yao. "Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results". In: *RUTCOR Research Reports* 32 (2012), p. 3.

[34] 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.

[35] 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.

[36] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. "Analytics for an online retailer: Demand forecasting and price optimization". In: *Manufacturing & Service Operations Management* 18.1 (2015), pp. 69–88.

[37] 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.

[38] Xiaoling Fu et al. *Block-wise alternating direction method of multipliers with Gaussian back substitution for multiple-block convex programming*. Tech. rep. Technical Report, 2014.

[39] Daniel Gabay and Bertrand Mercier. "A dual algorithm for the solution of nonlinear variational problems via finite element approximation". In: *Computers & Mathematics with Applications* 2.1 (1976), pp. 17–40.

[40] Rui Gao and Anton J Kleywegt. "Distributionally robust stochastic optimization with Wasserstein distance". In: *arXiv preprint arXiv:1604.02199* (2016).

[41] René Gélinas and Pierre Lefrançois. "On the estimation of time-series quantiles using smoothed order statistics". In: *International Journal of Forecasting* 9.2 (1993), pp. 227–243.

[42] Roland Glowinski and A Marroco. "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires". In: *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique* 9.R2 (1975), pp. 41–76.

[43] Gregory A Godfrey and Warren B Powell. "An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution". In: *Management Science* 47.8 (2001), pp. 1101–1112.

[44] Siddharth Gopal and Yiming Yang. "Distributed training of large-scale logistic models". In: *International Conference on Machine Learning*. 2013, pp. 289–297.

[45] Stephen C Graves. "A single-item inventory model for a nonstationary demand process". In: *Manufacturing & Service Operations Management* 1.1 (1999), pp. 50–61.

[46] Guoyong Gu, Bingsheng He, and Xiaoming Yuan. "Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: a unified approach". In: *Computational Optimization and Applications* 59.1-2 (2014), pp. 135–161.

[47] Meihui Guo, Zhidong Bai, and Hong Zhi An. "Multi-step prediction for nonlinear autoregressive models based on empirical distributions". In: *Statistica Sinica* (1999), pp. 559–570.

[48] Deren Han and Xiaoming Yuan. "A note on the alternating direction method of multipliers". In: *Journal of Optimization Theory and Applications* 155.1 (2012), pp. 227–238.

[49] Grani A Hanasusanto et al. "Distributionally robust multi-item newsvendor problems with multimodal demand distributions". In: *Mathematical Programming* 152.1-2 (2015), pp. 1–32.

[50] 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.

[51] Fumio Hayashi. *Econometrics*. 2000.

[52] Bing-Sheng He. "PPA-like contraction methods for convex optimization: a framework using variational inequality approach". In: *Journal of the Operations Research Society of China* 3.4 (2015), pp. 391–420.

[53] Bingsheng He. "A class of projection and contraction methods for monotone variational inequalities". In: *Applied Mathematics and optimization* 35.1 (1997), pp. 69–76.

[54] Bingsheng He. "Modified alternating directions method of multipliers for convex optimization with three separable functions". In: *Operations Research Transactions* 19.3 (2015), pp. 57–70.

[55] Bingsheng He, Min Tao, and Xiaoming Yuan. "A splitting method for separable convex programming". In: *IMA Journal of Numerical Analysis* 35.1 (2014), pp. 394–426.

[56] Bingsheng He, Min Tao, and Xiaoming Yuan. "Alternating direction method with Gaussian back substitution for separable convex programming". In: *SIAM Journal on Optimization* 22.2 (2012), pp. 313–340.

[57] Bingsheng He and Xiaoming Yuan. "On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers". In: *Numerische Mathematik* 130.3 (2015), pp. 567–577.

[58] Bingsheng He and Xiaoming Yuan. "On the O(1/n) Convergence Rate of the Douglas–Rachford Alternating Direction Method". In: *SIAM Journal on Numerical Analysis* 50.2 (2012), pp. 700–709.

[59] Bingsheng He, Xiaoming Yuan, and Wenxing Zhang. "A customized proximal point algorithm for convex minimization with linear constraints". In: *Computational Optimization and Applications* 56.3 (2013), pp. 559–572.

[60] Bingsheng He et al. "A new inexact alternating directions method for monotone variational inequalities". In: *Mathematical Programming* 92.1 (2002), pp. 103–118.

[61] BS He and XM Yuan. "A contraction method with implementable proximal regularization for linearly constrained convex programming". In: *Optimization Online* (2010), pp. 1–14.

[62] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2 (1991), pp. 251–257.

[63] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[64] Woonghee Tim Huh and Paat Rusmevichientong. "A nonparametric asymptotic analysis of inventory planning with censored demand". In: *Mathematics of Operations Research* 34.1 (2009), pp. 103–123.

[65] Woonghee Tim Huh et al. "Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator". In: *Operations Research* 59.4 (2011), pp. 929–941.

[66] L.V. Kantorovich and G.S. Rubinshtein. "On a space of totally additive functions". In: *Vestn.Leningr.Univ.* 13.52-59 (1958).

[67] Moutaz Khouja. "The single-period (news-vendor) problem: literature review and suggestions for future research". In: *Omega* 27.5 (1999), pp. 537–553.

[68] Rüdiger Kiesel. *Foundations of Modern Probability: Probability and Its Applications*. 1999.

[69] Tae Yoon Kim et al. "Artificial neural networks for non-stationary time series". In: *Neurocomputing* 61 (2004), pp. 439–447.

[70] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[71] Keith Knight. "Limiting Distributions for L1 Regression Estimators under General Conditions". In: *Annals of Statistics* (1998), pp. 755–770.

[72] Roger Koenker and Gilbert Bassett Jr. "Regression quantiles". In: *Econometrica: journal of the Econometric Society* (1978), pp. 33–50.

[73] Roger Koenker and Zhijie Xiao. "Quantile autoregression". In: *Journal of the American Statistical Association* 101.475 (2006), pp. 980–990.

[74] Friedrich Leisch, Adrian Trapletti, and Kurt Hornik. "On the stationarity of autoregressive neural network models". In: (1998).

[75] 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.

[76] 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.

[77] Tatsiana Levina et al. "Weak aggregating algorithm for the distribution-free perishable inventory problem". In: *Operations Research Letters* 38.6 (2010), pp. 516–521.

[78] 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.

[79] Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. "Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity". In: *Journal of Scientific Computing* 69.1 (2016), pp. 52–81.

[80] Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. "On the global linear convergence of the admm with multiblock variables". In: *SIAM Journal on Optimization* 25.3 (2015), pp. 1478–1497.

[81] Lanchao Liu and Zhu Han. "Multi-block ADMM for big data optimization in smart grid". In: *Computing, Networking and Communications (ICNC), 2015 International Conference on*. IEEE. 2015, pp. 556–561.

[82] Xiaochun Liu. "Markov switching quantile autoregression". In: *Statistica Neerlandica* 70.4 (2016), pp. 356–395.

[83] 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.

[84] Shiqian Ma. "Alternating proximal gradient method for convex minimization". In: *preprint* (2012).

[85] Gonzalo Mateos, Juan Andrés Bazerque, and Georgios B Giannakis. "Distributed sparse linear regression". In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5262–5276.

[86] Brian R Mitchell and B Bruce Bare. "A separable goal programming approach to optimizing multivariate sampling designs for forest inventory". In: *Forest Science* 27.1 (1981), pp. 147–162.

[87] Whitney K Newey and Daniel McFadden. "Large sample estimation and hypothesis testing". In: *Handbook of econometrics* 4 (1994), pp. 2111–2245.

[88] Rong Pan. "Holt–Winters Exponential Smoothing". In: *Wiley Encyclopedia of Operations Research and Management Science* (2010).

[89] Georgia Perakis and Guillaume Roels. "Regret in the newsvendor model with partial information". In: *Operations Research* 56.1 (2008), pp. 188–203.

[90] Robert S Pindyck and Daniel L Rubinfeld. *Econometric models and economic forecasts*. Vol. 4. Irwin/McGraw-Hill Boston, 1998.

[91] Min Qi and G Peter Zhang. "Trend time–series modeling and forecasting with neural networks". In: *IEEE Transactions on neural networks* 19.5 (2008), pp. 808–816.

[92] 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.

[93] Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de-Mello. "Distributionally Robust Newsvendor Problems with Variation Distance". In: 2017.

[94] R Tyrrell Rockafellar. "Augmented Lagrangians and applications of the proximal point algorithm in convex programming". In: *Mathematics of operations research* 1.2 (1976), pp. 97–116.

[95] R Tyrrell Rockafellar. "Monotone operators and the proximal point algorithm". In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.

[96] Anna-Lena Sachs. "The data-driven newsvendor with censored demand observations". In: *Retail Analytics*. Springer, 2015, pp. 35–56.

[97] Herbert Scarf. "A min-max solution of an inventory problem". In: *Studies in the mathematical theory of inventory and production* (1958).

[98] Chuen-Teck See and Melvyn Sim. "Robust approximation to multiperiod inventory management". In: *Operations research* 58.3 (2010), pp. 583–594.

[99] Ron Shefi and Marc Teboulle. "Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization". In: *SIAM Journal on Optimization* 24.1 (2014), pp. 269–297.

[100] Cong Shi, Weidong Chen, and Izak Duenyas. "Nonparametric Data-Driven Algorithms for Multiproduct Inventory Systems with Censored Demand". In: *Operations Research* 64.2 (2016), pp. 362–370.

[101] Lawrence V Snyder and Zuo-Jun Max Shen. *Fundamentals of supply chain theory*. John Wiley & Sons, 2011.

[102] Masashi Sugiyama and Amos J Storkey. "Mixture regression for covariate shift". In: *Advances in Neural Information Processing Systems*. 2007, pp. 1337–1344.

[103] James W Taylor. "A quantile regression neural network approach to estimating the conditional density of multiperiod returns". In: *Journal of Forecasting* 19.4 (2000), pp. 299–311.

[104] Adrian Trapletti, Friedrich Leisch, and Kurt Hornik. "Stationary and integrated autoregressive neural network processes". In: *Neural Computation* 12.10 (2000), pp. 2427–2450.

[105] Fang-Mei Tseng, Hsiao-Cheng Yu, and Gwo-Hsiung Tzeng. "Combining neural network model with seasonal time series ARIMA model". In: *Technological Forecasting and Social Change* 69.1 (2002), pp. 71–87.

[106] Arthur F Veinott Jr. "Optimal policy for a multi-product, dynamic, nonstationary inventory problem". In: *Management Science* 12.3 (1965), pp. 206–222.

[107] Grace Wahba et al. "Adaptive tuning of numerical weather prediction models: Randomized GCV in three-and four-dimensional data assimilation". In: *Monthly Weather Review* 123.11 (1995), pp. 3358–3370.

[108] Charles X Wang, Scott Webster, and Sidong Zhang. "Robust price-setting newsvendor model with interval market size and consumer willingness-to-pay". In: *International Journal of Production Economics* 154 (2014), pp. 100–112.

[109] Jian Wang et al. "Convergence of gradient method for double parallel feedforward neural network". In: *Int J Numer Anal Model* 8 (2011), pp. 484–495.

[110] Junxiang Wang and Liang Zhao. "Nonconvex generalizations of ADMM for nonlinear equality constrained problems". In: *arXiv preprint arXiv:1705.03412* (2017).

[111] Xiangfeng Wang et al. "Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers". In: *arXiv preprint arXiv:1308.5294* (2013).

[112] Yu Wang, Wotao Yin, and Jinshan Zeng. "Global convergence of ADMM in nonconvex nonsmooth optimization". In: *Journal of Scientific Computing* (2015), pp. 1–35.

[113] 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.

[114] Steven C Wheelwright, Spyros G Makridakis, et al. *Forecasting methods for management*. Wiley, 1973.

[115] Halbert White. "Nonparametric estimation of conditional quantiles using neural networks". In: *Computing Science and Statistics*. Springer, 1992, pp. 190–199.

[116] Linwei Xin and David A Goldberg. "Distributionally robust inventory control when demand is a martingale". In: *arXiv preprint arXiv:1511.09437* (2015).

[117] Qifa Xu et al. "Quantile autoregression neural network model with applications to evaluating value at risk". In: *Applied Soft Computing* 49 (2016), pp. 1–12.

[118] Zheng Xu et al. "An empirical study of ADMM for nonconvex problems". In: *arXiv preprint arXiv:1612.03349* (2016).

[119] Jia Zhai, Hui Yu, and Caihong Sun. "Robust optimization for the newsvendor problem with discrete demand". In: *Mathematical Problems in Engineering* 2018 (2018).

[120] G Peter Zhang. "Time series forecasting using a hybrid ARIMA and neural network model". In: *Neurocomputing* 50 (2003), pp. 159–175.

[121] G Peter Zhang and Min Qi. "Neural network forecasting for seasonal and trend time series". In: *European journal of operational research* 160.2 (2005), pp. 501–514.

[122] Yong Zhang, Vladimir Vovk, and Weiguo Zhang. "Probability-free solutions to the non-stationary newsvendor problem". In: *Annals of Operations Research* 223.1 (2014), pp. 433–449.

[123] Mingqiang Zhu and Tony Chan. "An efficient primal-dual hybrid gradient algorithm for total variation image restoration". In: *UCLA CAM Report* 34 (2008).

[124] Zhisu Zhu, Jiawei Zhang, and Yinyu Ye. "Newsvendor optimization with limited distribution information". In: *Optimization methods and software* 28.3 (2013), pp. 640–667.