Online and Offline Learning Algorithms in Operations Management

by

Jingwen Tang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering and Scientific Computing)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Cong Shi, Co-Chair
Professor Xiuli Chao, Co-Chair
Professor Izak Duenyas
Professor Siqian Shen

Jingwen Tang

tjingwen@umich.edu

ORCID iD: 0009-0008-5612-3313

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Cong Shi, who offered me the opportunity to explore the research world five years ago and has guided me through the journey with his professionalism and enthusiasm ever since. Words fall short of expressing my appreciation for his role in shaping my research philosophy, wholeheartedly advising me and enhancing my professional skills, and patiently encouraging me during challenging times. His inspiring way of thinking, dedication to scientific truth, and generous support make him the greatest researcher and the best advisor I can think of. I am fortunate to have him as my role model in the future academic career.

My profound gratitude also goes to Prof. Izak Duenyas, whose diligence often inspires me to progress, while his comments are always incisive and enlightening. His willingness to invest time and effort into my growth has been instrumental to my career development. I am also grateful to Prof. Xiuli Chao for his pivotal role in refining my review skills and sharing his invaluable feedback and sage advice on my research. I would also like to thank Prof. Siqian Shen for her precious time in providing me helpful feedback and continued support.

I greatly appreciate Prof. Boxiao Chen, Prof. Zhengling Qi, Prof. Ethan Fang and Prof. Yuqian Xu, with whom I have the privilege to collaborate. Their inspiring guidance have immensely enriched my experiences, fostering a collaborative environment contributing to my professional growth. Thanks to all Professors above and Dr. Mohammad Zhalechian, Dr. Xian Yu, Dr. Haoming Shen, Prof. Albert Berahas, Prof. Marina Epelman, Dr. Esmaeil Keyvanshokooh, Dr. Xubo Yue for their kind helping hands in my job searching process, and Dr. Chun Ye for his generous mentorship during my internship at Amazon.

I feel fortunate to have friends' companionship rendering the journey colorful and vivid. Credits go to Jiacheng, Xuefei, Xinyu, Moyan, Zhongzhu, Ziyi, Zichen, Jianhao, Guanhua, Guanru, Huiwen, Yaohui, Arlen, Hessa, Lauren, Kati. Memories with them have been stars shining in the time river.

Last but not least, I cannot thank my family enough especially my parents Yijun Tang and Baoling Zhong for their unconditional support. The kindness, integrity and resilience they taught me are the eyes through which I view the world and the arms with which I embrace life. This dissertation is dedicated to them for their grand sacrifice and endless love.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ALGORITHMS

ALGORITHM

LIST OF APPENDICES

APPENDIX

ABSTRACT

The primary focus of this dissertation is to develop both online and offline algorithmic frameworks for data-driven decision-making in the context of operations management problems including supply chain management and revenue management. In a data-poor environment, online learning algorithms can be developed to utilize streaming data to help decision-making sequentially balancing exploration and exploitation. On the other hand, when there is already massive logged data available, offline learning algorithms can be developed for stochastic optimization and policy making.

Many real-world operations management problems have complex system dynamics, abundant operational constraints, as well as varying qualities of accessible data (e.g., missing or censored data). These features, unique to operations management problems, become major challenges of utilizing data in the process of optimization and better decision-making, despite the existence of numerous learning frameworks developed by researchers from other disciplines such as computer science. To overcome these challenges, we carry out research on representative problems arising in the context of operations management. First, we formulate mathematical models capturing the main trade-offs based on specific domain knowledge. Second, we derive structural properties of the optimal policies that provide a foundation for the design of algorithms. Third, we propose a learning algorithm to solve the incomplete information problem, along with a theoretical analysis of its performance guarantee. Finally, we carry out extensive numerical or empirical studies for validation of the method and the discovery of managerial insights.

This dissertation first investigates the online algorithm for a multi-variate optimization problem within a multi-product system with general upgrading. Then still focusing on online learning algorithms, the dissertation proposes two distinct two-layer methodological frameworks designed to solve joint optimization problems that encompass two distinct decision variables. One is based on the setting where the underlying dynamics form a Markov chain in a dual-sourcing system. A learning algorithm combining Successive Elimination and Sample Average Approximation is proposed and demonstrates an optimal convergence rate of regret. The other one is designed for scenarios in two-sided markets where the decision maker makes no parametric assumptions on underlying functions. By integrating Bisection

Search with the Upper Confidence Bound (UCB) algorithm in bandit control, the proposed framework guides the sequential decision-making incurring regret with a provably tight upper bound, which is optimal for any learning algorithm. Finally, the dissertation studies the offline learning algorithm for the problem of feature-based pricing with an offline dataset containing information on historical decisions, covariates, and censored outcomes. An offline algorithm incorporating supervised learning techniques and survival analysis in the language of causal inference is proposed, whose profit is proven to converge to the optimum as the sample size goes to infinity.

CHAPTER 1

Introduction

1.1 Background

Data-driven approaches to decision-making in business have received tremendous attention from both academia and industry in the era of big data (Feng and Shanthikumar 2018a). Due to the increased complexity of business contexts and uncertainty in market trends, learning from proper data becomes necessary to retrieve information on the underlying dynamics.

The design and analysis of learning algorithms depend not only on system dynamics as well as the clairvoyant optimal policy structures, but also on the form of the data available to the decision maker. Specifically, when there is little information on the underlying environment (e.g., demand distribution, demand function) and the data is revealed over time, "learning-while-doing" online learning algorithms can be developed to help sequential decision-making. When there is a set of offline data (e.g., sales data) available and online exploration is expensive or infeasible due to practical issues (e.g., expensive real-time computation resources, data access limitations, concerns about fairness), "predict-then-optimize" types of offline algorithms can serve to yield a static policy for the decision maker to adopt.

1.1.1 Existing Learning Theories

The computer science and statistics communities have proposed learning methods that can be used to guide decision-making in both online and offline manners. It is noteworthy that while these methodological frameworks were originally developed from different perspectives, the methods and ideas can intersect with and be integrated to enhance each other.

Online learning algorithms offer the system a dynamic way of updating based on incoming data, consisting of various methods. This includes Online Reinforcement Learning Theory, which is characterized by its interaction with environments through methods such as UCRL2 (Auer et al. 2008), Q-learning (Watkins 1989), SARSA (Rummery and Niranjan 1994), and Actor-Critic (Konda and Tsitsiklis 1999). The field also includes Multi-Armed

Bandits, focusing on the exploration-exploitation balance in a state-less context, with notable algorithms including Explore-Then-Commit (Robbins 1952), UCB (Auer et al. 2002a), Successive Elimination (Even-Dar et al. 2002), Thompson Sampling (Thompson 1933), Zooming Algorithm (Kleinberg et al. 2008), Exp3/Exp4 (Auer et al. 2002b), LinUCB (Li et al. 2010), and others. Online Convex Optimization introduces methods like Stochastic Gradient Descent (Robbins and Monro 1951), Stochastic Newton (Martens 2010), and Regularization Techniques (Mirror Descent, FTRL), among others. Additionally, the domain encompasses other Stochastic Optimization and Online Optimization tools, such as Sample Average Approximation, Stochastic Approximation, Coordinate Descent, Primal-Dual algorithms, and Potential-Function methods.

In contrast, offline learning algorithms utilize pre-collected datasets for decision-making without real-time interaction with the environment. This approach is evident in Offline Reinforcement Learning with methods like Batch-Constrained Q-learning (BCQ) (Fujimoto et al. 2019), Conservative Q-Learning (CQL) (Kumar et al. 2020), BEAR (Kumar et al. 2019), BRAC (Wu et al. 2019), and others. Supervised Learning for decision-making employs models trained on historical data, including Decision Trees, Random Forest (Breiman 2001), Gradient Boosting Machines such as XGBoost (Chen and Guestrin 2016) and LightGBM (Ke et al. 2017), Support Vector Machines (Cortes and Vapnik 1995), Neural Networks (Rumelhart et al. 1986), and more. Bayesian Methods also contribute to this domain with Bayesian Neural Networks (Neal 2012), Gaussian Processes (Rasmussen and Williams 2003), Bayesian Optimization (Mockus 1994), and additional techniques.

The analysis and performance evaluation of these algorithms are supported by developed theories such as Statistical Learning Theory (Vapnik 2013), Empirical Processes (Sen 2018, Wellner et al. 2013), and High-Dimensional Statistics (Boucheron et al. 2013, Wainwright 2019). These theories provide a robust framework for understanding the mathematical and statistical principles underlying learning algorithms, enabling continued research and development in the field.

1.1.2 Motivations

Despite the existence of previously established learning algorithms, there remains a significant gap between the theories developed above and the resolution of problems in Operations Management (OM). In particular, the OM problems confronted by decision-makers are typically more complex than the models in the general-purpose study in CS and OR.

The system dynamics, particularly in supply chain systems, are usually complicated, leading to challenges rendering existing learning methods not applicable or invalid. The complexity arises from various operational constraints that add to the intricacy of optimization problem structures, even with complete information. For instance, the joint replenishment and allocation decisions in Chapter 2 must adhere to practical restrictions instead of being treated as purely unconstrained optimization problems. Moreover, the huge state spaces associated with real-life system dynamics pose additional challenges, particularly in capturing state transitions, which can lead to the curse of dimensionality when attempting to apply standard Reinforcement Learning frameworks. This issue is exemplified in the dual sourcing system in Chapter 3, which involves a multi-dimensional state variable that needs careful analysis to make the problem tractable. Additionally, the delay in feedback from certain actions complicates the evaluation of decision performance, as the full impact of decisions may not become apparent until the system reaches a steady state. This necessitates a convergence analysis of stochastic processes, especially evident when changes to order-up-to levels in the dual sourcing system incur a non-stationary phase.

In general, OM researchers must strike a balance between capturing key tradeoffs and ensuring tractability in modeling. This challenge is evident, for instance, when examining a two-sided revenue management problem in Chapter 4. The model needs to be comprehensive enough to accurately reflect real-world scenarios while also facilitating computationally efficient algorithm design and the interpretation of practical insights. Furthermore, the access to and quality of data collected in real life are crucial considerations in the design of estimation and optimization processes. For example, demand censoring, resulting from limited inventory, poses a common issue in demand estimation, as seen in an offline data setting for pricing optimization in Chapter 5. Additionally, the analysis of consumer behaviors, integrating insights from psychology, economics, and marketing with mathematical modeling, poses unique challenges in OM studies. This multidisciplinary approach is essential for deriving practical implications for decision-making.

The specific challenges mentioned above underscore the need for a comprehensive and in-depth study of learning algorithms, encompassing their design, analysis, and managerial insights in decision-making within the context of operations management.

1.1.3 Learning Algorithms in Operations Management

The research on learning algorithms in OM aims to propose realistic models for specific real-life problems and suggest implementable algorithms with provable performance guarantees. The contributions of such research are twofold: providing practitioners with probably-good decision support and fostering technical discoveries for methodological innovation in disciplines such as operations research, computer science, and applied probabilities.

The salient feature of the design of both online and offline learning algorithms in OM is to relax the assumption that the demand distribution or the demand-price function is known to the decision-maker, and then “learn” the underlying environment by combining optimization and machine learning techniques. The performance measure of any algorithm is **regret** of the decision maker, defined as the absolute difference between the total expected cost/revenue incurred by implementing the algorithm and that of the (clairvoyant) optimal decisions if the demand distribution or the demand function was known.

Practical online learning algorithms have been developed in recent years to address two main problems in operations management: supply chain management with demand learning and dynamic pricing with demand learning. These include Sample Average Approximation (SAA) methods (Cheung and Simchi-Levi 2019, Levi et al. 2015, 2007a, Lin et al. 2022, Qin et al. 2022), types of Stochastic Gradient Descent (SGD) algorithms (Chen and Shi 2023, Huh et al. 2009, Huh and Rusmevichientong 2009, Shi et al. 2016, Zhang et al. 2018, 2020), and their combination with Bandit algorithms, resulting in a two-layer algorithm for joint optimization of two decision variables simultaneously (Chen and Shi 2020, Chen et al. 2022a, Gong and Simchi-Levi 2023, Wang et al. 2021b, Yuan et al. 2021). While the performance of these algorithms is case-dependent and relies on specific system dynamics, the study of online decision-making for various business models with different structures and contexts continues to be an unresolved area.

Compared to online learning algorithms, offline learning algorithms in OM settings have been much less explored (Ban and Rudin 2019, Bu et al. 2023, Elmachtoub and Grigas 2022). This leaves ample opportunities for utilizing offline datasets in operations management, allowing for efficient decision-making.

This dissertation aims to contribute to the methodology framework of online and offline learning algorithms design and analysis, for problems in supply chain management and revenue management with demand learning, through a detailed analysis of four specific business settings: multi-product systems, dual sourcing systems, two-sided online platforms, and feature-based pricing problems.

1.2 Dissertation Overview

The dissertation consists of six chapters in total.

Chapter 1 introduces the background and motivation of studying learning in OM.

Chapter 2 considers the joint optimization of ordering and upgrading decisions in a dynamic multiproduct system over a finite time horizon of T periods. Multiple types of demand arrive in each period stochastically that can be satisfied with the supply of the same type or some higher type (upgrading). The firm does not know the demand distributions *a priori* and makes adaptive inventory replenishment and upgrading decisions based on historical demand observations. The structure of the clairvoyant optimal joint ordering and allocation policy is first characterized, based on which a new online learning algorithm termed stochastic gradient descent with perturbed gradient (SGD-PG for short) is proposed. The algorithm admits a cumulative regret upper bound of $O(\sqrt{T})$, which matches the lower bound for any learning algorithms. This project proposes an easy-to-implement and provably-good algorithm for the joint replenishment and allocation decisions in a multi-product system that allows general upgrading.

Chapter 3 considers a periodic-review dual-sourcing inventory system with a regular source (lower unit cost but longer lead time) and an expedited source (shorter lead time but higher unit cost), under carried-over supply and backlogged demand. The firm does not have access to the demand distribution *a priori* and relies solely on past demand realizations. Even with complete information on the demand distribution, it is well known in the literature that the optimal inventory replenishment policy is complex and state-dependent. Therefore, the focus of the chapter is on a class of popular, easy-to-implement, and near-optimal heuristic policies called the dual-index policy. The performance measure is the regret, defined as the cost difference of any feasible learning algorithm against the full-information optimal dual-index policy. We develop a nonparametric online learning algorithm that admits a regret upper bound of $O(\sqrt{T \log T})$, which matches the regret lower bound for any feasible learning algorithms up to a logarithmic factor, which provides practitioners with an easy-to-implement, robust, and provably-good online decision support system for managing a dual-sourcing inventory system.

Chapter 4 introduces a new model, the "remunerating newsvendor" problem, which extends the classical price-setting newsvendor problem to incorporate remuneration decisions in two-sided markets. This model has practical applications in modern business contexts, such as service platforms that connect clients with independent contractors or content creators. The platform aims to optimize both pricing for customers and remuneration for providers to maximize expected revenue. The demand is a (random) function of pricing, while the supply is a (random) function of remuneration, and the problem seeks to find an optimal match between them. In the case of complete information, the expected revenue function of the remunerating newsvendor problem is shown to be concave in remuneration given a posted price and Lipschitz continuous with respect to price. A new algorithm called Bandit Bisection Search (BBS) to solve the incomplete information problem. Matching upper and lower regret bounds are established for the algorithm BBS. Moreover, another new algorithm named Double Bisection Search (DBS) is specifically designed for the linear demand case, which leads to improved regret. Numerical experiments provide validation for the effectiveness of the proposed methods.

Chapter 5 studies a feature-based pricing problem with demand censoring in an offline data-driven setting. In this problem, a firm is endowed with a finite amount of inventory, and faces a random demand that is dependent on the offered price and the covariates (from products, customers, or both). Any unsatisfied demand that exceeds the inventory level is lost and unobservable. The firm does not know the demand function but has access to an offline dataset consisting of quadruplets of historical covariates, inventory, price, and potentially censored sales quantity. The objective is to use the offline dataset to find the optimal feature-based pricing rule so as to maximize the expected profit. Through the lens of causal inference, a novel data-driven algorithm is proposed, which is motivated by survival analysis and doubly robust estimation. A finite sample regret bound to justify the proposed offline learning algorithm and prove its robustness. Extensive numerical experiments demonstrate the robust performance of the proposed algorithm in accurately estimating optimal prices on both training and testing data. Furthermore, these experiments highlight the value of considering demand censoring in the context of feature-based pricing.

Chapter 6 offers a summary of the Chapters 2 to 5, pointing out some future directions along the direction of online and offline learning algorithms in OM.

1.3 Main Contributions

In Chapter 2, we studied the joint inventory control and demand allocation problem for a multi-product system with upgrading. When the firm knows the demand distributions *a priori*, we derive the optimal policy for joint inventory replenishment and allocation decisions in each period. When the firm does not know the demand distributions *a priori*, based on a myopic case, we propose the first nonparametric online learning algorithm based on stochastic gradient descent with perturbed gradient (SGD-PG for short), to solve the problem with demand learning. There are two key new ideas underlying SGD-PG. First, we propose a non-trivial subroutine to compute a valid sample-path gradient of the profit with respect to the inventory levels after replenishment. The main idea is to perturb the inventory levels after replenishment by an infinitesimal amount and compute the increment in profit. A key challenge is to correctly identify and quantify the "chaining effect" brought by a local perturbation at some supply node. Second, SGD-PG keeps track of the "real-time imbalance" between supply and demand the same way as in the clairvoyant optimal allocation policy, and carries out the first and second rounds of allocation via an order. Then if the empirical ordering levels are approaching the true optimal ordering levels, the allocation process will also converge to optimality, because we are following the exact structure of the clairvoyant optimal allocation policy. Then we show an upper bound on the regret of the proposed online learning algorithm using a different and simpler approach other than traditional proof using queueing methods (Huh and Rusmevichientong 2009). Specifically, the key is that the length of time for the (overshooting) inventory levels to drop below the target level is of a constant order, based on which, the total regret incurred can be efficiently bounded.

Chapter 3 investigates the stochastic process of the dual-sourcing system formed under the dual-index policy with backlogged demand. Specifically, the vector of pipeline inventory and inventory position forms a uniformly ergodic Markov chain and we can derive a simple expression for the long-run average cost of the dual-index policy. Then when the demand distribution is not known *a priori*, we propose a nonparametric learning algorithm to approach the optimal (z_r, z_e) of the dual-index policy, which admits a regret upper bound of $O(\sqrt{T \log T})$, matching the regret lower bound for any feasible learning algorithms up to a logarithmic factor. There are three key ideas underlying the algorithm design and regret analysis. First, we propose a two-layer learning algorithm that updates the two parameters (Δ, z_e) simultaneously, where the outer layer discretizes the set of Δ to obtain a grid and treats each point on the grid as an arm of the bandit problem, and the inner layer updates the expedited order-up-to level z_e using the empirical quantile for each Δ choice in the active set. This is the first attempt in the literature to integrate bandits with sample average approximation methods, which also provably achieves a tight regret bound. Second, Because of the special structure of the inventory system, after pulling one arm, we are able to evaluate the performance of all the arms based on the realized demand data, which enables us to achieve high estimation accuracies of all arms without extensive exploration of each of them. Third, due to the dependency between two consecutive samples, we study the concentration behavior of the estimation based on data from a Markov chain and utilize the ergodicity of the system variable to achieve a tight regret bound for the regret analysis. A key step in our proofs leverages a specific version of McDiarmid's inequality for Markov chains (Ortner 2020, Paulin 2015).

In Chapter 4, we introduce a novel newsvendor model called the "remunerating newsvendor" which incorporates remuneration as a decision variable to address the supply side's uncertainty. To the best of our knowledge, we are one of the first to expand the price-setting newsvendor model to incorporate remuneration in two-sided markets. Then we establish the concavity of the expected revenue function with respect to the remuneration decision and the concavity in price for any given remuneration choice. Leveraging the above structural properties of the full information problem, we propose an online algorithm BBS that integrates bandit control (specifically, Upper-Confidence-Bound) with a bisection search (specifically, a strictly quartering search) approach, and provide proof of an upper bound on the total regret which matches the lower bounds up to a logarithmic factor. In contrast to previous literature that uses bisection search and its variants for operations management problems (Agarwal et al. 2011, Chen et al. 2019a, Chen and Shi 2020, Chen et al. 2021c, Lei et al. 2014), our approach integrates query operations into the bisection search process with early termination criteria to bound the loss from suboptimal bandit selections. This allows us to update the algorithm at any time instead of restricting updates to only the end of each epoch, which would otherwise result in an inability to establish a tight regret bound. On the technical side, our approach leads to a novel concentration result for the regret of bisection search caused by any bandit choice up to any time. We also further modify the quartering search technique proposed by Agarwal et al. (2011) to improve efficiency.

Chapter 5 proposes a novel data-driven offline learning algorithm that gives the optimal feature-based pricing strategy based on customer/product covariates under demand censoring. To the best of our knowledge, we are the first to model this feature-based pricing problem under censored demand through the lens of causal inference. We model the relationship between demand and price under the celebrated potential outcome framework (Rubin 1974). This framework gives natural identification results on the effect of price on demand, which makes it amenable for offline learning. A novel aspect of our model is to factor in demand censoring. In order to estimate the profit function, we propose to borrow the tool from survival analysis to recover the expected true (conditional) demand. We also propose a doubly robust estimation procedure to further achieve the robustness of our estimation result (Bang and Robins 2005). Specifically, we leverage state-of-the-art supervised learning techniques in estimating the potential profit function and the propensity scores (Rosenbaum and Rubin 1983) as well as in optimizing the feature-based prices. Compared with most existing approaches using parametric models in the literature of profit management and pricing, all the aforementioned components are modeled non-parametrically, thus more robust to model mis-specification. Furthermore, our proposed algorithm is backed up by theoretical and empirical evidence. Theoretically, we provide a finite sample regret analysis of our offline learning algorithm showing that the expected profit of the estimated pricing strategy converges to the profit under the optimal pricing strategy asymptotically as the sample size of the offline data increases. Empirically, we conduct thorough numerical experiments to demonstrate that our proposed algorithm performs robustly well in estimating the optimal prices on both training and testing datasets. We also demonstrate the value of factoring in demand censoring in decision-making.

Overall, the contribution of this dissertation is as follows: Across four fundamental problems in supply chain management and revenue management, we formulate models capturing the system dynamics and decision-making tradeoffs, followed by a detailed analysis of the problem structure. Subsequently, we propose either online or offline learning algorithms based on the specific data environment and operational constraints. We then establish the performance guarantees of the proposed algorithms from both theoretical and empirical perspectives. Finally, the proposed methodology framework offers practitioners a provably effective method or support for decision-making not only in specific settings but also for other problems with similar structures.

CHAPTER 2

Online Learning for Multiproduct Systems with Upgrading

One common challenge practitioners encounter when applying first-order learning methods to practical problems is the inaccessibility of subgradient information for the objective function, a situation often complicated by intricate system dynamics and constraints. This obstacle underscores the importance of exploring and devising strategies to efficiently derive first-order information with complex system dynamics.

In this chapter, we consider a multi-product system allowing general upgrading and investigate the joint inventory replenishment and allocation rules, where the allocation procedure adds complexity to the problem structure. Our analysis is dedicated to identifying the optimal replenishment and allocation policy, and deriving first-order information of the expected profit with respect to inventory control with demand learning.

2.1 Introduction

Modern businesses often expand their product offerings to meet the diverse needs of their customers and drive additional profit streams. However, managing the inventory and fulfillment of a large portfolio of products can be extremely operationally challenging, especially in matching supply with demand across multiple product types over multiple periods. Upgrading, by using a higher-quality product to fulfill the demand for a lower-quality item, at the manager's discretion, is a popular and effective operational strategy. According to Yu et al. (2015), adopting this approach can increase profit without the need to produce additional products, improve the customer experience, reduce lost sales, retain customers, and improve their satisfaction. Additionally, it can reduce inventory holding costs, pool the risk of understocking, improve supply chain efficiency, and boost profitability.

This chapter studies a dynamic inventory system with n types of products. By "type", we mean that the products satisfy the same overall need, but vary in characteristics such as unit ordering cost, inventory holding cost, quality, and price. We will refer to the products as supply and the customers as demand hereafter. Without loss of generality, the higher the supply quality is, the lower the index. Demand i can only be met by supply j with $j \leq i$ (e.g., customer demand can be met using the same quality supply or higher), but only with profit r_i for type i. The process of satisfying one unit of demand i using a unit of supply j is also referred to as allocating (or matching) one unit of supply j to demand i. Essentially, our problem has a downward substitution structure. The firm needs to make two operational decisions in each period, namely, inventory replenishment decisions (for all supplies) and allocation or matching decisions (to match demand with the same type of supply or potential upgrades). Unmet demand is lost. Purchasing one unit of supply i will incur an ordering cost of c_i and each unit of supply carried over to the next period will incur a holding cost h_i. The overall goal is to find an optimal joint ordering and matching policy that maximizes the total expected revenue less the ordering and holding cost.

It is common in many industries to use higher priced or higher quality products to substitute for products that are out of stock. For instance, in the car rental and hotel industries, customers are often given an upgrade to a more expensive room or a larger vehicle when the firm has run out of the types of rooms or vehicles that the customer has ordered. Recently, we collaborated with a distributor of generators who had a large number of SKUs of generators in its inventory. Typically, different standby generators were distinguished from one another based on their rated wattage, such as 22kw for \$4900 or 24kw for \$5100. The distributor worked with a number of contractors who had strict schedules for installing a particular generator in a house and whose installation schedules would be negatively affected if the generator they had ordered from the distributor was unavailable. As a result, the distributor would often offer a slightly higher wattage generator to a contractor for the same price as the one the contractor had originally ordered if the originally ordered generator was out of stock.

While Hu and Zhou (2021) have previously investigated the matching decisions between supply and demand and Yu et al. (2015) have studied a joint capacity and allocation problem with backlogged demand with no replenishment, many firms need to make both inventory replenishment and allocation decisions instead of passively being offered supply. Prior studies on the joint optimization problem (see, e.g., Duenyas and Tsai (2000), Hu et al. (2008)) have predominantly focused on the "two-by-two" case or under specific demand distributions. Here we study the more general structure of the optimal policy of the joint inventory ordering and allocation decisions and propose the first online learning algorithm (when the firm does not know the demand distributions *a priori*). Specifically, we would like to address the following research questions.

(1) When the demand distributions are known, what is the (clairvoyant) optimal policy for making replenishment and allocation decisions of various types of supply in each period?

(2) When the demand distributions are unknown, how can we leverage the structure of optimal policies to design effective online learning algorithms and establish tight regret analysis?

(3) How can we extend our model and findings to a nested censored demand scenario?

2.1.1 Main Results and Contributions

We summarize our main results and key contributions as follows.

Clairvoyant Optimal Joint Ordering and Allocation Policy. When the firm knows the demand distributions *a priori*, we derive the optimal policy for joint inventory replenishment and allocation decisions in each period. There is limited literature on joint ordering and allocation problems (mostly focusing on simple two-by-two cases). When the demand distributions are i.i.d. across time periods, we show in Theorem 2.3.1 that the optimal ordering policy is an order-up-to policy with a well-specified order of allocation. This is, to the best of our knowledge, the first structural result for such joint inventory replenishment and allocation problems with general n products, and this compact structure makes it amenable for practitioners to adopt and implement in practice and for online learning for models with demand learning.

We establish the above result in three key steps. (a) We first show that the optimal allocation policy is greedy (according to the order specified in Theorem 2.3.1) given any inventory levels and realized demand. (b) Then based on this greedy allocation policy, we establish the preservation of the concavity of value functions. (c) Finally, based on the preservation of concavity, we show that the optimal replenishment policy is an order-up-to policy.

Online Learning Algorithms via Infinitesimal Perturbation. When the firm does not know the demand distributions *a priori*, we propose the first nonparametric online learning algorithm (Algorithm 1) based on stochastic gradient descent with perturbed gradient (SGD-PG for short), to solve the problem with demand learning. We use the notion of cumulative regret as our performance measure, which is the profit difference between running a clairvoyant optimal policy (that has access to the demand distributions) and SGD-PG (that has to learn the demand distributions based on past demand realizations). We prove that SGD-PG admits a cumulative regret upper bound of $O(\sqrt{T})$, which matches the regret lower bound for any online learning algorithms.

There are two key new ideas underlying SGD-PG. First, we propose a non-trivial subroutine (Algorithm 2) to compute a valid sample-path gradient of the profit with respect to the inventory levels after replenishment. The main idea is to perturb the inventory levels after replenishment by an infinitesimal amount and compute the increment in profit. A key challenge is to correctly identify and quantify the "chaining effect" brought by a local perturbation at some supply node. Second, SGD-PG keeps track of the "real-time imbalance" between supply and demand the same way as in the clairvoyant optimal allocation policy, and carries out greedy allocation via an order of pairs. Then if the empirical ordering levels are approaching the true optimal ordering levels, the allocation process will also converge to optimality, because we are following the exact structure of the clairvoyant optimal allocation policy.

We show that the SGD-PG algorithm also works in the nested censored demand case. We also demonstrate the numerical efficacy of the proposed algorithms.

2.1.2 Relevant Literature

Our work is closely related to the following streams of literature.

Multiproduct inventory systems with upgrading or substitution. Allowing for the allocation of multiple types of supply to multiple types of demand, such as product upgrading or substitution, can provide firms with increased flexibility in inventory and revenue management (see, for example, Jain et al. (2015), Parker and Olsen (2010). The key operational management decisions for centralized firms include inventory replenishment and allocation choices. We categorize the relevant research into three areas: (a) ordering decisions, (b) substitution decisions, and (c) joint ordering and substitution decisions in a dynamic setting.

In the context of inventory ordering decisions, Mahajan and van Ryzin (2001) studied a one-period inventory model in which customers dynamically substitute among product variants within a retail assortment when inventory is depleted. The customer choice decisions are based on a utility maximization criterion. Faced with such substitution behavior, the retailer must choose initial inventory levels for the assortment to maximize expected profits. They proposed a stochastic gradient algorithm to solve this problem (without demand learning). Several papers have considered multi-period models with two substitutable products. Pasternack and Drezner (1991) proved that a base-stock policy is optimal for two products with substitution in a dynamic newsvendor network. Nagarajan and Rajagopalan (2008) proved that a partially decoupled base-stock policy is optimal for the setting where the two products are partial substitutes and their demands are negatively correlated. As far as multiple substitution choices are concerned, Gallego et al. (2006) proposed two simple heuristics to determine ordering quantities in a semiconductor inventory system with downgrading, assuming that a myopic allocation is used. Schlapp and Fleischmann (2018) derived the optimal (capacity-dependent) ordering policy for a capacity-constrained firm selling multiple partially substitutable products over a finite season in a market with stockout-based customer substitution.

In the context of inventory allocation decisions, Hu and Zhou (2021) developed a two-way substitution model that showed the optimal allocation policy for both horizontally and vertically differentiated rewards is to match the pairs (down to some threshold levels) in the descending order of priorities. They proposed heuristics to attain these optimal threshold levels. Meanwhile, Baccara et al. (2020) derived the closed-form of optimal threshold levels in dynamic matching under Poisson arrivals. Elmachtoub et al. (2019) considered the multiproduct inventory problem with opaque products and derived a balancing policy that is asymptotically optimal. However, to the best of our knowledge, none of these studies have addressed the joint optimization problem of both inventory replenishment and product allocation decisions. In this chapter, we aim to address this gap and propose a novel approach to jointly optimize inventory replenishment and product allocation. We first review the literature on joint optimization starting from the single-period model to the multi-period settings with two or more products.

In regard to the joint optimization of inventory replenishment and allocation decisions, Bassok et al. (1999) investigated a single-period multiproduct inventory problem with downward substitution with a constant marginal cost of substitution and penalty cost. They demonstrated that the base-stock policy is optimal for stocking, and greedy allocation is optimal under certain conditions. In another study, Rao et al. (2004) addressed a single-period multiproduct inventory problem that involved decisions such as which products to produce, the production quantities, and the allocation. They formulated the problem as a mixed-integer program (MIP) and proposed effective heuristic algorithms for each part. It is worth noting that our work differs from these studies in that we consider a multi-period model.

For multi-period models, there is a significant amount of literature focusing on the case where there are two products. For instance, Chen (1997) studied the joint replenishment and allocation problem of two products with downward substitution, while Xu et al. (2011) studied the optimal policy for the one-time replenishment and substitution decisions between two mutually substitutable products, under Poisson arrivals. They demonstrated that the optimal substitution followed a threshold rule. In the context of chip substitutions in the semiconductor industry, Duenyas and Tsai (2000) studied the joint optimal policy for production and substitution policies and proposed a heuristic algorithm. Moreover, Hu et al. (2008) addressed the optimal joint control of inventory and transshipment for a firm that produces in two locations and faces capacity uncertainty. The optimal transshipment policy was found to be floor-rationing and the production policy was proved to be a state-dependent produce-up-to threshold policy. Similarly, Yu et al. (2017) studied optimal production, pricing, and substitution policies for a continuous-review production inventory system with two products under Poisson arrivals. They demonstrated that the optimal production policy for each product is a state-dependent base-stock policy and the optimal substitution policy consists of state-dependent thresholds. However, the main difference between these papers and ours is that we do not limit ourselves to the special case of two products.

Finally, in the context of the multi-period model with multiple products, Shumsky and Zhang (2009) studied a capacity allocation model with one-level upgrading, while Yu et al. (2015) extended this model to allow for general upgrading and showed that the optimal allocation policy involves a greedy allocation of supply to demand of the same type, followed by sequential rationing. However, their models assume that capacity is decided only once at the beginning of the horizon and that allocation of the capacity is needed in each period without replenishment, whereas in our model, joint inventory replenishment and allocation decisions are made in each period. Moreover, all the aforementioned papers assume that the firm knows the demand distributions, while we consider a more realistic setting with demand learning.

Learning algorithms for inventory management. Learning algorithms can be broadly categorized into two groups based on the information structure of the firm: *parametric* and *nonparametric*. Parametric algorithms involve the firm forming a prior belief about the demand distribution and updating the parameters of the distribution with new demand information. This Bayesian approach has been widely adopted in the literature, with early contributions from Iglehart (1964), Murray and Silver (1966), Scarf (1959), and more recent work by Chen and Plambeck (2008), Lu et al. (2005, 2008), Wang and Mersereau (2017).

On the other hand, nonparametric approaches have gained popularity in recent years. Instead of assuming a prior distribution, these methods rely on the empirical distribution formed by uncensored samples drawn from the true distribution. One such approach is sample average approximation (SAA), which has been widely used in the inventory literature (Kleywegt et al. (2002), Levi et al. (2015, 2007b)). Another popular nonparametric approach is concave adaptive value estimation, which approximates the objective cost function with a sequence of piecewise linear functions (Godfrey and Powell (2001), Powell et al. (2004)).

Our work belongs to the general class of gradient-based methods, which is arguably the most popular nonparametric approach, especially for inventory systems. There has been a growing literature on developing gradient-based methods for various inventory models, including capacitated inventory systems (Chen et al. (2020c), Shi et al. (2016)), perishable inventory systems (Zhang et al. (2018)), lost sales inventory systems with lead times (Agrawal and Jia (2022), Huh et al. (2009), Zhang et al. (2020)), dual-sourcing inventory systems (Chen and Shi (2020)), joint pricing and inventory control (Chen et al. (2019a, 2021a, 2022a, 2020b)), inventory systems in the presence of fixed costs (Ban (2020), Yuan et al. (2021)), and inventory control under non-stationary demand (Mao et al. (2020)). Nonparametric approaches have also been applied to the newsvendor problem with contextual information (Ban and Rudin (2019)). Our work differs from the above literature by considering not only ordering decisions but also allocation decisions for multiple products over multiple periods. Technically, we develop a new perturbed gradient estimation approach.

It is worth mentioning that more recently, Chen and Chao (2020a) considered a multi-product inventory control problem with stockout substitution and demand learning, which seems to be close to our problem setting. In their setting, customers attempt to substitute once when facing stockout, and the substitution choices are made by customers, depending on substitution probabilities. Hence, the firms in their settings only need to make inventory ordering decisions. In contrast, in our setting, the substitution is controlled by the firm, i.e., it is an *active* operational decision to make. To the best of our knowledge, our work represents the first attempt in the literature that studies the joint ordering and allocation problem with demand learning.

2.1.3 Chapter Organization and General Notation

The remainder of the chapter is organized as follows. We formulate the dynamic multiproduct inventory system with general upgrading in §2.2. We characterize the (clairvoyant) optimal joint ordering and allocation policy for this system in §2.3. We study the problem with demand learning and propose the first online learning algorithm termed SGD-PG with a theoretical regret analysis in §2.4. As an extension, we show the algorithm can handle censored demand in §2.5. The performance of the algorithm is evaluated in the numerical experiments in §2.6. Finally, we conclude the chapter in §2.7.

We introduce the general notation used in this chapter. For any real number x, we denote $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. For event A, the indicator function $\mathbb{1}(A)$ takes value 1 if A is true and 0 otherwise. The projection function is defined as $\mathbf{P}_{[a,b]}(x) = \min[b, \max(x, a)]$ for any real numbers x, a, and b. For integer $n \geq 1$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. The maximum operator functioning on vectors means taking the maximum component-wise, i.e., $\max\{\mathbf{x}, \mathbf{y}\} = \mathbf{z}$ where $z_i = \max\{x_i, y_i\}$. The symbol $\succeq$ ($\preceq$) denotes vector inequality or componentwise inequality, i.e., for any $\mathbf{y}, \mathbf{x} \in \mathbb{R}^n$, if $\mathbf{y} \succeq (\preceq)\mathbf{x}$, then $y_i \geq (\leq) x_i$ for $i = 1, \ldots, n$. We often use upper-case characters to denote random variables and corresponding lower-case characters to denote the realization of the random variables. The term i.i.d. means independent and identically distributed.

2.2 Problem Formulation

We formally describe the multiproduct inventory system with general upgrading. There are n products, each facing a stochastic demand. Without loss of generality, we index these n products from 1 to n according to their intrinsic qualities with 1 being the best and n being the poorest. We assume that the demand for product i can be satisfied using any product j from $[i] := \{1, \ldots, i\}$. (Note that with only slight modifications, our results continue to hold with satisfying demand i using product j from a strict subset of $[i]$.)

Figures 2.1 and 2.2 give schematic illustrations of direct allocation (perfect matching) and upgrading (imperfect matching), respectively. The arrow from supply j to demand i indicates that we can meet one unit of demand i using one unit of supply j, which generates revenue r_i *independent* of the supply type. For instance, the firm only receives r_2 if it chooses to satisfy demand 2 using supply 1 (i.e., upgrading). Purchasing one unit of supply j will incur an ordering cost of c_j, and each unit of supply j carried over to the next period will incur a holding cost h_j.

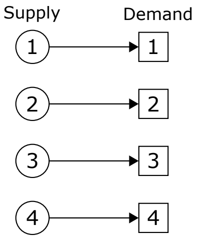


Figure 2.1: Direct Allocation (Perfect Matching)

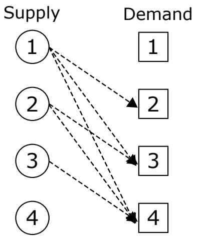


Figure 2.2: Upgrading (Imperfect Matching)

Any remaining supply upon completion of the allocation process is kept in inventory and incurs a holding cost, while any remaining unmatched demand is lost. Let $t \in \{1, 2, \ldots\}$ be the time period, which is indexed forward. For each product $i \in [n]$, we denote its demand in period t by D_i^t.

Assumption 2.2.1 *For each product $i \in [n]$, the demands D_i^t are i.i.d. random variables across $t \in [T]$. Denote D_i as a random variable with the same distribution as $D_i^t, \forall t \in [T]$, then the expectation $\mathbb{E}[D_i] = \alpha_i > 0$ and variance $Var[D_i] = \sigma_i^2 < \infty$.*

Note that we do not require the i.i.d. assumption across products. A salient feature of our setting is that the firm has no prior knowledge about the true underlying demand distributions. The firm can observe the realized demand over time and make adaptive decisions.

Assumption 2.2.2 *The revenue and cost parameters satisfy the following conditions.*

(a) $r_1 \geq r_2 \geq \ldots \geq r_n$.

(b) $c_1 \geq c_2 \geq \ldots \geq c_n$.

(c) $h_i = h_0 + \beta c_i, \forall\ i \in [n]$ *where* $\beta \in [0, 1]$.

Assumptions 2.2.2a and 2.2.2b indicate that the unit revenue and ordering cost are non-increasing in the index (or equivalently, non-decreasing in the product quality). Assumption 2.2.2c asserts that the holding cost consists of two parts, with the first part being the fixed physical holding cost h_0 and the second part is the financial cost proportional to the order cost.

2.2.1 System Dynamics

In each period $t \in [T]$, we use $\mathbf{x}^t = (x_1^t, \ldots, x_n^t)$ to denote the inventory levels at the beginning of any period t (before ordering). After an order is received, we use $\mathbf{y}^t = (y_1^t, \ldots, y_n^t)$ to denote the inventory level (after ordering). We use u_{ij}^t for all $1 \leq j \leq i \leq n$ to denote the amount of product j used to satisfy demand i in period t. For simplicity of notation, we define $u_{ij}^t \equiv 0$ when the index i or j is out of bound. With the realization of demand i denoted by d_i^t, we must have $\sum_{i=j}^n u_{ij}^t \leq y_j^t$ and $\sum_{j=1}^i u_{ij}^t \leq d_i^t$. For any policy π, the sequence of events in each period t, $t = 1, 2, \ldots$, is as follows. (Note that $\mathbf{x}^{t,\pi}$ and $\mathbf{y}^{t,\pi}$ depend on the policy π and the sample path ω, but we make the dependency implicit for notational convenience.)

(a) At the beginning of period t, the firm observes the on-hand inventory levels $\mathbf{x}^t = (x_1^t, \ldots, x_n^t)$.

(b) The firm places an order so that the after-replenishment inventory level of product i is $y_i^t \geq x_i^t$. The order arrives instantaneously, incurring an ordering cost $\sum_{i=1}^n c_i \left(y_i^t - x_i^t\right)$.

(c) The random demand $\mathbf{D}^t$ is realized to be $\mathbf{d}^t$. Then the firm makes an allocation decision u_{ij}^t, $\forall 1 \leq j \leq i \leq n$, generating a revenue of $\sum_{i=1}^n r_i \sum_{j=1}^i u_{ij}^t$.

(d) The remaining supply is carried over to the next period and the unmet demand is *lost*. The excess supply also incurs a holding cost $\sum_{j=1}^n h_j \left(y_j^t - \sum_{i=j}^n u_{ij}^t \right)$.

(e) The system proceeds to the next period with $\mathbf{x}^{t+1}$ given by $x_j^{t+1} = \left(y_j^t - \sum_{i=j}^n u_{ij}^t \right)$, $\forall j \in [n]$.

Given the (after-replenishment) inventory levels $\mathbf{y}^t$ and the allocation quantity U^t, the effective profit in period t can be written as

$$\sum_{j=1}^n \left(-c_j y_j^t + \sum_{i=j}^n r_i u_{ij}^t + (c_j - h_j) \left(y_j^t - \sum_{i=j}^n u_{ij}^t \right) \right).$$

The objective is to find a policy π that maximizes the total expected T-period profit.

2.3 (Clairvoyant) Optimal Joint Ordering and Allocation Policy

Before designing an effective online learning algorithm, we first characterize the clairvoyant optimal joint ordering and allocation policy (if the firm had access to the demand distribution $\mathbf{D}$ *a priori*).

2.3.1 Structural Results

We use $J^t(\mathbf{x}^t), t \in [T]$ to denote the optimal expected profit starting from period t till the end of the horizon given the starting inventory level is $\mathbf{x}^t$. Given it is a Markov decision process, the optimal policy maximizing the total expected T-period profit can be attained by solving the following dynamic program.

$$\begin{aligned}
J^t(\mathbf{x}^t) &= \max_{\mathbf{y}^t \succeq \mathbf{x}^t} V^t\left(\mathbf{y}^t\right) + \mathbf{c}^\intercal \mathbf{x}^t, \\
V^t\left(\mathbf{y}^t\right) &:= -\mathbf{c}^\intercal \mathbf{y}^t + \mathbb{E}_D \left\{ \max_{\substack{U^{t\intercal}\mathbf{1} \preceq \mathbf{y}^t, \\ U^t \mathbf{1} \preceq \mathbf{D}^t, \\ U^t \succeq \mathbf{0}}} \left\{ \mathbf{r}^\intercal U^t \mathbf{1} - h\left(\mathbf{y}^t - U^{t\intercal}\mathbf{1}\right)^\intercal \mathbf{1} + J^{t+1}\left(\mathbf{y}^t - U^{t\intercal}\mathbf{1}\right) \right\} \right\}, \\
J^{T+1}\left(\mathbf{x}^{T+1}\right) &= \mathbf{c}^\intercal \mathbf{x}^{T+1}.
\end{aligned}$$

where $U_{ij}^t = [u_{ij}^t]$ is the allocation matrix in period t. Specifically, in each period, the decision maker first needs to determine the optimal after-replenishment inventory levels $\mathbf{y}^t$ for all products. Then after the demand is realized, allocation decisions need to be made subject to the constraints of not exceeding the supply and demand quantities. Excess supply is carried over to period $t+1$ as the starting inventory level and unmet demand is lost. Finally, at the end of the horizon, any excess inventory j can be salvaged at the price of c_j, and any unmet demand i will incur a penalty cost c_i.

Proposition 2.3.1 *For any period $t \in [T]$, $J^t(\mathbf{x}^t)$ is concave in $\mathbf{x}^t$ and $V^t(\mathbf{y}^t)$ is concave in $\mathbf{y}^t$.*

2.3.2 Myopic Optimal Policy with Zero Starting Inventory

We first study the optimal policy given the starting inventory of the system is zero. Figure 2.3 shows the structure of the "priority" of pairs considered in the allocation policy π^{*A}. Row i consists of pairs with demand i and column j consists of pairs with supply j. The possible pairs are divided into n layers where layer m contains $n-m+1$ pairs where the difference between the indices of supply and demand is $m-1$.

Layer 1:	$(1,1)$	$(2,2)$	$(3,3)$	$\ldots$	$(n-2,n-2)$	$(n-1,n-1)$	(n,n)
Layer 2:	$(2,1)$	$(3,2)$	$(4,3)$	$\ldots$	$(n-1,n-2)$	$(n,n-1)$	
Layer 3:	$(3,1)$	$(4,2)$	$(5,3)$	$\ldots$	$(n,n-2)$		
$\ldots$							
Layer n:	$(n,1)$						

We denote the optimal allocation policy by π^{*A} and the optimal replenishment policy by π^{*R}.

Theorem 2.3.1 *Suppose that Assumptions 2.2.1 and 2.2.2 hold, and the system starts with zero inventory. Under the boundary condition $J^{T+1}\left(\mathbf{x}^{T+1}\right) = \mathbf{c}^\intercal \mathbf{x}^{T+1}$, the optimal allocation policy for each period π^{*A} is to use supply j to satisfy demand i until either the supply or the demand runs out, for pair (i,j) with $r_i - c_j + h_j \geq 0$ in increasing order of the index of its layer as shown in Figure 2.3. The order of pairs within the same layer is arbitrary.*

The optimal replenishment policy π^{*R} is to order up to $\mathbf{y}^* = \arg\max_{\mathbf{y} \succeq \mathbf{0}} \mathbf{R}(\mathbf{y})$ in each period with

$$\mathbf{R}(\mathbf{y}) := \mathbb{E}_D \sum_{j=1}^{n} \left(-c_j y_j + \sum_{i=j}^{n} r_i u_{ij}^*(\mathbf{y}, \mathbf{D}) + (c_j - h_j) \left(y_j - \sum_{i=j}^{n} u_{ij}^*(\mathbf{y}, \mathbf{D}) \right) \right), \tag{2.1}$$

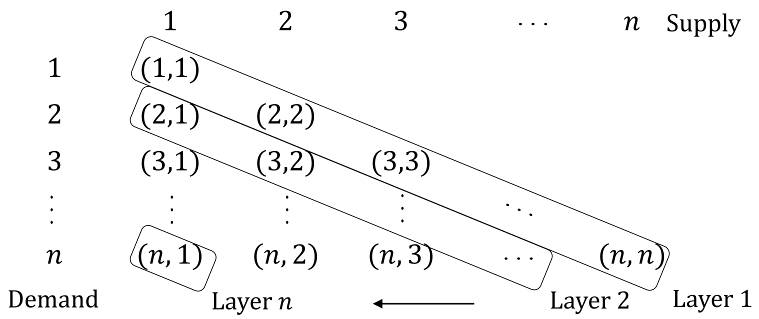


Figure 2.3: Allocation Structure

where $u^*_{ij}(\mathbf{y}, \mathbf{D})$ *is the quantity of supply* j *used to satisfy demand* i *given the on-hand inventory levels* $\mathbf{y}$ *and the demand* $\mathbf{D}$ *following* π^{*A} *specified above.*

2.3.3 Discussions about Theorem 2.3.1

Theorem 2.3.1 asserts that for our multiproduct system with general upgrading, it is optimal to follow an order-up-to policy for each supply (where the base-stock level is the maximizer of (2.1)) and to match supply with demand to the maximum extent in non-decreasing order of the difference of the supply and demand indices. For example, suppose $n = 3, \mathbf{c} - \mathbf{h} = [3, 2, 1]^\intercal, \mathbf{r} = [6, 4, 2]^\intercal$. Then there are in total 6 possible pairs with only one pair $(3, 1)$ has $r_3 - c_1 + h_1 < 0$. So Theorem 1 states that one of the optimal allocation policies for each period is to greedily use supply j to satisfy demand i in the order of $(1, 1), (2, 2), (3, 3), (2, 1), (3, 2)$ as shown in Figure 2.3 since the order of the pairs is distributed in non-decreasing order of layer index.

To the best of our knowledge, this is the first result to compactly characterize the structure of joint replenishment and allocation policy for a multiperiod multiproduct inventory system with upgrading. In contrast, the existing literature predominantly focused on 2-product settings (e.g., Duenyas and Tsai (2000), Hu et al. (2008), Xu et al. (2011), Yu et al. (2017)). There have been relatively limited results on multiproduct settings (see Shumsky and Zhang (2009), Yu et al. (2015) without replenishment decisions and Chen (1997), Shanthikumar et al. (2003) with with certain properties of fixed substitution decisions instead of specifying the allocation rule).

2.4 Stochastic Gradient Descent with Perturbed Gradient (SGD-PG)

Now suppose that the firm does not know the true underlying demand distribution D_i, $\forall i \in [n]$ *a priori*, and the system starts with zero inventory. Our goal is to find a provably-good online learning algorithm such that the average profit converges to that of the clairvoyant optimal policy.

2.4.1 The SGD-PG Algorithm

We describe our online learning algorithm π via stochastic gradient descent with perturbed gradient (SGD-PG) in Algorithm 1. The motivation of the design is as follows.

Let π^* denote the clairvoyant optimal policy. Based on the structural properties of Theorem 2.3.1, if our policy π follows the optimal (greedy) allocation policy π^{*A}, we can decompose the cumulative regret

$$\mathbf{Regret}(\pi, T) = \sum_{t=1}^{T} \mathbf{R}\left(\mathbf{y}^*\right) - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{R}\left(\mathbf{y}^t\right)\right],$$

where $\mathbf{y}^*$ is the optimal order-up-to level of π^* and $\mathbf{y}^t$ is the inventory level following our policy π. By the analysis of Theorem 2.3.1, we know that $\mathbf{R}(\cdot)$ has a nice concavity property.

Lemma 2.4.1 $\mathbf{R}\left(\mathbf{y}\right)$ *defined in* (2.1) *is concave in* $\mathbf{y}$.

Thus, it is natural to run a stochastic gradient descent scheme to iteratively update $\mathbf{y}^t$. However, the updated base-stock level upon each gradient descent step may not always be implementable, if the starting inventory level is already higher than the desired base-stock level. To resolve this issue, the algorithm π needs to maintain a pair of sequences $(\hat{\mathbf{y}}^t : t \geq 1)$ and $(\mathbf{y}^t : t \geq 1)$, where $\hat{\mathbf{y}}^t$ is the desired target inventory levels for period t and $\mathbf{y}^t$ is the actual implemented inventory levels in period t. Suppose the decision maker is aware of an upper bound of the optimal order-up-to levels $\bar{\mathbf{y}}$. The sequences are generated in the following way:

$$\hat{\mathbf{y}}^{t+1} = \mathbf{P}_{[\mathbf{0}, \bar{\mathbf{y}}]}\left(\hat{\mathbf{y}}^t + \epsilon^t \mathbf{G}^t\left(\hat{\mathbf{y}}^t\right)\right), \tag{2.2}$$

$$\mathbf{y}^{t+1} = \max\left(\hat{\mathbf{y}}^{t+1}, \mathbf{x}^{t+1}\right), \tag{2.3}$$

where the maximum operator is taken component-wise. Combined with the optimal allocation rules as specified in Theorem 2.3.1, this naturally gives rise to the following stochastic gradient descent based Algorithm 1, where $\gamma > 0$ and $\boldsymbol{\theta} = \max\{r_1 - c_n, c_1 - c_n + h\}\mathbf{1}$.

Algorithm 1 Stochastic Gradient Descent Algorithm with Perturbed Gradient (SGD-PG)

Initialization. Let $\mathcal{L}$ be an ordered list of pairs as stated in Theorem 2.3.1. Randomly initialize $\hat{y}_i^1 \in [0, \bar{y}_i]$, $\forall i \in [n]$. Let $\gamma > 0$ and $\boldsymbol{\theta} = \max\{r_1 - c_n, c_1 - c_n + h\}\mathbf{1}$.

Learning. For each period $1 \leq t \leq T$, carry out the following procedures.

Phase 1: Make a replenishment decision with the target base-stock level $\hat{\mathbf{y}}^t$:

$$\mathbf{y}^t = \max\left(\hat{\mathbf{y}}^t, \mathbf{x}^t\right).$$

Then carry out perfect matching, i.e., satisfying demand using the same type of supply.

Phase 2: Initialize the real-time imbalance between type i supply and demand l_i^t as $y_i^t - d_i^t$. For each pair of demand i and supply j in the ordered list $\mathcal{L}$, carry out upgrading as follows.

1. Allocate $u_{ij}^t = \min\{(l_i^t)^-, (l_j^t)^+\}$ quantity of supply j to demand i;
2. Update the real-time imbalance by $l_i^t = l_i^t + u_{ij}^t$ and $l_j^t = l_j^t - u_{ij}^t$.

Phase 3:

1. If $\mathbf{x}^t \preceq \hat{\mathbf{y}}$, call a subroutine (Algorithm 2) to obtain an unbiased gradient estimator $\mathbf{G}^t\left(\hat{\mathbf{y}}^t\right)$. Update the base-stock levels for period $t+1$:

$$\hat{\mathbf{y}}^{t+1} = \mathbf{P}_{[\mathbf{0},\bar{\mathbf{y}}]}\left(\hat{\mathbf{y}}^t + \epsilon^t \mathbf{G}^t\left(\hat{\mathbf{y}}^t\right)\right),$$
$$\epsilon^t = \frac{\|\bar{\mathbf{y}}\|\gamma}{\|\boldsymbol{\theta}\|\sqrt{t}},$$

and $\mathbf{P}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$ means (component-wise) projecting x_i onto $[a_i, b_i]$, $\forall i \in [n]$.
2. Otherwise, $\hat{\mathbf{y}}^{t+1} = \hat{\mathbf{y}}^t$.

Carry the excess supply i to the next period as $x_i^{t+1} = (l_i^t)^+$, $\forall i \in [n]$.

End. In period $T+1$, the excess supply i is salvaged at the unit price of c_i.

2.4.2 Subroutine for Perturbed Gradient

Up to now, the algorithmic template for Algorithm 1 seems natural in that (a) we follow the exact greedy allocation rule specified in Theorem 2.3.1 and (b) we leverage the fact that $\mathbf{R}(\cdot)$ has a nice concavity property. That said, there is an important missing piece (which is also a key contribution). That is, we need to compute a stochastic (sample-path) gradient for $\mathbf{R}(\cdot)$. We denote such a sample-path gradient by $\mathbf{G}^t\left(\hat{\mathbf{y}}^t\right)$ and we develop a new subroutine (Algorithm 2) for computing it.

Algorithm 2 Subroutine for Computing Sample-Path Gradient $\mathbf{G}^t\left(\hat{\mathbf{y}}^t\right)$

```
Initialize $\mathbf{e}^t = [M, \ldots, M]$, $\forall t \in [T]$. Now create an exact copy of the system and do the
following without modifying the original system. Carry out perfect matching, i.e., satis-
fying demand using the same type of supply. Initialize the real-time imbalance between
type $i$ supply and demand $l_i^t$ as $\hat{y}_i^t - d_i^t$. For each pair of demand $i$ and supply $j$ in the
ordered list $\mathcal{L}$, do:
  1. If $(l_j^t)^+ < (l_i^t)^-$, $e_j^t = i$; else if $(l_j^t)^+ > (l_i^t)^-$, $e_i^t = j$.
  2. Allocate $u_{ij}^t = \min\{(l_i^t)^-, (l_j^t)^+\}$ quantity of supply $j$ to demand $i$;
  3. Update the real-time imbalance by $l_i^t = l_i^t + u_{ij}^t$ and $l_j^t = l_j^t - u_{ij}^t$.
for $i = 1, \ldots, n$ do
    if $l_i^t = 0$ then              ▷ supply (demand) $i$ is depleted by demand (supply): $e_i^t$
        let $a = e_i^t$               ▷ use a temporary node $a$ to track the chain
        while $e_a^t != M$ do
            $a = e_a^t$               ▷ find where the chain of upgrading ends
        end while
        if $l_a^t > 0$ then           ▷ the chain ends at supply $a$
            $\mathbf{G}_i^t = -c_i + c_a - h_a$   ▷ one additional order means one more excess supply $a$
        else                          ▷ the chain ends at demand $a$
            $\mathbf{G}_i^t = r_a - c_i$          ▷ one additional order means one more met demand $a$
        end if
    else if $l_i^t > 0$ then         ▷ $e_i^t = M$, excess supply $i$ at the end of period $t$
        $\mathbf{G}_i^t = -h_i$       ▷ one more unit of excess supply $i$
    else                              ▷ $e_i^t = M$, unmet demand $i$ at the end of period $t$
        $\mathbf{G}_i^t = r_i - c_i$  ▷ one more perfect pair $i$
    end if
end for
Output $\mathbf{G}^t$.
```

2.4.2.1 An Illustrative Example.

The main idea of this subroutine is to hypothetically increase the inventory level of each product for an infinitesimal amount δ and compute the profit increment under this sample path. Note that a perturbation in the inventory level y_i^t of supply i not only affects demand i but also demand $j > i$ as well as other supplies $k < i$. We call it a "*chaining effect*". A key factor in this subroutine is to identify where a particular chain (starting with some δ perturbation in some supply) ends, thereby quantifying the profit increment.

To better illustrate this idea, we define the following graphical notation. In Figure 2.4, on the left hand side, an arrow is drawn from supply j to demand i, which means that supply j is depleted by demand i and there is some unmet demand i at the end. Similarly, on the right hand side, an arrow is drawn from demand i to supply j, which means that demand i is depleted by supply j and there is excess supply j at the end. Essentially, for any two nodes connected by an arrow, upon allocation, there are some positive units sitting on the node pointed by an arrowhead and nothing on the other node.

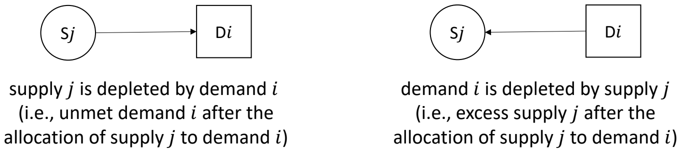

S*j*
D*i*
supply *j* is depleted by demand *i*
(i.e., unmet demand *i* after the
allocation of supply *j* to demand *i*)
S*j*
D*i*
demand *i* is depleted by supply *j*
(i.e., excess supply *j* after the
allocation of supply *j* to demand *i*)

Figure 2.4: Definition of Arrows

With the definition of these arrows in place, consider the following example shown in Figure 2.5. Upon perfect matching, there are excess supplies at supply nodes 2, 3, 6 and there are unmet demands at demand nodes 1, 4, 5. Let us focus on an instance of upgrading or substitution on the right hand side of Figure 2.5. There is excess supply at supply 3, which is used to satisfy demand 4. Then supply 3 is depleted by demand 4 and there is still unmet demand at demand 4. One could use excess supply at supply 2 to satisfy demand 4. Then demand 4 is depleted by supply 2 and there is excess supply at supply 2. One could use excess supply 2 to satisfy demand 5. Then supply 2 is depleted by demand 5 and there is still unmet demand at demand 5. However, at this moment, there are no more supplies from supply nodes $1, 2, 3, 4, 5$, and thus this "chain" ends at demand node 5. This means if we perturb supply 3 by δ, we can observe a chaining effect, i.e., supply 3 → demand 4 → supply 2 → demand 5 where the terminal node is demand 5.

We use Figure 2.6 (the second subfigure) to illustrate that the gradient with respect to y_3^t is

$$\frac{\partial \mathbf{R}\left(\mathbf{y}^t; \mathbf{d}^t\right)}{\partial y_3^t} = (r_5 - c_3).$$

Suppose that we increase y_3^t to $y_3^t + \delta$ and this δ perturbation forms a chain (as discussed earlier). It is clear that the non-terminal nodes of this chain do not affect the profit and only

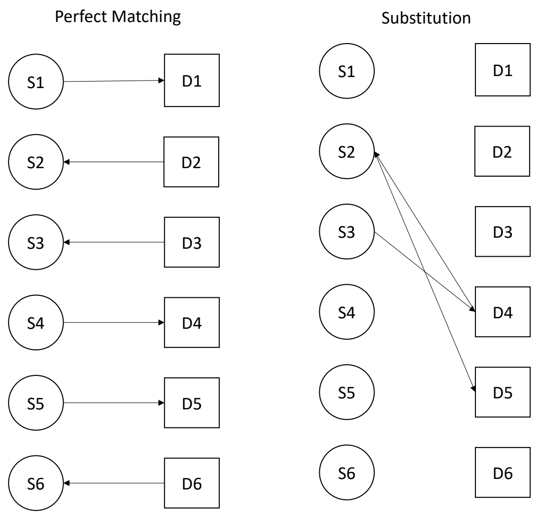


Figure 2.5: A Chaining Example

the terminal node does. As a result, there will be δ less unmet demand at demand node 5, resulting in a profit change of $(r_5 - c_3)\delta$. This implies a sample-path gradient with respect to y_3^t is $(r_5 - c_3)$.

2.4.2.2 Validity of Algorithm 2.

Recall that $\mathbf{R}(\mathbf{y})$ is defined in (2.1). We slightly abuse the notation and let $\mathbf{R}(\mathbf{y};\mathbf{d})$ denote the sample-path per-period profit given demand realization $\mathbf{d}$:

$$\mathbf{R}(\mathbf{y};\mathbf{d}) := \sum_{j=1}^{n} \left(-c_j y_j + \sum_{i=j}^{n} r_i u_{ij}^* (\mathbf{y},\mathbf{d}) + (c_j - h_j) \left(y_j - \sum_{i=j}^{n} u_{ij}^* (\mathbf{y},\mathbf{d}) \right) \right).$$

Proposition 2.4.1 *The output $\mathbf{G}^t(\mathbf{y})$ from Algorithm 2 is a valid gradient of $\mathbf{R}(\mathbf{y};\mathbf{d}^t)$ in $\mathbf{y}$.*

Specifically, Algorithm 2 provides an easy-to-implement routine to calculate the subgradient of the profit with respect to inventory levels of multiple products. The underlying idea is to hypothetically increase the inventory level and apply the property of the greedy allocation rule.

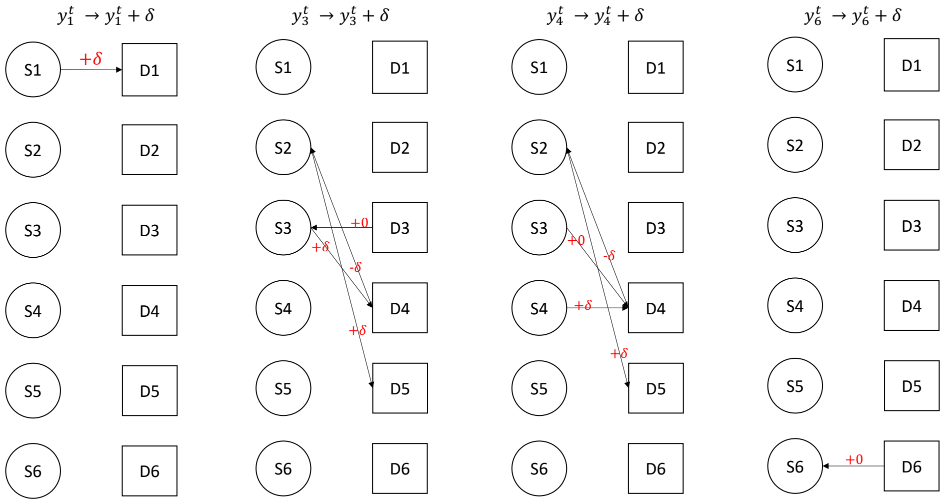

$y_1^t \to y_1^t + \delta$
+δ
$y_3^t \to y_3^t + \delta$
+0
+δ
-δ
+δ
$y_4^t \to y_4^t + \delta$
+0
-δ
+δ
+δ
$y_6^t \to y_6^t + \delta$
+0
S1
S2
S3
S4
S5
S6
D1
D2
D3
D4
D5
D6

Figure 2.6: A Perturbation Example

2.4.3 Regret Analysis

We decompose the average expected cumulative regret as

$$
\begin{aligned}
\mathbf{Regret}(\pi, T) =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbf{R}\left(\mathbf{y}^*\right)-\mathbf{R}\left(\mathbf{y}^t\right)\right)\right] \\
=& \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbf{R}\left(\mathbf{y}^*\right)-\mathbf{R}\left(\hat{\mathbf{y}}^t\right)\right)\right]+\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbf{R}\left(\hat{\mathbf{y}}^t\right)-\mathbf{R}\left(\mathbf{y}^t\right)\right)\right] \\
:=& \Lambda_1(T)+\Lambda_2(T).
\end{aligned}
$$

With the valid stochastic gradient established above, bounding $\Lambda_1(T)$ calls for an analysis of online convex optimization. With inventory carryover, the desired inventory target levels (resulting from gradient descent updates) may not be achieved. Bounding $\Lambda_2(T)$ requires mapping the "inventory overshoot" onto a queuing system with increasing service levels. The key to bound both parts is to analyze the length of cycles in Algorithm 1 and to construct an unbiased gradient estimator via infinitesimal perturbation (Algorithm 2 and Proposition 2.4.1).

Theorem 2.4.1 *There is a constant C_2 such that for any $T \geq 1$, the cumulative regret of running the SGD-PG algorithm is upper bounded as*

$$\mathbf{Regret}(\pi, T) \leq C_2\sqrt{T}.$$

The following proposition establishes that our cumulative regret upper bound matches the lower bound (for any online learning algorithms), up to a logarithmic factor.

Proposition 2.4.2 *Suppose that $T > 5$. The expected regret for any learning algorithm for our joint replenishment and allocation problem is lower bounded by $\Omega(\sqrt{T})$.*

Proposition 1 in Zhang et al. (2020) provides an example of a demand scenario under which any learning algorithm incurs regret of at least $\Omega(\sqrt{T})$. This proposition examines a *single-product* newsvendor problem that only involves ordering decisions and not allocation decisions. The proof of Proposition 2.4.2 is omitted.

2.5 Extension to the Setting with Nested Censored Demand

We consider a setting with nested censored demand. There are two salient features of the nested censored demand model (especially in the multiproduct setting). First, the censoring is nested, i.e., whether demand i is censored depends on the supply and demand of type $1, \ldots, i$. Second, censoring is dynamic in the sense that in each period, the customers arrive sequentially (not at once) and take action based on the available remaining inventory. There are three cases when a customer arrives and demands a type i product.

(1) If there is still supply i left in the inventory, the customer will be provided with supply i.

(2) If supply i is out of stock, but there is still (higher type) supply from $1, \ldots, i-1$ left in the inventory, the customer informs the manager of the stockout and asks to be considered for a potential upgrade. However, the allocation decision only happens at the end of that period. This is because the firm may choose to upgrade to a higher type of customer (for better revenue management). For instance, imagine a scenario in which there is only one unit of supply 1 left, and there are two customers (not yet satisfied), one from demand 2 and the other one from demand 3. Although demand 3 may have come in before demand 2, the firm allocates supply 1 to demand 2 at the end of the day.

(3) If supply i is out of stock and all potential upgrades, namely, supply $1, \ldots, i-1$ are out of stock, the customer will simply leave the system without notifying the manager. This is the portion of customers that are being *censored.*

Note that our model and results continue to hold with only slight modification if in Case (2) only an unknown fraction of customers inform the manager of a stockout and ask to be considered for a potential upgrade. That means the remaining fraction gets censored. For ease of presentation, we only focus on the setting with all eligible customers requesting upgrades.

2.5.1 Model Dynamics

We use a 3-type instance to explain the model dynamics. Let types 1, 2, and 3 be high, medium, and low quality, respectively. The optimal allocation policy is to first greedily satisfy demand i with product i, $i = 1, 2, 3$, and then greedily satisfy demand 2 with supply 1, demand 3 with supply 2 and demand 3 with supply 1 if possible. The sequence of events is described as follows.

1. At the beginning of the period, the inventory levels of three types in the store are x_1^t, x_2^t, and x_3^t, respectively. The firm makes the replenishment decisions such that the inventory levels after replenishment are y_1^t, y_2^t, and y_3^t. An ordering cost of $\sum_{i=1}^{3} c_i(y_i^t - x_i^t)$ is incurred.

2. The store starts to satisfy the demand. A random customer may face one of the three cases.

    (a) Customer finds the same type of supply and gets fulfilled. For example, customer A in Figure 2.7 will be satisfied with one unit of supply 1.

    (b) Customer finds the same type of supply out of stock but there is still supply for potential upgrade. Then the customer informs the store manager and gets the name recorded, for a potential upgrade at the end of the day. For example, customers C and D in Figure 2.7 belong to this category. Note that the recorded demand may not be upgraded in the end, as we can see that customer C is upgraded with supply 2 while customer D is unfulfilled.

    (c) Customer finds the same type of supply out of stock, and all potential upgrades are also out of stock. Then the customer simply leaves the system, and this piece of arrival information is *censored*. For example, customers B and E in Figure 2.7 belong to this category.

    From the store's perspective, the process above is to allocate the supply in stock to the demand as much as possible, receiving $\sum_{i=1}^{3} r_i \min(y_i^t, d_i^t)$ revenue. Also, the store records the unmet requests which might be upgraded at the end of the period.

3. At the end of the period, the store starts to upgrade the unmet demand recorded. The firm will allocate excess supply to unmet demand in the order demonstrated in

Theorem 2.3.1. Note that since the demand is only censored when there are no potential upgrades available, the demand censoring will not make a difference in this process.

4. At the end of the period, the excess inventory will be carried over to the next period with holding cost $\sum_{i=1}^{3} hx_i^{t+1}$ where x_i^{t+1} is the inventory level of supply i at the end of period t.

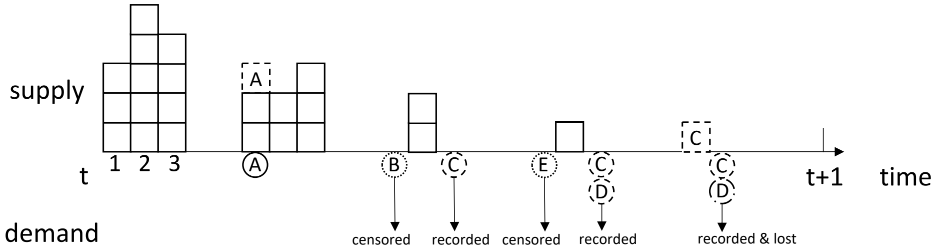


Figure 2.7: An Illustrative Example

Figure 2.7 illustrates an example: (1) The inventory levels for supply $1, 2, 3$ are $3, 5, 4$. (2) The inventory levels become $2, 2, 3$ after satisfying some customers. Customer A requesting one unit of supply 1 gets filled by one unit of supply 1. (3) There are only 2 units of supply 2 in stock. Then customer B asking for supply 1 sees no match and no potential upgrades and will leave without letting the store know. However, there is a potential upgrade available for customer C asking for supply 3 so C will be recorded. (4) There is only one unit of supply 2 left. Another customer D asking for supply 3 comes and is recorded while a customer E asking for supply 1 is censored. (5) At the end of the period, there are two recorded unmet demands for supply 3 which are C and D. While C is upgraded with supply 2, D is unfulfilled.

2.5.2 Online Learning Algorithms

Denote the observed censored demand vector in period t by $\tilde{\mathbf{d}}^t$. Algorithm 1 still works in the case of demand censoring. The only modification is to use the censored demand $\tilde{\mathbf{d}}^t$ in place of the fully realized $\mathbf{d}^t$ in Phase 2 and Subroutine 2.

Proposition 2.5.1 *Theorem 2.4.1 still holds in the case of demand censoring.*

Proof of Proposition 2.5.1. Consider Algorithm 1. We would like to show that the system dynamics are the same using $\tilde{\mathbf{d}}^t$ and $\mathbf{d}^t$ following Algorithm 1. For any period t, with the same initialization, the two systems are exactly the same until the end of Phase 1. As long as there is excess product i after Phase 1, the demand for product $i+1,\dots,n$ will be uncensored. That is to say, by initializing $k=n$ and letting $k=\min_{i\in[n]}\{i : y_i^t > \tilde{d}_i^t\}-1$ if any, the demand for products $1,\dots,k$ are all censored and the demand for products $k+1,\dots,n$ are all uncensored, i.e., $\tilde{d}_i^t \le d_i^t$, $\forall 1 \le i \le k$ and $\tilde{d}_i^t = d_i^t$, $\forall k+1 \le i \le n$. Consider Phase 2. Denote $\tilde{l}_i^t := y_i^t - \tilde{d}_i^t$. Then we have

$$\tilde{l}_i^t = \begin{cases} 0, & \forall 1 \le i \le k, \\ l_i^t, & \forall k+1 \le i \le n. \end{cases}$$

Thus,

$$\tilde{u}_{ij}^t = \begin{cases} 0, & \forall 1 \le j \le k, \\ \min\{(\tilde{l}_i^t)^-, (\tilde{l}_j^t)^+\} = \min\{(l_i^t)^-, (l_j^t)^+\}, & \forall k+1 \le j < i \le n, \end{cases} = u_{ij}^t,$$

where $\tilde{l}_i^t$ is the real-time imbalance between supply and censored demand of product i. Then in Phase 3, we show that the obtained $\mathbf{G}^t(\mathbf{y}^{t+1} \mid \tilde{\mathbf{d}}^t) = \mathbf{G}^t\left(\mathbf{y}^{t+1} \mid \mathbf{d}^t\right)$. Because $y_i^t \ge \hat{y}_i^t$, we have

$$\hat{y}_i^t - \tilde{d}_i^t \le y_i^t - \tilde{d}_i^t \le 0,\ \forall 1 \le i \le k,$$
$$\hat{y}_i^t - d_i^t \le y_i^t - d_i^t \le 0,\ \forall 1 \le i \le k.$$

Hence, following the same logic as in Phase 2, the allocation results are the same as using the fully realized demand $\mathbf{d}^t$. As a result, the output of the subroutine $\mathbf{G}^t(\mathbf{y}^{t+1} \mid \tilde{\mathbf{d}}^t) = \mathbf{G}^t\left(\mathbf{y}^{t+1} \mid \mathbf{d}^t\right)$. Then the unmet demand is lost and the excess supply of the same amount is carried over to the next period $t+1$. This shows that with the same initialization, the dynamics of the two systems using the censored demand and the fully realized demand are identical following Algorithm 1. The convergence result remains the same. **Q.E.D.**

2.6 Numerical Experiments

In our numerical experiments, we relax the somewhat restrictive Assumption 2.2.2c that the holding costs have to be the same for all products. We demonstrate the profit gap of the proposed policy in Theorem 2.3.1 with the optimal policy is very small, even when the holding costs are unequal. Then we implement Algorithm 1 for both the uncensored and censored demand cases.

Table 2.1: Previous Experiment Parameters

instance	c	r	h	$\mathbf{D}$	ϵ^t
1	$[5, 2, 1]$	$[10, 8, 6]$	$[0.5, 0.2, 0.1]$	$[\mathbf{N}(25, 3), \mathbf{N}(20, 5), \mathbf{N}(5, 1)]$	$\frac{1}{\sqrt{t}}$
2	$[5, 2, 1]$	$[10, 8, 6]$	$[0.5, 0.2, 0.1]$	$[\mathbf{U}[0, 50], \mathbf{U}[0, 40], \mathbf{U}[0, 10]]$	$\frac{1}{\sqrt{t}}$
3	$[5, 2, 1]$	$[10, 8, 6]$	$[0.5, 0.2, 0.1]$	$[\mathbf{N}(25, 10), \mathbf{N}(20, 8), \mathbf{N}(5, 2)]$	$\frac{1}{\sqrt{t}}$
4	$[10, 5, 2]$	$[30, 20, 10]$	$[2, 1, 0.4]$	$[\mathbf{N}(5, 1), \mathbf{N}(10, 2), \mathbf{N}(20, 4)]$	$\frac{2}{\sqrt{t}}$

2.6.1 Performance of Algorithm 1 with Uncensored Demand

Define the relative regret of a policy π as

$$\textbf{Relative Regret}(\pi) := \frac{T\mathbf{R}\left(\mathbf{y}^*\right) - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{R}\left(\mathbf{y}^t\right)\right]}{T\mathbf{R}\left(\mathbf{y}^*\right)},$$

where $\mathbf{y}^t$ are the inventory levels after replenishment in period t following some policy π.

We introduce a series of benchmark algorithms named "SAA-κ" for $\kappa = 100, 200, 300$. They belong to the category of "explore-then-exploit" policies since SAA-κ will first "explore" for κ periods with the order-up-to levels uniformly generated and then "exploit" the empirically best choice for the remaining horizon. The detailed procedure is given in Algorithm 3.

We run 1000 demand sets based on the instances specified in Table 2.1, where $\mathbf{N}[\mu, \sigma]$ denotes the truncated normal distribution truncated at $[0, 2\mu]$ with mean μ and standard deviation σ and $\mathbf{U}[a, b]$ denotes the uniform distribution with lower bound a and upper bound b. The average relative regret is shown in Figures 2.8–2.11. We also plot the cumulative regret in Figures 2.12–2.15, i.e.,

$$\textbf{Cumulative Regret } (\pi) := T\mathbf{R}\left(\mathbf{y}^*\right) - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{R}\left(\mathbf{y}^t\right)\right].$$

We can see that Algorithm 1 performs robustly well across all cases.

2.6.2 Performance of Algorithm 1 with Censored Demand

We implement Algorithm 1 under censored demand for the instances specified in Table 2.1.

Algorithm 3 SAA-κ (Sample-Average-Approximation-κ) Algorithm

Initialization. Let $\mathcal{L}$ be the ordered list of pairs specified in Theorem 2.3.1. Let $N' = 100$.

Exploration. For each period $1 \leq t \leq \kappa$, carry out the following procedures.

Phase 1: Make a replenishment decision with randomly selected base-stock level $\hat{y}_i^t \in [\underline{D_i}, \bar{D}_i]$, $\forall i \in [n]$.

$$\mathbf{y}^t = \max\left(\hat{\mathbf{y}}^t, \mathbf{x}^t\right).$$

Satisfy demand using the requested product as much as possible. Record the demand d_i^t, $i \in [n]$.

Phase 2: Initialize the real-time imbalance between type i supply and demand l_i^t as $y_i^t - d_i^t$, $\forall i \in [n]$. For each pair demand i and supply j in list $\mathcal{L}$:

1. Allocate $u_{ij}^t = \min\{(l_i^t)^-, (l_j^t)^+\}$ quantity of supply j to demand i;
2. Update the real-time imbalance by $l_i^t = l_i^t + u_{ij}^t$ and $l_j^t = l_j^t - u_{ij}^t$.

Carry the excess supply i to the next period as $x_i^{t+1} = (l_i^t)^+$, $\forall i \in [n]$.

Sample Average Approximation.

Step 1: For $i \in [n]$, discretize the order-up-to level and obtain a set $S_i := \{\underline{D_i} + \frac{\bar{D}_i - \underline{D_i}}{N'}, \ldots, \bar{D}_i\}$.

Step 2: Apply grid search for all combination of order-up-to levels. With the demand data of κ periods $\{\mathbf{d}^1, \ldots, \mathbf{d}^\kappa\}$, implement using each choice of order-up-to levels $\hat{y}_1, \ldots, \hat{y}_n \in S_1 \times \ldots \times S_n$ and denote the order-up-to levels with highest profit as $\hat{y}_1^{*SAA}, \ldots, \hat{y}_n^{*SAA}$.

Exploitation: For each period $\kappa + 1 \leq t \leq T$, carry out **Phase 1** and **Phase 2** with the order-up-to levels being $\hat{y}_1^{*SAA}, \ldots, \hat{y}_n^{*SAA}$ for each period.

End. In period $T + 1$, the excess supply i is salvaged at the unit price of c_i, $\forall i \in [n]$.

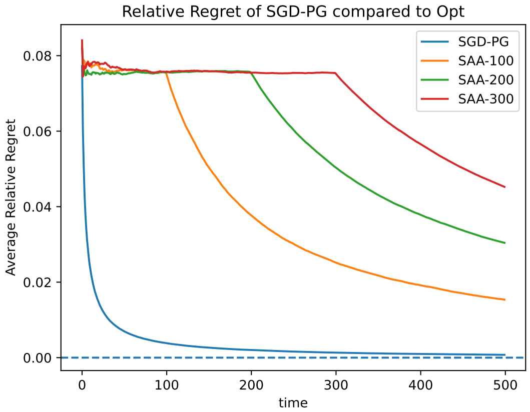


Figure 2.8: Instance 1

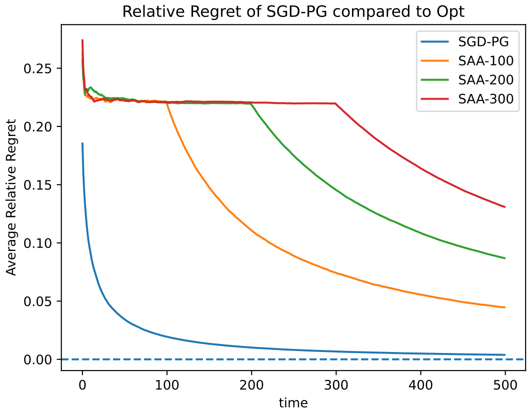


Figure 2.9: Instance 2

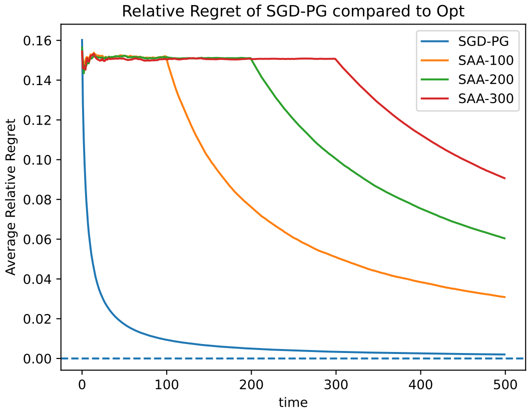


Figure 2.10: Instance 3

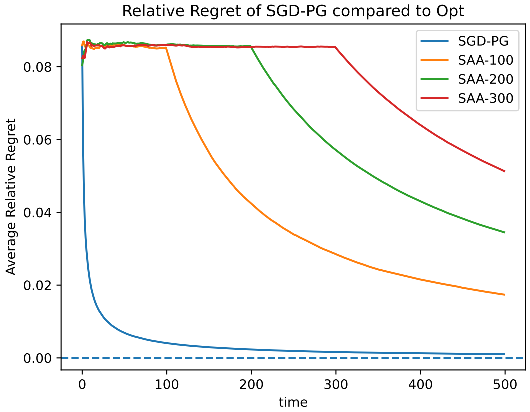


Figure 2.11: Instance 4

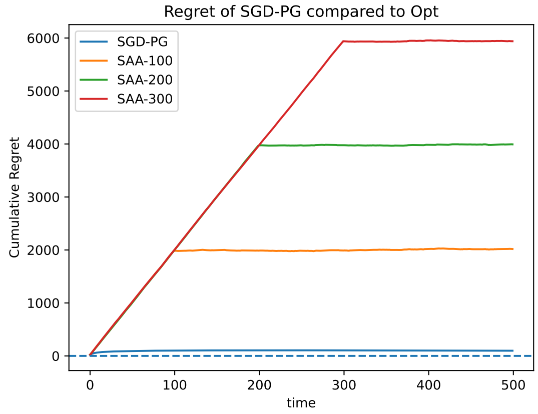


Figure 2.12: Instance 1 Cumulative Regret

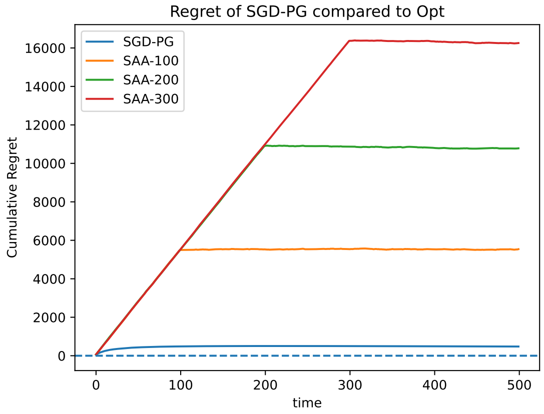


Figure 2.13: Instance 2 Cumulative Regret

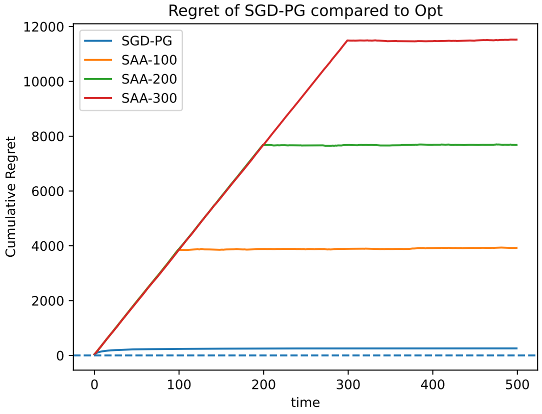


Figure 2.14: Instance 3 Cumulative Regret

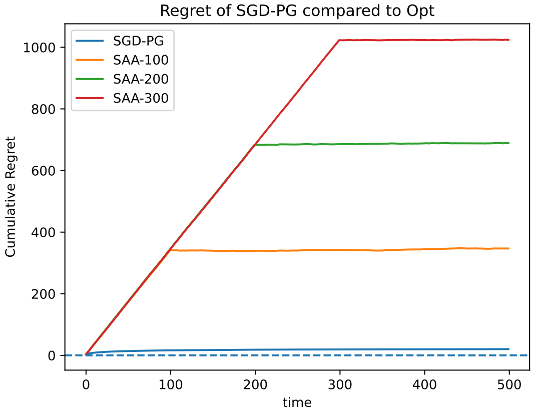


Figure 2.15: Instance 4 Cumulative Regret

Algorithm 4 KM-κ (Kaplan-Meier-κ) Algorithm

Initialization. Let $\mathcal{L}$ be the ordered list of pairs specified in Theorem 2.3.1. Let $N = N' = 100$. Initialize $k = n + 1$.

Exploration. For each period $1 \leq t \leq \kappa$, carry out the following procedures.

Phase 1: Make a replenishment decision with randomly selected base-stock level $\hat{y}_i^t \in [\underline{D}_i, \bar{D}_i]$, $\forall i \in [n]$.

$$\mathbf{y}^t = \max\left(\hat{\mathbf{y}}^t, \mathbf{x}^t\right).$$

Then satisfy demand using the requested product as much as possible. If any, let $k = \min_{i \in [n]}\{i : y_i^t > \tilde{d}_i^t\}$. Then we have the censoring indicator $\Delta_i^t = 0, \forall i \in [k-1]$ and $\Delta_i^t = 1, \forall k \leq i \leq n$. Record the observed demand and censoring indicator pairs $(\tilde{d}_i^t, \Delta_i^t), i \in [n]$.

Phase 2: Initialize the observed real-time imbalance between type i supply and demand l_i^t as $y_i^t - \tilde{d}_i^t$, $\forall i \in [n]$. For each pair demand i and supply j in list $\mathcal{L}$:

1. Allocate $u_{ij}^t = \min\{(l_i^t)^-, (l_j^t)^+\}$ quantity of supply j to demand i;
2. Update the real-time imbalance by $l_i^t = l_i^t + u_{ij}^t$ and $l_j^t = l_j^t - u_{ij}^t$.

Carry the excess supply i to the next period as $x_i^{t+1} = (l_i^t)^+, \forall i \in [n]$.

Kaplan-Meier Estimation.

For each product $i \in [n]$,

Step 1: Arrange $\{\tilde{d}_i^t, t \in [\kappa]\}$ values in increasing order denoted by $\{v_i^1, \ldots, v_i^\kappa\}$.

Step 2: For $j \in [\kappa]$, obtain $\tilde{m}_i^j = \sum_{t \in [\kappa]} \Delta_i^t \mathbb{1}_{\tilde{d}_i^t = v_i^j}$ and $\tilde{M}_i^j = \sum_{t \in [\kappa]} \mathbb{1}_{\tilde{d}_i^t \geq v_i^j}$.

Step 3: Obtain the estimated CDF as $\mathbb{P}[D_i \leq d] = 1 - \prod_{j : v_i^j \leq d}\left(1 - \frac{\tilde{m}_i^j}{\tilde{M}_i^j}\right)$.

Step 4: Generate N sets of T-period demands for product i following the empirical distribution obtained above.

Step 5: Discretize the order-up-to level and obtain a set $S_i := \{\underline{D_i} + \frac{\bar{D}_i - \underline{D_i}}{N'}, \ldots, \bar{D}_i\}$.

Apply grid search for all combinations of order-up-to levels. With the generated N sets of demand data, apply Monte-Carlo Simulation to each choice of order-up-to levels $\hat{y}_1, \ldots, \hat{y}_n \in S_1 \times \ldots \times S_n$ and denote the order-up-to levels with highest profit by $\hat{y}_1^{*KM}, \ldots, \hat{y}_n^{*KM}$.

Exploitation. For each period $\kappa + 1 \leq t \leq T$, carry out **Phase 1** and **Phase 2** with the order-up-to levels being $\hat{y}_1^{*KM}, \ldots, \hat{y}_n^{*KM}$ for each period.

End. In period $T + 1$, the excess supply i is salvaged at the unit price of c_i, $\forall i \in [n]$.

Similar to the uncensored case, we also introduce a series of benchmark algorithms named "KM-κ" for $\kappa = 100, 200, 300$. KM-κ will first choose the order-up-to levels uniformly for κ periods, and then use grid search to obtain the optimal order-up-to level with respect to the empirical distribution based on the celebrated *Kaplan-Meier estimator* (see Huh et al. (2009)). Then the empirically best order-up-to levels are applied for the remaining of the horizon. The detailed algorithm is given in Algorithm 4.

For each instance, we run 100 datasets. The average relative regret is shown in Figures 2.16–2.19 and the cumulative regret is shown in Figures 2.20–2.23. Again, we can see that Algorithm 1 performs robustly well across all cases.

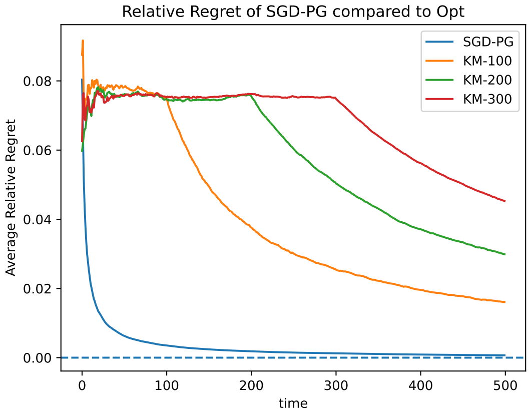


Figure 2.16: Censored Instance 1

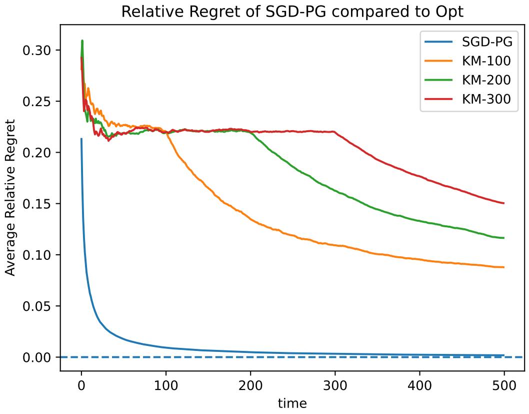


Figure 2.17: Censored Instance 2

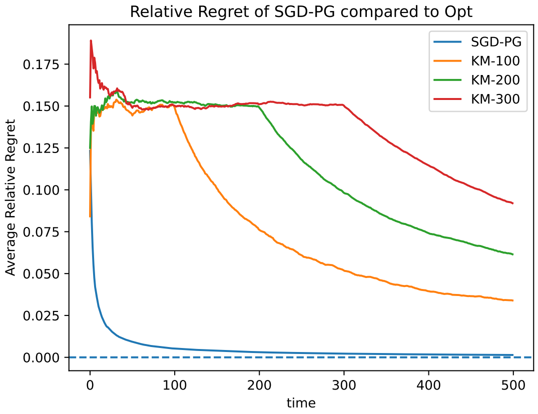


Figure 2.18: Censored Instance 3

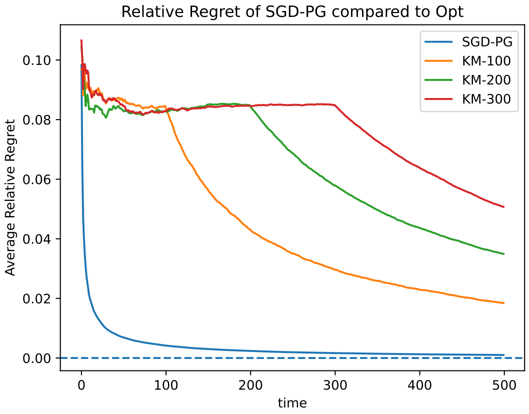


Figure 2.19: Censored Instance 4

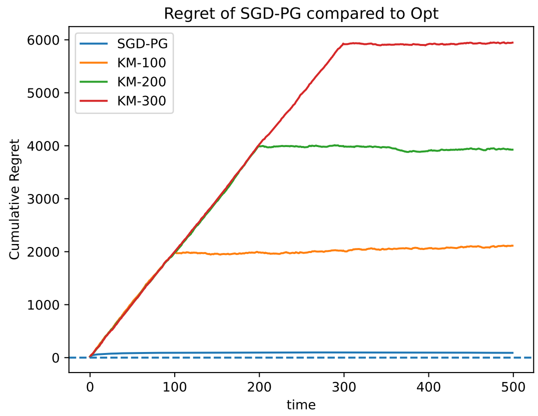


Figure 2.20: Censored Instance 1 Cumulative Regret

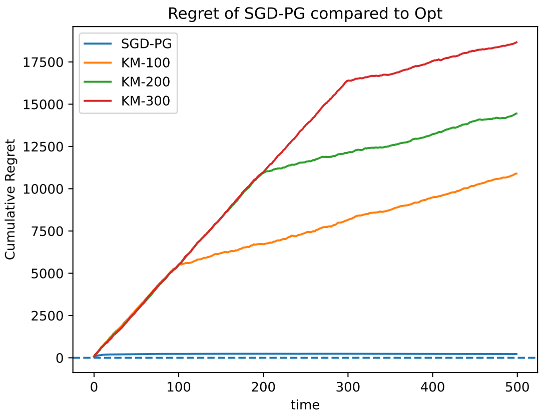


Figure 2.21: Censored Instance 2 Cumulative Regret

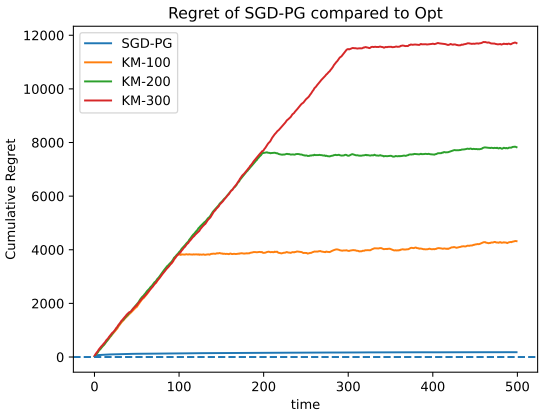


Figure 2.22: Censored Instance 3 Cumulative Regret

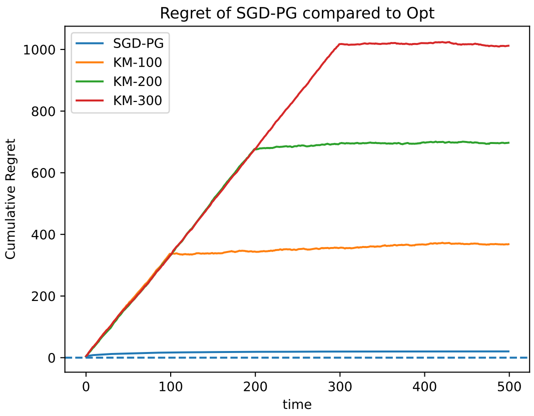


Figure 2.23: Censored Instance 4 Cumulative Regret

2.7 Concluding Remark

We have studied a dynamic multiproduct system with general upgrading. The demand arrives stochastically in each period. The firm needs to make both inventory replenishment decisions and allocation decisions to match demand and supply (of different types). We first characterized the optimal joint ordering and matching policy, had the firm known the true demand distributions *a priori*. When the demand information is incomplete, we proposed an online learning algorithm SGD-PG, to solve the joint learning and optimization problem under both uncensored demand and (nested) censored demand settings (see §2.5). We gave provably optimal regret bounds and also demonstrated the efficacy of the proposed algorithms in numerical experiments (see in §2.6).

To conclude this chapter, we would like to highlight three potential research areas for future study. First, the current model assumes an uncapacitated supply, but in practical scenarios, the supply may have fixed or stochastic constraints. Therefore, it would be valuable to extend the current model to include capacitated supply cases, which is an important research direction. Second, while product upgrading is prevalent in practice, there are also situations where two-way substitution, involving both upgrading and downgrading, is possible. Investigating the optimal structure for models with two-way substitution, as well as developing learning algorithms for the incomplete information problem, would be a worthwhile undertaking. Third, one may consider models with non-stationary demand. Addressing this direction would require significant innovations in modeling and the design and analysis of algorithms. Overall, these research avenues have the potential to significantly advance the field, but they will require substantial efforts in modeling and algorithmic design.

CHAPTER 3

Online Learning in Dual Sourcing Systems

The Sample Average Approximation (SAA) and Bandit Control techniques stand as pivotal methodologies within the realms of stochastic optimization and online learning algorithms. One fundamental assumption for the methods to work is the independence of the sample data. However, the complexity of operational dynamics or the delay of rewards undermines this independence, raising critical questions about the applicability and integration of these techniques in optimizing problems, as well as their impact on the regret convergence rate.

Dual sourcing systems are one of the important topics in supply chain management, which is known for its complicated dynamics and state-dependent optimal replenishment policy. This chapter delves into the feasibility of merging Bandit Control techniques with SAA in the application of dual-sourcing inventory systems, by analyzing underlying dynamics and algorithm performances, providing insights into data-driven methods for this complex yet crucial topic.

3.1 Introduction

The dual-sourcing inventory system has received tremendous attention from both industry and academia for several decades. The system's capability to order from two sources, one cheaper but slower, and the other one faster but more expensive, provides more flexibility for a firm's supply chain management and facilitates maintaining satisfactory customer services while controlling costs. Dual-sourcing systems are ubiquitous in practice. For example, Dell sources the majority of its computer components sold in the US from Asia, which offers a lower per-unit purchasing cost but suffers from a long lead time. To complement the supplier in Asia, Dell also sources from Mexico for the benefit of a shorter lead time, although the per-unit purchasing cost for the latter is higher (Xin and Van Mieghem (2021)). Dual sourcing also provides opportunities for firms to reduce supply chain risk, which is a primary concern for firms and has become even more prominent ever since the outbreak of the COVID-19 pandemic. The devastating impact of the pandemic reveals enormous risks of an inflexible supply chain and pushes firms worldwide to seek alternative sourcing to reduce supply chain risk (Svoboda et al. (2021)).

The formal study of dual-sourcing inventory systems with backlogged demand dates back as early as six decades ago, and it has become an important area in inventory management ever since. Even with a known demand distribution, the optimal inventory replenishment policy is known to be hard to compute. For the very special case where the lead times differ by exactly one, the optimal solution can be fully characterized (Bulinskaya (1964), Fukuda (1964)). For general lead times, however, the optimal policy is proved to be highly state-dependent and adopts no simple structure (Whittemore and Saunders (1977)). The unattainability of the optimal policy calls for the need to develop reasonable heuristic policies, among which one important class of policies is the dual-index policy proposed by Veeraraghavan and Scheller-Wolf (2008). The dual-index policy is characterized by two critical parameters, (z_r, z_e), where z_r is the order-up-to level for the regular source, and z_e is the order-up-to level for the expedited source. The dual-index policy is intuitive and convenient to implement, and the existing literature has shown that it is easy to compute and performs near-optimally in extensive simulation studies (Veeraraghavan and Scheller-Wolf (2008)). Li and Yu (2014) and Hua et al. (2015) also find that the dual-index policy exhibits excellent performance in many applications. Besides, the dual-index policy inspires the development of variants of other heuristic policies for the dual sourcing model such as the vector-based policy by Sheopuri et al. (2010) and the capped dual-index policy by Sun and Van Mieghem (2019).

When the demand distribution is not known *a priori*, the dual sourcing inventory control problem becomes more challenging to solve. In this chapter, by adopting the optimal dual-index policy as our benchmark, we develop online learning algorithms that learn the demand distribution while minimizing total costs over the planning horizon. Our algorithm integrates stochastic bandits and sample average approximation techniques in an innovative way, and we prove that the online learning algorithm converges to the optimal dual-index policy with a provable rate.

3.1.1 Main Results and Contributions

We summarize our main results and key contributions as follows.

Proving ergodicity of the supply chain formed under the dual-index policy. We investigate the stochastic process of the dual-sourcing system formed under the dual-index policy with backlogged demand. In Lemma 3.3.1, we demonstrate that the vector of pipeline inventory and inventory position forms a Markov chain. In Theorem 3.3.1, we prove that the Markov chain is ergodic under mild conditions and converges exponentially fast to a stationary distribution. By analyzing the stationary distribution, we can derive a simple expression for the long-run average cost of the dual-index policy. This expression serves as the foundation for defining and analyzing the regret of our learning algorithm.

Developing online learning algorithms and proving (tight) regret. When the demand distribution is not known *a priori*, we propose a nonparametric learning algorithm to approach the optimal (z_r, z_e) of the dual-index policy. The performance measure is regret, which is the cost difference between a feasible learning algorithm and the clairvoyant (full-information) optimal dual-index policy. We develop a nonparametric online learning algorithm and show that it admits a regret upper bound of $O(\sqrt{T \log T})$ in Theorem 3.5.1, which matches the regret lower bound for any feasible learning algorithms up to a logarithmic factor.

Let $\Delta = z_r - z_e$. Next, we explain the three key ideas in our learning algorithm.

1. We propose a two-layer learning algorithm that updates the two parameters (Δ, z_e) simultaneously. The outer layer discretizes the set of Δ to obtain a grid and treats each point on the grid as an arm of the bandit problem. The algorithm maintains an active set for Δ and approximates the clairvoyant optimal value by pruning the poorly behaved choices in the active set. The inner layer updates the expedited order-up-to level z_e using the empirical quantile to approximate the optimal expedited order-up-to level $z_e^*(\Delta)$ according to Proposition 3.3.2 for each Δ choice in the active set. This is the first attempt in the literature to integrate bandits with sample average approximation methods, which also provably achieves a tight regret bound.

2. In the pruning process to search for the optimal Δ, we utilize the same data set to evaluate each arm within the current active set. Because of the special structure of the inventory system, after pulling one arm, we are able to evaluate the performance of all the arms based on the realized demand data, which enables us to achieve high estimation accuracies of all arms without extensive exploration of each of them. This property of the problem enables our algorithm to achieve a tight regret convergence rate because it maximizes the usage of any realized demand information.

3. In the estimation procedure, we apply SAA for both the empirical quantile solution for $z_e^*(\Delta)$ and the average period cost for each Δ choice. Due to the dependency between two consecutive samples, we study the concentration behavior of the estimation based on data from a Markov chain and utilize the ergodicity of the system variable to achieve a tight regret bound for the regret analysis. A key step in our proofs leverages a specific version of McDiarmid's inequality for Markov chains (Ortner 2020, Paulin 2015).

3.1.2 Literature Review

Our work is related to the following streams of literature.

Dual-sourcing systems. There has been a considerable amount of literature devoted to studying dual-sourcing inventory systems for the past six decades due to their practical importance. However, analyzing such systems is challenging due to the multi-dimensional inventory pipeline vector that needs to be tracked. Earlier research focused on finding exact optimal policies. For systems with consecutive lead times, where the lead times of the expedited and regular sources are k and $k+1$, respectively, for some non-negative k, Fukuda (1964) showed that the optimal policy is a Single-Index Dual-Base-Stock policy. For $k = 0$, Bulinskaya (1964) derived the explicit form of the optimal parameters. However, when the lead times are arbitrary, Whittemore and Saunders (1977) pointed out that the optimal policy would be state-dependent and difficult to obtain, a finding that was confirmed by Sheopuri et al. (2010).

Since deriving the optimal policy for dual-sourcing systems with arbitrary demands is complex and difficult (Janakiraman and Seshadri 2017), recent literature has focused on developing effective heuristic policies that are reasonable and easy to implement. For example, the vector base-stock policy (Sheopuri et al. 2010), the capped dual-index policy (Sun and Van Mieghem 2019), and the tailored base-surge (TBS) policy (Allon and Van Mieghem 2010, Janakiraman et al. 2015, Xin and Goldberg 2018) have been proposed. The dual-index policy is one of the most important heuristic policies and was first introduced by Veeraraghavan and Scheller-Wolf (2008). The optimal dual-index policy mimics the behavior of the complex optimal state-dependent policy found via dynamic programming and thus performs nearly optimally for most cases. The dual-index policy has also been shown to exhibit excellent performance in Li and Yu (2014) and Hua et al. (2015). Furthermore, the vector-based policy proposed by Sheopuri et al. (2010) and the capped dual-index policy proposed by Sun and Van Mieghem (2019) are both variants of the basic dual-index policy.

Nonparametric learning algorithms in inventory management. There has been an ongoing research effort focused on developing nonparametric online learning algorithms for inventory systems. For instance, Shi et al. (2016) and Chen et al. (2020c) have studied the capacitated inventory system, while Zhang et al. (2018) has investigated the inventory system with perishable products. The lost sales inventory system with lead times has been examined by Agrawal and Jia (2022), Huh et al. (2009), Zhang et al. (2020). Cheung et al. (2022) explored the general resource allocation problem. One popular method for tackling these problems is the sample average approximation (SAA) approach (see, e.g., Kleywegt et al. (2002), Levi et al. (2015, 2007a)), which leverages the empirical distribution derived from *uncensored* samples drawn from the actual distribution. This approach has also been embedded in part of our algorithm design. More recently, Gong and Simchi-Levi (2023) leveraged Q-learning techniques for online decision-making in inventory systems with cyclic demands and proposed two algorithms for subsets of problems with full feedback and one-sided feedback of demand.

There has been a stream of online learning algorithms in which there are two decision variables to determine in each period. Our work is a sequel to Chen and Shi (2020), who proposed an online algorithm that converges to the optimal TBS policy for the dual-sourcing system. Our work differs from Chen and Shi (2020) in two main aspects. Firstly, we adopt the optimal dual-index policy as the clairvoyant benchmark, and the system dynamics are very different. Secondly, we consider general lead times for both regular and expedited sources. These two differences result in a completely new algorithmic design and analysis. For other applications, Chen et al. (2019a), Chen et al. (2021a), Chen et al. (2022a) and Chen et al. (2020a) studied the joint pricing and inventory control problem. Yuan et al. (2021) studied inventory management with the fixed cost where the two decision variables are the reorder point and the order-up-to level. To the best of our knowledge, this chapter is the first to consider learning the optimal dual-index policy with general lead times.

3.1.3 Organization and Notation

The remainder of the chapter is organized as follows. We formulate the periodic-review dynamic stochastic dual sourcing inventory model in §3.2. We identify a sufficient condition for the Markov chain associated with a dual-index policy to be ergodic in §3.3. Based on the uniform ergodicity of the Markov chain, we propose our online learning algorithm in §3.4 with regret analyzed in §3.5. Finally, we conclude the chapter and point out several future research directions in §3.7.

We introduce the general notation used in this chapter. For any real number x, we denote $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. For event A, the indicator function $\mathbb{1}(A)$ takes value 1 if A is true and 0 otherwise. The event $A^{\complement}$ denotes the complement of event A. For integer $n \geq 1$, $[n]$ means the set $\{1, \ldots, n\}$. For some of the random variables of interest denoted by upper-case characters, we use corresponding lower-case characters to denote their realizations.

3.2 Model Formulation

We formally describe the periodic-review dual-sourcing system with backlogged demand. Let $t \in \{1, 2, \ldots\}$ represent the time period indexed forward. We denote the demand in period t

by D^t and its realization by d^t. We assume that D^t, $t = 1, \dots, T$, are i.i.d. continuous random variables across periods with cumulative distribution function $H(\cdot)$ and mean μ. Let D be a (time) generic random variable of D^t, and for notational convenience, we shall use them interchangeably unless there is ambiguity. For any $t \in [T],\ k \in \mathbb{Z}$, denote $D_k^t := \sum_{i=0}^{k} D^{t+i}$, which is the cumulative demand from period t to period $t + k$ with k being an integer either positive or negative. Note that when $k < 0$, $D_k^t := \sum_{i=k}^{0} D^{t+i}$. We use d_k^t to denote the realization of D_k^t.

When making replenishment in each period, the firm can either order through the regular channel at the unit cost c_r or through the expedited channel with a shorter lead time at some premium cost c_e per unit. Let q_e^t and q_r^t be the ordering quantities from the expedited source and the regular source, respectively. Regular orders arrive after l_r periods and the lead time for expedited orders is l_e with $l_e < l_r$, i.e., the lead time difference is $l := l_r - l_e \geq 1$. At the beginning of period t, the on-hand inventory level is denoted by I^t. Demands are satisfied as much as possible by the on-hand inventory. Any demand that is not satisfied is backlogged and incurs a per-unit penalty cost of b. If there are leftover inventories at the end of each period, they incur a per-unit holding cost of h.

3.2.1 Sequence of Events under Dual-Index Policy

Now we present the sequence of events under a dual-index policy with parameters (z_r, z_e). Define $\Delta = z_r - z_e$ and denote the lower bound of Δ as $\underline{\Delta}$. Note that $\underline{\Delta} > 0$; otherwise, the problem would be reduced to a single-sourcing system. Instead of using (z_r, z_e) to represent a dual-index policy, we use (Δ, z_e) which is more amenable for later algorithmic design and analysis. Period t begins with on-hand inventory I^t. Note that if we define "the next x periods" to be period $t,\ t+1,\ \dots,\ t+x$, before the firm places any order in this period, there are already expedited orders $q_e^{t-l_e}, \dots, q_e^{t-1}$ due to arrive in the next $l_e - 1$ periods and regular orders $q_r^{t-l_r}, \dots, q_r^{t-1}$ due to arrive in the next $l_r - 1$ periods. These are also referred to as pipeline inventories. At the beginning of period t, the expedited inventory position IP_e^t equals the on-hand inventory plus the orders due to arrive in the next $l_e - 1$ periods while the regular inventory position IP_r^t equals the on-hand inventory plus the orders due to arrive in the next $l_r - 1$ periods. More specifically, one has

$$\begin{aligned} IP_e^t &= I^t + (q_e^{t-l_e} + \ldots + q_e^{t-1}) + (q_r^{t-l_r} + \ldots + q_r^{t-l-1}), \\ IP_r^t &= I^t + (q_e^{t-l_e} + \ldots + q_e^{t-1}) + (q_r^{t-l_r} + \ldots + q_r^{t-1}) \\ &= IP_e^t + q_r^{t-l} + \ldots + q_r^{t-1}. \end{aligned}$$

At the beginning of period t, the firm places an expedited order q_e^t to raise the expedited inventory position IP_e^t to the target level z_e, and after that, the firm places a regular order q_r^t to raise the regular inventory position IP_r^t up to z_r. Note that at the beginning of period t, the regular order q_r^{t-l} enters the consideration time window of the expedited source, because it will arrive in period $t + l_e$. After the addition of q_r^{t-l}, the expedited inventory position may be already larger than the expedited order up to level z_e, in which case an "overshoot" happens. The quantity of overshoot is denoted by O^t, which can be represented as $O^t = (IP_e^t + q_r^{t-l} - z_e)^+$. The realization of O^t is denoted by o^t. Therefore, the firm places an expedited order

$$q_e^t = (z_e - IP_e^t - q_r^{t-l})^+.$$

Then the firm takes the expedited order made into consideration and places the regular order

$$\begin{aligned} q_r^t &= z_r - IP_r^t - q_e^t \\ &= D^{t-1} - (z_e - IP_e^t - q_r^{t-l})^+. \end{aligned}$$

Next, we summarize the sequence of events:

1. In period t, the firm begins with on-hand inventory level $I^t \in \mathbb{R}$.

2. The expedited inventory position is IP_e^t and q_r^{t-l} enters the time window, based on which the firm makes q_e^t expedited orders at the cost of c_e per unit.

3. The regular inventory position is IP_r^t, and the firm makes q_r^t regular order at the unit cost c_r.

4. The orders $q_r^{t-l_r}$ and $q_e^{t-l_e}$ physically arrive. The demand D^t realizes to be d^t, which is satisfied by the on-hand inventory to the maximum extent; any excess demand is backlogged. By Lemmas 4.2 and 4.3 in Veeraraghavan and Scheller-Wolf (2008), the on-hand inventory for the next period can be represented as

$$\begin{aligned} I^{t+1} &= z_e + o^{t-l_e} - d_{-l_e}^t \\ &= z_r - (q_r^{t-l_e} + q_r^{t-l_e-1} + \ldots + q_r^{t-l_r+1}) - (d^{t-l_e} + \ldots + d^t), \end{aligned}$$

   incurring a holding and penalty cost of $h(I^{t+1})^+ + b(I^{t+1})^-$.

The total cost incurred in period t, which consists of ordering costs from both two channels as well as holding and penalty costs, is the following,

$$C^t(\Delta, z_e) = c_e q_e^t + c_r q_r^t + h(I^{t+1})^+ + b(I^{t+1})^-. \tag{3.1}$$

The firm would like to minimize the total expected cost over the planning horizon of T periods.

3.3 Ergodicity of Underlying Markov Chain and Notion of Regret

3.3.1 Markov Chain

Before introducing the problem objective and the notion of regret, we first analyze the underlying stochastic process of this problem. Given the two order-up-to levels z_e, z_r in the dual-index policy, define the state variable

$$W^t(z_e, z_r) := (q_r^{t-1}, \ldots, q_r^{t-l+1}, IP_e^t + q_r^{t-l}) \in \mathbb{R}_+^{l-1} \times \mathbb{R}.$$

As $q_e^t = \left(z_e - IP_e^t - q_r^{t-l}\right)^+$, the last component of $W^t(z_e, z_r)$ captures the information of the expedited order to make. The first $l-1$ components are the regular order from period $t-1$ to period $t-(l-1)$, which capture the regular orders to arrive from period $t+l-1$ to period $t+1$. The next result shows that $(W^t(z_e, z_r), t \geq 1)$ forms a Markov chain.

Lemma 3.3.1 *$(W^t(z_e, z_r), t \geq 1)$ forms a Markov chain.*

Proof of Lemma 3.3.1. We would like to show that $W^{t+1}(z_e, z_r) = (q_r^t, \ldots, q_r^{t-l+2}, IP_e^{t+1} + q_r^{t-l+1})$ only depends on $W^t(z_e, z_r) = (q_r^{t-1}, \ldots, q_r^{t-l+1}, IP_e^t + q_r^{t-l})$. First note that $q_r^{t-1}, \ldots, q_r^{t-l+2}$ are included in $W^t(z_e, z_r)$. Next we tackle q_r^t and $IP_e^{t+1} + q_r^{t-l+1}$.

By the system dynamics, we have

$$IP_r^t = IP_r^{t-1} + q_e^{t-1} + q_r^{t-1} - D^{t-1} = z_r - D^{t-1}.$$

Hence we have the following relationship,

$$q_r^t = z_r - (IP_r^t + q_e^t) = D^{t-1} - q_e^t = D^{t-1} - (z_e - IP_e^t - q_r^{t-l})^+,$$

which only depends on W^t. In addition,

$$IP_e^{t+1} + q_r^{t-l+1} = \max(z_e, IP_e^t + q_r^{t-l}) - D^t + q_r^{t-l+1},$$

which also only depends on W^t. **Q.E.D.**

Define $\bar{W}^t(z_e, z_r) := (q_r^{t-1}, \ldots, q_r^{t-l+1}, \max(z_e, IP_e^t + q_r^{t-l})) \in \mathbb{R}_+^l$.

3.3.2 Ergodicity

As shown above, $(W^t(z_e, z_r), t \geq 1)$ forms a Markov chain, which may not necessarily be ergodic due to the complicated structure of the process. Next, we provide a sufficient condition under which the Markov chain can be proved to be ergodic. For any (z_e, z_r), define

$$\gamma(z_e, z_r) := \mathbb{P}\left(D \leq \frac{z_r - z_e}{l_r + 1}\right).$$

Assumption 3.3.1 *The following two conditions hold.*

(a) $\left(1 - \mathbb{P}^{2l_r}\left(D \leq \frac{\underline{\Delta}}{l_r+1}\right)\right)^{\log T} \leq \frac{1}{\sqrt{e}}$,

(b) $\mathbb{P}\left(D \leq \frac{\bar{Z}}{l_r+1}\right) \leq \lambda$ *with* $\lambda < 1$ *being a known constant.*

Because $\underline{\Delta} > 0$, Assumption 3.3.1a is readily satisfied for large values of T. Assumption 3.3.1b is also very mild because $l_r + 1 \geq 2$. Recall that $\underline{\Delta}$ is the lower limit of $\Delta = z_r - z_e$ and $\bar{Z}$ is the upper limit of z_r (and Δ).

Next we show that under Assumption 3.3.1a, the Markov chain $W^t(z_e, z_r)$ is ergodic. Note that the Markov chain $\{W^t(z_e, z_r) : t \geq 1\}$ is ergodic if there exists a random variable $W^\infty(z_e, z_r)$ such that for any initial state w^1,

$$\lim_{t \to \infty} \delta_t\left(z_e, z_r, w^1\right) = 0,$$

where for any $t \geq 1$,

$$\begin{aligned} &\delta_t\left(z_e, z_r, w^1\right) \qquad (3.2)\\ =&\sup\Bigg\{\left|\mathbb{P}\left(W^t(z_e, z_r) \in \Omega \mid W^1(z_e, z_r) = w^1\right) - \mathbb{P}\left(W^\infty(z_e, z_r) \in \Omega\right)\right| :\\ &\text{measurable set } \Omega \subseteq \mathbb{R}_+^{l-1} \times \mathbb{R}\Bigg\}. \end{aligned}$$

In such case, we say $W^\infty(z_e, z_r)$ is the steady-state vector of $\{W^t(z_e, z_r) : t \geq 1\}$. With the definitions above, we have the following result for the Markov chain $\{W^t(z_e, z_r) : t \geq 1\}$. The proof of Theorem 3.3.1 is relegated to Appendix B.2.1.

Theorem 3.3.1 *Let z_e and z_r be the base stock levels for expedited and regular orders respectively. Under Assumption 3.3.1a, the Markov chain $\{W^t(z_e, z_r) : t \geq 1\}$ is ergodic with a steady-state random variable $W^\infty(z_e, z_r)$. Furthermore, for any initial inventory vector $w^1 \in \mathbb{R}_+^{l-1} \times \mathbb{R}$ and $t \geq 2l_r + 1$,*

$$\begin{aligned}
&\delta^{t+1}\left(z_e, z_r, w^1\right) \\
&\leq \begin{cases} (1-\gamma(z_e,z_r)^{2l_r})^{t/4l_r} + H(\bar{w}^1 \cdot \mathbf{1}^l - z_r)^{t/2-l_r}, & \textit{if infinite support}, \\ (1-\gamma(z_e,z_r)^{2l_r})^{t/4l_r} + \exp\left(\dfrac{4(\bar{w}^1 \cdot \mathbf{1}^l - z_r)}{\bar{D}} - \dfrac{2\mu^2(t/2-l_r)}{\bar{D}^2}\right), & \textit{if } D \leq \bar{D} \textit{ w.p. } 1, \end{cases}
\end{aligned}$$

where $\bar{D}$ and $H(\cdot)$ are the upper bound and cumulative distribution function of demand D.

Theorem 3.3.1 states that the Markov chain $W^t(z_e, z_r)$ not only is ergodic but also converges to its steady state $W^\infty(z_e, z_r)$ exponentially fast regardless of the demand distribution or the initial state. This result holds for both D having an infinite support and D upper bounded with probability 1. For the rest of the analysis, we will focus on the case where $D \leq \bar{D}$ with probability 1.

The main idea of the proof of Theorem 3.3.1 is that all sample paths couple after a certain demand pattern. If demand is small enough, say $D^t \leq \dfrac{z_r - z_e}{l_r + 1}$ consecutively for $2l_r$ periods, then in the second l_r periods during the pattern, the expedited order is zero and the regular order only depends on the demand realizations. In this way, all the sample paths will meet regardless of the state of the inventory vector before the demand pattern occurs. This coupling argument builds the foundation for proving the ergodicity of the Markov chain. Note that Huh et al. (2009) proved the ergodicity of a single-sourcing inventory system with lost sales. Because the system dynamics of the dual-sourcing system are completely different, our analysis is also different and new.

3.3.3 Regret Definition

Because the overshoot $O^t = (IP_e^t + q_r^{t-l} - z_e)^+$ only depends on W^t, its steady state distribution exists and is denoted by O^∞. Moreover, the per-period cost $C^t(\Delta, z_e)$ defined in (3.1) also only depends on W^t. This is because

$$\begin{aligned}
q_e^t &= (z_e - IP_e^t - q_r^{t-l})^+, \\
q_r^t &= D^{t-1} - (z_e - IP_e^t - q_r^{t-l})^+, \\
I^{t+1} &= z_e + O^{t-l_e} - D_{-l_e}^t = \max\left\{z_e, IP_e^t + q_r^{t-l}\right\} - D_{-l_e}^t.
\end{aligned}$$

Therefore, there exists a stationary distribution for $C^t(\Delta, z_e)$, denoted by $C^\infty(\Delta, z_e)$, which can be represented as

$$\begin{aligned} C^\infty(\Delta, z_e) &= c_e q_e^\infty + c_r q_r^\infty + h\mathbb{E}\left[(I^\infty)^+\right] + b\mathbb{E}\left[(I^\infty)^-\right] \\ &= c_e q_e^\infty + c_r q_r^\infty + h\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^+\right] + b\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^-\right] \\ &= (c_e - c_r) q_e^\infty + c_r\mathbb{E}\left[D\right] + h\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^+\right] + b\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^-\right], \end{aligned} \tag{3.3}$$

where D_{l_e} denotes the sum of $l_e + 1$ random variable D.

Let (z_e^*, z_r^*) be the clairvoyant optimal dual-index policy that maximizes $\mathbb{E}\left[C^\infty(z_e, z_r)\right]$ under complete information about the demand distribution, and let $\Delta^* = z_r^* - z_e^*$. To solve for (z_e^*, z_r^*), it is equivalent to solve for (Δ^*, z_e^*). We aim to develop a learning algorithm **ALG** that only makes use of historical demand data. Using the clairvoyant optimal dual-index policy (Δ^*, z_e^*) as a natural benchmark, we define the regret by

$$\mathcal{R}_T^{\mathbf{ALG}} := \mathbb{E}\left[\sum_{t=1}^{T} C_{\mathbf{ALG}}^t - TC^\infty(\Delta^*, z_e^*)\right],$$

where $\sum_{t=1}^{T} C_{\mathbf{ALG}}^t$ is the total cost by running the learning algorithm **ALG** when the demand distribution is unknown. Note again that our regret is defined as the cost difference between a feasible learning algorithm and the clairvoyant (full-information) optimal dual-index policy instead of the true optimal policy.

To solve for the optimal (Δ^*, z_e^*), we first quote the following critical result directly from Veeraraghavan and Scheller-Wolf (2008):

Proposition 3.3.1 (Proposition 4.1 and Lemma 5.1 in Veeraraghavan and Scheller-Wolf (2008)). *Both the overshoot O^t and the expedited order quantity q_e^t are functions of Δ, independent of z_e.*

Based on Proposition 3.3.1, to solve for z_e^* in (3.3), one needs to minimize $h\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^+\right] + b\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^-\right]$, which is the newsvendor problem. For any Δ, let F_Δ be the CDF of $D_{l_e} - O^\infty(\Delta)$, the following proposition states that z_e^* is the newsvendor quantile solution.

Proposition 3.3.2 (Theorem 4.1 in Veeraraghavan and Scheller-Wolf (2008)) *The optimal level of z_e given Δ is $z_e^*(\Delta) = F_\Delta^{-1}\left(\frac{b}{b+h}\right)$.*

Note that although the stationary distribution was mentioned in Veeraraghavan and Scheller-Wolf (2008), there were no results in Veeraraghavan and Scheller-Wolf (2008) regarding how fast the system converges to the stationary distribution, which is a critical result for developing online learning algorithms. Therefore, we provide Theorem 3.3.1 confirming that the Markov chain converges to its stationary distribution exponentially fast.

After optimizing over z_e under a given Δ, a natural question to ask is how to solve for the optimal Δ. Unfortunately, there are no structural results of the objective as a function of Δ even with known demand distribution, and one needs to conduct a one-dimensional search to solve for the optimal Δ.

3.4 Online Learning Algorithm

We propose an online learning algorithm called (Δ, z_e) algorithm that converges to the clairvoyant optimal dual-index policy with a provably fast rate.

The development of the (Δ, z_e) algorithm is based on Propositions 3.3.1 and 3.3.2. Next, we present a high-level overview of it. At the beginning of the algorithm, the algorithm discretizes Δ into J arms and initializes the demand dataset $\mathcal{D}^0$ to be empty. The algorithm consists of two layers, with the outer layer learning the optimal Δ via adaptively eliminating inferior arms and the inner layer approximating the optimal z_e under a given Δ using empirical quantiles. The algorithm proceeds in epochs. Epoch n starts with an active set $\mathcal{A}^n \subset [J]$ that contains all the "good performing" arms of Δ evaluated by the empirical average cost. For each $j \in \mathcal{A}^n$, the algorithm has estimated the empirical optimal expedited order-up-to level z_{ej}^n using historical data. During epoch n, the algorithm randomly selects an arm Δ_{j^n} from the active set $\mathcal{A}^n$ and operates the system for B^n periods under $(\Delta_{j^n}, z_{ej^n}^n)$. After demands during epoch n realize, the demand dataset is updated from $\mathcal{D}^{n-1}$ to $\mathcal{D}^n$ with the addition of new demand data. The updated dataset $\mathcal{D}^n$ is then applied to simulate the dual-index policy for each $j \in \mathcal{A}^n$. Specifically, based on $\mathcal{D}^n$, the algorithm recomputes the empirical quantile as the estimator for the optimal expedited order-up-to level and updates it from z_{ej}^n to z_{ej}^{n+1}. In addition, the algorithm also updates the empirical average cost under Δ_j. By comparing the empirical average costs under Δ_j, $j \in \mathcal{A}^n$, the algorithm prunes the active set $\mathcal{A}^n$ according to a confidence size ε^n and obtains $\mathcal{A}^{n+1}$.

In the algorithm, let $J = \lfloor \sqrt{T} \rfloor, B^n = \lceil \frac{2^n}{\log T} \rceil$, where J is the number of arms for Δ after discretization and B^n is the length of epoch n. Let $\underline{\Delta}$ be the lower limit of $\Delta = z_r - z_e$ and $\bar{Z}$ be the upper limit of z_r (and Δ). Denote $\underline{\mu}$ as the known lower limit of the demand mean μ. Let $\hat{F}_{t,\Delta_j}(x) = \hat{P}(D_{-l_e}^t - O^{t-l_e}(\Delta_j) \leq x)$ be the empirical cumulative distribution function of $D_{-l_e}^t - O^{t-l_e}(\Delta_j)$ under Δ_j at time t. We let all variables with superscripts not in $[T]$ be 0. The detailed pseudo-code for the (Δ, z_e) algorithm is presented in Algorithm 5.

As the simulation process is embedded in Algorithm 5, we discuss its runtime and space complexity here. The initialization step takes $O(\sqrt{T})$ time. For each epoch n, the dual-index policy with selected indices $(z_e^t, z_r^t) = (z_{ej^n}^n, z_{ej^n}^n + \Delta_{j^n})$ has a time complexity of

Algorithm 5 The (Δ, z_e) learning algorithm (π for short) for the dual-index policy

```
Let $J = \#$ discrete $\Delta$'s and $N = \min\left\{n : \sum_{i=1}^{n} \lceil \frac{2^i}{\log T} \rceil \geq T\right\}$ the number of epochs.
Let $B^i = \lceil \frac{2^i}{\log T} \rceil$ be the $i$-th epoch length.
Let $L^n = \sum_{i=1}^{n} B^i$, $\forall n \in [N-1]$ with $L^0 = 0, L^N = T$.          ▷ Parameters
Initialize the active set $\mathcal{A}^1 = \{1, \ldots, J\}$, $\mathcal{D}^0 = \varnothing$.          ▷ Initialization
For $j \in \mathcal{A}^1$, define $\Delta_j = \underline{\Delta} + \frac{j}{J}\left|\bar{Z} - \underline{\Delta}\right|$ and assign $z_{ej}^1 \in [0, \bar{Z} - \Delta_j]$ arbitrarily.
for $n = 1, 2, \ldots, N$ do          ▷ Outer Loop
    Randomly select $j^n \in \mathcal{A}^n$. Let demand set $\mathcal{D}^n = \mathcal{D}^{n-1}$.
    for $t = L^{n-1} + 1, \ldots, \min\{L^n, T\}$:  do
        Apply the dual-index policy $(z_e^t, z_r^t) = (z_{ej^n}^n, z_{ej^n}^n + \Delta_{j^n})$.
        Append the realized demand $d^t$ into $\mathcal{D}^n$.

            $q_e^t = (z_{ej^n}^n - IP_e^t - q_r^{t-l})^+, q_r^t = (z_{ej^n}^n + \Delta_{j^n} - IP_r^t - q_e^t)^+,$
            $IP_e^{t+1} = IP_e^t + q_e^t - d^t + q_r^{t-l}, IP_r^{t+1} = IP_r^t + q_e^t + q_r^t - d^t,$
            $o^t = (IP_e^t + q_r^{t-l} - z_{ej^n}^n)^+, I^{t+1} = I^t + q_e^{t-l_e} + q_r^{t-l_r} - d^t.$

    end for
    for $j \in \mathcal{A}^n$ do          ▷ Inner Loop
        Simulate the policy $(z_{ej}^n, z_{ej}^n + \Delta_j)$ for $\min\{L^n, T\}$ periods using $\mathcal{D}^n$ and denote the
    state variables of this simulation as $\hat{W}_j^t := (\hat{q}_{rj}^{t-1}, \ldots, \hat{q}_{rj}^{t-l+1}, \widehat{IP}_{ej}^t + \hat{q}_{rj}^{t-l}) \in \mathbb{R}_+^{l-1} \times \mathbb{R}$ for
    $t = 1, \ldots, L^n$.
        Obtain the estimated average period cost:

            $\hat{G}_j^n = \frac{1}{L^n} \sum_{t \in [L^n]} c_e \hat{q}_{ej}^t + c_r \hat{q}_{rj}^t + h(\hat{I}_j^{t+1})^+ + b(\hat{I}_j^{t+1})^-.$

        Let $\mathcal{X}_j^n = \left\{d_{l_e}^t - \hat{o}_j^t, t \in [L^n - l_e]\right\}$.
        Let $\hat{F}_{\Delta_j}^n(\cdot)$ be the empirical CDF of $X_j^t = D_{l_e}^t - \hat{O}_j^t(\Delta_j)$ with data sample $\mathcal{X}_j^n$.
        Update $z_{ej}^{n+1} = \hat{F}_{\Delta_j}^{n^{-1}}(\frac{b}{b+h})$.          ▷ Inner Layer Optimization
    end for
    Update and prune the active set          ▷ Outer Layer Optimization

            $\mathcal{A}^{n+1} = \left\{j \in \mathcal{A}^n : \hat{G}_j^n - \min_{j' \in \mathcal{A}^n} \hat{G}_{j'}^n \leq \varepsilon^n\right\}.$

end for
```

$O(B^n)$. Then, the simulation process iterates over all $O(\sqrt{T})$ arms in $\mathcal{A}^n$, with each arm having a time complexity of $O(L^n)$. Thus, the total time complexity of Algorithm 5 is $O(\sum_{n=1}^{N}(B^n + \sqrt{T}L^n))$, which is $O(T^{\frac{3}{2}})$ since $N \leq \log_2(T\log T + 2) - 1$. For Algorithm 5 to work, we require $O(\sqrt{T})$ space to store the information for the active arm set $\mathcal{A}^n$, including the Δ_j, z_{ej}, and $\hat{G}_j^n$ for arm j. In the implementation of the dual-index policy, we need to store the information of q_e^t, q_r^t, q_r^{t-l}, IP_e^t, IP_r^t, O^t, and I^t, which is of $O(1)$ space complexity. For the simulation process, we need the same $O(1)$ vector to store the information for each arm and $O(T)$ to store the demand information. Thus, the total space complexity of Algorithm 5 is $O(T)$.

3.5 Regret Analysis

We have the following convergence result for the regret of our (Δ, z_e) algorithm. For ease of notation, we refer to our (Δ, z_e) learning algorithm as Algorithm π for the remainder of this chapter.

Theorem 3.5.1 *Set the parameters $J = \lfloor\sqrt{T}\rfloor$ and $\varepsilon^n = 2\,(b+h)\,\bar{Z}\alpha^n + 2\beta^n$, where*

$$\alpha^n = \frac{3}{2}\sqrt{\frac{3T_0\log T}{L^{n-1}}} \quad \text{and} \quad \beta^n = \frac{3\bar{C}}{2}\sqrt{\frac{T_0\log T}{L^n}},$$

and T_0 being a constant defined in (3.9) and $\bar{C} = (c_e + c_r + h)\bar{Z} + b(l_e+1)\bar{D}$ being the upper limit of the cost per period. Then the regret of our learning algorithm $\mathcal{R}_T^\pi = O\big(\sqrt{T\log T}\big)$.

The regret upper bound in Theorem 3.5.1 is based on our regret definition in §3.3.3, which uses a "weaker" clairvoyant benchmark, namely the optimal dual-index policy, rather than the true optimal policy for the dual-sourcing system, which is state-dependent and difficult to obtain.

The following proposition establishes the lower bound (for any online learning algorithms).

Proposition 3.5.1 *Suppose that $T > 5$. The expected regret for any learning algorithm for our dual-sourcing problem is lower bounded by $\Omega(\sqrt{T})$.*

Proposition 1 in Zhang et al. (2020) provides an example of a demand scenario in which any learning algorithm incurs a regret of at least $\Omega(\sqrt{T})$. This proposition examines a single source problem, which is a special case of our problem with $l_e = 0$ and $l_r = \infty$. We thus omit the proof of Proposition 3.5.1 here. By combining the results from Theorem 3.5.1 and Proposition 3.5.1, we assert that the regret upper and lower bounds match, up to a logarithmic factor.

Now, the remainder of this section will be devoted to establishing Theorem 3.5.1. Denote $z_e^*(\Delta)$ as the optimal expedited order-up-to level given Δ and j^* as the index of the optimal arm of Δ among all the discretized values. Recall that C_π^t is the cost in period t by running our learning algorithm π. To prove Theorem 3.5.1, we decompose the regret of our algorithm π into four parts as below:

$$\begin{aligned}
\mathcal{R}_T^\pi =& \mathbb{E}\left[\sum_{t=1}^{T} C_\pi^t - TC^\infty(\Delta^*, z_e^*)\right] \\
=& \mathbb{E}\left[\sum_{t=1}^{T} \left(C_\pi^t - C^\infty\left(\Delta^t, z_e^t\right)\right)\right] && \textbf{(Nonstationarity Loss)} \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \left(C^\infty\left(\Delta^t, z_e^t\right) - C^\infty\left(\Delta^t, z_e^*(\Delta^t)\right)\right)\right] && \textbf{(Empirical Suboptimality Loss)} \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \left(C^\infty(\Delta^t, z_e^*(\Delta^t)) - C^\infty(\Delta_{j^*}, z_e^*(\Delta_{j^*}))\right)\right] && \textbf{(Bandit Pruning Loss)} \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \left(C^\infty\left(\Delta_{j^*}, z_e^*(\Delta_{j^*})\right) - C^\infty\left(\Delta^*, z_e^*\right)\right)\right]. && \textbf{(Discretization Loss)}
\end{aligned}$$

Then, to establish Theorem 3.5.1, it suffices to bound the above four terms. We provide an explicit upper bound for each of them as follows.

Proposition 3.5.2 (Nonstationary Loss Bound)

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right] = O((\log T)^3).$$

Proposition 3.5.3 (Empirical Suboptimality Loss Bound)

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(C^\infty(\Delta^t, z_e^t) - C^\infty(\Delta^t, z_e^*(\Delta^t))\right)\right] = O\left(\sqrt{T \log T}\right).$$

Proposition 3.5.4 (Bandit Pruning Loss Bound)

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(C^\infty\left(\Delta^t, z_e^*\left(\Delta^t\right)\right) - C^\infty\left(\Delta_{j^*}, z_e^*\left(\Delta_{j^*}\right)\right)\right)\right] = O\left(\sqrt{T \log T}\right).$$

Proposition 3.5.5 (Discretization Loss Bound)

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(C^\infty\left(\Delta_{j^*}, z_e^*(\Delta_{j^*})\right) - C^\infty\left(\Delta^*, z_e^*\right)\right)\right] = O(\sqrt{T}).$$

To help illustrate the structure of the proof of Theorem 3.5.1, we present a roadmap showing the main steps in regret analysis for the algorithm in Figure 3.1.

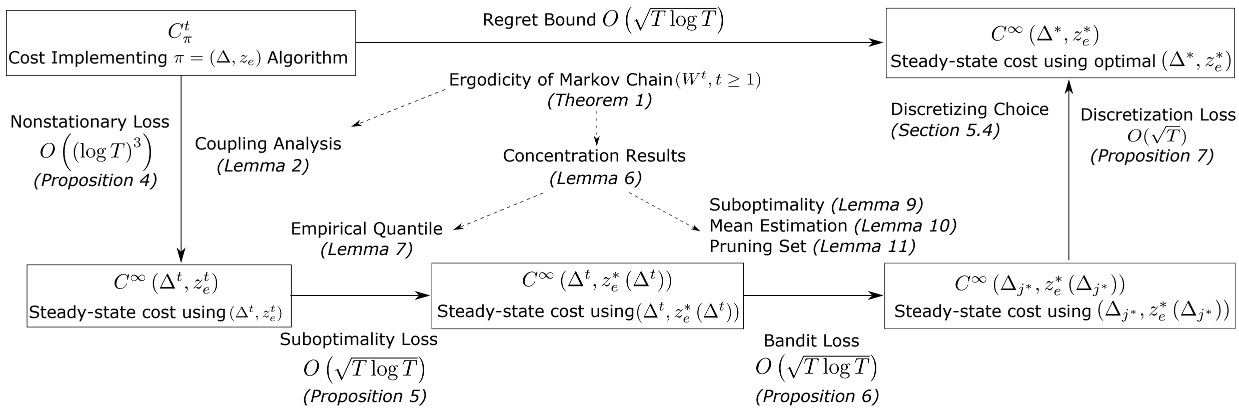

C_π^t
Cost Implementing $\pi = (\Delta, z_e)$ Algorithm
Regret Bound $O\left(\sqrt{T \log T}\right)$
$C^\infty(\Delta^*, z_e^*)$
Steady-state cost using optimal (Δ^*, z_e^*)
Ergodicity of Markov Chain $(W^t, t \geq 1)$
(Theorem 1)
Nonstationary Loss
$O\left((\log T)^3\right)$
(Proposition 4)
Coupling Analysis
(Lemma 2)
Concentration Results
(Lemma 6)
Discretizing Choice
(Section 5.4)
Discretization Loss
$O(\sqrt{T})$
(Proposition 7)
Empirical Quantile
(Lemma 7)
Suboptimality (Lemma 9)
Mean Estimation (Lemma 10)
Pruning Set (Lemma 11)
$C^\infty(\Delta^t, z_e^t)$
Steady-state cost using (Δ^t, z_e^t)
$C^\infty(\Delta^t, z_e^*(\Delta^t))$
Steady-state cost using $(\Delta^t, z_e^*(\Delta^t))$
$C^\infty(\Delta_{j^*}, z_e^*(\Delta_{j^*}))$
Steady-state cost using $(\Delta_{j^*}, z_e^*(\Delta_{j^*}))$
Suboptimality Loss
$O\left(\sqrt{T \log T}\right)$
(Proposition 5)
Bandit Loss
$O\left(\sqrt{T \log T}\right)$
(Proposition 6)

Figure 3.1: High-Level Roadmap for Regret Analysis

3.5.1 Proof of Proposition 3.5.2 - Bound the Nonstationarity Loss

To bound $\mathbb{E}\left[\sum_{t=1}^T \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right]$, we first propose a result of the mixing property of the process. Note that mixing properties are important for the analysis of Markov chains, especially in non-episodic reinforcement learning such as Azizzadenesheli et al. (2016).

Lemma 3.5.1 *Consider the process $(W^t(z_e, z_r), t \geq 1)$ under the dual-index policy (z_e, z_r) with arbitrary $\bar{W}^1 \cdot \mathbf{1}^l$. Define an auxiliary process $(\tilde{W}^t(z_e, z_r), t \geq 1)$ where $\tilde{W}^1(z_e, z_r)$ is randomly sampled from the steady-state $W^\infty(z_e, z_r)$, and $(\tilde{W}^t(z_e, z_r), t \geq 1)$ is driven in the same way by the same demand process as $(W^t(z_e, z_r), t \geq 1)$. Then under Assumption 3.3.1a, we have*

$$\mathbb{P}\left(W^{\tau+1} = \tilde{W}^{\tau+1}\right) \geq 1 - \frac{K_0 + 1}{T^{\frac{5}{2}}},$$

where $\tau := \lceil \frac{5\bar{D}^2}{4\mu^2} \log T \rceil + 2l_r \lceil 5 (\log T)^2 \rceil$.

Lemma 3.5.1 states that the processes with two different starting states will couple after $O\left((\log T)^2\right)$ periods with high probability. The intuition for proving Lemma 3.5.1 is the following.

1. Regardless of the initial state, after implementing (z_e, z_r) for $O(\log T)$ periods for any given (z_e, z_r), the regular inventory position will not exceed z_r with high probability.

2. When there is no overshoot over z_r, suppose that we create an auxiliary process starting with a state sampling from the steady-state $W^\infty(z_e, z_r)$ and sharing the same demand sample path with $W^t(z_e, z_r)$. Then after $O\left((\log T)^2\right)$ periods, the process $(W^t(z_e, z_r), t \geq 1)$ will couple with the auxiliary process with high probability.

Before introducing the formal proof for Lemma 3.5.1, we first introduce the following two lemmas that support the intuitions explained above.

Lemma 3.5.2 *Consider the process* $(W^t(z_e, z_r), t \geq 1)$ *under the dual-index policy* (z_e, z_r)*, for any initial state* $W^1(z_e, z_r)$ *we have*

$$\mathbb{P}\left(\bar{W}^{\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r\right) \geq 1 - \frac{K_0}{T^{\frac{5}{2}}},$$

where $K_0 = e^{\frac{4(\bar{Z}-z_r)}{\bar{D}} + \frac{2\mu^2 l_r}{\bar{D}^2}}$.

Lemma 3.5.3 *Consider the process* $(W^t(z_e, z_r), t \geq 1)$ *under dual-index policy* (z_e, z_r) *with* $\bar{W}^1 \cdot \mathbf{1}^l \leq z_r$*. Define an auxiliary process* $(\tilde{W}^t(z_e, z_r), t \geq 1)$ *where* $\tilde{W}^1(z_e, z_r)$ *is randomly sampled from the steady-state* $W^\infty(z_e, z_r)$*. Then under Assumption 3.3.1a, we have*

$$\mathbb{P}\left(W^{2l_r\lceil 5\log T\rceil+1} = \tilde{W}^{2l_r\lceil 5\log T\rceil+1}\right) \geq 1 - \frac{1}{T^{\frac{5}{2}}}.$$

Proof of Lemma 3.5.2. Note that the sum of the components of $\bar{W}^t(z_e, z_r)$ is the regular inventory position before making the regular order, i.e., $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l = IP_r^t + q_e^t$. Denote events $S = \{\bar{W}^{\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r\}$ and

$$\tilde{S}(\bar{w}^1) = \left\{\bar{W}^{\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r | \bar{W}^1(z_e, z_r) = \bar{w}^1\right\},$$

and the latter means that the regular inventory position before making regular order is no larger than z_r in period $\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil$ given the regular inventory position in period 1 is $\bar{w}^1$.

By Lemma B.2.3 in Appendix B.2.1, we have

$$\mathbb{P}\left(\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l > z_r | \bar{W}^1(z_e, z_r) = \bar{w}^1\right) \leq e^{4(\bar{w}^1 \cdot \mathbf{1}^l - z_r)/\bar{D}} \cdot e^{-2\mu^2(t-l_r)/\bar{D}^2}.$$

Therefore, when $t = \lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil \geq \frac{5\bar{D}^2}{4\mu^2}\log T$, we have

$$\mathbb{P}\left(\tilde{S}(\bar{w}^1)\right) \geq 1 - \frac{e^{\frac{4(\bar{w}^1 \cdot \mathbf{1}^l - z_r)}{\bar{D}} + \frac{2\mu^2 l_r}{\bar{D}^2}}}{T^{\frac{5}{2}}}.$$

As $\bar{w}^1 \cdot \mathbf{1} \leq \bar{Z}$, which is the upper limit of the regular inventory position, we have

$$\mathbb{P}\left(\tilde{S}(\bar{w}^1)\right) \geq 1 - \frac{K_0}{T^{\frac{5}{2}}},\ \forall \bar{w}^1.$$

Hence, we have $\mathbb{P}(S) \geq 1 - \dfrac{K_0}{T^{\frac{5}{2}}}$. **Q.E.D.**

Proof of Lemma 3.5.3. In the auxiliary stochastic process $\{\tilde{W}^t(z_e, z_r), t \geq 1\}$, note that $\tilde{W}^1(z_e, z_r)$ is a realization of the steady state variable $W^\infty(z_e, z_r)$:

$$\tilde{W}^1(z_e, z_r) = (\tilde{q}_r^\infty, \ldots, \tilde{q}_r^\infty, \tilde{IP}_e^\infty + \tilde{q}_r^\infty).$$

For $t \geq 1$, the state variables follow the same system dynamics as $W^t(z_e, z_r)$, i.e., the two processes share the same demand realizations:

$$\begin{aligned}
&\tilde{q}_e^t = (z_e - \tilde{IP}_e^t - \tilde{q}_r^{t-l})^+, \quad \tilde{q}_r^t = z_e + \Delta - \tilde{IP}_r^t - \tilde{q}_e^t, \\
&\tilde{IP}_e^{t+1} = \tilde{IP}_e^t + \tilde{q}_e^t - d^t + \tilde{q}_r^{t-l}, \quad \tilde{IP}_r^{t+1} = \tilde{IP}_r^t + \tilde{q}_e^t + \tilde{q}_r^t - d^t, \\
&\tilde{o}^t = (\tilde{IP}_e^t + \tilde{q}_r^{t-l} - z_e)^+, \quad \tilde{I}^{t+1} = \tilde{I}^t + \tilde{q}_e^{t-l_e} + \tilde{q}_r^{t-l_r} - d^t.
\end{aligned}$$

After a specific demand pattern, the two processes $\tilde{W}^t(\Delta, z_e)$ and $\hat{W}^t(\Delta, z_e)$ will couple, implying that the states in $\tilde{W}^t(\Delta, z_e)$ will be independent of its initial state, given that the initial regular inventory position is at most z_r. This demand pattern is the following:

$$D^t \leq \frac{z_r - z_e}{l_r + 1}, \ \forall 1 \leq t \leq 2l_r. \tag{3.4}$$

After $2l_r$ periods of such a pattern where the demand in each period is small enough, then for $t = l_r + 1, \ldots, 2l_r$, the overshoot is always positive, and the regular order $q_r^t = d^{t-1}$. Hence, if such a pattern ends in period t', then according to (B.1) in Appendix B.2.1, the state in period $t' + 1$ is $W^{t'+1} = (d^{t'-1}, \ldots, d^{t'-l+1}, d^{t'})$ which only depends on the demand realizations.

Thus, we denote event $\tilde{U}$ as the occurrence of such a demand pattern during periods 1 to $2l_r\lceil 5(\log T)^2\rceil$, which means that there are $2l_r$ consecutive periods where demands satisfy $D^t \leq \dfrac{z_r - z_e}{l_r + 1}$. If we divide the $2l_r\lceil 5(\log T)^2\rceil$ periods into $\lceil 5(\log T)^2\rceil$ fragments of length $2l_r$ and consider each fragment independent of the others, the probability that one segment does not follow the pattern specified in (3.4) is at most $1 - \gamma^{2l_r}$. Hence, we have

$$\mathbb{P}\left(\tilde{U}^\complement\right) \leq \left(1 - \gamma^{2l_r}\right)^{\lceil 5(\log T)^2\rceil} \leq \left(1 - \gamma^{2l_r}\right)^{5(\log T)^2},$$

and $1 - \gamma^{2l_r} \in [0, 1)$.

By Assumption 3.3.1a, we have $\left(1-\gamma^{2l_r}\right)^{\log T} \leq \frac{1}{\sqrt{e}}$, which implies

$$\mathbb{P}\left(\tilde{U}^{\complement}\right) \leq \frac{1}{\left(\sqrt{e}\right)^{5\log T}} = \frac{1}{T^{\frac{5}{2}}}.$$

Q.E.D.

Combining Lemma 3.5.2 and 3.5.3, Lemma 3.5.1 naturally holds.

Proof of Lemma 3.5.1. Recall that event $S = \bar{W}^{\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil}(z_e, z_r)\cdot \mathbf{1}^l \leq z_r$, which means that the inventory position drops below z_r after $\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil$ periods. We further denote event U as the demand pattern described in (3.4) occurring during periods $\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil + 1$ to period $\tau = \lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil + 2l_r\lceil 5\left(\log T\right)^2\rceil$. By Lemma 3.5.2, we have

$$\mathbb{P}\left(S\right) \geq 1 - \frac{K_0}{T^{\frac{5}{2}}} \quad \text{with} \quad K_0 = e^{\frac{4(\bar{Z}-z_r)}{\bar{D}} + \frac{2\mu^2 l_r}{\bar{D}^2}},$$

and by Lemma 3.5.3, we have $\mathbb{P}\left(U\right) \geq 1 - \dfrac{1}{T^{\frac{5}{2}}}$. Because event S only depends on the demand realizations during the first $\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil$ periods while event U depends on the demand pattern from period $\lceil \frac{5\bar{D}^2}{4\mu^2}\log T\rceil + 1$ to period τ, we know that event S is independent of event U. Consider the event $\tilde{V} := \left\{W^{\tau+1} = \tilde{W}^{\tau+1}\right\}$, we have $S \cap U \subseteq \tilde{V}$ (if the regular inventory level drops below z_r and the demand pattern in (3.4) appears afterwards, then event $\tilde{V}$ happens), thus $\tilde{V}^{\complement} \subseteq (S\cap U)^{\complement}$. Therefore,

$$\mathbb{P}\left(\tilde{V}^{\complement}\right) \leq \mathbb{P}\left(S^{\complement}\cup U^{\complement}\right) \leq \mathbb{P}\left(S^{\complement}\right) + \mathbb{P}\left(U^{\complement}\right) \leq K_0\frac{1}{T^{\frac{5}{2}}} + \frac{1}{T^{\frac{5}{2}}} = (K_0+1)\frac{1}{T^{\frac{5}{2}}}.$$

Q.E.D.

Now we are ready to apply the mixing property of the process as indicated in Lemma 3.5.1. Let

$$N_0 = \log_2\log T + \log_2\left(10l_r(\log T)^2 + \frac{5\bar{D}^2}{4\mu^2}\log T + 2l_r + 1\right).$$

Then when $n \geq N_0$, we have that the length of the epoch n is at least τ, i.e.,

$$B^n = \left\lceil \frac{2^n}{\log T}\right\rceil \geq \left\lceil \frac{2^{N_0}}{\log T}\right\rceil = 10l_r(\log T)^2 + \frac{5\bar{D}^2}{4\mu^2}\log T + 2l_r + 1 \geq \tau.$$

Specifically, we can decompose the regret in epoch $n \geq N_0$ in the following way:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=L^{N_0}+1}^{T} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right] \\
=&\sum_{n=N_0+1}^{N} \mathbb{E}\left[\sum_{t=L^{n-1}+1}^{L^n} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right] \\
=&\sum_{n=N_0+1}^{N} \mathbb{E}\left[\sum_{t=L^{n-1}+1}^{L^{n-1}+\tau} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right) + \sum_{t=L^{n-1}+\tau+1}^{L^n} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right].
\end{aligned}$$

For each epoch $n \geq N_0$, define an auxiliary process $\{\tilde{W}^t(z_e, z_r)\}_{L^{n-1}+1 \leq t \leq L^n}$ where $\tilde{W}^{L^{n-1}+1}(z_e, z_r)$ is randomly sampled from the steady-state $W^\infty(z_e, z_r)$. Denote event $V^n := \left\{W^{L^{n-1}+\tau+1}(z_e, z_r) = \tilde{W}^{L^{n-1}+\tau+1}(z_e, z_r)\right\}$, which means that the process following the algorithm will couple with the auxiliary process. Then based on Lemma 3.5.1, we have

$$\mathbb{P}\left(V^n\right) \geq 1 - \frac{K_0 + 1}{T^{\frac{5}{2}}}.$$

Also, when event V^n occurs, it means that the event $\left\{W^t(z_e, z_r) = \tilde{W}^t(z_e, z_r),\ \forall L^{n-1} + \tau + 1 \leq t \leq L^n\right\}$ also occurs. Because $\tilde{W}^{L^{n-1}+1}(z_e, z_r)$ is randomly sampled from the steady-state $W^\infty(z_e, z_r)$, we have that $W^t(z_e, z_r)$ also follows the steady-state distribution for periods $L^{n-1} + \tau + 1$ to L^n. Furthermore, the per-period cost $C^t\left(\Delta, z_e\right)$ also depends on $W^t(\Delta, z_e)$ as shown in §3.3.3, we have $\mathbb{E}\left[\sum_{t=L^{n-1}+\tau+1}^{L^n} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right) \mid V^n\right] = 0$.

Therefore, the total nonstationary loss satisfies

$$\begin{aligned}
&\sum_{n=1}^{N} \mathbb{E}\left[\sum_{t=L^{n-1}+1}^{L^{n-1}+\tau} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right) + \sum_{t=L^{n-1}+\tau+1}^{L^n} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right] \\
=&\sum_{t=1}^{L^{N_0}} \mathbb{E}\left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right) + \sum_{n=N_0+1}^{N} \mathbb{E}\left[\sum_{t=L^{n-1}+1}^{L^{n-1}+\tau} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right)\right] \\
&+ \sum_{n=N_0+1}^{N} \mathbb{E}\left[\sum_{t=L^{n-1}+\tau+1}^{L^n} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right) \mid V^n\right] \mathbb{P}\left(V^n\right) \\
&+ \sum_{n=N_0+1}^{N} \mathbb{E}\left[\sum_{t=L^{n-1}+\tau+1}^{L^n} \left(C_\pi^t - C^\infty(\Delta^t, z_e^t)\right) \mid V^{n\complement}\right] \mathbb{P}\left(V^{n\complement}\right) \\
\leq&L^{N_0}\bar{C} + N\tau\bar{C} + 0 \cdot \mathbb{P}\left(V^n\right) + T\bar{C}\mathbb{P}\left(V^{n\complement}\right) \qquad (3.5)
\end{aligned}$$

$$\leq L^{N_0}\bar{C} + N\tau\bar{C} + 0\cdot 1 + T\frac{K_0+1}{T^{\frac{5}{2}}}\bar{C} = O((\log T)^3), \tag{3.6}$$

where (3.5) is because, for the first L^{N_0} periods, the cost difference is at most $\tau\bar{C}$ with $\bar{C}$ being the upper bound of the cost per period. For each epoch $n \geq N_0$, when $t \leq \tau$, the cost difference per period is still bounded by $\bar{C}$ and for $t > \tau$, the cost difference is calculated using the law of total expectation. Conditional on event V^n, we have that the difference is 0 and conditional on event $V^{n\complement}$, the difference is also bounded by $\bar{C}$ for each period. The last inequality (3.6) is due to $L^{N_0} = \sum_{i=1}^{N_0} B^i \leq N_0\lceil\frac{2N_0}{\log T}\rceil = O(\log T)^3$, $N \leq \log_2(T\log T + 2) - 1$ as well as $\mathbb{P}(V^n) \leq 1$.

3.5.2 Proof of Proposition 3.5.3 - Bound the Empirical Suboptimal Loss

To show the bound for the cost difference between the estimated quantile and the theoretical quantile, i.e., $\mathbb{E}\left[\sum_{t=1}^{NL}\left(C^\infty\left(\Delta^t, z_e^t\right) - C^\infty\left(\Delta^t, z_e^*(\Delta^t)\right)\right)\right]$, first we show the following lemma on the relationship between the accuracy of the empirical quantile and the expected cost difference.

Lemma 3.5.4 *Let $F_\Delta(\cdot)$ be the CDF of $D_{l_e} - O^\infty(\Delta)$ and $z_e^* = F_\Delta^{-1}\left(\frac{b}{b+h}\right)$. For any z_e, if $|F(z_e) - F(z_e^*)| \leq \alpha$ for some $\alpha > 0$, then we have*

$$\mathbb{E}\left[C^\infty(\Delta, z_e) - C^\infty(\Delta, z_e^*)\right] \leq \alpha(b+h)\left|z_e - z_e^*\right|.$$

Proof of Lemma 3.5.4. Note that the limiting average cost of the system under the dual-index policy (Δ, z_e) is

$$\begin{aligned}\mathbb{E}\left[C^\infty(\Delta, z_e)\right] =&(c_e - c_r)\mathbb{E}\left[q_e^\infty\right] + c_r\mathbb{E}\left[D\right] + h\mathbb{E}\left[(I^\infty)^+\right] + b\mathbb{E}\left[(I^\infty)^-\right] \\ =&(c_e - c_r)\mathbb{E}\left[q_e^\infty\right] + c_r\mathbb{E}\left[D\right] + h\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^+\right] + b\mathbb{E}\left[(z_e + O^\infty - D_{l_e})^-\right],\end{aligned}$$

Based on Proposition 3.3.1, we know the expedited ordering quantity q_e^t and the overshoot O^t are functions of Δ but not z_e. Hence, we have

$$\begin{aligned}&\mathbb{E}\left[C^\infty(\Delta, z_1) - C^\infty(\Delta, z_2)\right] \\ =&h\mathbb{E}\left[(z_1 + O^\infty - D_{l_e})^+ - (z_2 + O^\infty - D_{l_e})^+\right] + b\mathbb{E}\left[(z_1 + O^\infty - D_{l_e})^- - (z_2 + O^\infty - D_{l_e})^-\right],\end{aligned} \tag{3.7}$$

where (3.7) is the difference of the newsvendor cost with demand variable $D_{l_e} - O^\infty$. Therefore, by Lemma 2.1 in Levi et al. (2007a), we have that (3.7) can be bounded as follows:

$$\mathbb{E}\left[C^{\infty}(\Delta, z_e) - C^{\infty}(\Delta, z_e^*)\right] \leq \alpha(b+h)\left|z_e - z_e^*\right|.$$

Lemma 3.5.4 is thus proved. **Q.E.D.**

Now we focus on the difference between the empirical quantile and the theoretical quantile solution. The main idea is based on the concentration result for uniformly ergodic Markov chains. We define $d_{\mathrm{TV}}(P,Q) := \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$ for two distributions P and Q defined on the same $(\mathcal{S},\mathcal{F})$ where $\mathcal{S}$ is a state space and $\mathcal{F}$ is a σ-algebra defined on $\mathcal{S}$. Recall that a Markov chain with stationary distribution ψ, state space $\mathcal{S}$, and transition kernel $P(x,dy)$ is uniformly ergodic if there exists some $\rho < 1$ and $M < \infty$ such that for all $n \in \mathbb{Z}_+$,

$$\sup_{x\in\mathcal{S}} d_{\mathrm{TV}}\left(P^n(x,\cdot),\psi\right) \leq M\rho^n.$$

By Theorem 3.3.1, we have the Markov chain $\{W^t(z_e,z_r) : t \geq 1\}$ is uniformly ergodic, which bears the following version of McDiarmid's inequality.

Lemma 3.5.5 (Lemma 1 in Ortner (2020)) *Consider a uniformly ergodic Markov chain $X_1,\ldots,X_n$ on state space $\mathcal{S}$ with stationary distribution ψ and transition kernel $P(x,dy)$. The mixing time t_{mix} is defined by $t_{\mathrm{mix}} := \min\left\{t : \sup_{x\in\mathcal{S}} d_{\mathrm{TV}}\left(P^t(x,\cdot),\psi\right) \leq \frac{1}{4}\right\}$ where $d_{\mathrm{TV}}(P,Q) := \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$ for two distributions P and Q defined on the same $(\mathcal{S},\mathcal{F})$. Let $f : \mathcal{S}^n \to \mathbb{R}$ with*

$$f\left(s_1,\ldots,s_n\right) - f\left(s_1',\ldots,s_n'\right) \leq \sum_{i=1}^{n} \iota_i \mathbb{1}\left[s_i \neq s_i'\right], \forall\left(s_1,\ldots,s_n\right),\left(s_1',\ldots,s_n'\right) \in \mathcal{S}^n. \quad (3.8)$$

Then for any $\theta \geq 0$,

$$\mathbb{P}\left\{\left|f\left(X_1,\ldots,X_n\right) - \mathbb{E}\left[f\left(X_1,\ldots,X_n\right)\right]\right| \geq \theta\right\} \leq 2\exp\left(-\frac{2\theta^2}{9\|\iota\|_2^2 t_{\mathrm{mix}}}\right).$$

As Theorem 3.3.1 established the uniform ergodicity of Markov chain $W^t(z_e,z_r)$, we have the following corollary.

Corollary 3.5.1 *Consider the Markov chain $\left(W^t\left(z_e,z_r\right), t \geq 1\right)$, we have $t_{\mathrm{mix}} \leq T_0$ where*

$$T_0 := \min_{\eta\in\left(0,\frac{1}{4}\right)} \max\left\{4l_r \cdot \exp\left(\log\eta\log\left(1-\lambda\right)^{2l_r}\right), -\frac{\bar{D}^2}{\underline{\mu}^2}\log\left(\frac{1}{4}-\eta\right) + \frac{4\bar{D}\left(\bar{w}^1\cdot\mathbf{1}^l\right)}{\underline{\mu}^2} + 2l_r\right\}. \quad (3.9)$$

Proof of Corollary 3.5.1. Denote the minimizer of (3.9) by η^*. For the Markov chain $\left(W^t\left(z_e, z_r\right), t \geq 1\right)$, we have $t_{\text{mix}} = \min\left\{n \mid \delta_t\left(z_e, z_r, w^1\right) \leq \frac{1}{4}\right\}$ where $\delta_t\left(z_e, z_r, w^1\right)$ is defined in (3.2). Consider when $t \geq T_0$, by Assumption 3.3.1b, we have

$$t \geq 4l_r \cdot \exp\left(\log \eta^* \cdot \log\left(1-\gamma(z_e,z_r)^{2l_r}\right)\right), \quad t \geq -\frac{\bar{D}^2}{\mu^2}\log\left(\frac{1}{4}-\eta^*\right)+\frac{4\bar{D}\left(\bar{w}^1\cdot \mathbf{1}^l\right)}{\mu^2}+2l_r,$$

which means

$$\left(1-\gamma\left(z_e,z_r\right)^{2l_r}\right)^{t/4l_r} \leq \eta^*, \quad \exp\left(\frac{4\left(\bar{w}^1\cdot\mathbf{1}^l - z_r\right)}{\bar{D}} - \frac{2\mu^2\left(t/2-l_r\right)}{\bar{D}^2}\right) \leq \frac{1}{4}-\eta^*.$$

Hence, by the definition of t_{mix}, the corollary is proved. **Q.E.D.**

Now we utilize Lemma 3.5.5 to achieve a concentration result for the empirical quantile of the variable $D_{l_e} - O^\infty$ obtained from the simulation. Consider the Markov chain $\left\{\hat{W}_j^t\right\}_{t=1}^{L^n}$, $\forall n \in [N],\ j \in [J]$ which is the simulated process if the system follows the dual-index policy (Δ_j, z_{ej}^n) using the demand data from period 1 to L^n. Recall from Proposition 3.3.1, the overshoot O^t only depends on Δ. Hence, for the Markov chain $\left\{\hat{W}_j^t\right\}_{t=1}^{L^n}$ following the dual-index policy $\left(\Delta_j, z_{ej}^n\right)$, we have that the steady-state overshoot $O^\infty\left(\Delta_j\right)$ only depends on Δ_j. For simplicity of notation, we denote $O^\infty\left(\Delta_j\right)$ by O_j^∞ in the following analysis, in accordance with the notation scheme in Algorithm 5. For any $j \in [J]$, let $F_{\Delta_j}(\cdot)$ denote the CDF of $D_{l_e} - O_j^\infty$. Recall that $F_{\Delta_j}\left(z_e^*\left(\Delta_j\right)\right) = \frac{b}{b+h}$.

Lemma 3.5.6 *For any Δ index $j \in [J]$ in any epoch $n \in [N-1]$, we have for any $\theta \geq 0$*

$$\mathbb{P}\left(\left|F_{\Delta_j}\left(z_{e_j}^{n+1}\right) - \frac{b}{b+h}\right| \geq \theta\right) \leq 4\exp\left\{-\frac{2L^n\theta^2}{9T_0}\right\}.$$

Proof of Lemma 3.5.6. Consider the function $f^n\left(\hat{W}_j^1, \ldots, \hat{W}_j^{L^n}\right) := \frac{1}{L^n}\sum_{t=1}^{L^n}\mathbb{1}(D_{l_e}^t - \hat{O}_j^t \leq z_e^*\left(\Delta_j\right) + \zeta)$ where $F_{\Delta_j}\left(z_e^*\left(\Delta_j\right)+\zeta\right) = \frac{b}{b+h}+\theta$. Note that condition (3.8) holds with $\iota_i = \frac{1}{L^n}$, $\forall i \in [L^n]$. For any epoch $n \in [N]$, recall that $\hat{F}_{\Delta_j}^n(\cdot)$ denote the empirical CDF of $D_{l_e} - O_j^\infty$ based on dataset $\mathcal{X}_j^n = \left\{d_{l_e}^t - \hat{o}_j^t\right\}_{t=1}^{L^n - l_e}$. Thus, we have

$$\begin{aligned}\mathbb{P}\left(F_{\Delta_j}\left(z_{e_j}^{n+1}\right) > \frac{b}{b+h}+\theta\right) &= \mathbb{P}\left(z_{e_j}^{n+1} > z_e^*\left(\Delta_j\right)+\zeta\right)\\ &= \mathbb{P}\left(\hat{F}_{\Delta_j}^n\left(z_{e_j}^{n+1}\right) > \hat{F}_{\Delta_j}^n\left(z_e^*\left(\Delta_j\right)+\zeta\right)\right)\\ &= \mathbb{P}\left(\frac{b}{b+h} > \frac{1}{L^n}\sum_{t=1}^{L^n}\mathbb{1}(D_{l_e}^t - \hat{O}_j^t \leq z_e^*\left(\Delta_j\right)+\zeta)\right)\end{aligned}$$

$$= \mathbb{P}\left(\mathbb{E}\left[f^n\left(\hat{W}_j^1,\ldots,\hat{W}_j^{L^n}\right)\right] - \theta > f^n\left(\hat{W}_j^1,\ldots,\hat{W}_j^{L^n}\right)\right) \tag{3.10}$$

$$\leq 2\exp\left\{-\frac{2L^n\theta^2}{9T_0}\right\} \tag{3.11}$$

where (3.10) is from ergodicity and (3.11) is by Lemma 3.5.5. Also $\mathbb{P}\left(F_{\Delta_j}\left(z_{e_j}^{n+1}\right) < \frac{b}{b+h} + \theta\right) \leq 2\exp\left\{-\frac{2L^n\theta^2}{9T_0}\right\}$ holds following similar analysis which completes the proof. **Q.E.D.**

For any epoch $n \in [N]$, denote event $A_j^n = \left\{\left|F_{\Delta_j}\left(z_{e_j}^n\right) - \frac{b}{b+h}\right| \leq \alpha^n\right\}$ where $\alpha^0 := 1$ and $\alpha^n := \frac{3}{2}\sqrt{\frac{3T_0\log T}{L^{n-1}}}, \forall n \geq 2$. Note that event A_j^n denotes the event that the empirical estimator $z_{e_j}^n$ is accurate enough for $z^*\left(\Delta_j\right)$. Define event $\bar{A} := \cap_{n\in[N]} A_{j^n}^n$. Then for any epoch $n \in [N]$, by Lemma 3.5.4 we have

$$\mathbb{E}\left[\left(C^\infty\left(\Delta_{j^n}, z_{ej^n}^n\right) - C^\infty\left(\Delta_{j^n}, z_e^*(\Delta_{j^n})\right)\right) \mid \bar{A}\right] \leq \alpha^n(b+h)\left|z_{ej^n}^n - z_e^*(\Delta_{j^n})\right|.$$

By Lemma 3.5.6, we have $\mathbb{P}\left(A_{j^n}^{n\complement}\right) \leq \frac{4}{T^{\frac{3}{2}}}$, $\forall n \in [N]$, which further implies $\mathbb{P}\left(\bar{A}^\complement\right) \leq \sum_{n\in[N]}\mathbb{P}\left(A_{j^n}^{n\complement}\right) \leq \frac{4\lfloor\sqrt{T}\rfloor}{T^{\frac{3}{2}}} \leq \frac{4}{T}$ as $N \leq 2\log_2 T \leq \sqrt{T}$. Given the fact that the estimator is accurate enough with high probability, we apply the law of total expectation and obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T\left(C^\infty\left(\Delta^t, z_e^t\right) - C^\infty\left(\Delta_{j^n}, z_e^*(\Delta_{j^n})\right)\right)\right] \\
=&\mathbb{E}\left[\sum_{t=1}^T\left(C^\infty\left(\Delta_{j^n}, z_{ej^n}^n\right) - C^\infty\left(\Delta_{j^n}, z_e^*(\Delta_{j^n})\right)\right) \mid \bar{A}\right]\mathbb{P}\left(\bar{A}\right) \\
&+\mathbb{E}\left[\sum_{t=1}^T\left(C^\infty\left(\Delta_{j^n}, z_{ej^n}^n\right) - C^\infty\left(\Delta_{j^n}, z_e^*(\Delta_{j^n})\right)\right) \mid \bar{A}^\complement\right]\mathbb{P}\left(\bar{A}^\complement\right) \\
\leq& (b+h)\,\bar{Z}\sum_{n=1}^N B^n\alpha^n\mathbb{P}\left(\bar{A}\right) + \sum_{n=1}^N B^n\bar{C}\mathbb{P}\left(\bar{A}^\complement\right) \\
\leq& (b+h)\,\bar{Z}\sum_{n=1}^N B^n\alpha^n + T\cdot\frac{2\bar{C}}{T} = O(\sqrt{T\log T}),
\end{aligned}$$

where the last inequality is from Lemma 3.5.7 below with proof in Appendix B.3.1.

Lemma 3.5.7 $\sum_{n=1}^N B^n\alpha^n = O(\sqrt{T\log T})$.

3.5.3 Proof of Proposition 3.5.4 - Bound the Bandit Pruning Loss

To bound $\mathbb{E}\left[\sum_{t=1}^{T}\left(C^{\infty}\left(\Delta^{t}, z_{e}^{*}\left(\Delta^{t}\right)\right)-C^{\infty}\left(\Delta_{j^{*}}, z_{e}^{*}\left(\Delta_{j^{*}}\right)\right)\right)\right]$, which is the loss due to the suboptimality of the choice of the arm Δ^{n} selected for epoch n, we provide upper bounds for the following three components:

1. gap between the steady-state average cost with the optimal expedited order-up-to level and with estimated expedited order-up-to level for any arm j, i.e., $\left|\mathbb{E}\left[C^{\infty}\left(\Delta_{j}, z_{e}^{*}(\Delta_{j})\right)\right]-\mathbb{E}\left[C^{\infty}\left(\Delta_{j}, z_{j}^{n}\right)\right]\right|$, denoted as **Suboptimality Loss**;

2. gap between the steady-state average cost and the estimated average cost both with the estimated expedited order-up-to level, i.e., $\left|\left(\mathbb{E}\left[C^{\infty}\left(\Delta_{j}, z_{j}^{n}\right)\right]-\hat{G}_{j}^{n}\right)\right|$, denoted as **Mean Estimation Loss**;

3. difference between the estimated average cost of the optimal arm j^{*} with its corresponding estimated expedited order-up-to level and the estimated average cost of the selected arm j^{n} with its corresponding estimated expedited order-up-to level, i.e., $\hat{G}_{j^{n}}^{n}-\hat{G}_{j^{*}}^{n}$, denoted as **Pruning Set Loss**.

The upper bounds for the three components above are established in the three lemmas below.

Lemma 3.5.8 (Suboptimality loss) *We have*

$$\mathbb{P}\left(\left|\mathbb{E}\left[C^{\infty}\left(\Delta_{j}, z_{e}^{*}(\Delta_{j})\right)\right]-\mathbb{E}\left[C^{\infty}\left(\Delta_{j}, z_{j}^{n}\right)\right]\right| \leq (b+h)\, \bar{Z} \alpha^{n},\ \textit{for any } n \in [N],\ j \in \mathcal{A}^{n}\right) \geq 1-\frac{2}{\sqrt{T}}.$$

Lemma 3.5.9 (Mean Estimation Loss) *We have*

$$\mathbb{P}\left(\left|\mathbb{E}\left[C^{\infty}\left(\Delta_{j}, z_{j}^{n}\right)\right]-\hat{G}_{j}^{n}\right| \leq \beta^{n},\ \textit{for any } n \in [N], j \in \mathcal{A}^{n}\right) \geq 1-\frac{2}{\sqrt{T}}.$$

Lemma 3.5.10 (Pruning Set Loss) *With probability of at least $1-\frac{4}{\sqrt{T}}$, we have that for any $n \in [N] \setminus \{1\}$, the optimal arm $j^{*} \in \mathcal{A}^{n}$ and*

$$\hat{G}_{j^{n}}^{n-1}-\hat{G}_{j^{*}}^{n-1} \leq 2 \beta^{n-1}+2(b+h)\, \bar{Z} \alpha^{n-1}.$$

The proof of Lemma 3.5.8 is based on the accuracy of the empirical quantile. To establish Lemma 3.5.9, we apply the concentration bound for an empirical mean of the data sampled from an ergodic Markov chain. The proof for Lemma 3.5.10 relies on proving that with high probability the optimal arm will remain in the active set.

Proof of Lemma 3.5.8. Define the event that the estimated expedited order-up-to level based on empirical quantile is close to the true optimal expedited order-up-to level for any arm $j \in \mathcal{A}^n$ in epoch n as A, i.e.,

$$A = \left\{ \left| F_{\Delta_j}\left(z^n_{e_j}\right) - \frac{b}{b+h} \right| \leq \alpha^n, \ \forall n \in [N], \ j \in \mathcal{A}^n \right\} = \bigcap_{j \in \mathcal{A}^n, n \in [N]} A^n_j.$$

By Lemma 3.5.6, we have

$$\mathbb{P}\left(A^{\complement}\right) \leq \sum_{j \in [J], n \in [N]} \mathbb{P}\left(A^{n\complement}_j\right) \leq \lfloor \sqrt{T} \rfloor \lfloor \sqrt{T} \rfloor \frac{2}{T^{\frac{3}{2}}} \leq \frac{2}{\sqrt{T}}. \tag{3.12}$$

Conditional on event A and Lemma 3.5.4, we have that for any arm j and any epoch n,

$$\left| \mathbb{E}\left[C^{\infty}\left(\Delta_j, z^*_e(\Delta_j)\right)\right] - \mathbb{E}\left[C^{\infty}\left(\Delta_j, z^n_j\right)\right] \right| \leq (b+h)\, \bar{Z} \alpha^n. \tag{3.13}$$

Q.E.D.

Before the proof of Lemma 3.5.9, we first present a concentration result of the estimated mean cost, the proof of which is similar to that of Lemma 3.5.6 and thus is relegated to Appendix B.3.2.

Lemma 3.5.11 *For any Δ index $j \in [J]$ in any epoch $n \in [N]$, we have for any $\theta \geq 0$,*

$$\mathbb{P}\left(\left| \hat{C}^n_j - C^{\infty}\left(\Delta_j, z^n_j\right) \right| \geq \theta \right) \leq 2 \exp\left\{ -\frac{2L^n \theta^2}{9T_0 \left(\bar{C}\right)^2} \right\}.$$

Proof of Lemma 3.5.9. Let $\hat{C}^t(\Delta_j, z^n_{ej})$ be the simulated cost for arm j in round n in the algorithm (corresponding to the process $\hat{W}^t_j(\Delta_j, z^n_{ej})$) and $\tilde{C}^t(\Delta^t, z^t_e)$ to be the cost of the auxiliary process sampling from the steady state for arm j in round n. Recall that the cost $C^t\left(\Delta, z_e\right)$ in period t of a system following the dual-index policy only depends on $W^t(\Delta, z_e)$ as shown in §3.3.3.

Denote event that the estimated cost for arm j in epoch n is accurate enough as M^n_j, i.e.,

$$M^n_j := \left\{ \left| \hat{G}^n_j - \mathbb{E}\left[C^{\infty}(\Delta_j, z^n_{ej})\right] \right| \leq \beta^n \right\},$$

where $\beta^n = \frac{3\bar{C}}{2}\sqrt{\frac{T_0 \log T}{L^n}}$. Then by Lemma 3.5.11, we have

$$\mathbb{P}\left(M^{n\complement}_j\right) \leq \frac{2}{T^{\frac{3}{2}}}.$$

Therefore, let event $M = \bigcap_{j\in[J],n\in[N]} M_j^n$ and we have

$$\mathbb{P}\left(M^{\complement} \mid V\right) \leq \sum_{j\in[J],n\in[N]} \frac{2}{T^{\frac{3}{2}}} \leq \frac{2}{\sqrt{T}}. \tag{3.14}$$

Conditional on event M, for any arm $j \in [J]$ and any epoch $n \in [N]$, we have

$$\left|\mathbb{E}\left[C^{\infty}\left(\Delta_j, z_j^n\right)\right] - \hat{G}_j^n\right| \leq \beta^n. \tag{3.15}$$

Q.E.D.

Proof of Lemma 3.5.10. We would like to show that with high probability, the optimal arm j^* remains in the active set $\mathcal{A}^n$ and the difference between the estimated costs of the selected arm j^n and of j^* is small.

First, we show that with high probability, the optimal arm j^* will remain in the active set $\mathcal{A}^n$, which is equivalent to showing that $\hat{G}_{j^*}^k - \min_j \hat{G}_j^k \leq \varepsilon^k$, $\forall 1 \leq k \leq n$. Note that for any arm j, we have:

$$\begin{aligned}
&\hat{G}_{j^*}^n - \hat{G}_j^n && (3.16)\\
\leq&\hat{G}_{j^*}^n - \mathbb{E}\left[C^{\infty}(\Delta_{j^*}, z_e^*(\Delta_{j^*}))\right] + \mathbb{E}\left[C^{\infty}(\Delta_j, z_e^*(\Delta_j))\right] - \hat{G}_j^n && (3.17)\\
=&\frac{1}{L^n}\sum_{t=1}^{L^n} C^t(\Delta_{j^*}, z_{ej^*}^n) - \mathbb{E}\left[C^{\infty}(\Delta_{j^*}, z_{ej^*}^n)\right] && (3.18)\\
&+ \mathbb{E}\left[C^{\infty}(\Delta_{j^*}, z_{ej^*}^n)\right] - \mathbb{E}\left[C^{\infty}(\Delta_{j^*}, z_e^*(\Delta_{j^*}))\right] && (3.19)\\
&+ \mathbb{E}\left[C^{\infty}(\Delta_j, z_e^*(\Delta_j))\right] - \mathbb{E}\left[C^{\infty}(\Delta_j, z_{ej}^n)\right] && (3.20)\\
&+ \mathbb{E}\left[C^{\infty}(\Delta_j, z_{ej}^n)\right] - \frac{1}{L^n}\sum_{k=1}^{L^n} C^t(\Delta_j, z_{ej}^n). && (3.21)
\end{aligned}$$

Conditional on M, which means that the estimation for the average cost is well enough for all arms and epochs, by (3.15), we have

$$(3.18) \leq \beta^n, \qquad (3.21) \leq \beta^n.$$

Conditional on A, which indicates the estimation for the expedited order-up-to level is accurate enough for all arms and epochs, by (3.13), we have

$$(3.19) \leq (b+h)\,\bar{Z}\alpha^n, \qquad (3.20) \leq (b+h)\,\bar{Z}\alpha^n.$$

Conditional on $A \cap M$, we have $\forall j \in [J]$,

$$\begin{aligned}\hat{G}^n_{j^*} - \hat{G}^n_j \leq & \beta^n + (b+h)\,\bar{Z}\alpha^n + (b+h)\,\bar{Z}\alpha^n + \beta^n \\ \leq & 2\beta^n + 2\,(b+h)\,\bar{Z}\alpha^n,\end{aligned}$$

which implies that

$$\hat{G}^n_{j^*} - \min_j \hat{G}^n_j \leq 2\beta^n + 2\,(b+h)\,\bar{Z}\alpha^n = \varepsilon^n.$$

Thus, conditional on event $A \cap M$, the optimal bandit will remain in the active set $\mathcal{A}^n$, $\forall n \in [N]$.

Now we can bound the difference between the estimated cost for j^n and that for j^*. Since $j^n \in \mathcal{A}^n$, j^n is not removed in the $(n-1)$th iteration. Hence, conditional on event $A \cap M$, we have

$$\begin{aligned}& \hat{G}^{n-1}_{j^n} - \hat{G}^{n-1}_{j^*} \\ \leq & \hat{G}^{n-1}_{j^n} - \min_j \hat{G}^{n-1}_j \\ \leq & 2\beta^{n-1} + 2\,(b+h)\,\bar{Z}\alpha^{n-1}.\end{aligned} \tag{3.22}$$

Note that by (3.12) and (3.14), we have

$$\begin{aligned}\mathbb{P}\left((A \cap M)^{\complement}\right) &= \mathbb{P}\left(M^{\complement} \cup A^{\complement}\right) \\ &\leq \mathbb{P}\left(M^{\complement}\right) + \mathbb{P}\left(A^{\complement}\right) \\ &\leq \frac{4}{\sqrt{T}}.\end{aligned}$$

Lemma 3.5.10 is thus proved. **Q.E.D.**

Equipped with Lemmas 3.5.8, 3.5.9, and 3.5.10, now we are ready to prove Proposition 3.5.4.

$$\mathbb{E}\left[C^\infty\left(\Delta^t, z^*_e(\Delta^t)\right) - C^\infty\left(\Delta_{j^*}, z^*_e(\Delta_{j^*})\right)\right] \tag{3.23}$$

$$= \mathbb{E}\left[C^\infty\left(\Delta_{j^n}, z^*_e(\Delta_{j^n})\right) - C^\infty\left(\Delta_{j^*}, z^*_e(\Delta_{j^*})\right)\right] \tag{3.24}$$

$$= \left(\mathbb{E}\left[C^\infty\left(\Delta_{j^n}, z^*_e(\Delta_{j^n})\right)\right] - \mathbb{E}\left[C^\infty\left(\Delta_{j^n}, z^{n-1}_{j^n}\right)\right]\right) \tag{3.25}$$

$$+ \left(\mathbb{E}\left[C^\infty\left(\Delta_{j^n}, z^{n-1}_{j^n}\right)\right] - \hat{G}^{n-1}_{j^n}\right) \tag{3.26}$$

$$+ \left(\hat{G}^{n-1}_{j^n} - \hat{G}^{n-1}_{j^*}\right) \tag{3.27}$$

$$+\left(\hat{G}_{j^*}^{n-1}-\mathbb{E}\left[C^\infty\left(\Delta_{j^*},z_{j^*}^{n-1}\right)\right]\right) \tag{3.28}$$

$$+\left(\mathbb{E}\left[C^\infty\left(\Delta_{j^*},z_{j^*}^{n-1}\right)\right]-\mathbb{E}\left[C^\infty\left(\Delta_{j^*},z_e^*(\Delta_{j^*})\right)\right]\right). \tag{3.29}$$

By Lemma 3.5.8, we have

$$\begin{aligned}(3.25)&\leq(b+h)\,\bar{Z}\alpha^{n-1}+\frac{2}{\sqrt{T}}\bar{C},\\(3.29)&\leq(b+h)\,\bar{Z}\alpha^{n-1}+\frac{2}{\sqrt{T}}\bar{C}.\end{aligned}$$

By Lemma 3.5.9, we have

$$\begin{aligned}(3.26)&\leq\beta^{n-1}+\frac{2}{\sqrt{T}}\bar{C},\\(3.28)&\leq\beta^{n-1}+\frac{2}{\sqrt{T}}\bar{C}.\end{aligned}$$

By Lemma 3.5.10, we have

$$(3.27)\leq 2\beta^{n-1}+2\,(b+h)\,\bar{Z}\alpha^{n-1}+\frac{4}{\sqrt{T}}\bar{C}.$$

Combining all the results above, we have

$$\begin{aligned}&\mathbb{E}\left[C^\infty\left(\Delta^t,z_e^*(\Delta^t)\right)-C^\infty\left(\Delta_{j^*},z_e^*(\Delta_{j^*})\right)\right]\\\leq&4\beta^{n-1}+4\,(b+h)\,\bar{Z}\alpha^{n-1}+\bar{C}\frac{12}{\sqrt{T}}.\end{aligned}$$

Summing over $\sum_{n=1}^N B^n$ periods, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T\left(C^\infty(\Delta^t,z_e^t)-C^\infty(\Delta^t,z_e^*(\Delta_{j^*}))\right)\right]\\\leq&12\bar{C}\sqrt{T}+\sum_{n=2}^N B^n\left(4\beta^{n-1}+4\,(b+h)\,\bar{Z}\alpha^{n-1}\right)\\=&O\left(\sqrt{T\log T}\right),\end{aligned}$$

where the last inequality is due to Lemma 3.5.7.

3.5.4 Proof of Proposition 3.5.5 - Bound the Discretization Loss

Finally, to bound the loss due to the discretization of the specified Δ choices, i.e., $\mathbb{E}\left[\sum_{t=1}^{T}\left(C^{\infty}\left(\Delta_{j^*}, z_e^*(\Delta_{j^*})\right) - C^{\infty}\left(\Delta^*, z_e^*\right)\right)\right]$, we prove that $\mathbb{E}\left[C^{\infty}(\Delta, z_e^*(\Delta))\right]$ is Lipschitz in Δ in Lemma 3.5.12 below. The proof is relegated to Appendix B.3.3.

Lemma 3.5.12 *The minimum steady state cost* $\mathbb{E}\left[C^{\infty}(\Delta, z_e^*(\Delta))\right]$ *is Lipschitz in* Δ*. Specifically, for any* Δ_1, Δ_2 *we have* $\left|\mathbb{E}\left[C^{\infty}(\Delta_1, z_e^*(\Delta_1))\right] - \mathbb{E}\left[C^{\infty}(\Delta_2, z_e^*(\Delta_2))\right]\right| \leq (c_e + c_r + h + b)\left|\Delta_1 - \Delta_2\right|$.

Denote $k = \arg\min_{j\in[J]} |\Delta^* - \Delta_j|$. As we initialize the active set with size $\lfloor\sqrt{T}\rfloor$, we have $\Delta_k - \Delta^* \leq \left(\bar{Z} - \underline{\Delta}\right)/\lfloor\sqrt{T}\rfloor$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(C^{\infty}\left(\Delta_{j^*}, z_e^*(\Delta_{j^*})\right) - C^{\infty}\left(\Delta^*, z_e^*\right)\right)\right] \\
\leq & T(c_e + c_r + h + b)\left(\bar{Z} - \underline{\Delta}\right)/\lfloor\sqrt{T}\rfloor \\
= & O(\sqrt{T}).
\end{aligned}$$

3.6 Numerical Results

Consider a dual-sourcing system starting with zero inventory, with the total number of periods to consider $T = 1600$. The unit prices for the order from the expedited and regular source are $c_e = 20$ and $c_r = 15$, respectively. The lead time of the expedited inventory is $l_e = 2$. We run the algorithm for 24 instances as below.

Table 3.1: Experimental Parameters

Parameters	Group 1				Group 2				Group 3				Group 4				Group 5				Group 6			
l_r	4				6				8				4				6				8			
d	$\mathbf{N}[50,10]$				$\mathbf{N}[50,10]$				$\mathbf{N}[50,10]$				$\mathbf{N}[50,20]$				$\mathbf{N}[50,20]$				$\mathbf{N}[50,20]$			
b	5	5	10	10	5	5	10	10	5	5	10	10	5	5	10	10	5	5	10	10	5	5	10	10
h	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4

In Table 3.1, $\mathbf{N}[\mu, \sigma]$ denotes the truncated normal distribution here truncated at $[0, 100]$. We test the performances under short, medium, and long lead time difference scenarios, respectively, with various values of holding and penalty cost. The performance measure we use for each instance is the average relative regret defined as the cumulative difference between the total cost of (Δ, z_e) algorithm and the cost if we operate the system under the dual-index policy using the clairvoyant optimal order-up-to levels z_e^*, z_r^*, divided by time, i.e.,

$$\textbf{Relative Regret} := \frac{\sum_{t=1}^{T} C_{\pi}^{t} - \sum_{t=1}^{T} C^{t}(z_e^*, z_r^*)}{\sum_{t=1}^{T} C^{t}(z_e^*, z_r^*)}.$$

For each instance, we run 1000 times and take the average of the relative regret. Figures 3.2–3.7 are the figures of the relative regret of Group 1 to 6 averaged over the instances of different values of b and h, respectively. The algorithm converges robustly under various settings. While the average relative regret level slightly increases as the lead time difference increases, the performances of the algorithm for demand with different coefficients of variation are comparable.

The dual sourcing literature commonly assumes independent and identically distributed (i.i.d.) demand (Allon and Van Mieghem 2010, Sheopuri et al. 2010), and the dual-index policy was developed specifically for stationary demand (Veeraraghavan and Scheller-Wolf 2008). However, in real-world scenarios, nonstationary demand is often encountered. To this end, we propose the Restart Learning Algorithm, a modified approach that incorporates the "restart" idea from Besbes et al. (2015) to handle non-i.i.d. demand environments. This modified algorithm involves restarting the learning process every τ_T periods, where τ_T is determined by the upper bound of the variation budget of the stationary per-period cost, denoted as V_T. We refer interested readers to Appendix B.4 for more details. We demonstrate that this modified approach based on restarting the algorithm from time to time performs well in non-i.i.d. demand environments.

3.7 Conclusions

We studied the dual-sourcing system with backlogged demand, where the decisions are the inventory replenishment quantities from expedited and regular channels, respectively. Specifically, we considered the celebrated dual-index policy and provided a sufficient condition for the inventory system to be ergodic. Based on this key ergodic result, we proposed an online learning algorithm and analyzed its theoretical regret bounds.

We close this chapter by pointing out three promising future research avenues. First, according to Svoboda et al. (2021), of all publications on this topic, 70% considered dual sourcing while the remaining 30% looked at multiple sourcing with more than two suppliers. One possible direction would be investigating the optimal or near-optimal replenishment policy for systems with more than two sources (Feng et al. 2005). Second, this chapter only considered the case without capacity constraints for expedited and regular orders. It would be worthwhile investigating the settings with capacity constraints (Federgruen et al. 2020, 2022). Third, dual-sourcing systems with lost sales and general lead times remain a long-standing open problem. Any progress on the lost sales counterpart model would be a substantial contribution to the literature.

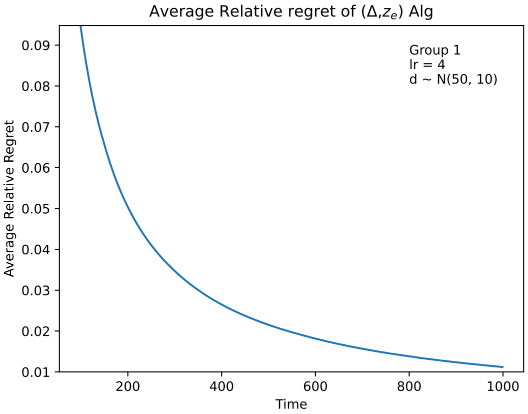


Figure 3.2: Computational Performance for Group 1

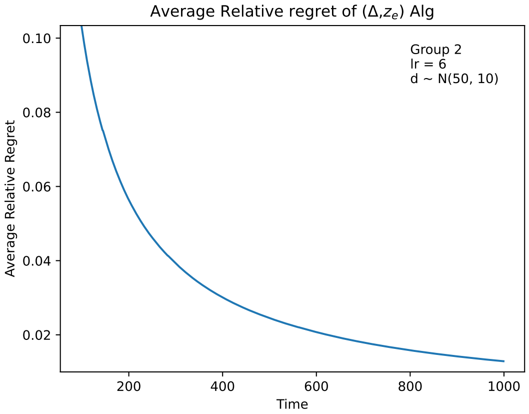


Figure 3.3: Computational Performance for Group 2

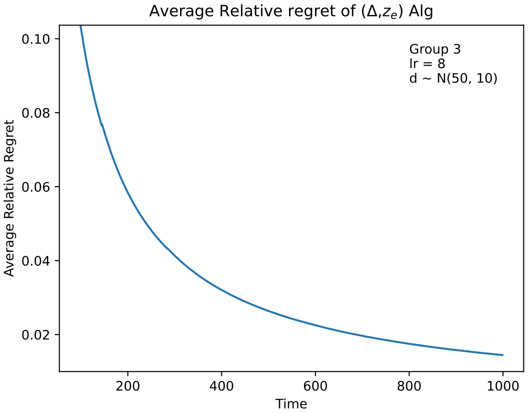


Figure 3.4: Computational Performance for Group 3

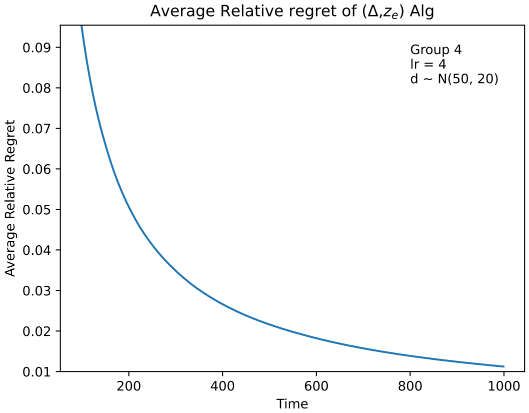


Figure 3.5: Computational Performance for Group 4

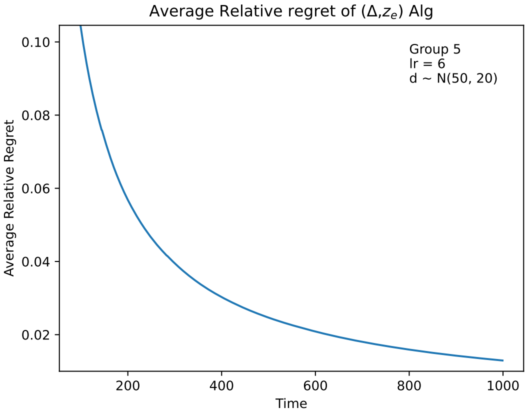


Figure 3.6: Computational Performance for Group 5

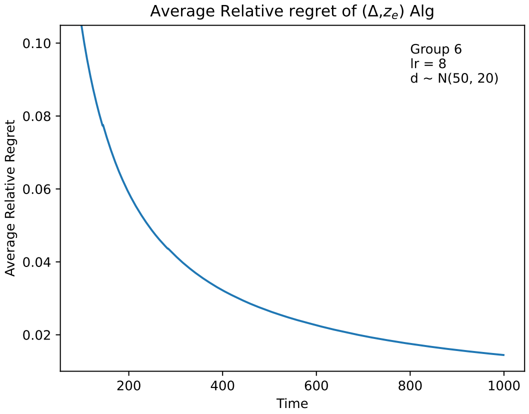


Figure 3.7: Computational Performance for Group 6

CHAPTER 4

Online Learning in Two-Sided Markets

While the development and application of a two-layer online learning algorithm have seen considerable success across various studies (Chen et al. 2019a, 2021a, Chen and Shi 2020, Chen et al. 2022a, Yuan et al. 2021), the frameworks are often based on assumptions about the parametric forms of objective functions. However, the increasing uncertainty in markets prompts a relaxation of these parametric assumptions where only bandit feedback is available for optimization.

One of the applications in recent years with highly uncertain markets is online platforms, which serve as a connection between demand and supply. The decision-maker needs to make decisions for both sides without the form of the demand and supply functions. Given the limited information on these functions, this chapter introduces a nonparametric online learning algorithm designed to navigate the joint optimization of two decision variables.

4.1 Introduction

The classical price-setting newsvendor problem has been a hallmark in operations management (Whitin 1955). This problem concerns finding the optimal price and ordering quantity of a product when demand for the product is a (random) function of the posted price (Chou et al. 2012). The fundamental concept of balancing supply and demand through the use of pricing and inventory levers is embodied in this problem, which has many practical applications across various industries, such as airlines, hospitality, brick-and-mortar retailing, and online retailing.

The rapid expansion of platform-based businesses, driven by the proliferation of mobile and other information technology, is profoundly transforming the retail and service sectors (Chen et al. 2020d). These dual-sided market intermediaries, distinct from conventional inventory-based business models, employ information and communication technologies to foster interactions between users (EuroCommission 2022). Revenues are generated from the differential between the prices paid by demand-side users and the remuneration allocated to supply-side providers. In such two-sided markets, demand is contingent upon established prices while supply is influenced by the platform's remuneration offerings. A centralized decision-maker must jointly optimize pricing and remuneration strategies to maximize profits. However, the traditional price-setting newsvendor model fails to capture the random and endogenous nature of the supply side in these markets. This is precisely our main motivation for proposing a novel newsvendor model that incorporates *both pricing and remuneration decisions*, and for studying it in-depth.

4.1.1 Brief Problem Statement and Motivating Applications

We refer to our main problem as the "remunerating newsvendor" problem arising in the context of two-sided markets, which significantly expands upon the conventional price-setting newsvendor problem. This problem examines a scenario in which a platform aims to maximize its total expected revenue by jointly establishing pricing for demand and remuneration for supply. Specifically, the platform sets both the price p for customers and the remuneration w for providers. The demand from customers is given by $D(p) = \lambda(p) + \varepsilon$, and the supply from providers is given by $S(w) = \mu(w) + \delta$, where $\lambda(p)$ and $\mu(w)$ are the demand and supply functions, respectively, with random noise terms ε and δ. The goal is to maximize the expected revenue of a "repeated" remunerating newsvendor problem over a finite horizon of T periods, where a single-period time-generic objective function can be defined as follows:

$$\begin{aligned} \max_{p,\, w} \; & R(p, w) := (p - w)\mathbb{E}\left[\min(\lambda(p) + \varepsilon, \mu(w) + \delta)\right] \qquad \textbf{(Remunerating Newsvendor)} \\ \text{s.t. } & 0 \leq w \leq p. \end{aligned}$$

We are particularly interested in solving the incomplete information problem, where $D(p)$ and $S(w)$, as well as the distributions of ε and δ, are not known *a priori* and have to be learned over time. Our objective is to propose online learning algorithms with provably tight regret.

This problem setting is prevalent in contemporary business contexts, particularly in service platforms that facilitate connections between clients seeking on-demand services and independent contractors, known as "gig workers", who provide these services (Xu et al. 2023). By setting prices for clients and remuneration for providers, the platform profits from the difference between clients' payments and remuneration for suppliers. For example, digital platforms such as Handy and Helping provide home cleaning services by connecting clients with professional house cleaning and other home services. Clients are presented with predetermined rates for various services based on factors such as the size of the property, the nature of the service, and the location. A portion of the total service fee is disbursed to the professionals as remuneration, while the platform profits from the remaining amount. It is worth noting that the participation of clients and service providers depends on their agreement with the prices and remunerations determined by the platform (Benjaafar and Hu 2020), which can indirectly balance demand and supply and subsequently improve the profit.

Analogous applications are evident in service platforms that provide a range of other services. For example, Rover is a digital platform that connects pet owners with pet care service providers, including pet sitters, dog walkers, and boarding facilities. The platform establishes a suggested price range for each service type based on market data, location, and other factors, and providers can set their prices within this range. However, the platform maintains control over the general pricing structure. When a client books a service, Rover retains a commission from the fee paid by pet owners. Another example is HelloTech, which focuses on in-home and online tech support services that span a variety of areas, from smart home device installation and setup to TV mounting, computer repair, and more. The prices for different services are predetermined by the platform itself, providing predictability and transparency to customers which is crucial for customer acquisition and retention in the digital platform business. On the provider side, the techs are independent contractors compensated based on the rates set by the platform, providing a consistent and relatively predictable income stream.

Our model provides a concise and stylized representation of the newsvendor problem faced in many two-sided markets. We believe that our approaches can offer valuable insights for devising pricing and remuneration strategies in all of the above examples.

4.1.2 Main Results and Our Contributions

This chapter makes two key contributions.

Modeling. We introduce a novel newsvendor model called the "remunerating newsvendor" which incorporates remuneration as a decision variable to address the supply side's uncertainty. To the best of our knowledge, we are one of the first to expand the price-setting newsvendor model to incorporate remuneration in two-sided markets. Previous literature on price-setting newsvendor models treats supply uncertainty as exogenous, assuming random yield (e.g., Huh and Nagarajan (2010), Kazaz and Webster (2015), Kouvelis et al. (2018), Li and Zheng (2006)). However, in contemporary two-sided business platforms such as those mentioned above, these formulations fail to capture the random and endogenous nature of the supply side. Therefore, we model demand and supply as unspecified functions of price and remuneration, respectively, with corresponding price and remuneration independent randomness modeled in an additive manner. Note that the formulation of additive randomness in the price-setting newsvendor model is widely used in the literature (e.g., Chen and Simchi-Levi (2004a,b), Federgruen and Heching (1999), Petruzzi and Dada (1999) and numerous more recent works). We believe that our concise newsvendor-type model in a two-sided market serves as one of the fundamental models for contemporary businesses.

Structural Results and Online Learning Algorithms. We first analyze the complete information problem. The main structural properties are stated (informally) below.

Theorem 1 (Informal) *For the complete information problem (where the demand and supply functions as well as the noise distributions are known a priori), the expected revenue function $R(p, w)$ of the remunerating newsvendor problem has the following two properties:*

(a) $R(p, w)$ is concave and Lipschitz continuous in w for a given p.

(b) $R(p, w^*(p))$ is Lipschitz continuous in p.

Leveraging the above structural properties of the full information problem, we devise a new online learning algorithm that combines the powers of bandit control and bisection search. We provide the following *matching* upper and lower regret bounds.

Theorems 2 and 3 (Informal) *For the incomplete information problem (where the demand and supply functions as well as the noise distributions are not known a priori),*

*(a) Our bandit bisection search algorithm (**BBS** for short) attains $\mathbf{Regret}^T_{\mathbf{BBS}} = \tilde{O}(T^{\frac{2}{3}})$.*

*(b) For any algorithm **ALG**, there exist problem instances such that $\mathbf{Regret}^T_{\mathbf{ALG}} = \Omega(T^{\frac{2}{3}})$.*

Note that the regret upper and lower bounds match up to a logarithmic factor. The high-level ideas and novelties of the proposed algorithm **BBS** are summarized as follows:

(i) We propose an online algorithm that integrates bandit control (specifically, Upper-Confidence-Bound) with a bisection search (specifically, a strictly quartering search) approach, and provide proof of an upper bound on the total regret. The revenue function $R(p, w)$ is not jointly concave in p and w, which rules out direct stochastic gradient descent methods. Instead, we adopt a two-layer algorithm. Given a fixed p, the inner layer searches for the optimal w using bisection search (strictly quartering search) based on the concavity result of $R(p, w)$ stated in Theorem 4.2.1. Note that we adopt bisection search since there is no gradient information available. The outer layer then searches for the best p using the Lipschitz bandit approaches. However, the online nature of the intertwined outer and inner layers requires careful handling of intricacies that are not captured by one-dimensional bisection methods (Agarwal et al. 2011).

(ii) In contrast to previous literature that uses bisection search and its variants for operations management problems (Agarwal et al. 2011, Chen et al. 2019a, Chen and Shi 2020, Chen et al. 2021c, Lei et al. 2014), our approach integrates query operations into the bisection search process with early termination criteria to bound the loss from suboptimal bandit selections. This allows us to update the algorithm at any time instead of restricting updates to only the end of each epoch, which would otherwise result in an inability to establish a tight regret bound. On the technical side, our approach leads to a novel concentration result for the regret of bisection search caused by any bandit choice up to any time (Lemma 4.4.4 and related results). We also further modify the quartering search technique proposed by Agarwal et al. (2011) to improve efficiency. Specifically, we define a "clean event" $\mathcal{E}$ to represent the case where the estimated revenue is sufficiently close to the true expected revenue at any time over the decision space (see §4.4.2 for more details). Conditional on this clean event, we simplify their proof (Lemma 4.4.2) and improve the efficiency of the procedure by reusing queries from previous rounds. This modification may provide insights for future usage of the quartering search technique.

Subsequently, we focus our analysis on a special case, in which we presume linear relationships between the demand (supply) and price (remuneration). This specific focus allows us to examine the nuances of these relationships with greater precision. Compared to the general case, it turns out that an additional concavity result concerning a transformed variable $\Delta := p - w$ which is the gap between price and remuneration, can be attained.

Theorem 4 (Informal) *For the complete information problem where the demand and supply functions are known linear functions and the noise distributions are known a priori, the expected revenue function $R(\Delta, w)$ of the remunerating newsvendor problem has the following two properties:*

(a) $R(\Delta, w)$ is concave and Lipschitz continuous in w for a given Δ.

(b) $R(\Delta, w^*(\Delta))$ is concave and Lipschitz continuous in Δ.

For this special case, we improve the regret convergence rate by proposing another algorithm, matching the lower bound of the regret up to logarithmic factors.

Theorem 5 and Proposition 1 (Informal) *For the incomplete information problem (where the linear demand and supply functions as well as the noise distributions are not known a priori),*

*(a) Our double bisection search algorithm (**DBS** for short) attains $\mathbf{Regret}_{\mathbf{DBS}}^{T} = \tilde{O}(T^{\frac{1}{2}})$.*

*(b) For any algorithm **ALG**, there exist problem instances such that $\mathbf{Regret}_{\mathbf{ALG}}^{T} = \Omega(T^{\frac{1}{2}})$.*

Given that linear relationships are frequently discussed in the operations management literature (Keskin and Zeevi 2014, Mills 1959, Petruzzi and Dada 1999, Taylor 1974), our extension to this specific case may offer valuable insights for analogous problem settings in this field.

4.1.3 Related Literature

Our work is closely related to the following streams of literature.

Price-Setting Newsvendor with Random Supplies. The joint production-pricing optimization problem in the newsvendor model has been extensively studied, with research dating back to Whitin (1955). Interested readers can refer to survey papers such as Petruzzi and Dada (1999) and Chen and Simchi-Levi (2012) for a comprehensive overview of this literature. Our work extends the price-setting newsvendor model by incorporating remuneration decisions that affect the supply function. To provide context for our research, we briefly discuss related work on models with random supplies. In the current literature, supply uncertainty can be modeled in two main ways: through random capacity and random yield. Random capacity models refer to situations where a firm's production capacity is uncertain due to various factors such as machine breakdowns, supplier delays, or labor shortages (see, e.g., Chen et al. (2020b, 2018), Ciarallo et al. (1994), Duenyas et al. (1997), Feng (2010), Güllü (1998)). In contrast, random yield models deal with situations where the amount of usable product that a firm can produce from a batch of raw materials is uncertain and subject to random variation. This means that the firm may start with a certain quantity of raw materials, but the proportion that will become finished goods is uncertain (see, e.g., Bu et al. (2020), Federgruen and Yang (2011), Feng and Shanthikumar (2018b), Henig and Gerchak (1990), Huh and Nagarajan (2010), Wang and Gerchak (1996)). In our research, we consider a different approach to supply uncertainty. Specifically, we model supply as an endogenous function of remuneration, which is more appropriate for two-sided markets. This differs from the random capacity and random yield models discussed above, which assume that supply is exogenously determined by external factors.

We also survey the literature on newsvendor-type models in two-sided markets, with a particular focus on joint optimization of price and remuneration. Such problems are relevant to pricing in on-demand service platforms (Hu and Liu 2023). Taylor (2018) and Bai et al. (2019) formulated the problem as a queueing model and investigated the impact of various factors on optimal per-service price and wage. Hu and Zhou (2020) studied the performance of the fixed commission contracts where the wage is equal to a fraction of the price, in comparison with the optimal expected profit. In their setting, the platform is aware of the scenario realization and corresponding deterministic demand or supply functions with respect to price and wage when making the decisions, while our model assumes there is no information on the demand and supply functions and we incorporate stochasticity via random noises with unknown distributions. In addition, Parker and Van Alstyne (2005) presented a formal model of two-sided network externalities for other types of two-sided markets, such as video game platforms, newspapers, and payment card systems. Armstrong (2006) further investigated three models of such markets. Closer to our work, Chou et al. (2012) studied the pricing problem of a platform intermediary to jointly determine the selling price of platforms (hardware) sold to consumers and the royalty charged to content developers for content (software), when the demands for content and for platforms are interdependent. Their model elucidated the impact of supply chain replenishment costs and demand uncertainty on the strategic issues of platform pricing in a two-sided market. Our model differs in that we study an incomplete information problem in which the underlying demand and supply functions as well as the random noise distributions are unknown.

Online Learning Algorithms in Related Problems. Our algorithm belongs to the research area of newsvendor type models with demand learning (Chen et al. 2022b). One popular approach is to apply the Sample Average Approximation (SAA) method, which leverages the sample average of historical data to solve stochastic optimization problems in newsvendor settings. The SAA approach has been widely used in inventory management (e.g., Cheung and Simchi-Levi (2019), Kleywegt et al. (2002), Levi et al. (2015, 2007a), Lin et al. (2022)) and pricing (e.g., Qin et al. (2022)) literature. Compared with the conventional SAA approaches, we adopt a novel bandit bisection search approach.

The design of our proposed online algorithms is closely related to the literature on dynamic pricing problems with demand learning and continuous price range. Besbes and Zeevi (2009) proposed an exploration-exploitation algorithm for a single product dynamic pricing problem, based on the structural result from the seminal work by Gallego and Van Ryzin (1994), and achieved a regret bound of $\tilde{O}(T^{\frac{3}{4}})$. Wang (2014) used a more adaptive learning-while-doing approach to close the gap of the problem and attained an asymptotic regret of $\tilde{O}(T^{\frac{1}{2}})$. The framework was further extended to network revenue management problems by Besbes and Zeevi (2012), and several follow-up works have used different techniques, such as spline approximation by Chen et al. (2019b), bisection search by Lei et al. (2014), primal-dual approach by Chen and Gallego (2022), online inverse batch gradient descent by Chen and Shi (2023), and robust ellipsoid method by Miao and Wang (2021). Our work distinguishes itself from the existing literature by introducing a remuneration mechanism into the newsvendor model and jointly optimizing price and remuneration to maximize the expected revenue, instead of only optimizing the price.

The optimization process for remuneration in our algorithm, which constitutes the inner layer, essentially involves solving a convex optimization problem with bandit feedback. In contrast to previous dynamic pricing settings, the convergence rate of the learning process is embedded in another optimization layer and therefore requires careful design. As the problem is, in essence, a convex optimization problem with bandit feedback, Flaxman et al. (2005) initiated a discussion on learning from bandit feedback under convex and Lipschitz reward functions. They approximated the gradient using a random sampling technique and achieved an $O(T^{\frac{3}{4}})$ regret bound using the Bandit Gradient Descent algorithm. Cope (2009) tackled the problem using Kiefer-Wolfowitz Stochastic Approximation techniques and achieved a $O(\sqrt{T})$ regret bound under strict conditions. Agarwal et al. (2011) proposed a bisection method that achieves an $O(\sqrt{T})$ regret bound, proven to be tight in Shamir (2013). Note that the term "bisection" is used here for consistency with the existing optimization literature. While the method in Lei et al. (2014) was actually a trisection search, Agarwal et al. (2011) used a quartering search, with the middle point out of the three trials each time serving as a sentinel benefiting the total regret. To the best of our knowledge, Chen and Shi (2020) is the only work adopting this approach in revenue management. The difference between the application of the bisection approach in our work and theirs is that, in order to bound the loss for total regret, we distribute the bisection search process to multiple outer loops, and therefore, the search process is no longer consecutively executed.

All aforementioned approaches require structural properties of the objective function, such as unimodality or concavity. Our algorithm is also related to the problem of continuum-armed bandits (Agrawal 1995) for price optimization, which is the outer loop of our algorithm and has no particular structural results besides Lipschitz continuity. The continuum-armed bandit problem was addressed by Kleinberg (2004) using uniform discretization, while the Lipschitz bandits problem was addressed in Kleinberg et al. (2008). In the context of dynamic pricing, Wang et al. (2021b) studied the single product pricing problem allowing the expected reward function to be multimodal. They proposed an algorithm combining the Upper-Confidence-Bound algorithm for the multi-armed bandit and local polynomial approximation. In contrast to optimizing only price under various smoothness conditions in Wang et al. (2021b), we study the joint optimization of price and remuneration with first-order smoothness conditions (see Assumption 4.2.1b).

4.1.4 Organization and Notation

The rest of the chapter is structured as follows. In §4.2, we present our problem formulation and discuss several essential structural properties of our revenue function. In §4.3, we propose an online algorithm for solving the incomplete information problem and analyze its convergence rate in §4.4, which is proved to be optimal up to logarithmic factors in §4.5. Then, in §4.6, we consider a special case with linear demand and supply functions and derive better regret bounds. Numerical experiments in §4.7 validate the performance of proposed algorithms. Finally, we conclude the chapter in §4.8.

We also introduce some general notation used in this chapter. The Big-O notation $\tilde{O}(\cdot)$ hides any logarithmic factors. For any scalar a, $|a|$ denotes its absolute value. For any set A, $|A|$ denotes its cardinality; if A is an interval, then $|A|$ is its length. The event $A^{\complement}$ denotes the complement of event A. For any integer $n \geq 1$, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

4.2 Problem Formulation and Structural Properties

We formally describe what we call the "remunerating newsvendor" problem, involving joint pricing and remuneration decisions. Suppose there is a firm that sells a single product with the goal of maximizing its expected revenue. The decisions to be made include the price p for customers and the remuneration or wage w for suppliers, while the supply S is a random variable depending on the remuneration, and demand D is a random variable depending on the price. Specifically, we assume additive random noises for demand and supply, i.e.,

$$D(p) = \lambda(p) + \varepsilon, \qquad S(w) = \mu(w) + \delta,$$

where $\lambda(p)$ and $\mu(w)$ represent the underlying demand and supply functions, while ε and δ denote random noises with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\delta] = 0$. Then the remunerating newsvendor problem for maximizing the expected revenue can be formulated as

$$\begin{aligned} \max_{p,\, w}\ & R(p, w) := (p - w)\mathbb{E}[\min(\lambda(p) + \varepsilon, \mu(w) + \delta)] \\ & \text{s.t. } 0 \leq w \leq p. \end{aligned} \tag{4.1}$$

Assumption 4.2.1 *We make the following modeling assumptions.*

(a) There are lower and upper bounds for the price p, denoted by $\underline{P}$ and $\bar{P}$, respectively. Additionally, for any $0 \leq w \leq p \leq \bar{P}$, we assume that $0 \leq D(p) \leq \bar{S}$ and $0 \leq S(w) \leq \bar{S}$ almost surely for some finite positive constant $\bar{S}$.

(b) The demand function $\lambda(p)$ is Lipschitz continuous in p. That is, there exists a constant K_1 such that $|\lambda(p_1) - \lambda(p_2)| \leq K_1|p_1 - p_2|$.

(c) The supply function $\mu(w)$ is Lipschitz continuous in w. That is, there exists a constant

K_2 such that $|\mu(w_1) - \mu(w_2)| \leq K_2|w_1 - w_2|$.

(d) The supply function $\mu(w)$ is non-decreasing and concave in w, i.e., $\mu'(w) \geq 0$ and $\mu''(w) \leq 0$.

Assumption 4.2.1a posits that the platform sets the price over a bounded set. Additionally, there is a cap on the maximum values for both demand and supply. Furthermore, we assume that the probability of negative demand or negative supply is negligible. This condition can be readily satisfied if the noises have finite supports. In cases where the noises are unbounded, it is more realistic to consider scenarios where the standard deviation of the noises is small enough to ensure non-negative demand and supply. Similar assumptions regarding the impact of random noises on the non-negativity of demand can be found in, for example, Le Guen (2008) §4.4.1, Keskin and Zeevi (2014) §2, and Snyder and Shen (2019) §4.2.

Assumptions 4.2.1b and 4.2.1c essentially suggest that the rates of change in demand and supply, $\lambda'(p)$ and $\mu'(w)$, are both bounded, which is standard in the revenue management literature (Besbes and Zeevi 2009, 2012, Lei et al. 2014, Wang 2014).

Assumption 4.2.1d introduces an additional regularity assumption for the supply function, which is also commonly assumed in the literature (Talluri et al. 2004). Additionally, it is assumed that the supply function is non-decreasing in remuneration, which is expected in practical applications. Furthermore, the assumption states that the supply curve is concave in $w \in [0, \bar{P}]$, indicating a saturation effect in the finite supply market where the rate of increase in supply slows down as the remuneration approaches $\bar{P}$. This assumption is also employed in Hu and Zhou (2020). Another interpretation is a case where the seller has a random utility U and will participate in the system only when the realized utility is less than or equal to the remuneration w. In this setting, the probability that the seller will join the market given the remuneration w is $\mu(w) = \mathbb{P}(U \leq w)$. Then Assumption d is satisfied when the cumulative distribution function of the random variable U is concave, which is considered a reasonable assumption (e.g., exponential distribution).

Based on the formulation and assumptions described above, in the case of complete information where the demand and supply functions, as well as the random noise distributions, are known *a priori*, we establish Theorem 4.2.1. The theorem states the concavity of the expected revenue with respect to the remuneration given a fixed price in a, and the smoothness condition of the optimal expected revenue as a function of the price in b. These structural results serve as the foundation for the design of Algorithm 6 in §4.3. The full proof of Theorem 4.2.1 is given in Appendix C.2.1.

Theorem 4.2.1 *In the case of complete information where the demand and supply func-*

tions, as well as the random noise distributions, are known a priori, the remunerating newsvendor problem has the following structural properties under Assumptions 4.2.1c and 4.2.1d.

(a) The expected revenue $R(p,w)$ is concave and $\max\{K_2,\bar{S}\}\bar{P}$-Lipschitz in the remuneration $w \in [0,p]$. That is, for any $w_1, w_2 \in [0,p]$,

$$|R(p,w_1) - R(p,w_2)| \leq \max\{K_2,\bar{S}\}\bar{P}\,|w_1 - w_2|\,.$$

(b) The optimal expected revenue as a function of price p, denoted by $R\left(p, w^*(p)\right)$, is Lipschitz continuous in p. That is, for any $p_1, p_2 \in [\underline{P}, \bar{P}]$,

$$\left|R\left(p_1, w^*(p_1)\right) - R\left(p_2, w^*(p_2)\right)\right| \leq K_3\,|p_1 - p_2|\,,$$

where $K_3 := \max\left\{K_1\bar{P}, \max\{K_2,\bar{S}\}\bar{P} + \bar{S}\right\}$.

Now, in the case of incomplete information where the demand and supply functions, as well as the random noise distributions, are not known *a priori*, we introduce the notion of regret as a performance measure for online learning algorithms. Consider a T-period problem in which the firm needs to jointly determine the prices for customers and the remunerations for suppliers for each period. Any excess demand or supply at the end of each period will be lost. Let the clairvoyant (or the complete information) optimal price be p^* and the clairvoyant optimal remuneration be w^*. Recall that the expected revenue of price and remuneration pair (p,w) is denoted by $R(p,w)$ as defined in (4.1). Define the regret of any learning algorithm **ALG** for a T-period finite horizon joint pricing and remuneration decision problem as

$$\mathbf{Regret}^T_{\mathbf{ALG}} = TR(p^*,w^*) - \mathbb{E}\left[\sum_{s=1}^{T} \breve{R}^s_{\mathbf{ALG}}\right],$$

where $\breve{R}^s_{\mathbf{ALG}}$ denotes the revenue collected in period s by running algorithm **ALG**.

4.3 Bandit Bisection Search Algorithm (BBS)

In the case of incomplete information where the demand and supply functions, as well as the random noise distributions, are not known *a priori*, we propose a new online learning algorithm, termed Bandit Bisection Search Algorithm (**BBS**), and also discuss its main ideas and novelties. Subsequently, each arm (bandit) refers to a discretized price value. Pulling or selecting an arm means choosing the corresponding value as the price to implement. Subscripts l, c, r stand for "left", "center", and "right", respectively.

4.3.1 A Two-Layered Design of BBS

Algorithm 6 is essentially a two-layered algorithm that integrates bandit control with a bisection search method. Let us explain why we need a two-layered design where the inner layer aims to optimize the remuneration w^* and the outer layer aims to optimize the price p^*. For the inner layer where price p is fixed, based on Theorem 4.2.1a, we know that the expected revenue $R\left(p, w\right)$ is concave in the remuneration w. One might naturally consider employing gradient descent to find $w^*(p)$. However, since we are unable to extract gradient information, our best option is to employ a bisection search approach to find $w^*(p)$ for each value of p.

Now let us move on to the outer layer. Suppose that we know $w^*(p)$ for any p (from running the inner layer). Determining the optimal price p that maximizes $R\left(p, w^*(p)\right)$ can be viewed as a continuum-armed bandit (CAB) problem, given the Lipschitz continuity property established in Theorem 4.2.1b. Note that the set of arms is the continuous price range $p \in [\underline{P}, \bar{P}]$. Consequently, we utilize a discretization technique in conjunction with the Upper-Confidence-Bound algorithm for multi-armed bandits in order to identify the optimal price p^*.

While the aforementioned two-layered design appears to be natural and feasible, executing them in the correct sequence and order presents significant challenges. If not executed carefully, the resulting regret could become multiplicative with respect to the inner and outer layers. It is worth noting that bisection search and bandit control individually yield a regret of $O(\sqrt{T})$, and therefore a naive implementation would result in a linear regret $(O(\sqrt{T}) \times O(\sqrt{T}))$. Thus, effectively intertwining these two methods remains a critical challenge.

4.3.2 Detailed Implementation of BBS and Technical Novelties

Based on the intuition above, we now give more details of implementation. For ease of presentation, we consider a time horizon of $3T$ periods. We then use the term "time t" hereafter to denote periods $3(t-1)+1$, $3(t-1)+2$, $3(t-1)+3$. The algorithm operates as depicted in Figure 4.1.

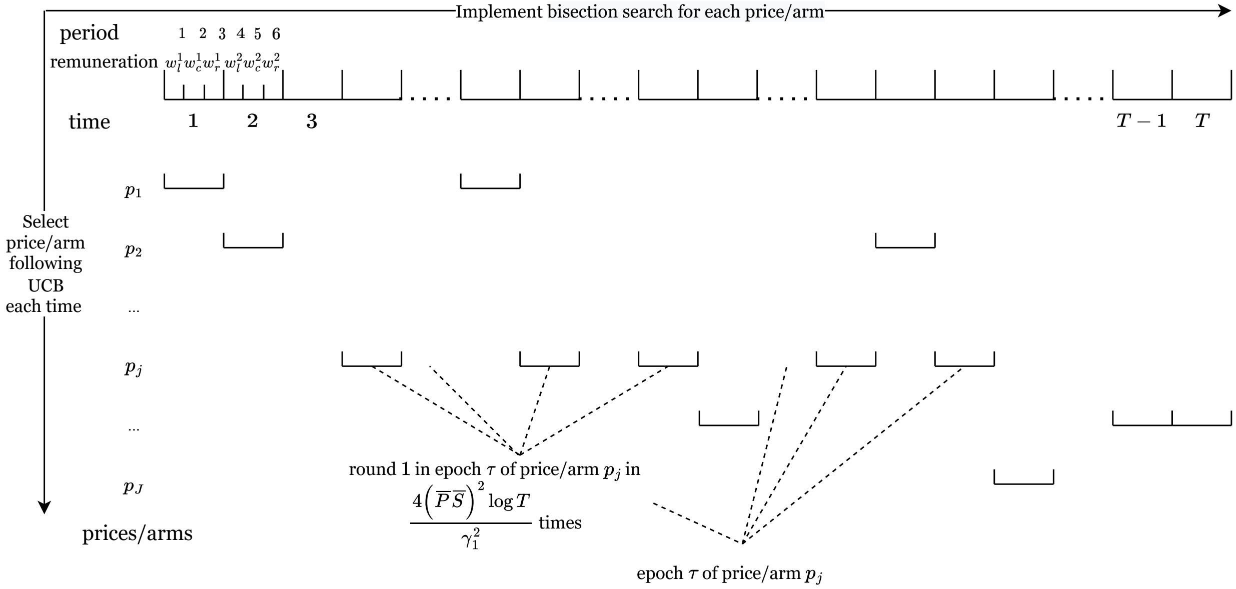

Implement bisection search for each price/arm
period
1 2 3 4 5 6
remuneration $w_l^1 w_c^1 w_r^1 w_l^2 w_c^2 w_r^2$
time
1 2 3
$T-1$ T
p_1
p_2
...
p_j
...
p_J
Select price/arm following UCB each time
prices/arms
round 1 in epoch τ of price/arm p_j in $\frac{4\left(\overline{PS}\right)^2 \log T}{\gamma_1^2}$ times
epoch τ of price/arm p_j

Figure 4.1: Illustration of Algorithm 6

The price range is discretized into J discrete values p_j for $j \in [J]$ with equal intervals. Each price corresponds to an "arm" in the outer loop algorithm, forming the vertical dimension. The top horizontal line in the figure represents time intervals with each consisting of 3 periods. Each subsequent horizontal line represents the time interval in which the price $p_j,\ j \in [J]$ gets implemented. Whenever the price is selected, three quartiles of the current interval for the optimal remuneration for this price are implemented as remuneration successively. In Figure 4.1, the interval corresponding to the price and three periods is solid. As shown in the figure, the price p_j is implemented during a subset of T times, which is not necessarily consecutive. Based on the value of remuneration chosen at each time, this subset of times consists of multiple epochs with each epoch comprising multiple rounds. The length of an epoch depends on both the shape of the function $R(p_j, w)$ concerning w and the separation of values in the inspected interval sample path.

At the beginning of each time t (i.e., periods $3(t-1)+1,\ 3(t-1)+2,\ 3(t-1)+3$), the algorithm selects the price with the highest upper confidence bound following the Upper-Confidence-Bound (UCB) algorithm. This selection is based on the empirical revenue obtained from historical data. The chosen arm j^t is associated with an interval $[l_{j^t}, r_{j^t}]$, which contains the optimal remuneration $w^*(p_{j^t})$ for the corresponding price with high probability. Subsequently, the algorithm implements the price p_{j^t} along with the remunerations $w_l^t = \frac{3}{4}l_{j^t} + \frac{1}{4}r_{j^t}$, $w_c^t = \frac{1}{2}l_{j^t} + \frac{1}{2}r_{j^t}$, $w_r^t = \frac{1}{4}l_{j^t} + \frac{3}{4}r_{j^t}$ for three consecutive periods.

The revenue realizations of $\left(p_{j^t}, w_x^t\right), x \in \{l, c, r\}$ are then used to obtain the following estimators:

(a) Empirical average of the revenue $R\left(p_{j^t}, w_x^t\right),\ x \in \{l, c, r\}$ for bisection search, i.e., $\frac{\hat{R}_{j^t,x}^{t+1}}{m_{j^t}^{t+1}}$.

(b) Estimation of the revenue $R\left(p_{j^t}, w^*(p_{j^t})\right)$ for the UCB procedure, i.e., $\frac{\alpha_{j^t}^{t+1}}{3n_{j^t}^{t+1}}$.

This process continues until the number of samples for $R\left(p_{j^t}, w_x^t\right),\ x \in \{l, c, r\}$ reaches $4\gamma_1^{-2}(\bar{P}\bar{S})^2 \log T$. Subsequently, we compare the empirical means of $R\left(p_{j^t}, w_x^t\right),\ x \in \{l, c, r\}$ to narrow down the interval. Lemma 4.4.1 assures that estimator a provides sufficient accuracy with a high probability for the truncation operation. If the interval can be successfully narrowed, the algorithm updates the interval for arm j^t and proceeds to another epoch. However, if one cannot discard a quarter interval based on the estimated revenue for the three points, the algorithm continues to query the points in the original interval for arm j^t and initiates a new round within the same epoch, where the term "query" applied to the process of bisection search means implementing the process in one period with the queried value as the remuneration. During this round, a comparison will occur on a smaller scale, and the cumulative number of samples required before comparison will increase. Regarding the loss incurred due to the suboptimality of the arm selected each time, specifically the estimation in b, Lemma 4.4.4 guarantees the accuracy of estimating $R\left(p_{j^t}, w^*(p_{j^t})\right)$. This accurate estimation helps in bounding the total regret.

In the following, we emphasize a crucial aspect in designing Algorithm 6 and the underlying technical innovations. One key consideration when designing the algorithm is whether we are compelled or restricted to updating the Confidence Radius (**Rad**) of the UCB algorithm and selecting a better arm solely until an epoch or a round of the bisection search is complete. Note that the total loss compared to the optimal expected revenue of the optimal arm can be decomposed into loss from suboptimality of the price p_j as in (4.2) and the loss from the remuneration given the price as in (4.3), i.e., the process of approximating $w^*(p_j)$.

$$
\begin{aligned}
&3TR\left(p_{j^*}, w^*(p_{j^*})\right) - \sum_{s=1}^{3T} \breve{R}_{\mathbf{BBS}}^s \\
=&3TR\left(p_{j^*}, w^*(p_{j^*})\right) - \sum_{j=1}^{J} 3n_j^{T+1} R\left(p_j, w^*(p_j)\right) \qquad (4.2)\\
&+\sum_{j=1}^{J}\left(3n_j^{T+1} R\left(p_j, w^*(p_j)\right) - \sum_{t\in H_j}\sum_{x\in\{l,c,r\}} \tilde{R}_x^t\right), \qquad (4.3)
\end{aligned}
$$

where $3n_j^{T+1}$ is the total number of times of selecting price p_j and H_j denotes the set of indices of time when the price is p_j. Part (4.3) can be bounded by the bisection search process with

Algorithm 6 Bandit Bisection Search Algorithm (**BBS**)

```
Let $J = \#$ discrete $p$'s and $p_j = \underline{P} + \frac{j}{J}\left|\bar{P} - \underline{P}\right|$ for $j = 1, \ldots, J$. Let $\gamma_i := 2^{-i}$, $i \geq 1$.   ▷ Input
for $j \in [J]$ do Initialize:   ▷ Initialization
    $l_j = 0$ and $r_j = p_j$ to be the left and right end-points of the interval for $w^*(p_j)$, respectively;
    $k_j = 1$ to be the current level of confidence of estimation for price $p_j$;
    $m_j^1 = 0$ and use $m_j^t$ to track the number of times of selecting price $p_j$ in the current epoch;
    $n_j^1 = 0$ and use $n_j^t$ to track the total number of times of selecting price $p_j$ until time $t$;
    $\alpha_j^1 = 0$ and use $\alpha_j^t$ to track the cumulative realized revenue using price $p_j$ until time $t$;
    $\hat{R}_{j,x}^1 = 0, x \in \{l, c, r\}$ and use $\hat{R}_{j,x}^t$ to track the cumulative realized revenue $\hat{R}_{j,x}^t$ using price $p_j$ in the current epoch of price $p_j$;
    $\mathbf{Rad}_j^1 = 0$ and use $\mathbf{Rad}_j^t$ to denote the confidence radius of estimation of $R\left(p_j, w^*(p_j)\right)$ for price $p_j$ until time $t$.
end for
for $t = 1, 2, \ldots, T$ do
    if $t \leq J$ then
        $j^t = t$,
    else
        Select $j^t = \arg\max_{j \in [J]} \left\{ \frac{\alpha_j^t}{3n_j^t} + \mathbf{Rad}_j^t \right\}$.   ▷ Outer Layer (Bandit Control)
    end if
    Let $u = r_{j^t} - l_{j^t}$ and $w_l^t = l_{j^t} + \frac{1}{4}u, w_c^t = l_{j^t} + \frac{1}{2}u, w_r^t = l_{j^t} + \frac{3}{4}u$.
    Implement $(p_{j^t}, w_x^t)$ and obtain realized revenue $\tilde{R}_x^t$ for $x \in \{l, c, r\}$.
    Update $\alpha_{j^t}^{t+1} = \alpha_{j^t}^t + \sum_{x \in \{l,c,r\}} \tilde{R}_x^t$,  $n_{j^t}^{t+1} = n_{j^t}^t + 1$,  $m_{j^t}^{t+1} = m_{j^t}^t + 1$,
    $\hat{R}_{j^t,x}^{t+1} = \hat{R}_{j^t,x}^t + \tilde{R}_x^t, x \in \{l, c, r\}$, $\mathbf{Rad}_{j^t}^{t+1} = \max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\frac{\log T}{n_{j^t}^{t+1}}} \log_{\frac{4}{3}} \frac{\bar{P}^2 \max\{K_2^2, \bar{S}^2\} n_{j^t}^{t+1}}{9\bar{S}^2 \log T}$.
    For $j \neq j^t$, update $\alpha_j^{t+1} = \alpha_j^t$,  $n_j^{t+1} = n_j^t$,  $m_j^{t+1} = m_j^t$,  $\hat{R}_{j,x}^{t+1} = \hat{R}_{j,x}^t$,  $x \in \{l, c, r\}$, $\mathbf{Rad}_j^{t+1} = \mathbf{Rad}_j^t$.
    if $m_{j^t}^{t+1} = \left\lceil 4\gamma_{k_{j^t}}^{-2}(\bar{P}\bar{S})^2 \log T \right\rceil$ then        ▷ Inner Layer (Bisection Search)
        Arm $j^t$ enters a new round.
        For $x \in \{l, c, r\}$, $\mathrm{UB}_{k_{j^t}}(w_x^t) = \frac{\hat{R}_{j^t,x}^{t+1}}{m_{j^t}^{t+1}} + \frac{\gamma_{k_{j^t}}}{2}$, $\mathrm{LB}_{k_{j^t}}(w_x^t) = \frac{\hat{R}_{j^t,x}^{t+1}}{m_{j^t}^{t+1}} - \frac{\gamma_{k_{j^t}}}{2}$.
        if $\min\left\{\mathrm{UB}_{k_{j^t}}(w_l^t),\ \mathrm{UB}_{k_{j^t}}(w_r^t)\right\} \le \max\left\{\mathrm{LB}_{k_{j^t}}(w_l^t),\ \mathrm{LB}_{k_{j^t}}(w_r^t)\right\} - \gamma_{k_{j^t}}$ then
            if $\mathrm{UB}_{k_{j^t}}(w_l^t) \le \mathrm{UB}_{k_{j^t}}(w_r^t)$ then $l_{j^t} = w_l^t, r_{j^t} = r_{j^t}$,
            else $l_{j^t} = l_{j^t},\ r_{j^t} = w_r^t$.
            end if
            $k_{j^t} = 1,\ m_{j^t}^{t+1} = 0,\ \hat{R}_{j^t,x}^{t+1} = 0, x \in \{l, c, r\}$, arm $j^t$ enters a new epoch.
        else if $\min\left\{\mathrm{UB}_{k_{j^t}}(w_l^t), \mathrm{UB}_{k_{j^t}}(w_r^t)\right\} \le \mathrm{LB}_{k_{j^t}}(w_c^t) - \gamma_{k_{j^t}}$ then
            if $\mathrm{UB}_{k_{j^t}}(w_l^t) \le \mathrm{UB}_{k_{j^t}}(w_r^t)$ then $l_{j^t} = w_l^t,\ r_{j^t} = r_{j^t}$,
            else $l_{j^t} = l_{j^t},\ r_{j^t} = w_r^t$.
            end if
            $k_{j^t} = 1,\ m_{j^t}^{t+1} = 0,\ \hat{R}_{j^t,x}^{t+1} = 0, x \in \{l, c, r\}$, arm $j^t$ enters a new epoch.
        else $k_{j^t} = k_{j^t} + 1$.
        end if
    end if
end for
```

high probability. For any price p_j, the number of periods in one epoch of bisection search depends on the separation of the queried remuneration values and cannot be bounded by a small enough value. If we keep selecting price p_j until one epoch finishes in the inner layer bisection search, the number of periods becomes unpredictable. Consequently, Part (4.2), incurred from the product of the gap between the optimal revenue of p_{j^*} and selected p_{j^t} and the total number of times price p_{j^t} is selected, cannot be bounded efficiently. Therefore, to address the suboptimality of the arm, Algorithm 6 enhances the price choice at each time, and the bisection search process for each arm does not need to be consecutive. This unique characteristic differentiates our method from previous literature employing either bisection or trisection search (Agarwal et al. 2011, Chen et al. 2019a, Chen and Shi 2020, Chen et al. 2021c, Lei et al. 2014). Essentially, this "smart" adaptive querying operation in the bisection search procedure enables us to update the algorithm at any time, rather than restricting updates to only the end of each epoch. This flexibility is essential to establish a tight regret bound based on the fact that the estimation of the optimal expected revenue using price $p_1, \ldots, p_J$ at any time t will be sufficiently accurate with high probability (see Lemma 4.4.4).

We can further improve the efficiency of the bisection search procedure by altering the sampling process to *reuse* samples from previous rounds. Specifically, for the quartering search technique, both Agarwal et al. (2011) and Chen and Shi (2020) query exactly $\Theta\left(\frac{\log T}{\gamma_i^2}\right)$

in round i consecutively. In Algorithm 6, however, round $i \geq 2$ incorporates $\Theta\left(\frac{\log T}{\gamma_i^2} - \frac{\log T}{\gamma_{i-1}^2}\right)$ samples, i.e., we will reutilize the samples from previous rounds for estimation in the current round. Although this modification cannot strictly improve the order of the final regret bound, it potentially accelerates the bisection search process and also simplifies the proof for regret bound for each epoch (see the proof of Lemma 4.4.2 in comparison with the proof of Lemma 3 in Agarwal et al. (2011)).

In summary, the unique querying approach enables the algorithm to continuously update the confidence radius (**Rad**) and select a new arm at any point in time, rather than only at the end of an epoch or a complete round of the bisection search. This prevents the regret from being adversely affected by excessive exploration of suboptimal arms. Furthermore, efficiency is enhanced by reusing queries from previous rounds within each epoch.

4.4 Regret Analysis

We first state our main result in Theorem 4.4.1. Together with Theorem 4.5.1 established later, we obtain *matching* upper and lower regret bounds, up to a logarithmic factor.

Theorem 4.4.1 *Under Assumption 4.2.1, by setting the price discretization parameter* $J = T^{\frac{1}{3}}$*, Algorithm 6 achieves* $\mathbf{Regret}_{\mathbf{BBS}}^{3T} = \tilde{O}(T^{\frac{2}{3}})$ *for a finite horizon of* $3T$ *periods.*

The remainder of this section is devoted to proving Theorem 4.4.1. For notational convenience, for any $j \in [J]$, we define the following notations. We denote the set of time indices contained in price p_j's epoch τ as $L_{j,\tau}$ and those when the price is p_j as set H_j. Let the index of the epoch of price p_j at time t be τ_j^t. Additionally, we denote the left and right end-points of the interval of price p_j in epoch τ as l_j^τ and r_j^τ. Recall that we employ w_x^t, $x \in \{l, c, r\}$ to denote the remuneration values implemented at time t.

4.4.1 High Level Roadmap

To present a concise overview of the regret upper bound for the **BBS** Algorithm, we provide a high-level demonstration using Figure 4.2 as a visual roadmap. In particular, for any given sample path, the total loss of Algorithm 6 can be decomposed into the following components:

$$
\begin{aligned}
&3TR(p^*, w^*) - \sum_{s=1}^{3T} \breve{R}_{\mathbf{BBS}}^{s} \\
=&3TR\left(p^*, w^*\right) - 3TR\left(p_{j^*}, w^*(p_{j^*})\right) && \textbf{(Discretization Loss)} \quad (4.4) \\
&+ 3TR\left(p_{j^*}, w^*(p_{j^*})\right) - 3\sum_{j=1}^{J} \sum_{t \in H_j} R\left(p_j, w^*(p_j)\right) && \textbf{(Bandit Selection Loss)} \quad (4.5)
\end{aligned}
$$

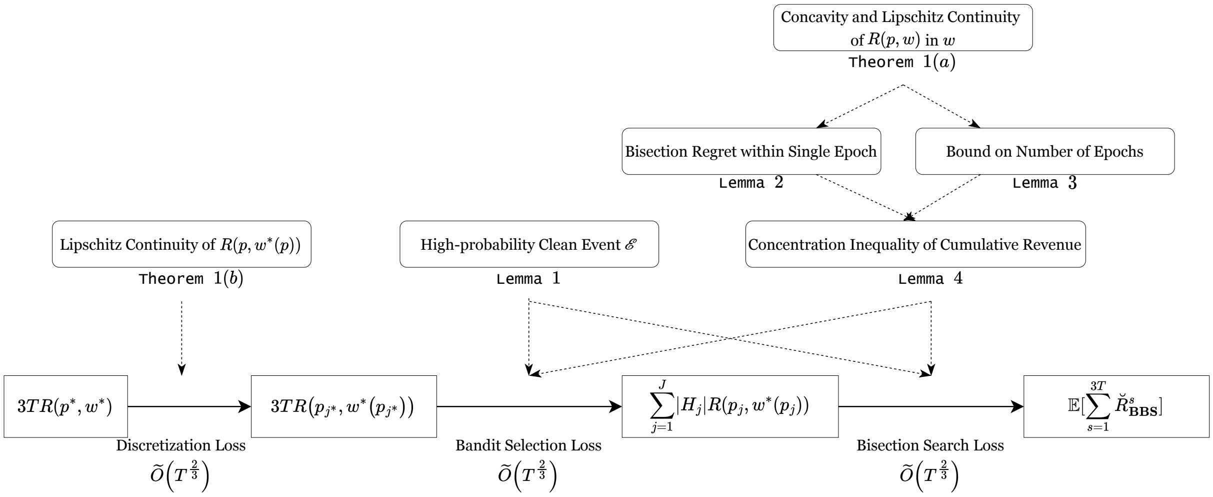

Concavity and Lipschitz Continuity of $R(p,w)$ in w
Theorem $1(a)$
Bisection Regret within Single Epoch
Lemma 2
Bound on Number of Epochs
Lemma 3
Lipschitz Continuity of $R(p,w^*(p))$
Theorem $1(b)$
High-probability Clean Event $\mathscr{E}$
Lemma 1
Concentration Inequality of Cumulative Revenue
Lemma 4
$3TR(p^*,w^*)$
$3TR(p_{j^*},w^*(p_{j^*}))$
$\sum_{j=1}^{J}|H_j|R(p_j,w^*(p_j))$
$\mathbb{E}[\sum_{s=1}^{3T}\breve{R}^s_{\mathbf{BBS}}]$
Discretization Loss
$\tilde{O}\big(T^{\frac{2}{3}}\big)$
Bandit Selection Loss
$\tilde{O}\big(T^{\frac{2}{3}}\big)$
Bisection Search Loss
$\tilde{O}\big(T^{\frac{2}{3}}\big)$

Figure 4.2: Proof of Theorem 4.4.1 Roadmap

$$+\sum_{j=1}^{J}\sum_{t\in H_j}\left(3R\left(p_j,w^*(p_j)\right)-\sum_{x\in\{l,c,r\}}\tilde{R}_x^t\right), \qquad \textbf{(Bisection Search Loss)} \tag{4.6}$$

where (4.4) is the loss from the discretization over the price range, (4.5) is the loss from selecting the suboptimal arms of price, and (4.6) is the loss incurred during the process of the bisection search for the optimal remuneration for each arm.

For the upper bound of the Discretization Loss (4.4), since $R\left(p,w^*(p)\right)$ is Lipschitz in p as shown in Theorem 4.2.1b, the difference between the expected revenue $R\left(p^*,w^*(p^*)\right)$ and $R\left(p_{j^*},w^*(p_{j^*})\right)$ can be bounded by the difference between p^* and p_{j^*} by controlling the discretization interval to be $O(T^{-\frac{1}{3}})$.

Recall that we select the price arm with the highest upper confidence bound estimated by the average realized revenue. Loss (4.5) from selecting a suboptimal price at time t can be bounded by the confidence radius at time t in estimating the optimal expected revenue using the price. The confidence radius can be further bounded as below. For any price $p_j,\ j\in[J]$, the set of times when the price is p_j until time t consist of multiple epochs, where the loss in each epoch originating from the suboptimality of the remuneration can be bounded by the property of bisection search in Lemma 4.4.2. Furthermore, the number of rounds in any epoch for any $j\in[J]$ is upper bounded based on the number of times of selecting price p_j up to time t. Then the length of the interval for remuneration in time t is bounded below. Because for any $j\in[J]$ and any k, the interval after round k will contain all w values satisfying $R\left(p_j,w^*\left(p_j\right)\right)-R\left(p_j,w\right)\leq\gamma^k$. Thus, if the number of rounds is upper bounded, by Lipschitz continuity of $R(p,w)$ in w, the length of the interval for w is bounded below. In addition, the length of the interval for searching $w^*(p_j)$ will shrink by a factor of $\frac{3}{4}$ after each epoch. Therefore, the number of epochs can be bounded as in Lemma 4.4.3. Combined with the bound on the loss incurred from each epoch, the confidence radius can be bounded.

Part (4.6) incurred from the suboptimality of remuneration in the bisection process for each pulled arm, can be bounded in the same way as discussed above.

4.4.2 High Probability Event

We establish a high probability result for the "clean event" as opposed to "bad events" which will incur a large regret. Specifically, we present the following lemma.

Lemma 4.4.1 *We define an event $\mathcal{E}$ (the so-called clean event) as*

$$\mathcal{E} := \left\{ \left| \frac{\hat{R}^{t+1}_{j^t,x}}{m^{t+1}_{j^t}} - R\left(p_{j^t}, w^t_x\right) \right| \leq \bar{P}\bar{S}\sqrt{\frac{\log T}{m^{t+1}_{j^t}}},\ \forall t \in [T],\ x \in \{l, c, r\} \right\}. \tag{4.7}$$

Then we have $\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \frac{6}{T}$. Moreover, conditional on event $\mathcal{E}$, for any calculated $\mathrm{UB}_{k_{j^t}}\left(w^t_x\right)$ and $\mathrm{LB}_{k_{j^t}}\left(w^t_x\right)$, we have $R\left(p_j, w^t_x\right) \in \left[\mathrm{LB}_{k_{j^t}}\left(w^t_x\right), \mathrm{UB}_{k_{j^t}}\left(w^t_x\right)\right]$ for any $x \in \{l, c, r\}$.

Proof of Lemma 4.4.1. For any $t \in [T]$ and $x \in \{l, c, r\}$, we define the event

$$\mathcal{E}^t_x := \left\{ \left| \frac{\hat{R}^{t+1}_{j^t,x}}{m^{t+1}_{j^t}} - R\left(p_{j^t}, w^t_x\right) \right| \leq \bar{P}\bar{S}\sqrt{\frac{\log T}{m^{t+1}_{j^t}}} \right\},$$

indicating that the average of $m^{t+1}_{j^t}$ realizations of revenue using (p_{j^t}, w^t_x) in the current round of pulled arm j^t sufficiently approximates $R\left(p_{j^t}, w^t_x\right)$. By Hoeffding's inequality, we have $\mathbb{P}\left(\left(\mathcal{E}^t_x\right)^{\complement}\right) \leq \frac{2}{T^2}$. Note that the event $\mathcal{E} = \cap_{t=1}^{T} \cap_{x\in\{l,c,r\}} \mathcal{E}^t_x$ denotes the event that at any time t and $x \in \{l, c, r\}$ the estimate of $R\left(p_{j^t}, w^t_x\right)$ is sufficiently accurate. Then by union bound, we have $\mathbb{P}\left(\mathcal{E}^{\complement}\right) \leq \frac{6}{T}$.

Recall that

$$\mathrm{UB}_{k_{j^t}}\left(w^t_x\right) = \frac{\hat{R}^{t+1}_{j^t,x}}{m^{t+1}_{j^t}} + \frac{\gamma_{k_{j^t}}}{2}$$

is calculated when $m^{t+1}_{j^t}$ reaches $\left\lceil \frac{4(\bar{P}\bar{S})^2 \log T}{\gamma^2_{k_{j^t}}} \right\rceil$. Consequently,

$$\mathrm{UB}_{k_{j^t}}\left(w^t_x\right) \geq \frac{\hat{R}^{t+1}_{j^t,x}}{m^{t+1}_{j^t}} + \bar{P}\bar{S}\sqrt{\frac{\log T}{m^{t+1}_{j^t}}}.$$

Similar analysis can be applied to $\mathrm{LB}_{k_{jt}}\left(w_x^t\right)$ and we have

$$\mathrm{LB}_{k_{jt}}\left(w_x^t\right) \leq \frac{\hat{R}_{jt,x}^{t+1}}{m_{jt}^{t+1}} - \bar{P}\bar{S}\sqrt{\frac{\log T}{m_{jt}^{t+1}}}.$$

Therefore conditional on event $\mathcal{E}$, we have $R\left(p_j, w_x^t\right) \in \left[\mathrm{LB}_{k_{jt}}\left(w_x^t\right), \mathrm{UB}_{k_{jt}}\left(w_x^t\right)\right]$ for any $x \in \{l, c, r\}$ and any $t \in [T]$. **Q.E.D.**

4.4.3 New Concentration Bound for Estimation

We analyze the regret conditional on the event $\mathcal{E}$, as defined in (4.7), which occurs with a probability of at least $1 - \frac{6}{T}$. Specifically, we first bound the regret for any price p_j in any epoch τ (§4.4.3.1), then we bound the number of epochs until time t (§4.4.3.2). Finally, we provide the result to bound the deviation of the total realized revenue for any price choice until time t from the true optimal revenue (§4.4.3.3).

4.4.3.1 Regret within Single Epoch.

We have the following lemma that bounds the deviation of total realized revenue $\sum_{t \in L_{j,\tau}} \tilde{R}_x^t$ from the optimal revenue $\left|L_{j,\tau}\right| R\left(p_j, w^*(p_j)\right)$, for any price p_j in its epoch τ.

Lemma 4.4.2 *For any arm $j \in [J]$, suppose it is in its epoch τ which ends in round k. Then for any $x \in \{l, c, r\}$, we have*

$$\left|\sum_{t \in L_{j,\tau}} \left(R\left(p_j, w^*(p_j)\right) - \tilde{R}_x^t\right)\right| \leq \max\{8\bar{P}\bar{S}, 98\}\frac{(\bar{P}\bar{S})^2 \log T}{\gamma_k}.$$

Proof of Lemma 4.4.2. We analyze various scenarios based on different values of k.

1. If $k = 1$, we have

$$\left|\sum_{t \in L_{j,\tau}} \left(R\left(p_j, w^*(p_j)\right) - \tilde{R}_x^t\right)\right| \leq \frac{4(\bar{P}\bar{S})^2 \log T}{\gamma_1^2}\bar{P}\bar{S} = \frac{8\left(\bar{P}\bar{S}\right)^3 \log T}{\gamma_1}.$$

2. If $k \geq 2$, note that for any epoch τ of price p_j, we have

$$\left|\sum_{t \in L_{j,\tau}} \left(R\left(p_j, w^*(p_j)\right) - \tilde{R}_x^t\right)\right| \leq \underbrace{\sum_{t \in L_{j,\tau}} \left(R\left(p_j, w^*(p_j)\right) - R\left(p_j, w_x^t\right)\right)}_{A_1} + \underbrace{\left|\sum_{t \in L_{j,\tau}} \left(R\left(p_j, w_x^t\right) - \tilde{R}_x^t\right)\right|}_{A_2}.$$

Recall that $L_{j,\tau}$ is the set of time indices contained in price p_j's epoch τ. We know $|L_{j,\tau}| \leq \frac{4(\bar{P}\bar{S})^2 \log T}{\gamma_k^2}$ because epoch τ ends in round k. Therefore, we have

$$
\begin{aligned}
(A_1) = &\sum_{t \in L_{j,\tau}} \left(R\left(p_j, w^*(p_j)\right) - R\left(p_j, w_x^t\right)\right) \\
\leq &12\gamma_{k-1} \left|L_{j,\tau}\right| = 24\gamma_k \left|L_{j,\tau}\right| \leq \frac{96(\bar{P}\bar{S})^2 \log T}{\gamma_k},
\end{aligned}
$$

where the first inequality is due to the fact that for any price p_j, if epoch τ ends at round $k \geq 2$, then $R\left(p_j, w^*(p_j)\right) - R\left(p_j, w_x^t\right) \leq 12\gamma_{k-1},\ \forall t \in L_{j,\tau},\ x \in \{l, c, r\}$, which is Lemma 2 in Agarwal et al. (2011). Since we condition on event $\mathcal{E}$, we have

$$
(A_2) \leq |L_{j,\tau}| \bar{P}\bar{S} \sqrt{\frac{\log T}{|L_{j,\tau}|}} = \bar{P}\bar{S}\sqrt{\log T |L_{j,\tau}|} \leq \frac{2(\bar{P}\bar{S})^2 \log T}{\gamma_k},
$$

because $|L_{j,\tau}| \leq \dfrac{2\left(\bar{P}\bar{S}\right)^2 \log T}{\gamma_k}$. Thus, for $k \geq 2$, the total loss can be bounded by

$$
\left| \sum_{t \in L_{j,\tau}} \left(R\left(p_j, w^*(p_j)\right) - \tilde{R}^t\left(p_j, w_{j,x}\right)\right) \right| \leq \frac{98(\bar{P}\bar{S})^2 \log T}{\gamma_k}.
$$

Then we sum up both situations and the proof is completed. **Q.E.D.**

4.4.3.2 Bound on the Number of Epochs.

We bound the total number of epochs for any arm at any time t.

Lemma 4.4.3 *At any time t, for any arm $j \in [J]$, the total number of epochs*

$$
\tau_j^t \leq \frac{1}{2} \log_{\frac{4}{3}} \frac{\bar{P}^2 \max\{K_2^2, \bar{S}^2\} n_j^{t+1}}{9\bar{S}^2 \log T},
$$

where n_j^{t+1} is the number of times price p_j is selected by the end of time t.

Proof of Lemma 4.4.3. At any time t and for any $j \in [J]$, define $\gamma_j^{t\,\min} := 2\bar{P}\bar{S}\sqrt{\frac{\log T}{n_j^{t+1}}}$. Note that for any epoch τ of price p_j up to the end of time t, assuming epoch τ ends in round k, we have $\gamma_k \geq \gamma_j^{t\,\min}$ because otherwise price p_j will be selected for more than n_j^{t+1} times in total by the end of time t, which is a contradiction. Define interval $IN_j^t := \left[w^*(p_j) - \frac{\gamma_j^{t\,\min}}{\max\{K_2, \bar{S}\}\bar{P}}, w^*(p_j) + \frac{\gamma_j^{t\,\min}}{\max\{K_2, \bar{S}\}\bar{P}}\right]$. Then because $R(p, w)$ is $\max\{K_2, \bar{S}\}\bar{P}$-Lipschitz in w given p by Theorem 4.2.1a, for any $w \in IN_j^t$ we have

$$R\left(p_j, w^*(p_j)\right) - R\left(p_j, w\right) \leq \max\{K_2, \bar{S}\}\bar{P}\left|w - w^*(p_j)\right| \leq \gamma_j^{t\min}.$$

Suppose $\tau_j^t \geq 2$, suppose epoch $\tau_j^t - 1$ ends in round k. With a minor abuse of notation, we denote the interval during epoch τ of price p_j as $\left[l_j^\tau, r_j^\tau\right]$. Note that by Lemma 1 in Agarwal et al. (2011), for any price p_j, if epoch τ ends at round k, then the interval $\left[l_j^{\tau+1}, r_j^{\tau+1}\right]$ contains every $w \in \left[l_j^\tau, r_j^\tau\right]$ such that $R\left(p_j, w\right) \geq R\left(p_j, w^*(p_j)\right) - \gamma_k$. In particular, $w^*(p_j) \in \left[l_j^\tau, r_j^\tau\right]$ for all epochs τ. Consequently,

$$IN_j^t \subseteq \{w : R\left(p_j, w\right) \geq R\left(p_j, w^*(p_j)\right) - \gamma_k\} \subseteq \left[l_j^{\tau_j^t}, r_j^{\tau_j^t}\right],$$

which implies

$$\left|IN_j^t\right| = \frac{2\gamma_j^{t\min}}{\max\{K_2, \bar{S}\}\bar{P}} \leq \left|\left[l_j^{\tau_j^t}, r_j^{\tau_j^t}\right]\right| = \bar{P}\left(\frac{3}{4}\right)^{\tau_j^t - 1}.$$

Therefore, we have

$$\tau_j^t \leq \frac{1}{2}\log_{\frac{4}{3}} \frac{\bar{P}^2 \max\{K_2^2, \bar{S}^2\} n_j^{t+1}}{16\bar{S}^2 \log T} + 1 = \frac{1}{2}\log_{\frac{4}{3}} \frac{\bar{P}^2 \max\{K_2^2, \bar{S}^2\} n_j^{t+1}}{9\bar{S}^2 \log T}.$$

Note that if $\tau_j^t = 1$, the result holds trivially. **Q.E.D.**

4.4.3.3 Bound of the Regret for Fixed Arm.

With Lemmas 4.4.2 and 4.4.3 established, we now provide a bound on the total regret of any given arm at any time.

Lemma 4.4.4 *At any time t, for any price p_j, recall that H_j^t is the set of time indices by the end of time t when value p_j is selected for price, we have*

$$\left|R\left(p_j, w^*(p_j)\right) - \frac{1}{3n_j^{t+1}} \sum_{s \in H_j^t} \sum_{x \in \{l,c,r\}} \tilde{R}^s\left(p_j, w_x^s\right)\right| \leq \max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\frac{\log T}{n_j^{t+1}}} \log_{\frac{4}{3}} \frac{\bar{P}^2 \max\{K_2^2, \bar{S}^2\} n_j^{t+1}}{9\bar{S}^2 \log T}$$

$$= \mathbf{Rad}_j^{t+1}.$$

Proof of Lemma 4.4.4. Until time t, for any epoch τ of price p_j which ends in round k, we have

$$\left|\sum_{s\in L_{j,\tau}} R\left(p_j, w^*(p_j)\right) - \tilde{R}^s\left(p_j, w_x^s\right)\right| \le \max\{8\bar{P}\bar{S}, 98\}\frac{(\bar{P}\bar{S})^2\log T}{\gamma_k} \le \max\{8\bar{P}\bar{S}, 98\}\frac{(\bar{P}\bar{S})^2\log T}{\gamma_j^{t\,\min}}$$
$$= \max\{4\bar{P}\bar{S}, 49\}\bar{P}\bar{S}\sqrt{\log T n_j^{t+1}},$$

where the first inequality is by Lemma 4.4.2 and the second inequality is due to the fact that the number of times of selecting price p_j until time t should be no larger than n_j^{t+1}. In addition, by Lemma 4.4.3, we have a bound on the number of epochs of price p_j until time t, i.e.,

$$\tau_j^t \le \frac{1}{2}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}n_j^t}{9\bar{S}^2\log T}.$$

Therefore, we have

$$\left|\sum_{s\in H_j^t}\left(3R\left(p_j, w^*(p_j)\right) - \sum_{x\in\{l,c,r\}}\tilde{R}^s\left(p_j, w_x^s\right)\right)\right|$$
$$\le \max\{6\bar{P}\bar{S}, \frac{147}{2}\}\bar{P}\bar{S}\sqrt{\log T n_j^{t+1}}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}n_j^{t+1}}{9\bar{S}^2\log T}.$$

Then the proof is complete because $\left|H_j^t\right| = n_j^{t+1}$. **Q.E.D.**

Lemma 4.4.4 presents an important concentration inequality for any given arm at any given point in time, which provides an essential framework for constraining the overall regret. In particular, it handles the key challenge in designing the online algorithm for the remunerating newsvendor problem, which is the tradeoff between the loss incurred by the exploration of price (Part (4.5)) and the loss attributable to the suboptimality of remuneration for any price choice (Part (4.6)). These two parts of losses, incurred in the process of searching for the optimal price and the corresponding optimal remuneration, are intrinsically determined by the duration devoted to the exploration of a price's optimal revenue prior to the exploitation of this data in pursuit of better price options. This balancing act between exploration and exploitation drives the efficacy of the online algorithm designed for two decision variables and needs to be constructed carefully.

It is observed that the naive combination of Bisection Search and Bandit Control, which involves implementing a phase of Bisection Search (on remunerations) under a single price choice and then using the obtained revenue to estimate the performance of that price, leads to a total regret of order $O(T)$. Under this naive design, let us assume that the algorithm makes a total of K price selections, denoted by p^k for the k-th choice. While Part (4.6) can still be effectively bounded, Part (4.5) poses a problem. We can express Part (4.5) as follows:

$$3TR\left(p_{j^*}, w^*(p_{j^*})\right) - 3\sum_{k=1}^{K} N^k R\left(p^k, w^*(p^k)\right) = 3\sum_{k=1}^{K} N^k \left(R\left(p_{j^*}, w^*(p_{j^*})\right) - R\left(p^k, w^*(p^k)\right)\right),$$

where N^k denotes the total number of times for kth price choice. While $R\left(p_{j^*}, w^*(p_{j^*})\right) - R\left(p^k, w^*(p^k)\right)$ does not depend on N^k, the length of a round or epoch in the Bisection Search process with respect to w is determined by the specific shape (predominantly the first-order derivatives) of the function $R\left(p^k, w\right)$. Since we do not impose any assumption on the function's curvature, determining a conclusive bound for N^k is difficult. This renders Part (4.5) to be instance dependent. An extreme example is when the function $R\left(p^1, w\right), w \in [0, p^1]$ is flat enough such that $K = 1$ and $N^1 = T$, this naive design adopts a total loss of order $O(T)$.

To address such pathological circumstances, it is necessary to reduce the Bandit Selection Loss (4.5) by truncating the exploratory phase of suboptimal price choices compared to the naive design. To this end, we propose updating the price choice at each time instead of at the end of each epoch, and distributing the Bisection Search process for each price into different outer layers. This design allows the algorithm to spend less time on suboptimal prices and enables faster learning of the optimal remuneration corresponding to superior price choices. Specifically, Lemma 4.4.4 quantifies the concentration inequality for the estimation of each price choice at any time, supporting the concurrent price updates. Based on the design and the concentration result, the regret of the algorithm can be shown to be $\tilde{O}(T^{\frac{2}{3}})$. The details are specified below in §4.4.4.

4.4.4 Total Regret Bound

We are now ready to put everything together to obtain the total regret bound.

Proof of Theorem 4.4.1. First, we condition on the event $\mathcal{E}$ defined in (4.7). By Theorem 4.2.1b,

$$\left|R\left(p_1, w^*(p_1)\right) - R\left(p_2, w^*(p_2)\right)\right| \leq K_3 \left|p_1 - p_2\right|,$$

it can be seen that (4.4) $\leq 3T \cdot \dfrac{K_3 \bar{P}}{J}$.

For the loss from bandit selection, it can be shown

$$(4.5) \le 3\bar{P}\bar{S}J + 3\sum_{t=J+1}^{T}\left[\left(\frac{\alpha^t_{j^*}}{3n^t_{j^*}} + \mathbf{Rad}^t_{j^*}\right) - \left(\frac{\alpha^t_{j^t}}{3n^t_{j^t}} + \mathbf{Rad}^t_{j^t}\right) + \left(\frac{\alpha^t_{j^t}}{3n^t_{j^t}} + \mathbf{Rad}^t_{j^t}\right) - R\left(p_{j^t}, w^*(p_{j^t})\right)\right] \tag{4.8}$$

$$\le 3\bar{P}\bar{S}J + 3\sum_{t=J+1}^{T}\left(\left(\frac{\alpha^t_{j^t}}{3n^t_{j^t}} + \mathbf{Rad}^t_{j^t}\right) - R\left(p_{j^t}, w^*(p_{j^t})\right)\right) \tag{4.9}$$

$$\le 3\bar{P}\bar{S}J + 6\sum_{t=J+1}^{T}\mathbf{Rad}^t_{j^t} \tag{4.10}$$

$$\le 3\bar{P}\bar{S}J + 6\sum_{j=1}^{J}\sum_{i=1}^{n_j^{T+1}}\max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\frac{\log T}{i}}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}i}{9\bar{S}^2\log T}$$

$$\le 3\bar{P}\bar{S}J + 12\max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\log T}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}T}{9\bar{S}^2\log T}\sum_{j=1}^{J}\sqrt{n_j^{T+1}},$$

where (4.8) and (4.10) are by Lemma 4.4.4, and (4.9) is due to the selection rule in Algorithm 6.

For the loss caused by the bisection search process, it can be shown that

$$(4.6) \le \sum_{j=1}^{J}\mathbf{Rad}_j^{T+1}3n_j^{T+1} = \sum_{j=1}^{J}3\max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\log T n_j^{T+1}}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}n_j^{T+1}}{9\bar{S}^2\log T}$$

$$\le 3\max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\log T}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}T}{9\bar{S}^2\log T}\sum_{j=1}^{J}\sqrt{n_j^{T+1}},$$

where the first inequality is by applying Lemma 4.4.4 to any price p_j at the end of time T.

Note that $\sum_{j=1}^{J} n_j^{T+1} = T$, so we have $\sum_{j=1}^{J}\sqrt{n_j^{T+1}} \le \sqrt{JT}$ by Jensen's inequality. In sum, by Lemma 4.4.1,

$$\begin{aligned}\mathbf{Regret}^{3T}_{\mathbf{BBS}} =& 3TR(p^*, w^*) - \mathbb{E}\left[\sum_{s=1}^{3T}\breve{R}^s_{\mathbf{BBS}}\right] \\ =& \mathbb{E}\left[3TR(p^*, w^*) - \sum_{s=1}^{3T}\breve{R}^s_{\mathbf{BBS}} \mid \mathcal{E}\right]\mathbb{P}\left(\mathcal{E}\right) + \mathbb{E}\left[3TR(p^*, w^*) - \sum_{s=1}^{3T}\breve{R}^s_{\mathbf{BBS}} \mid \mathcal{E}^{\complement}\right]\mathbb{P}\left(\mathcal{E}^{\complement}\right) \\ \le& 15\max\{2\bar{P}\bar{S}, \frac{49}{2}\}\bar{P}\bar{S}\sqrt{\log T}\log_{\frac{4}{3}}\frac{\bar{P}^2\max\{K_2^2, \bar{S}^2\}T}{9\bar{S}^2\log T}\sqrt{JT} + 3\bar{S}\bar{P}J + 3K_3\bar{P}TJ^{-1} \\ & + \frac{6}{T}3T\bar{P}\bar{S}.\end{aligned}$$

Finally, when $J = T^{\frac{1}{3}}$, we have $\mathbf{Regret}^{3T}_{\mathbf{BBS}} = \tilde{O}(T^{\frac{2}{3}})$ and Theorem 4.4.1 is proved. **Q.E.D.**

We would like to highlight two challenges that our algorithm has overcome, which did not arise in the work of Agarwal et al. (2011) that focused on a single-dimensional bisection search. First, we significantly extend the existing methods to a framework that encompasses two decision variables (without having joint convexity). To tackle this problem, we integrate carefully designed Bandit Control techniques into the Bisection Search method. This combination leverages the strengths of both methodologies and enables an innovative approach to handle the complexities introduced by the inclusion of an additional decision variable and the absence of a joint convex structure. Second, we introduce a crucial modification to the querying process within Bisection Search, which distributes the search process over multiple outer loops and non-consecutive steps. This tailored modification plays a key role in ensuring that the regret incurred from the bisection search process, for any price selection at any given time, is bounded with high probability, as given in Lemma 4.4.4. Consequently, the learning algorithm can achieve a tight regret bound.

4.5 Regret Lower Bound

We derive the following lower bound for the regret of a T-period finite horizon joint pricing and remuneration decision problem, which implies that the proposed Algorithm 6 achieves the optimal regret bound up to a logarithmic factor for any admissible pricing strategy.

Theorem 4.5.1 *Under Assumption 4.2.1, for any admissible pricing strategy* **ALG***, there exist problem instances such that* $\mathbf{Regret}^T_{\mathbf{ALG}} = \Omega(T^{\frac{2}{3}})$*.*

To prove Theorem 4.5.1, we invoke the following result from Wang et al. (2021b) without proof.

Lemma 4.5.1 (Theorem 2 in Wang et al. (2021b)) *Fix any integer $k \geq 1$ and $p_{\min} = 1$, $p_{\max} = 2, C = 1$. There exists a constant $C_k > 0$ depending only on k, such that for any admissible dynamic pricing strategy π,*

$$\sup_{f \in \Sigma^k([p_{\min},\, p_{\max}];c)} \mathbb{E}^\pi \left[T \times \max_{p \in [p_{\min},\, p_{\max}]} pf(p) - \sum_{t=1}^T p_t f\left(p_t\right) \right] \geq C_k \times T^{(k+1)/(2k+1)},$$

where $\{p_t\}_{t=1}^T$ are the prices set by the pricing strategy π.

Proof of Theorem 4.5.1. Consider the case where $\mu(w) = \bar{S}$ (the upper bound for demand and supply) and $\varepsilon = \delta = 0$, satisfying Assumption 4.2.1. In this case, the problem simplifies to:

$$\max_{p^t,\, w^t} \sum_{t=1}^T (p^t - w^t)\lambda(p^t)$$

$$\text{s.t. } 0 \leq w^t \leq p^t, \ \forall t \in [T].$$

For any p, we have $\dfrac{\partial R(p,w)}{\partial w} = -\lambda(p) \leq 0$. Therefore, the optimal remuneration price $w^*(p) = 0$ for any p, and the problem simplifies to:

$$\begin{aligned} \max_{p^t} \ & \sum_{t=1}^{T} p^t \lambda(p^t) \\ \text{s.t. } & 0 \leq p^t \leq \bar{P}, \ \forall t \in [T]. \end{aligned} \tag{4.11}$$

Because $\lambda(\cdot)$ is Lipschitz continuous by Assumption 4.2.1b, the problem (4.11) corresponds to the dynamic pricing problem studied in Wang et al. (2021b) when $k = 1$. In particular, they constructed J revenue functions where instance j achieves its maximum in the interval $[1 + (j-1)/J, 1 + j/J]$. For any policy π, there exists an instance such that the regret is of order $\Omega\left(T^{(k+1)/(2k+1)}\right)$. By Lemma 4.5.1, we can conclude that Theorem 4.5.1 holds. **Q.E.D.**

4.6 Special Case: Linear Model

We now turn our attention to a specific instance of the remunerating newsvendor problem where the underlying demand and supply functions are presumed to have a linear relationship with price and remuneration, respectively. We improve the regret upper bound for this special case by imposing another online algorithm.

Specifically, with the assumption of additive noises for the demand $D(p) = \lambda(p) + \varepsilon$ and supply $S(w) = \mu(w) + \delta$, we further assume a linear relationship between demand (supply) and price (remuneration), which is common in operations management literature (Keskin and Zeevi 2014, Mills 1959, Petruzzi and Dada 1999, Taylor 1974).

Assumption 4.6.1 *The demand and supply are decreasing and increasing linear functions with respect to price and remuneration respectively, each with an additive zero-mean noise, i.e.,*

$$\begin{aligned} & \lambda(p) = c_1 - a_1 p, \quad \mu(w) = c_2 + a_2 q, \quad a_1, c_1, a_2, c_2 > 0, \\ & \mathbb{E}[\varepsilon] = 0, \quad \mathbb{E}[\delta] = 0. \end{aligned}$$

We term the difference between the price and remuneration as a "gap" defined as $\Delta := p - w$. With slight abuse of notation, we denote $R(\Delta, w) := \Delta \mathbb{E}[\min(\lambda(\Delta + w) + \varepsilon, \mu(w) + \delta)]$.

Once again, we posit that the likelihood of negative supply or demand can be disregarded, as discussed in §4.2. Then the problem can be written as:

$$\begin{aligned} \max_{\Delta,\, w}\ & R(\Delta, w) = \Delta \mathbb{E}\left[\min(c_1 - a_1(\Delta + w) + \varepsilon, c_2 + a_2 w + \delta)\right] \\ \text{s.t. } & \Delta, w \geq 0 \end{aligned}$$

Subsequently, we obtain the following structural results with proof provided in Appendix C.2.2.

Theorem 4.6.1 *In the case of complete information where the demand and supply functions, as well as the random noise distributions, are known a priori, the remunerating newsvendor problem has the following structural properties under Assumptions 4.2.1a and 4.6.1,*

(a) For any gap Δ, the expected revenue $R(\Delta, w)$ is concave and Lipschitz continuous in remuneration w. In particular, for any $w_1, w_2 \in [0, \bar{P}]$,

$$\left|R\left(\Delta, w_1\right) - R\left(\Delta, w_2\right)\right| \leq K_4 \left|w_1 - w_2\right|,$$

where $K_4 = \bar{P}\bar{a}$ and $\bar{a} = \max\{a_1, a_2\}$.

(b) The optimal expected revenue $R\left(\Delta, w^*(\Delta)\right)$ as a function of gap Δ is concave and Lipschitz continuous in Δ. In particular, for any $\Delta_1, \Delta_2 \in [0, \bar{P}]$,

$$\left|R\left(\Delta_1, w^*(\Delta_1)\right) - R\left(\Delta_2, w^*(\Delta_2)\right)\right| \leq K_5 \left|\Delta_1 - \Delta_2\right|,$$

where $K_5 = \max\left(c_1, \frac{2a_1a_2}{a_1+a_2}\bar{P} + \frac{2a_1+a_2}{a_1+a_2}\bar{S}, \bar{P} + \bar{S}\right)$.

4.6.1 Double Bisection Search Algorithm (DBS)

Given a context of incomplete information, where a_1, a_2, c_1, c_2 and the distributions of ε and δ are unknown, we propose an alternative online algorithm. This algorithm, called Double Bisection Search Algorithm (**DBS**), offers an enhanced upper bound for regret. The algorithm applies Bisection Search in both outer and inner layers and is detailed in Algorithm 7.

4.6.1.1 Main Ideas and High-Level Analysis.

The main idea of Algorithm 7 is a Bisection Search procedure implemented to optimize Δ, with the standard query process within each epoch supplanted by another **BS** routine (see

Algorithm 7 Double Bisection Search Algorithm (**DBS**)

```
Define $\gamma_i := \frac{1}{2^i}, i \geq 1$ and $\alpha(T) = \frac{1}{2}\log_{\frac{4}{3}} \frac{1}{16\bar{S}^2}\frac{T}{\log T}$. Initialize $l_1 = 0, r_1 = \bar{P}$.
Let $K_6 := \max(24K_4, 512)\bar{P}\bar{S}$.                                  ▷ Parameters
for epoch $\tau = 1, 2, \ldots$ do                                            ▷ Outer Bisection Search
    Let $u := r_\tau - l_\tau$ and $\Delta_l := l_\tau + u/4, \Delta_c := l_\tau + u/2, \Delta_r := l_\tau + 3u/4$.
    Initialize $m_x = 0$ and $q_x = \bar{P} - \Delta_x$ for $x \in \{l, c, r\}$.
    for round $i = 1, 2, \ldots$ do
        for $x \in \{l, c, r\}$ do
            Update $\hat{R}_x, m_x, q_x = \mathbf{BS}(\Delta_x, \frac{4K_6^2 \log T}{\gamma_i^2}, m_x, q_x)$.   ▷ Inner Bisection Search
        end for
        Let $\mathrm{LB}_{\gamma_i}(\Delta_x) = \hat{R}_x - \frac{\gamma_i}{2}\alpha(T)$ and $\mathrm{UB}_{\gamma_i}(\Delta_x) = \hat{R}_x + \frac{\gamma_i}{2}\alpha(T)$ for $x \in \{l, c, r\}$.
        if $\min\{\mathrm{UB}_{\gamma_i}(\Delta_l), \mathrm{UB}_{\gamma_i}(\Delta_r)\} \leq \max\{\mathrm{LB}_{\gamma_i}(\Delta_l), \mathrm{LB}_{\gamma_i}(\Delta_r)\} - \gamma_i\alpha(T)$ then
            if $\mathrm{UB}_{\gamma_i}(\Delta_l) \leq \mathrm{UB}_{\gamma_i}(\Delta_r)$ then let $l_{\tau+1} = \Delta_l$ and $r_{\tau+1} = r_\tau$.
            else let $l_{\tau+1} = l_\tau$ and $r_{\tau+1} = \Delta_r$.
            end if
            Continue to epoch $\tau + 1$.
        else if $\min\{\mathrm{UB}_{\gamma_i}(\Delta_l), \mathrm{UB}_{\gamma_i}(\Delta_r)\} \leq \mathrm{LB}_{\gamma_i}(\Delta_c) - \gamma_i\alpha(T)$ then
            if $\mathrm{UB}_{\gamma_i}(\Delta_l) \leq \mathrm{UB}_{\gamma_i}(\Delta_r)$ then $l_{\tau+1} = \Delta_l$ and $r_{\tau+1} = r_\tau$.
            else let $l_{\tau+1} = l_\tau$ and $r_{\tau+1} = \Delta_r$.
            end if
            Continue to epoch $\tau + 1$.
        end if
    end for
end for
```

Algorithm 8 **BS**(Δ, n, m, q)

```
Define $\gamma_i := \frac{1}{2^i}, i \geq 1$ and let $l_1 := m$ and $r_1 := q$.
for epoch $\tau = 1, 2, \ldots$ do
    Let $u := r_\tau - l_\tau$ and $w_l := l_\tau + u/4, w_c := l_\tau + u/2, w_r := l_\tau + 3u/4$. Initialize $S = 0$.
    for round $i = 1, 2, \ldots$ do
        $S_x = 0$ for $x \in \{l, c, r\}$.
        for $t = 1, \ldots, \frac{4\bar{P}^2\bar{S}^2}{\gamma_i^2} \log T$ do
            For $x \in \{l, c, r\}$, implement with $\Delta, w_x$ and obtain the realized revenue $\breve{R}_x^t$.
            Update $S = S + \breve{R}_x^t$ and $S_x = S_x + \breve{R}_x^t$.
        end for
        Let $\mathrm{LB}_{\gamma_i}(w_x) = \frac{S_x\gamma_i^2}{4\bar{P}^2\bar{S}^2 \log T} - \frac{\gamma_i}{2}$ and $\mathrm{UB}_{\gamma_i}(w_x) = \frac{S_x\gamma_i^2}{4\bar{P}^2\bar{S}^2 \log T} + \frac{\gamma_i}{2}$ for $x \in \{l, c, r\}$.
        if $\min\{\mathrm{UB}_{\gamma_i}(w_l), \mathrm{UB}_{\gamma_i}(w_r)\} \leq \max\{\mathrm{LB}_{\gamma_i}(w_l), \mathrm{LB}_{\gamma_i}(w_r)\} - \gamma_i$ then
            if $\mathrm{UB}_{\gamma_i}(w_l) \leq \mathrm{UB}_{\gamma_i}(w_r)$ then let $l_{\tau+1} = w_l$ and $r_{\tau+1} = r_\tau$.
            else let $l_{\tau+1} = l_\tau$ and $r_{\tau+1} = w_r$.
            end if
            Continue to epoch $\tau + 1$.
        else if $\min\{\mathrm{UB}_{\gamma_i}(w_l), \mathrm{UB}_{\gamma_i}(w_r)\} \leq \mathrm{LB}_{\gamma_i}(w_c) - \gamma_i$ then
            if $\mathrm{UB}_{\gamma_i}(w_l) \leq \mathrm{UB}_{\gamma_i}(w_r)$ then $l_{\tau+1} = w_l$ and $r_{\tau+1} = r_\tau$.
            else let $l_{\tau+1} = l_\tau$ and $r_{\tau+1} = w_r$.
            end if
            Continue to epoch $\tau + 1$.
        end if
    end for
end for
Output $\frac{S}{n}, l_\tau, r_\tau$.
```

Algorithm 8). The Bisection Search procedure focuses on remuneration w and is *consecutively* executed within each outer epoch. It is noteworthy that all inner **BS** procedures during the same epoch in Algorithm 7 should be consecutively executed, i.e., the ending interval of the **BS** procedure in the preceding round should be "memorized" to be used for the initial interval in the subsequent round. This is crucial because a naive implementation that invokes the **BS** subroutine independently in each round would result in an unbounded cumulative number of inner **BS** executions, and consequently yield a suboptimal total regret bound.

Lemma 4.6.1 establishes a concentration inequality (Boucheron et al. 2013) that bounds the deviation of the average revenue generated by the **BS** process from $R\left(\Delta, w^*(\Delta)\right)$. This guarantees the validity of the outer Bisection Search process over Δ. Consequently, the loss from the suboptimality of Δ (i.e., Part (4.12) in §4.6.2) can be bounded efficiently with high probability. The loss from the suboptimality of w and random noise (i.e., Part (4.13) in §4.6.2) can be bounded by Lemma 4.6.1 together with Proposition 4.6.2, which bounds the total number of **BS** cycles by a logarithmic factor.

4.6.1.2 Improved Regret Bounds.

We present the subsequent theorem that establishes the upper bound of the regret for Algorithm 7. The proof for this theorem is provided in §4.6.2.

Theorem 4.6.2 *Under Assumptions 4.2.1a and 4.6.1, if we choose* $\eta^t = \frac{1}{\sqrt{t}}$*, we have* $\mathbf{Regret}^T_{\mathbf{DBS}} = \tilde{O}(T^{\frac{1}{2}})$.

It's noteworthy that the marked improvement of the regret from $\tilde{O}(T^{\frac{2}{3}})$ for the general function case to $\tilde{O}(T^{\frac{1}{2}})$ for the linear case principally stems from the concavity result of $R\left(\Delta, w^*(\Delta)\right)$, which enables us to apply the Bisection Search Algorithm instead of Bandit Algorithm. Coupled with the succeeding proposition, we demonstrate that the proposed Algorithm 7 attains the optimal convergence rate in regret, up to logarithmic factors.

Proposition 4.6.1 (Theorem 1 in Keskin and Zeevi (2014)) *There exists a finite constant* c *such that* $\mathbf{Regret}^T_{\mathbf{ALG}} \geq c\sqrt{T}$ *for any policy* **ALG** *and time horizon* $T \geq 3$.

Notably, when $\mu(w) = \bar{S}$ and $\varepsilon = \delta = 0$, our problem transforms into a dynamic pricing problem with a linear underlying correlation between demand and price. This is the problem setting studied in Keskin and Zeevi (2014). Therefore, we omit the proof.

4.6.2 Proof of Theorem 4.6.2

The total loss under any sample path can be decomposed into two parts as below.

$$
\begin{aligned}
&TR\left(\Delta^*, w^*\right) - \sum_{t=1}^{T} \breve{R}^t_{\mathbf{DBS}} \\
=&TR\left(\Delta^*, w^*\right) - \sum_{t=1}^{T} R\left(\Delta^t, w^*\left(\Delta^t\right)\right) && \textbf{(Outer Bisection Search Loss)} \quad (4.12)\\
&+ \sum_{t=1}^{T} R\left(\Delta^t, w^*\left(\Delta^t\right)\right) - \sum_{t=1}^{T} \breve{R}^t_{\mathbf{DBS}}. && \textbf{(Inner Bisection Search Loss)} \quad (4.13)
\end{aligned}
$$

The component (4.13), which is bounded as per §4.6.2.1, signifies the loss resulting from the **BS** Subroutine 8. Then in §4.6.2.2, we analyze the regret bound for (4.12), which denotes the loss from the suboptimal choice of Δ that is optimized over the Bisection Search process. We summarize the total loss in §4.6.2.3. It is noteworthy that the rounds or epochs within both Bisection Search procedures are now consecutively executed without the reuse of past rounds' realizations, in order to directly incorporate the existing outcomes of the Bisection Search and enhance the conciseness of the proof.

4.6.2.1 Inner Bisection Loss.

It is noted that the inner **BS** in Algorithm 8 is a standard quartering search procedure with starting interval $[m, q]$ and horizon length n. Consequently, we can apply similar proof with different Lipschitz constants and Hoeffding's inequality for bounded variables to the proof of Theorem 1 in Agarwal et al. (2011) and obtain the following lemma for the **BS** procedure.

Lemma 4.6.1 *For any Δ, n, m, q, if we run Algorithm 8 $\mathbf{BS}(\Delta, n, m, q)$, then with probability at least $1 - \frac{2}{T^2}$, we have*

$$
\begin{aligned}
\left| R\left(\Delta, w^*(\Delta)\right) - \frac{1}{n}\sum_{i=1}^{n} \tilde{R}^i \right| \leq & \max(12K_4(q-m), 216)\bar{P}\bar{S}\sqrt{\frac{\log T}{n}} \log_{\frac{4}{3}} \frac{\left(q-m\right)^2}{16(\bar{P}\bar{S})^2}\frac{n}{\log T} \\
\leq & K_6\sqrt{\frac{\log T}{n}}\alpha(T),
\end{aligned}
$$

where $K_4 = \bar{P}\bar{a}$, $K_6 = \max(24K_4\bar{P}, 432)\bar{P}\bar{S}$, and $\alpha(T) = \frac{1}{2}\log_{\frac{4}{3}} \frac{1}{16(\bar{S})^2}\frac{T}{\log T}$ as in Algorithm 7.

We denote the number of periods in epoch τ in Algorithm 7 as n_τ and the total number of epochs is N. For any epoch τ in Algorithm 7, we define the inspected gap values as $\Delta_{\tau,x}, x \in \{l, c, r\}$. Define $\hat{R}^i_{\tau,x}$ as the output from **BS** subroutine for estimation of $\Delta_{\tau,x}$ in round i. With the i-th realized revenue using gap $\Delta_{\tau,x}$ during epoch τ denoted as $\tilde{R}^i_{\tau,x}$ for $i \in [n_\tau]$, we define the following events where $B^i_{\tau,x}$ denotes the event that the loss incurred in round i of epoch τ with respect to $R\left(\Delta_{\tau,x}, w^*\left(\Delta_{\tau,x}\right)\right)$ is small enough and $C_{\tau,x}$ denotes that the estimation for $R\left(\Delta_{\tau,x}, w^*\left(\Delta_{\tau,x}\right)\right)$ from epoch τ is accurate enough. That is,

$$\begin{aligned}
B^i_{\tau,x} &= \left\{ \left| \hat{R}^i_{\tau,x} - R\left(\Delta_{\tau,x}, w^*(\Delta_{\tau,x})\right) \right| \leq \frac{\gamma_i}{2}\alpha(T) \right\}. \\
C_{\tau,x} &= \left\{ \left| \frac{1}{n_\tau}\sum_{i=1}^{n_\tau} \tilde{R}^i_{\tau,x} - R\left(\Delta_{\tau,x}, w^*(\Delta_{\tau,x})\right) \right| \leq K_6\sqrt{\frac{\log T}{n_\tau}}\alpha(T) \right\}.
\end{aligned}$$

By Lemma 4.6.1 with $\Delta = \Delta_{\tau,x}$ and $n = \frac{4K_6^2 \log T}{\gamma_i^2}$, for any $x \in \{l, c, r\}$ and any round i in any epoch τ, we have $\mathbb{P}\left(B^{i\complement}_{\tau,x}\right) \leq \frac{2}{T^2}$. In addition, note that for each epoch τ, the **BS** procedure is implemented consecutively and consequently is itself a **BS** procedure of length n_τ. By Lemma 4.6.1 with $\Delta = \Delta_{\tau,x}$ and $n = n_\tau$, we have $\mathbb{P}\left(C^{\complement}_{\tau,x}\right) \leq \frac{2}{T^2}$ for any $\tau \in [N]$ and $x \in \{l, c, r\}$.

Following this, for the events $B := \cap_{\tau,x,i} B^i_{\tau,x}$ and $C := \cap_{\tau,x} C_{\tau,x}$, through the application of union bounds, we deduce the following

$$\mathbb{P}(B) \geq 1 - \frac{6}{T}, \quad \mathbb{P}(C) \geq 1 - \frac{6}{T}. \tag{4.14}$$

Prior to analyzing the cumulative loss from all iterations of **BS**, we introduce the subsequent lemma concerning the total number of epochs, that is, the number of a complete cycle of inner **BS**. The proof follows a similar logic to the proof of Lemma 4 in Agarwal et al. (2011) with the differences lying in the Lipschitz constant and the number of queries in each round. We thus omit the proof of this lemma.

Lemma 4.6.2 *Conditional on event B, the total number of epochs N in the outer Bisection Search performed by Algorithm 7 is bounded as* $N \leq \frac{1}{2}\log_{\frac{4}{3}}\left(\frac{K_5^2 \bar{P}^2 T}{16 K_6^2 \log T}\right)$.

Conditional on event $B \cap C$, we have

$$\begin{aligned}
(4.13) =& \sum_{t=1}^{T} R\left(\Delta^t, w^*\left(\Delta^t\right)\right) - \sum_{t=1}^{T} \breve{R}^t_{\mathbf{DBS}} \\
=& \sum_{\tau=1}^{N} \sum_{x \in \{l,c,r\}} R\left(\Delta_{\tau,x}, w^*(\Delta_{\tau,x})\right) - \frac{1}{n_\tau}\sum_{i=1}^{n_\tau} \tilde{R}^i_{\tau,x}\left(\Delta_{\tau,x}, w^i_\tau\right) \\
\leq& \sum_{\tau=1}^{N} K_6\sqrt{n_\tau \log T}\alpha(T) && (4.15) \\
\leq& K_6\sqrt{TN\log T}\alpha(T) && (4.16)
\end{aligned}$$

$$\leq K_6\sqrt{T\log T}\alpha(T)\frac{1}{2}\log_{\frac{4}{3}}\left(\frac{K_5^2\bar{P}^2T}{16K_6^2\log T}\right), \tag{4.17}$$

where (4.15) is by conditioning on event C; (4.16) is by Jensen's inequality and (4.17) is by Lemma 4.6.2.

4.6.2.2 Outer Bisection Search Loss.

In the context of Theorem 4.6.1b, the concavity and Lipschitz continuity of $R(\Delta, w^*(\Delta))$ allow us to establish the following proposition concerning the Bisection Search Procedure. As the proof mirrors that of the proof of standard bisection search with the confidence intervals multiplied by $\alpha(T)$, it is omitted for brevity.

Proposition 4.6.2 *Conditional on event B, the total regret resulting from the outer Bisection Search can be expressed as*

$$\begin{aligned}(4.12) =&TR\left(\Delta^*, w^*\right) - \sum_{t=1}^{T} R\left(\Delta^t, w^*\left(\Delta^t\right)\right)\\ \leq&\max\left\{6K_5\bar{P}, 108\alpha(T)\right\}K_6\sqrt{T\log T}\log_{\frac{4}{3}}\left(\frac{K_5^2\bar{P}^2T}{16K_6^2\log T}\right).\end{aligned}$$

4.6.2.3 Total Loss.

Proof of Theorem 4.6.2. Given the above intermediate results, we can then bound the total regret.

$$\begin{aligned}&\mathbf{Regret}^T_{\mathbf{DBS}}\\ =&\mathbb{E}\left[TR\left(\Delta^*, w^*\right) - \sum_{t=1}^{T} R\left(\Delta^t, w^*\left(\Delta^t\right)\right) + \sum_{t=1}^{T} R\left(\Delta^t, w^*\left(\Delta^t\right)\right) - \sum_{t=1}^{T}\breve{R}^t_{\mathbf{DBS}} \mid B\cap C\right]\mathbb{P}\left(B\cap C\right)\\ &+\mathbb{E}\left[TR\left(\Delta^*, w^*\right) - \sum_{t=1}^{T}\breve{R}^t_{\mathbf{DBS}} \mid B^{\mathsf{C}}\cup C^{\mathsf{C}}\right]\mathbb{P}\left(B^{\mathsf{C}}\cup C^{\mathsf{C}}\right)\\ \leq&K_6\sqrt{T\log T}\frac{1}{2}\alpha(T)\log_{\frac{4}{3}}\left(\frac{K_5^2\bar{P}^2T}{16K_6^2\log T}\right) + \max\left\{6K_5\bar{P}, 108\alpha(T)\right\}K_6\sqrt{T\log T}\log_{\frac{4}{3}}\left(\frac{K_5^2\bar{P}^2T}{16K_6^2\log T}\right)\\ &+\bar{P}\bar{S}T\frac{12}{T}\\ =&\tilde{O}\left(\sqrt{T}\right),\end{aligned}$$

where the inequality is by results (4.17) and Proposition 4.6.2. **Q.E.D.**

From a methodological standpoint, while the Bisection Search technique proposed by Agarwal et al. (2011) was tailored for convex optimization over a single decision variable, our approach introduces two key innovations. First, our proposed algorithm is specifically designed to incorporate two decision variables through a two-layered application of the technique, while maintaining a tight upper bound for the total regret. Second, we emphasize the importance of maintaining a continuous inner bisection search procedure within each outer epoch. This sequential design ensures that the starting interval of the subsequent round aligns with the ending interval of the current round. Such a structural design is essential for effectively bounding the total number of inner bisection searches by the total number of epochs in the outer layer. Without adhering to this sequential design, if the inner Bisection Search were to be independently invoked during each iteration, the total regret would fail to achieve the optimal rate.

4.7 Numerical Experiments

We conduct simulation-based numerical experiments on the proposed algorithms in this section.

4.7.1 General Case

We consider a remunerating newsvendor problem with a finite horizon defined by $T = 1000$ and a lower price bound of $\underline{P} = 0$. We implement Algorithm 6 for the 6 instances in Table 4.1, where $\mathbf{N}[\mu, \sigma]$ represents a truncated normal distribution with mean μ, standard deviation σ and a bounded support $[-10, 10]$, while $\mathbf{U}[a, b]$ represents a uniform distribution with a and b as the lower and upper bounds, respectively.

Instance	$\lambda(p)$	$\mu(w)$	ε	δ	$\bar{P}$	$\bar{S}$
1	$110 - p$	$10 + w$	$\mathbf{N}[0, 5]$	$\mathbf{N}[0, 5]$	100	120
2	$110 - p$	$10 + w$	$\mathbf{N}[0, 10]$	$\mathbf{N}[0, 10]$	100	120
3	$110 - p$	$10 + w$	$\mathbf{U}[-5, 5]$	$\mathbf{U}[-5, 5]$	100	120
4	$110 - p^2$	$-w^2 + 20w + 10$	$\mathbf{N}[0, 5]$	$\mathbf{N}[0, 5]$	10	120
5	$110 - p^2$	$-w^2 + 20w + 10$	$\mathbf{N}[0, 10]$	$\mathbf{N}[0, 10]$	10	120
6	$110 - p^2$	$-w^2 + 20w + 10$	$\mathbf{U}[-5, 5]$	$\mathbf{U}[-5, 5]$	10	120

Table 4.1: Experiment Parameters

We assess the performance of the model under different scenarios, involving linear and quadratic functions, and normal and uniform noise distributions with varying degrees of variance. We denote the policy that uses the clairvoyant optimal price p^* and remuneration w^* for each period as π^*. The performance measure we use is the average relative regret, which is the average ratio of the cumulative difference between the revenue of implementing

π^* and the revenue generated by any algorithm **ALG**, i.e.,

$$\textbf{Relative Regret}_{\textbf{ALG}} := \frac{\sum_{s=1}^{3T} \breve{R}_{\pi^*}^s - \sum_{s=1}^{3T} \breve{R}_{\textbf{ALG}}^s}{\sum_{s=1}^{3T} \breve{R}_{\pi^*}^s}. \tag{4.18}$$

For each instance, we conduct 1000 iterations and compute the average relative regret for Algorithm 6. Figures 4.3–4.8 display the relative regret for instances 1 through 6, respectively. The algorithm converges robustly under various settings.

4.7.2 Linear Case

We also implement Algorithm 7 for the linear case and compare its performance with another type of algorithm which follows the "Explore Then Commit" framework as shown in Algorithm 9. We implement **ECO**-n for $n = 10, 25, 50$ on Instance $1, 2, 3$ respectively, each with $N = 100$, and conducted 100 iterations.

The average relative ratios of the algorithms as defined in (4.18) with $T = 1000$ are in Table 4.2. We can see that **DBS** outperforms **ECO**. Moreover, because **ECO** relies on the quality of the samples in the exploration phase, the performance of **DBS** is more stable as the comparison of the standard deviation of the relative ratios in Table 4.3 indicates.

Algorithm 9 Explore Then Commit using OLS Estimators (**ECO**-n)

```
  Denote set $S := \left\{0, \frac{\bar{P}}{N}, \dots, \bar{P}\right\}$. Set $p_1, p_2 \in [0, \bar{P}]$ and $w_1 \in [0, p_1], w_2 \in [0, p_2]$ arbitrarily.
  for $t = 1, \dots, 2n$ do                                                ▷ Explore
      If $t \leq n$, then $p^t = p_1$ and $w^t = w_1$; otherwise $p^t = p_2$ and $w^t = w_2$.
      Implement price $p^t$ and remuneration $w^t$. Observe the demand $d^t$ and supply $s^t$.
  end for
  Obtain the ordinary least squares (OLS) estimator $\hat{a}_1, \hat{c}_1$ using samples $\{d^t\}_{t=1}^{2n}$ and $\hat{a}_2, \hat{c}_2$
  using samples $\{s^t\}_{t=1}^{2n}$.
  Estimate the random noises by $\hat{\varepsilon}^t = d^t - \hat{c}_1 + \hat{a}_1 p^t$ and $\hat{\delta}^t = s^t - \hat{c}_2 - \hat{a}_2 w^t$ for $t = 1, \dots, 2n$.
  Calculate the cumulative revenue for each $(p, w)$ pair with $p, w \in S$ and $p \geq w$ using
  estimated parameters $\hat{a}_1, \hat{c}_1, \hat{a}_2, \hat{c}_2$ and random noise set $\{\hat{\varepsilon}^t\}_{t=1}^{2n}$ and $\{\hat{\delta}^t\}_{t=1}^{2n}$.
  Denote the pair with the largest calculated revenue as $(p_{ECO}^*, w_{ECO}^*)$.
  for $t = 2n + 1, \dots, T$ do                                            ▷ Commit
      Implement with price $p_{ECO}^*$ and remuneration $w_{ECO}^*$.
  end for
```

Instance	**ECO**-10	**ECO**-25	**ECO**-50	**DBS**
1	0.091	0.092	0.061	0.034
2	0.045	0.037	0.039	0.037
3	0.058	0.049	0.045	0.037

Table 4.2: Average Relative Regret at $t = 1000$

Instance	**ECO**-10	**ECO**-25	**ECO**-50	**DBS**
1	0.183	0.17	0.119	0.003
2	0.115	0.071	0.064	0.002
3	0.145	0.112	0.092	0.002

Table 4.3: Standard Deviation at $t = 1000$

4.8 Conclusion

We have introduced the "remunerating newsvendor" model, which extends the classical price-setting newsvendor model by incorporating remuneration decisions in a two-sided market. We have analyzed the optimal pricing and remuneration policy for fully-informed scenarios and proposed an online algorithm for situations where the demand and supply functions, relative to price and remuneration, as well as the noise distributions, are unknown. The algorithm achieves a provably tight regret bound. Furthermore, we improved the regret bound through an additional online algorithm, specifically for cases presenting a linear relationship between expected demand (or supply) and price (or remuneration).

Finally, we identify three potential avenues for future research. First, while demand models in much of the existing literature can be categorized as either additive or multiplicative, our approach handles the random noise additively. One possible extension would be to study the problem in a multiplicative or hybrid demand model (e.g., Chen and Simchi-Levi (2004a)). Second, we consider the scenario of lost sales and no carry-over inventory. It is natural to wonder how to solve the problem in the backlogged setting with excess inventory carried over to the next period. Third, while our model presumes that demand (or supply) is independent of remuneration (or price), it might be intriguing to consider cross-sided effects, thus capturing a broader array of characteristics inherent in real-world two-sided markets.

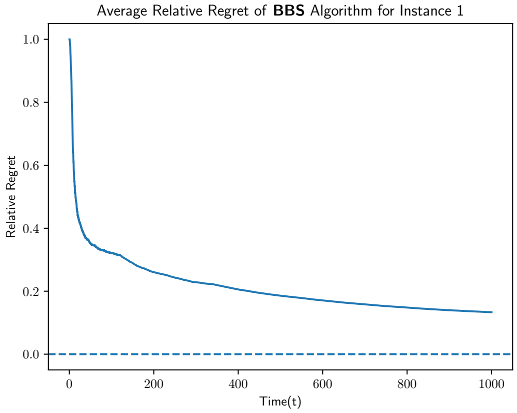


Figure 4.3: Computational Performance for Instance 1

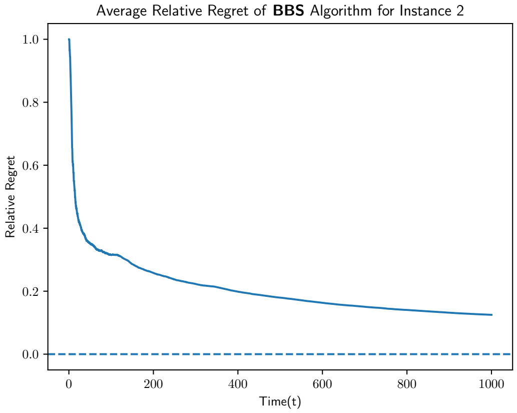


Figure 4.4: Computational Performance for Instance 2

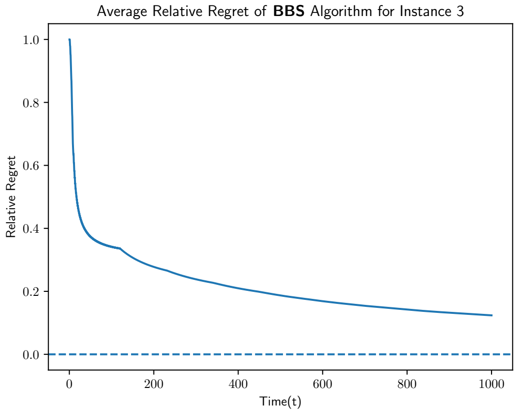


Figure 4.5: Computational Performance for Instance 3

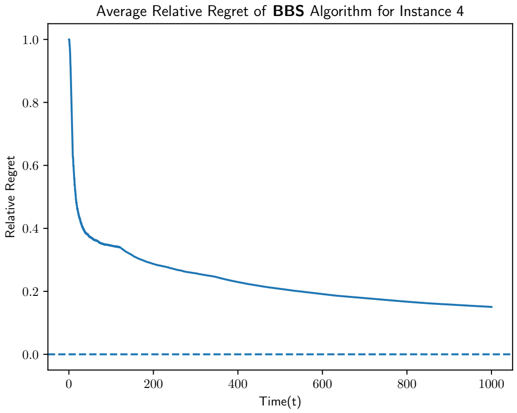


Figure 4.6: Computational Performance for Instance 4

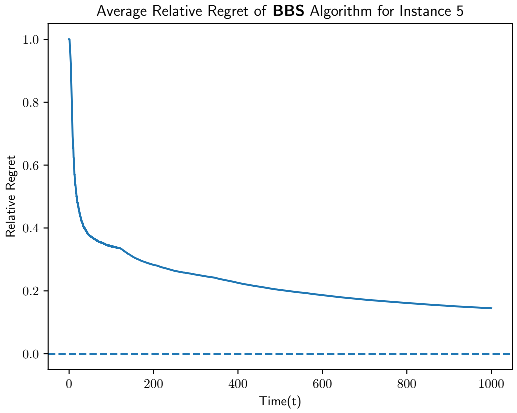


Figure 4.7: Computational Performance for Instance 5

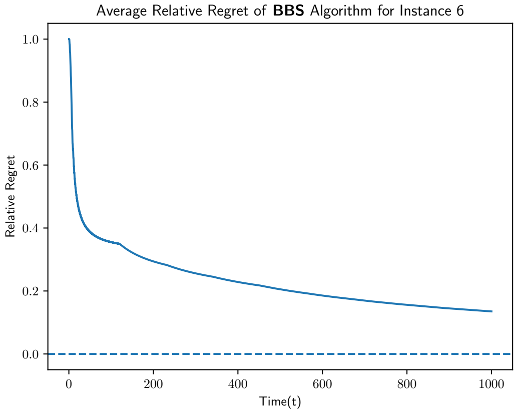


Figure 4.8: Computational Performance for Instance 6

CHAPTER 5

Offline Learning in Feature-Based Pricing

Previous chapters focus on the online learning algorithms in OM, highlighting their reliance on continuous online exploration, which may not always be practical due to operational constraints or regulatory considerations. For instance, contractual agreements with suppliers may prevent retailers from adjusting order quantities frequently while e-commerce platforms might take fairness in customer transactions into consideration when setting product prices. Furthermore, The implementation of online learning algorithms involving data collection, retrieval, pre-processing, and calculation in real time poses requirements for computational resources. Finally, in some large organizations, the procedural requisites for data access can impede timely algorithm implementation.

Given these considerations and the availability of extensive historical data, this chapter shifts focus to the exploration of offline learning algorithms. Specifically, it particularly addresses one critical issue firms face: setting prices based on historical data based on a range of features. This chapter presents an in-depth analysis of feature-based pricing with offline censored demand data, introducing an offline learning algorithm that is supported by both theoretical performance guarantees and empirical validation. In addition to unraveling the intricacies of feature-based pricing, this study establishes the first framework for tackling optimization challenges through the lens of causal inference.

5.1 Introduction

In today's data-rich business environment, the increased availability of customer data has fueled interest in feature-based pricing strategies. Firms (especially big techs) combine these covariates information with machine learning and optimization tools to predict a customer's willingness to pay and offer an attractive feature-based price (Elmachtoub et al. 2021).

For any price optimization model, understanding the demand function is the first and arguably the most critical component. In reality, this task is often very challenging especially when the demand function D is a function not just of price P offered but also of customer/product covariates X. Thanks to the rapid advances in information technology, firms have been collecting petabytes of historical data from past selling seasons, which is commonly referred to as *offline data* (Bu et al. 2023). A central task is to leverage the offline data to extract the demand information and ultimately make near-optimal *data-driven* pricing decisions.

A major challenge lies in that typical offline data only contains historical sales, which suffers from a well-known phenomenon called *demand censoring* (Huh and Rusmevichientong 2009). This is because when customers face stock-out, they will leave the system without any purchase records. Hence, the sales quantity is the minimum of the true demand and the available inventory. That is, the lost sales quantity is censored and unobservable. If demand censoring is not carefully factored in the design of offline policy learning algorithms, it could potentially lead to biased and inconsistent demand estimation and, consequently, suboptimal pricing decisions. Note that demand censoring is not limited to brick-and-mortar settings; it also occurs on online platforms. On these platforms, when a product is out of stock, the product page often displays an "out-of-stock" sign. As a result, interested customers may simply walk away without making any clicks or further engagement.

5.1.1 Brief Problem Statement and Motivating Applications

We consider an *offline learning* problem in which a firm with a finite amount of inventory aims to find the optimal feature-based pricing strategy based on customer/product covariates information. Note that in our setup, any unsatisfied demand that exceeds the inventory level is lost and unobservable. In particular, the firm does not know the demand function D but has access to an offline dataset consisting of quadruplets of historical covariates X, inventory Y, price P, and potentially censored sales quantity $S = \min\{D, Y\}$. We assume that all confounders are measured in the data. Our goal is to find the optimal feature-based pricing strategy π that maps any given covariate vector and available inventory (X, Y) into price P, so as to maximize the potential profit.

Online retailing is one key application area of our developed approach, including Amazon, Walmart, eBay, and Etsy, to assist sellers in pricing their products. For instance, Wayfair is an American e-commerce company that offers a wide range of home goods and furniture products, including sofas, whose pricing is affected by features such as the type of material used, design complexity, size, brand, and additional features like built-in recliners or pull-out beds. With the use of historical data, our approach can effectively price these differentiated products.

It can also be applied to online flash sales websites like Gilt, Rue La La, Belle & Clive, and

HauteLook. These websites receive luxury goods from various brands periodically and need to set prices for each item during a short time frame. Our model, which takes into account historical offline datasets and the highly differentiated products sold, can be applied to these scenarios. Moreover, the algorithmic approach may be useful for traditional markets, such as high-end art and premium wine, requiring pricing based on product features.

Our developed approach could also be applied to the hospitality industry. Hotel products have several features that differentiate them from each other, including their location, type of accommodation and amenities, reputation, seasonality, brand, and special events. The size and type of rooms, the amenities provided, as well as the time of year and special events in the area, can affect pricing. Our algorithm can assist in pricing these differentiated products, given the large amount of offline data that is available for analysis.

As highlighted in the above examples, our model primarily uses product covariates information. However, our framework is flexible enough to accommodate customer covariates as well. Nonetheless, this requires more scrutiny due to legal implications. According to Ban and Keskin (2021), price discrimination based on customer characteristics is a well-established legal practice, unless it involves "suspect categories" such as race or religion, or violates antitrust or price-fixing laws. Insurance companies, for instance, legally quote prices based on customers' credit, marriage status, and annual income, among other factors. Similarly, customized pricing strategies are widely adopted in various client-oriented industries, such as advertising and consulting. E-commerce giants like Amazon and Walmart also use customized pricing, for instance, by offering digital coupons and membership/student discounts, which are commonly used in online economics. We refer interested readers to Cohen et al. (2020), Miao et al. (2022) and Chen and Gallego (2021) for more application areas of feature-based pricing strategies, including our own.

5.1.2 Main Results and Contributions

We propose a novel data-driven offline learning algorithm that gives the optimal feature-based pricing strategy based on customer/product covariates under demand censoring.

Our key contribution is two-fold.

(a) **Modeling.** To the best of our knowledge, we are the first to model this feature-based pricing problem under censored demand through the lens of causal inference. We model the relationship between demand and price under the celebrated potential outcome framework (Rubin 1974). This framework gives natural identification results on the effect of price on demand, which makes it amenable for offline learning. A novel aspect of our model is to factor in demand censoring. In order to estimate the profit function, we propose to borrow the tool from survival analysis to recover the expected true (conditional) demand. We also propose a doubly robust estimation procedure to further achieve the robustness of our estimation result (Bang and Robins 2005). Specifically, we leverage state-of-the-art supervised learning techniques in estimating the potential profit function and the propensity scores (Rosenbaum and Rubin 1983) as well as in optimizing the feature-based prices. Compared with most existing approaches using parametric models in the literature of profit management and pricing, all the aforementioned components are modeled non-parametrically, thus more robust to model mis-specification.

(b) **Performance.** Our proposed algorithm is backed up by theoretical and empirical evidence. Theoretically, we provide a finite sample regret analysis of our offline learning algorithm showing that the expected profit of the estimated pricing strategy converges to the profit under the optimal pricing strategy asymptotically as the sample size of the offline data increases. Empirically, we conduct thorough numerical experiments to demonstrate that our proposed algorithm performs robustly well in estimating the optimal prices on both training and testing datasets. We also demonstrate the value of factoring in demand censoring in decision-making. Figure 5.1 shows that if one "mistakenly" used the sales as the uncensored demands, the resulting prices (represented by the pink line) would be drastically lower than the theoretical optimal prices (represented by the orange line), leading to profit degradation by up to 5%. Note that 5% improvement is significant on an e-commerce company's bottom line (Columbus 2020).

5.1.3 Literature Review

Our work is related to the following streams of literature.

Offline Learning for Pricing and Inventory Models. There has been increasing attention and interest in developing effective offline learning strategies for pricing and inventory models, thanks to the massive amount of historical data on customer and/or product information. This differs from online learning in that the entire dataset is available before the algorithm starts. Offline learning is especially useful when conducting online exploration can sometimes be very expensive or infeasible. Levi et al. (2007a) studied both single-period and multi-period inventory problems. They proposed sample average approximation (SAA) algorithms by approximating the true demand distribution with an empirical distribution, and developed sample complexity bounds. Levi et al. (2015) leveraged the notion of weighted mean spread to further improve this sample complexity bound for the newsvendor model. Cheung and Simchi-Levi (2019) gave both sample complexity upper and lower bounds for the

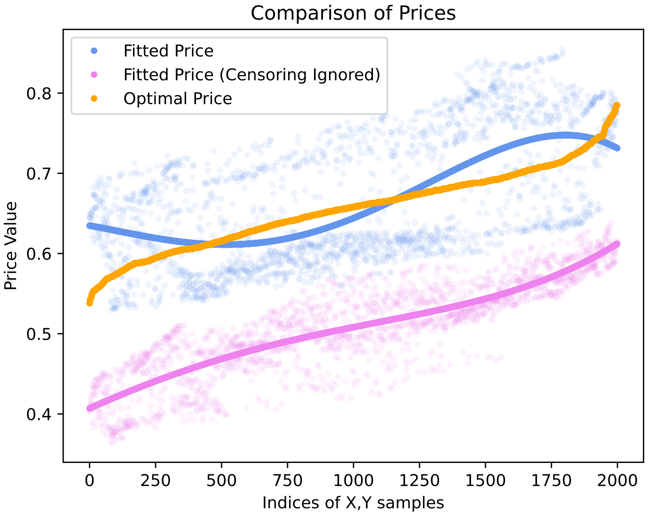


Figure 5.1: Theoretical optimal prices, our recommended prices, and prices without factoring in demand censoring

capacitated model. Qin et al. (2022) derived a sample complexity bound for the joint pricing and inventory control model. Ban and Rudin (2019) applied machine learning algorithms to the data-driven newsvendor with feature information. However, all aforementioned studies assume that the historical demand samples can be fully observed, whereas our approach considers demand censoring (i.e., only historical sales samples are available).

To the best of our knowledge, fewer than a handful of papers in the literature considered demand censoring for the offline learning setting. Ban (2020) studied a multi-period inventory system with fixed costs under censored demand, and developed a nonparametric estimation procedure for the (s, S) policy which is consistent and asymptotically normal. More closely related, Bu et al. (2023) studied a single product pricing problem under censored demand, and developed a necessary and sufficient condition for problem identifiability by relating to distributionally robust optimization (DRO) problems. They also proposed a data-driven algorithm that hedges against the distributional uncertainty arising from censored data, with provable finite-sample performance guarantees regardless of problem identifiability and offline data quality. There are two major distinctions between our work and their work. First, our work takes a completely different approach based on a canonical causal inference framework. Second, our work prescribes an optimal feature-based pricing strategy based on customer/product covariates, which is precisely consistent with the key research vision proposed in Feng and Shanthikumar (2018a) (in terms of how to leverage Big Data in production and operations management research). There also has been a growing interest in developing offline reinforcement learning approaches (Bu et al. 2022, Foster et al. 2021, Simchi-Levi and Xu 2022), where offline data is leveraged to learn a decision-making policy.

Online Learning for Pricing and Inventory Models. Most existing literature focuses on online learning for pricing and inventory models, where sales data are generated by sequential actions on the fly. There have been studies focusing on the repeated newsvendor problem with censored demand (Besbes and Muharremoglu 2013, Huh et al. 2011, Huh and Rusmevichientong 2009, Lugosi et al. 2017). Subsequently, more studies have been devoted to more involved systems with censored demand, e.g., the lost sales problem with positive lead times (Agrawal and Jia 2022, Huh et al. 2009, Zhang et al. 2020), perishable inventory control (Zhang et al. 2018), inventory control with stock substitutions (Chen and Chao 2020b), inventory control with fixed costs (Yuan et al. 2021), dynamic pricing with high dimensional features (Wang et al. 2020), joint pricing and inventory control (Chen et al. 2019a, 2021a, 2020a, 2022a). In contrast, our chapter focuses on offline learning with an available (and potentially massive) dataset. Our proposed framework is particularly useful when there has already been a plethora of historical data in the firm and conducting online experimentation could be very expensive both in terms of costs and time commitment (to actively explore). A well-performed pricing strategy learned from the historical data can nevertheless be an initial policy for promoting efficient online learning. The major challenge is that one needs to factor in demand censoring which is always inherent in the underlying dataset.

Feature-Based Pricing Strategies. There has been a huge body of literature on joint learning and pricing strategies (Den Boer 2015). The increased availability of customer/product information has led to advances in feature-based pricing that have clear advantages over traditional static pricing (Elmachtoub et al. 2021). Here we only discuss a list of papers (by no means exhaustive) that focus on feature-based pricing strategies. Cohen et al. (2020) proposed an ellipsoid method that admits a worst-case regret which is quadratic in the dimension of the feature space and logarithmic in the time horizon. Chen and Gallego (2021) gave an online learning algorithm based on adaptively splitting the covariate space into smaller bins and learning the optimal decision in each bin. Miao et al. (2022) considered a context-based dynamic pricing problem of online products which have low sales and gave an online clustering algorithm. Javanmard and Nazerzadeh (2019) considered a dynamic pricing problem with a binary choice model, and constructed a near-optimal policy in their setting. Xu and Wang (2021) proposed two algorithms for stochastic and adversarial feature settings respectively with logarithmic regret bounds. Fan et al. (2022) extended Javanmard and Nazerzadeh (2019) to a semiparametric demand model. Qiang and Bayati (2016) proposed a greedy iterative least squares approach that admits a logarithmic regret. Ban and Keskin (2021) also proposed an iterative least squares approach for a refined model that incorporates feature-dependent price sensitivity. Nambiar et al. (2019) proposed a "random price shock" algorithm that dynamically generates randomized price shocks to estimate price elasticity, and showed that this approach is robust to model mis-specification. Luo et al. (2021) studied the contextual dynamic pricing problem with unknown random noise in the valuation model. Wang et al. (2021a) proposed a simple pricing algorithm for the dynamic pricing problem with very few assumptions on the covariates. The above papers all assumed various parametric forms of demand functions (mostly linear or generalized linear functions) whereas we take a nonparametric approach based on a causal inference framework. There has also been a recent stream of literature considering feature-based pricing strategies with fairness (Chen et al. 2021c, Cohen et al. 2022, 2021) and differential privacy (Chen et al. 2021b, 2022c). To the best of our knowledge, we are the first in the literature to study *nonparametric* feature-based pricing strategies in an *offline* data-driven setting with censored demand through the lens of causal inference. Since the time this chapter was written, two subsequent studies Wang (2023) and Miao et al. (2023) have utilized the causal inference framework to explore feature-based pricing problems under approximate or invalid instrumental variables.

Policy Learning under Causal Inference. Policy learning under the framework of causal inference has also been well-studied in the statistics community. Here we review several papers related to our proposal. For a complete overview, we refer to Kosorok and Laber (2019) and the references therein. In particular, Chen et al. (2016) and Kallus and Zhou (2018) studied policy learning with continuous treatment in the clinical trial setting, where the generalized propensity score (Hirano and Imbens 2004) is known. Therefore they adopted an inverse probability weighting approach for evaluating and optimizing policies. Recently, Cai et al. (2021) proposed a deep Q-learning-typed approach for estimating an optimal interval-policy in the continuous treatment setting. Chernozhukov et al. (2019) and Schulz and Moodie (2021) developed doubly robust approaches for policy learning in the continuous treatment space by assuming some parametric component related to the treatment effect. Different from the aforementioned works, we propose to use kernel approximation and develop a different doubly robust estimator for evaluating and optimizing policies. All our models are non-parametric, thus enjoying more robustness of model mis-specification. More importantly, our proposal is able to handle the potential censoring in the outcome.

5.1.4 Organization and Notation

The rest of the chapter is organized as follows. In Section 5.2, we describe the mathematical model with first the case of fully observable demand and then the case of censored demand, through the potential outcome framework in causal inference. We also establish problem identifiability. In Section 5.3, we propose an offline data-driven algorithm based on survival analysis and a doubly robust estimation approach to computing the optimal feature-based pricing strategy. In Section 5.4, we give a theoretical regret analysis of our proposed offline learning algorithm. A numerical study is provided in Section 5.5, and we conclude the chapter and point out several future research avenues in Section 5.6. Technical proofs are presented in the Appendix.

Throughout this chapter, we distinguish between a random variable and its realizations using capital and lower-case letters, respectively. The function $x^+ = \max(x, 0)$. The indication function $\mathbb{1}(A)$ takes the value 1 if the event A is true and 0 otherwise. For generic sequences $\{\varpi(n)\}$ and $\{\theta(n)\}$, $\varpi(n) \lesssim \theta(n)$ means that there exists a sufficiently large constant $c_1 > 0$ such that $\varpi(n) \leq c_1\theta(n)$. We use "covariates", "features", and "contexts", interchangeably.

5.2 Model Formulation

5.2.1 Feature-Based Pricing with Fully Observable Demand

We first describe our model with fully observable demand (i.e., uncensored demand) for ease of presentation. We denote by P the price of some product which takes values in a compact space $\mathcal{P}$, i.e., $P \in \mathcal{P} = [p_1, p_2]$ with $0 \leq p_1 \leq p_2$. Let Y be the amount of inventory available for sales, which takes a non-negative value. In particular $Y \in \mathcal{Y} \subseteq [0, \infty)$. We note that here we consider Y to be continuous, but our framework can be easily extended to the case where Y is discrete. We model the relationship between demand and price under the celebrated *potential outcome* framework in causal inference (Rubin 1974). In particular, let $D(p)$ be the potential demand of a product if the treatment or price P is set as a (deterministic) value p. Let D be the observed demand. We can only observe $D = D(p)$ if we set the treatment or price $P = p$. We denote by X the observed q-dimensional covariates associated with the product that belongs to some covariate space $\mathcal{X} \subset \mathbb{R}^q$. Note that our model allows for either customer covariates (e.g., geographical information, past clicks, spending patterns) or product covariates (e.g., color, size, quality), or both. In summary, for each product, we can observe a random tuple (X, Y, P, D) if the demand is not censored.

Under this potential outcome framework of the price-demand setting, our goal is to find the optimal feature-based pricing strategy that maximizes the potential profit based on covariates. Specifically, let Π be the class of all pricing strategies where each strategy $\pi \in \Pi$ is a measurable function: $(\mathcal{X}, \mathcal{Y}) \to \mathcal{P}$, i.e., mapping from the covariate space $\mathcal{X}$ and the inventory space $\mathcal{Y}$ into the pricing space $\mathcal{P}$. Then the potential outcome under a pricing strategy $\pi \in \Pi$ is defined as $D\left(\pi\left(X, Y\right)\right)$. For ease of presentation, we write it as $D(\pi)$ hereafter.

Then the expected profit of a pricing strategy $\pi \in \Pi$ is defined as

$$V(\pi) \triangleq \mathbb{E}\left\{\pi(X, Y) \times \min\{D(\pi), Y\} - c \times (D(\pi) - Y)^{+}\right\}, \tag{5.1}$$

where c is the stockout cost per unit. For simplicity, we assume c is fixed and known. (Our framework can also treat the stockout cost as a random variable, which can be observed as a part of covariates X.) Note that $V(\pi)$ defined in (5.1) may not be identified by the observed data without any assumptions, since for each observation one can only observe a particular demand $D(p)$ under the current price p. Consistent with standard causal identification results such as Robins (1986), we make the following three standard assumptions.

Assumption 5.2.1 (Standard Causal Assumption)

(a) $D = D(P)$ *almost surely;*

(b) There exists some constant $f_{\min}$ *such that the conditional probability density of the price* $f(P = p \,|\, X = x, Y = y) \geq f_{\min} > 0$ *for every* $p \in \mathcal{P}, x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$*;*

(c) $D(p) \perp\!\!\!\perp P \mid (X, Y)$ *where* $\perp\!\!\!\perp$ *represents the statistical independence.*

Assumption 5.2.1(a) states that when the treatment or price P is set to p, the observed demand D is equal to the potential demand $D(p)$ at price p. This assumption also rules out interference among observations, meaning that there cannot be a scenario where the treatment or price P is set to p, but the observed demand D corresponds to the potential demand $D(p')$ at some other price $p' \neq p$ (Rubin 1974). This assumption is standard in the causal inference literature. To make it clearer for the general audience, we will illustrate it using a *special case* with a linear demand function, ignoring the impact of Y. We assume that the potential demand follows a linear form: $D(\pi(X)) = a^\intercal X - (b^\intercal X) \cdot \pi(X) + \varepsilon$, where a and b are constant vectors in $\mathbb{R}^q$, and ε represents random noise. Assumption 5.2.1(a) ensures that for any pricing strategy $\pi(X) \in \mathcal{P}$, the observed demand $D \,|\, X, P = \pi(X)$ is the same as the potential demand $D(\pi(X)) \,|\, X, P = \pi(X)$, which can be expressed as $a^\intercal X - (b^\intercal X) \cdot \pi(X) + \varepsilon$. Note again that our model is nonparametric, capable of handling any form of demand functions.

Assumption 5.2.1(b) essentially states that each price has at least some positive probability of being assigned for every covariate. In causal inference, $f(P \,|\, X, Y)$ is commonly referred to as the generalized propensity score (Hirano and Imbens 2004), which is an extension of the propensity score for use with quantitative exposures. In our setting, since offline data may include those collected from some pricing experiments run by companies before, it may be sensible to assume all prices are possibly observed, although $f_{\min}$ could be very small. Assumption 5.2.1(b) is used to establish the non-parametric identification result on $V(\pi)$ via Lemma 5.2.2.

Assumption 5.2.1(c) indicates that all confounders are measured in the data. In other words, by adjusting for confounders, i.e., covariates X and the inventory Y, we are able to fully identify the causal effect of the price P on the demand D. If Assumption 3 is violated such as there is a measurement error on X, we cannot identify the effect of P on the demand, which will lead to a bias. In this case, we can investigate the sensitivity of the estimates of causal effects to the choice of pre-treatment variables used for adjustments to assess unconfoundedness (see §21.5 in Imbens and Rubin (2015)). One can also use the instrumental variable approaches for identifying the optimal pricing under unmeasured confounding. See Wang and Tchetgen (2018) and references therein.

Hence, we have the following identification result, indicating that under the current data generating process, $V(\pi)$ is uniquely defined. In other words, there does not exist another $\widetilde{V}(\pi)$ that is consistent with the offline data distribution. This situation could arise if, conditional on the covariates, the observed demand cannot reflect the effect of price on demand, and the statistical estimates obtained from the offline dataset lead to another variable instead of the potential expected revenue under policy π. This is the building block for finding an optimal pricing strategy using the offline data.

Lemma 5.2.1 *Under Assumption 5.2.1, we can identify the value function $V(\pi)$ by*

$$V(\pi) = \mathbb{E}\left\{Q(X, Y, \pi(X, Y))\right\}, \tag{5.2}$$

where the expectation is taken over X, Y, and the Q-function is

$$Q(X, Y, P) = \mathbb{E}\left\{\left(P \times \min\{D, Y\} - c \times (D - Y)^{+}\right) \mid X, Y, P\right\}.$$

Note that the problem identifiability here refers to that we can uniquely determine the potential revenue. By comparison, the problem identifiability defined in Bu et al. (2023) pertains to whether the distribution of model parameters in the linear demand linear model can be learned by any algorithm from censored demand data, even with an infinite number of samples and a fixed observable boundary. Since we do not make any parametric assumptions regarding the demand model and we do not assume a uniform boundary on the upper bound of observable demand, this notion of identifiability defined within the distributionally robust framework does not apply to our problem. By maximizing $V(\pi)$ in (5.2) over the pricing strategy class Π, a *global* optimal pricing strategy is

$$\pi^* \in \operatorname*{arg\,max}_{\pi \in \Pi} \mathbb{E}\left\{Q(X, Y, \pi(X, Y))\right\}.$$

Since Π is the class of all pricing strategies, the optimal pricing strategy can be further shown as

$$\pi^*(X, Y) \in \operatorname*{arg\,max}_{p \in \mathcal{P}} \left\{\mathbb{E}\left[P \times \min\{D, Y\} - c \times (D - Y)^+ \,|\, X, Y, P = p\right]\right\}, \tag{5.3}$$

almost surely. Essentially, for each (X, Y), a price $p = \pi^*(X, Y)$ should be assigned so that the expected profit is maximized. Based on (5.2), one possible way is to directly apply supervised learning techniques to estimate the Q-function and then optimize the strategy over Π. However, such an approach may suffer from bias when the model for estimating Q-function is mis-specified. To account for mis-specification, in Section 5.3.1, we introduce a doubly robust estimator for learning the optimal pricing strategy. Before that, we introduce an alternative approach for identifying $V(\pi)$, which is a key step in the proposed doubly robust estimator. Specifically, one may employ an inverse probability weighting (IPW) approach (Robins 1986) to identify $V(\pi)$. In particular, consider

$$\mathbb{E}\left\{\frac{(P \times \min\{D, Y\} - c \times (D - Y)^+)\, \mathbb{1}(\pi(X, Y) = P)}{f(P \,|\, X, Y)}\right\}. \tag{5.4}$$

When the pricing space $\mathcal{P}$ is discrete, we can use the above formulation to identify $V(\pi)$ under Assumption 5.2.1(b), where the conditional probability density becomes a mass function. However, when P is continuously distributed in the observed data, given X and Y, $\mathbb{1}(\pi(X, Y) = P)$ may not be absolutely continuous with respect to the probability measure of the price P in the observed data. Hence the above formulation could be invalid for identifying $V(\pi)$.

In what follows, we provide a valid IPW-type approach to approximately identify $V(\pi)$, similar to that in Kallus and Zhou (2018), based on which we can combine the IPW-type and Q-function estimation methods for learning the optimal pricing strategy robustly. In particular, we adopt a kernel-based approach to approximate $V(\pi)$. Recall that a kernel function $K(u) : \mathbb{R} \to [0, \infty)$ satisfies $\int uK(u)du = 0$ and $\int K(u)du = 1$, and specific examples of kernel functions include uniform, triangular, and Gaussian kernels among many others. We adopt the following function with kernels to approximate $V(\pi)$:

$$V_h(\pi) = \mathbb{E}\left\{ \frac{\left(P \times \min\{D, Y\} - c \times (D - Y)^+\right) K(\frac{P - \pi(X,Y)}{h})}{hf(P \,|\, X, Y)} \right\}, \tag{5.5}$$

where h is the bandwidth used in kernel approximation. We refer interested readers to Parzen (1962) for more details of kernel methods.

We then show that $V(\pi)$ can be approximated by $V_h(\pi)$ up to arbitrary precision as $h \to 0$, where we first impose the following assumption.

Assumption 5.2.2 (Kernel Property and Approximation)

(a) $K(u) : \mathbb{R} \to [0, \infty)$ satisfies $\int K(u)du = 1$ and $\int |u|K(u)du \leq C_1$ for some constant C_1.

(b) There exists some universal constant C_2 such that

$$\mathbb{E}\left\{ \sup_{p_1 \leq p < p' \leq p_2} \left| \frac{Q(X, Y, p) - Q(X, Y, p')}{p' - p} \right| \right\} \leq C_2. \tag{5.6}$$

Assumption 5.2.2(a) is satisfied by a wide range of kernel functions including the aforementioned examples. Assumption 5.2.2(b) is a mild condition on the smoothness of Q-function defined in Lemma 5.2.1. Then the next lemma shows that when h is small, $V_h(\pi)$ well approximates $V(\pi)$.

Lemma 5.2.2 *Under Assumptions 5.2.1–5.2.2, there exists some constant C_3 such that for all $\pi \in \Pi$,*

$$|V_h(\pi) - V(\pi)| \leq C_3 h. \tag{5.7}$$

So far we have discussed (approximately) identifying $V(\pi)$ if one can fully observe the information (X, Y, P, D) for all products. However, when the inventory cannot fully satisfy the demand, D will be censored. In the following subsection, we address the issue of potential censoring in the demand. In Section 5.3, focusing on estimating $V_h(\pi)$, we combine these two approaches to enhance the robustness in terms of statistical estimation.

5.2.2 Feature-Based Pricing with Censored Demand

One major challenge of estimating $\pi^*(X, Y)$ using the observed data is that any extra demands over inventory level Y are lost and thus cannot be fully leveraged by retailers to infer the optimal feature-based prices. Ignoring demand censoring in the dataset will inevitably cause biases in estimating the value function, leading to suboptimal prices as we discussed in

Section 5.1 (See Figure 5.1). In this subsection, we address this potential issue. We denote by S the observed sales quantity, i.e., $S = \min\{D, Y\}$. Therefore instead of having (X, Y, P, D), we may only obtain realizations of (X, Y, P, S) in practice. In this case, we augment this random tuple (X, Y, P, S) with a censoring indicator defined as $\Delta = \mathbb{1}(D < Y)$. In particular, if the demand D is less than the inventory level Y, we let $\Delta = 1$ and otherwise $\Delta = 0$. Here we can observe the censoring indicator for either continuous or discrete demand case as we can observe $\mathbb{1}(D < Y)$ by observing $\mathbb{1}(S < Y)$. When demand or sales is continuously distributed, we can safely ignore the case when $Y = D$, where there is indeed no censoring. For a more general discussion, we refer interested readers to Besbes and Muharremoglu (2013). In order to identify $V(\pi)$ under censored demand, we make one additional assumption.

Assumption 5.2.3 (Censored Demand Identification)

(a) Demand D and inventory Y are conditionally independent given feature X and price P, i.e., $D \perp\!\!\!\perp Y \mid X, P$.

(b) There exists some known constant $D_{\max}$ such that $0 \leq D \leq D_{\max}$ almost surely.

Assumption 5.2.3(a) essentially states that demand is not affected by the inventory level given covariates X and the price information P. This is reasonable as the inventory level is often considered as the private information of retailers. In the literature, Assumption 5.2.3(a) can be called (conditional) non-informative censoring, which rules out the dependence between the demand and inventory level. If violated, standard methods of estimating (the probability distribution of) the survival outcome will fail and lead to biased estimation. In the literature on informative censoring in survival analysis, a specific model is often imposed for modeling the relationship between response (i.e., demand) and censor time (i.e., inventory level). See Diggle and Kenward (1994) for more details. Assumption 5.2.3(b) imposes a uniform upper bound for the demands across covariates, which is mainly used to simplify the theoretical analysis of estimating $\mathbb{E}\left[D|X, P, S, \Delta = 0\right]$.

We identify $V(\pi)$ under the censored demand in the next lemma, where we first define a surrogate profit (outcome) as

$$R(X, P, S, \Delta) = P \times S + c \times \mathbb{1}(\Delta = 0)\left(S - \mathbb{E}\left[D|X, P, S, \Delta = 0\right]\right). \tag{5.8}$$

Lemma 5.2.3 *Suppose Assumptions 5.2.1–5.2.3*(a) *hold, then for every $\pi \in \Pi$, we have*

$$V_h(\pi) = \mathbb{E}\left\{\frac{RK(\frac{P-\pi(X,Y)}{h})}{hf(P \mid X, Y)}\right\}, \tag{5.9}$$

where R is given in (5.8) *as a function of (X, P, S, Δ).*

For the remainder of this chapter, we write $R = R(X, P, S, \Delta)$ explicitly to indicate its dependency on X, P, S, and Δ, whenever needed. As seen from (5.8), R is a function of the observed data. Therefore if one can estimate R accurately, then $V_h(\pi)$ can be estimated properly under the setting of censored demand, after which we can optimize with respect to π to achieve the optimal feature-based pricing strategy.

5.3 Offline Feature-Based Pricing Strategy

5.3.1 Estimation Framework

To derive the optimal feature-based pricing strategy π^*, we first leverage the (offline) observed data to estimate the objective function $V_h(\pi)$, which is of crucial importance in achieving the optimal pricing strategy. Figure 5.2 shows the roadmap of our main estimation framework. For the plain estimation strategy, we need to estimate two quantities, namely, surrogate profit R and propensity scores $f(P \mid X, Y)$, which we shall discuss immediately. Later, we will also use an improved strategy called doubly robust estimation to further increase the robustness of our estimation. We first outline our estimation framework with key ideas, and defer the implementation details in Section 5.3.2.

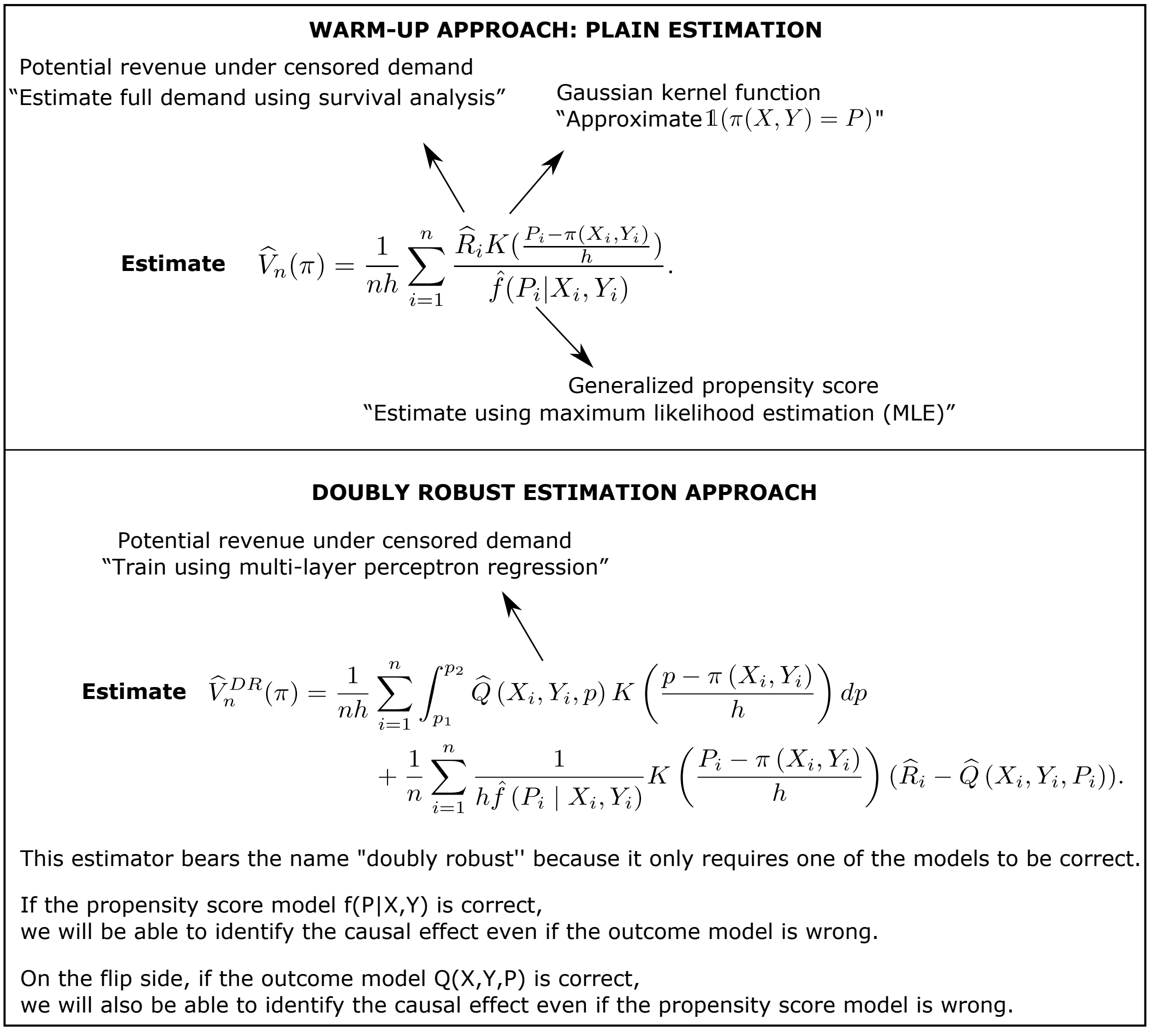
Estimation of Potential Profit R. We first discuss how we estimate the potential profit R in (5.8). Now suppose that we observe n independent and identically distributed (i.i.d.) samples

$$\mathcal{D}_n = \{(X_i, Y_i, P_i, S_i, \Delta_i)\}_{1 \le i \le n}\,,$$

where $\Delta_i = \mathbb{1}(D_i \le Y_i)$ denotes the censoring indicator for i-th product.

Note that the i.i.d. assumption on the offline dataset implies that the quintuples $(X_i, Y_i, P_i, S_i, \Delta_i)$ are independent across data entries. However, within each data entry, $(X_i, Y_i, P_i, S_i, \Delta_i)$ could exhibit arbitrary correlations. For instance, a firm may employ a complex strategy f to determine $Y_i = f(X_i)$, where inventory decisions are based on the given features, which is permissible within our model. Moreover, to further address the assumption of independence across data entries, we investigate a multi-center case in Appendix D.3 where dependence exists among the observations. We extend our method by incorporating stationary and exponential β-mixing processes, and also provide a theoretical guarantee.

In order to estimate π^* using the observed data, we first address the key unknown quantity $\mathbb{E}\left[D|X, P, S, \Delta = 0\right]$ in (5.8). Inspired by the notion of conditional survival function in survival analysis (see, e.g., Kleinbaum and Klein 2010 and Cui et al. 2017), we have the following lemma.



WARM-UP APPROACH: PLAIN ESTIMATION

Potential revenue under censored demand
“Estimate full demand using survival analysis”

Gaussian kernel function
“Approximate $\mathbb{1}(\pi(X,Y)=P)$"

Estimate $\widehat{V}_n(\pi)=\frac{1}{nh}\sum_{i=1}^{n}\frac{\widehat{R}_i K(\frac{P_i-\pi(X_i,Y_i)}{h})}{\hat{f}(P_i|X_i,Y_i)}.$

Generalized propensity score
“Estimate using maximum likelihood estimation (MLE)”

DOUBLY ROBUST ESTIMATION APPROACH

Potential revenue under censored demand
“Train using multi-layer perceptron regression”

Estimate

$$\widehat{V}_n^{DR}(\pi)=\frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2}\widehat{Q}\left(X_i,Y_i,p\right)K\left(\frac{p-\pi\left(X_i,Y_i\right)}{h}\right)dp$$
$$+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\hat{f}\left(P_i\mid X_i,Y_i\right)}K\left(\frac{P_i-\pi\left(X_i,Y_i\right)}{h}\right)(\widehat{R}_i-\widehat{Q}\left(X_i,Y_i,P_i\right)).$$

This estimator bears the name "doubly robust'' because it only requires one of the models to be correct.

If the propensity score model f(P|X,Y) is correct,
we will be able to identify the causal effect even if the outcome model is wrong.

On the flip side, if the outcome model Q(X,Y,P) is correct,
we will also be able to identify the causal effect even if the propensity score model is wrong.

Figure 5.2: Roadmap of Our Approach

Lemma 5.3.1 *Under Assumptions 5.2.3, we have*

$$\mathbb{E}\left[D|X,P,S,\Delta=0\right] = S + \int_S^{D_{\max}} \frac{H(t|X,P)}{H(S|X,P)}dt, \tag{5.10}$$

where $H(t|X,P) = \mathbb{P}(D > t \,|\, X, P)$.

By Lemma 5.3.1, we have that it is sufficient to estimate $H(t|X,P)$, which is called the conditional survival function in the literature (e.g., Kleinbaum and Klein (2010)). To achieve modeling robustness, we then propose to adopt the nonparametric random survival forests to estimate $H(t|X,P)$ (Ishwaran et al. 2008). Plugging the estimator of $H(t|X,P)$ into (5.10), we denote the estimator of $\mathbb{E}\left[D|X,P,S,\Delta=0\right]$ as $\widehat{\mathbb{E}}\left[D|X,P,S,\Delta=0\right]$. Then, we let the estimator for R be

$$\widehat{R} = P \times S + c \times \mathbb{1}(\Delta = 0)\left(S - \widehat{\mathbb{E}}\left[D|X,P,S,\Delta=0\right]\right). \tag{5.11}$$

Meanwhile, we denote the estimator for the potential profit of the i-th product as $\widehat{R}_i$ for $1 \le i \le n$ using the same estimation procedure described above. In what follows, we write $\widehat{R}$ and $\widehat{R}_i$ for $\widehat{R}(X,P,S,\Delta)$ and $\widehat{R}_i(X_i,P_i,S_i,\Delta_i)$, respectively.

Estimation of Propensity Scores $f(P\,|\,X,Y)$. Since the propensity score $f(P\,|\,X,Y)$, which is the conditional density function of the price, is generally unknown, we can estimate it using the observed data. For example, one can model $P\,|\,X,Y$ as a Gaussian random variable. Then it is sufficient to use the maximum likelihood method to estimate the mean $\mu_P(X,Y)$ and the variance $\mathbf{Var}(P\,|\,X,Y)$. One can also incorporate non-parametric models such as kernel density estimation or generative adversarial networks (Goodfellow et al. 2014) to obtain an estimate of $f(P\,|\,X,Y)$. The resulting estimator is denoted as $\widehat{f}(P\,|\,X,Y)$.

Plain Estimation of Objective Function $V_h(\pi)$. Given the observed data and two estimators, $\widehat{R}$ and $\widehat{f}(P\,|\,X,Y)$ described above, we plug them into (5.9) and get an estimator for $V_h(\pi)$. We then solve the following maximization problem to get a pricing policy $\widehat{\pi}$ that

$$\widehat{\pi} \in \arg\max_{\pi\in\Pi_0} \frac{1}{nh}\sum_{i=1}^n \frac{\widehat{R}_i K(\frac{P_i - \pi(X_i,Y_i)}{h})}{\widehat{f}(P_i|X_i,Y_i)} - \lambda_n J(\pi), \tag{5.12}$$

where Π_0 is some pre-specified class of pricing strategies, $J(\pi)$ is some regularization function on the policy π, and λ_n is a positive tuning parameter possibly depending on the sample size n. In this chapter, for ease of presentation, we consider Π_0 as a Hilbert space with norm $\|\cdot\|_{\Pi_0}$ and $J(\pi) = \|\pi\|^2_{\Pi_0}$. For example, if we consider a class of linear pricing strategies, i.e.,

$$\Pi_0 = \left\{\pi \,:\, \pi(X,Y) = \min(p_2, \max(p_1, \beta_0 + (X,Y)^\top\beta)) \quad \text{with} \quad \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^{q+1}\right\}, \tag{5.13}$$

then we let $J(\cdot)$ be the ridge penalty. Then the optimization problem in (5.12) becomes

$$\max_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^{q+1}} \frac{1}{nh}\sum_{i=1}^{n}\frac{\widehat{R}_i K(\frac{P_i-\min(p_2,\max(p_1,\beta_0+(X_i,Y_i)^\top\beta))}{h})}{\widehat{f}(P_i|X_i,Y_i)} - \lambda_n\|\beta\|_2^2. \tag{5.14}$$

We remark that plain estimation in (5.12) could incur large errors due to the possibly small propensity estimation $\widehat{f}(P\,|\,X,Y)$ in the denominator, especially when there is model misspecification. In the following, we consider a doubly robust estimation for $V_h(\pi)$.

Doubly Robust Estimation of Objective Function $V_h(\pi)$. To address the potential model mis-specification of the propensity score $f(P|X,Y)$ and possibly large errors, we adopt the doubly robust estimation idea in causal inference to estimate $V_h(\pi)$ (Bang and Robins 2005). The motivation for proposing the doubly robust estimator is bi-directional. As shown in Figure 5.2, the plain estimation involves the treatment model, while the direct estimation of potential revenue is based on an outcome model. The doubly robust estimator can be viewed as a correction of the outcome regression by a function that involves the treatment model, or as a correction of the inverse probability weighting (IPW) estimator by incorporating the outcome model. This approach allows us to demonstrate that the estimator will converge to the true value as long as at least one of the models is consistent. In the doubly robust framework, we first use some supervised learning techniques to estimate $\mathbb{E}[R\,|\,X,Y,P]$, and we denote the estimator as $\widehat{Q}(X,Y,P)$. Then we propose the following doubly robust estimator for estimating $V_h(\pi)$ that

$$\begin{aligned}\widehat{V}_n^{DR}(\pi) =&\frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2}\widehat{Q}(X_i,Y_i,p)K\left(\frac{p-\pi(X_i,Y_i)}{h}\right)dp\\ &+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\widehat{f}(P_i|X_i,Y_i)}K\left(\frac{P_i-\pi(X_i,Y_i)}{h}\right)(\widehat{R}_i-\widehat{Q}(X_i,Y_i,P_i)).\end{aligned} \tag{5.15}$$

Besides plain estimation, as seen from Lemma 5.2.1, one can also use $\widehat{Q}$ to construct an estimator for $V(\pi)$. However, due to the censoring issue, the estimator $\widehat{Q}$ could be biased. This indicates that estimating Q could also be hard especially when the censoring rate is high. A doubly robust estimator (5.15) naturally combines these two approaches together for achieving a robust estimation property. The double robustness means that as long as either $\widehat{Q}(X,Y,P)$ or $\widehat{f}(P\,|\,X,Y)$ consistently estimates the counterpart, $\widehat{V}_n^{DR}(\pi)$ is a consistent estimator for $V_h(\pi)$ and $V(\pi)$ (when $h\to 0$), which is shown below. Basically when the estimated outcome model is replaced by the true one, the bias term (last term in (5.15)) vanishes asymptotically. Thus we get the consistency result. When the propensity score is correctly specified, by the change of measure, quantities related to the estimated outcome

$\widehat{Q}$ disappear, which implies consistency as well. Hence, compared with plain estimation and directly optimizing Q function for π^*, our doubly robust estimator and the proposed learning algorithm provide additional robustness against the potential model mis-specification on the propensity score $f(P\,|\,X,Y)$ and Q function. More specifically, denote $\widehat{V}_n^{DR}(\pi)$ as $\widehat{V}_n^{DR}(\pi, Q, f, R)$ to indicate its dependency on Q, f, and R, we have the following property.

Theorem 5.3.1 *Let $\widetilde{Q}$ and $\widetilde{f}$ be estimators for Q and f, respectively. Suppose that Assumptions 5.2.1–5.2.2 and Assumption 5.4.3*(a) *in Section 5.4 hold. If either $\widetilde{Q}$ or $\widetilde{f}$ is consistent in terms of sup-norm, then for any given $\varepsilon > 0$, we have*

$$\lim_{h\to 0}\lim_{n\to\infty}\mathbb{P}\left(\left|\widehat{V}_n^{DR}(\pi,\widetilde{Q},\widetilde{f},\widehat{R}) - V(\pi)\right| \geq \varepsilon\right) = 0.$$

As seen from Theorem 5.3.1, as long as either Q or f can be estimated consistently, we can show that $\widehat{V}_n^{DR}(\pi)$ converges to the truth in probability.

Based on the doubly robust estimator, we propose to estimate π^* by solving the following optimization problem that

$$\widehat{\pi} = \arg\max_{\pi\in\Pi_0} \widehat{V}_n^{DR}(\pi) - \lambda_n J(\pi). \tag{5.16}$$

In practice, we may further implement cross-fitting technique (Bickel 1982, Chernozhukov et al. 2018) to remove the dependence between nuisance functions (i.e., $\widehat{Q}(X,Y,P)$ and $\widehat{f}(P\,|\,X,Y)$) and the estimated pricing strategy, so that the data efficiency can be guaranteed under less restrictive conditions on each nuisance function. See Chernozhukov et al. (2018) for more details. In particular, in cross-fitting, we randomly split data into M folds and apply the following procedure: first, for each fold $m = 1,\cdots,M$, we use the other $M-1$ folds to obtain the estimators $\widehat{Q}^{(-m)}(X,Y,P)$ and $\widehat{f}^{(-m)}(P\,|\,X,Y)$ for $Q(X,Y,P)$ and $f(P\,|\,X,Y)$ respectively; then we obtain an estimated optimal pricing strategy $\widehat{\pi}_n$ by solving the following optimization problem that

$$\begin{aligned}\widehat{\pi}_n \in \arg\max_{\pi\in\Pi_0}\Bigg\{ &\frac{1}{nh}\sum_{i=1}^n \int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(X_i,Y_i,p) K\left(\frac{p-\pi(X_i,Y_i)}{h}\right) dp \\ &+\frac{1}{n}\sum_{i=1}^n \frac{1}{h\widehat{f}^{(-m(i))}(P_i|X_i,Y_i)} K\left(\frac{P_i-\pi(X_i,Y_i)}{h}\right)(\widehat{R}_i - \widehat{Q}^{(-m(i))}(X_i,Y_i,P_i)) - \lambda_n J(\pi)\Bigg\},\end{aligned} \tag{5.17}$$

where $m(i)$ denotes the fold containing the i-th observation.

We outline the proposed algorithm in Table 10, and provide more details in the next subsection. A brief description of the random survival forests method is in Appendix D.4.4. We remark that while the random survival forests method is used for estimating the surrogate outcome R, this task can also be solved by other methods such as the Kaplan-Meier estimator or Nelson-Aalen estimator.

Algorithm 10 (Offline) Learning Optimal Feature-Based Pricing Strategy

```
Input: Observed data $\{X_i, Y_i, P_i, S_i, \Delta_i\}_{i=1}^n$; kernel bandwidth $h$ and tuning parameter $\lambda_n$;
Divide $\mathcal{D}_n$ into $M$ folds and denote $\mathcal{D}_n^{(-m)}$ as the other $(M-1)$ folds except $m$.
Estimate $\mathbb{E}[D|X, S, Y, \Delta = 0]$ using $\mathcal{D}_n$ via random survival forests (Ishwaran et al. 2008).
for $m = 1, \ldots, M$ do
    Apply supervised learning techniques with response $\widehat{R}$ and covariates $(X, Y, P)$ to obtain estimates $\widehat{Q}^{(-m)}$ using $\mathcal{D}_n^{(-m)}$.
    Compute estimates $\widehat{f}^{(-m)}(P \mid X, Y)$ via kernel density estimation or maximum likelihood estimation using $\mathcal{D}_n^{(-m)}$.
end for
Solve the optimization problem in (5.17).
Output: $\widehat{\pi}_n(X, Y)$
```

5.3.2 Implementation Details of Algorithm 10

We discuss the implementation of Algorithm 10 in detail. Suppose that we are given an offline dataset of n records with censored demand. Each record $i \in [n]$ includes the i-th product's features X_i, its inventory level Y_i, its price $P_i \in [p_1, p_2]$, and the corresponding sales data S_i. We aim to find an optimal pricing strategy $\pi(X, Y)$ based on this dataset, i.e., to find the optimal price given a product's features and inventory level.

As we discussed in the previous subsection, the basic idea of our algorithm is to first estimate some nuisance functions, and then find a policy $\widehat{\pi}$ that maximizes the estimated profit as stated in (5.17). Our framework requires computing, $\widehat{R}_i$, $\widehat{Q}^{(-m(i))}(X_i, Y_i, p)$ for each $p \in \mathcal{P}$, and $\widehat{f}^{(-m(i))}(P_i|X_i, Y_i)$. In what follows, we discuss how we compute them in more detail.

(a) We first discuss how to compute $\widehat{R}(X, P, S, \Delta)$ in (5.11). First, we apply the random survival forests algorithm for estimating the conditional survival function $H(t \mid X, P) = \mathbb{P}(D > t \mid X, P)$ as in Ishwaran et al. (2008). Then according to (5.10), we can further obtain an $\widehat{\mathbb{E}}[D \mid X, P, S, \Delta = 0]$. Finally, we estimate the conditional expected reward

$$\widehat{R}(X, P, S, \Delta) = P \times S + c \times \mathbb{1}(\Delta = 0)(S - \widehat{\mathbb{E}}[D \mid X, P, S, \Delta = 0]). \tag{5.18}$$

(b) We then discuss how we compute $\widehat{Q}^{(-m(i))}(X_i, Y_i, p)$ and $\widehat{f}^{(-m(i))}(P_i \mid X_i, Y_i)$ using the cross-fitting technique. We randomly split data into M folds, and for each fold $m \in [M]$, we use the other $M-1$ folds as the training data to obtain estimator $\widehat{Q}^{(-m)}(X, Y, P)$ and $\widehat{f}^{(-m)}(P \mid X, Y)$ for estimates of $Q(X, Y, P)$ and $f(P \mid X, Y)$. In particular,

(i) For $\widehat{Q}^{(-m(i))}(X_i, Y_i, P)$, we use deep neural networks to model

$$\widehat{Q}(X, Y, P) = \mathbb{E}[\widehat{R} \mid X, Y, P]$$

and for each record i in the dataset, we obtain an estimate $\widehat{Q}(X_i, Y_i, p)$ for all $p \in \mathcal{P}$.

(ii) For $\widehat{f}^{(-m(i))}(P_i \mid X_i, Y_i)$, we adopt a Gaussian probabilistic model to estimate $f(P \mid X, Y)$ with deep neural networks to approximate the mean and covariance matrix. In particular, we assume that conditioning on X and Y, P follows a multi-variate Gaussian distribution $\mathcal{N}\left(\mu_{\phi_1}(X, Y), \sigma^2_{\phi_2}(X, Y)\right)$, where ϕ_1 and ϕ_2 are parameters of two neural networks to model the mean and the covariance matrix respectively. Then we apply maximum likelihood estimation (MLE) to obtain estimates $(\widehat{\phi}_1, \widehat{\phi}_2)$. In this way, we obtain estimates $\widehat{\mu}^{(-m(i))}(X_i, Y_i)$ and $\widehat{\sigma}^{(-m(i))}(X_i, Y_i)$, and then the estimated probability density function $\widehat{f}^{(-m(i))}(P_i \mid X_i, Y_i)$ for each record i in the dataset.

Upon computing all the estimators on the right-hand side of (5.17), we build a deep neural network for the pricing policy π with parameters ϕ_3. By minimizing the loss function customized to be the negative of the right-hand side of (5.17), we thus obtain the estimated network parameters $\widehat{\phi}_3$ maximizing the right-hand side of (5.17). After feeding the network with the offline dataset to train the pricing strategy, we can then output the near-optimal feature-based price for any given (X, Y). Note that one may use DC (difference of convex) programming (see, e.g., Cui et al. 2018) to solve for (5.17) if π uses a piecewise linear model (as a special case).

5.4 Regret Analysis and Double Robustness

We establish a finite sample regret bound for our estimated pricing strategy $\widehat{\pi}_n$ in terms of the sample size n, which shows that our estimated pricing strategy $\widehat{\pi}_n$ converges to the optimal one in terms of the regret as the sample size n goes to infinity. We define the regret of $\widehat{\pi}_n$ as the difference between the expected profit of the optimal strategy π^* and $\widehat{\pi}_n$ that

$$\mathbf{Regret}(\widehat{\pi}_n) = V(\pi^*) - V(\widehat{\pi}_n). \tag{5.19}$$

We first make the following technical assumptions.

Assumption 5.4.1 *There exists a constant $C_4 > 0$ such that $|Y| \leq C_4$.*

Assumption 5.4.2 *There exist constants $A > 0$ and $v > 0$ such that*

$$\sup_{\widetilde{Q}} N(\Pi_0, \widetilde{Q}, \varepsilon \|F\|_{\widetilde{Q},2}) \leq (A/\varepsilon)^v ,$$

for all $0 < \varepsilon \leq 1$, where $N(\Pi_0, \widetilde{Q}, \varepsilon \|F\|_{\widetilde{Q},2})$ denotes the covering number of the policy class Π_0, F is the envelope function of Π_0, $\| \cdot \|_{\widetilde{Q},2}$ denotes the L_2-norm under some finitely discrete probability measure $\widetilde{Q}$ on (X, Y), and the supremum is taken over all such probability measures.

Assumption 5.4.3 (Rate Conditions)

(a) There exists some constant $C_5(\varepsilon)$ depending on $\varepsilon \in (0, 1)$ such that

$$\sup_{x \in \mathcal{X}, p \in [p_1, p_2], 0 \leq s \leq C_4} |\widehat{R}(x, p, s, 0) - R(x, p, s, 0)| \leq C_5(\varepsilon) n^{-\delta},$$

with probability $1 - \varepsilon$ for some $\delta > 0$.

(b) The nuisance function estimators $\widehat{Q}^{(-m)}$ and $\widehat{f}^{(-m)}$ obtained from the other $(M - 1)$ folds of the data in the cross-fitting procedure in Algorithm 10 satisfy that there exist constants $\alpha > 0$ and $\beta > 0$ such that

$$\mathbb{E}\left[\|\widehat{Q}^{(-m)}(X, Y, P) - Q(X, Y, P)\|_2^2\right] = O(n^{-2\alpha}),$$

$$\mathbb{E}\left[\|1/\widehat{f}^{(-m)}(P \mid X, Y) - 1/f(P \mid X, Y)\|_2^2\right] = O(n^{-2\beta}),$$

uniformly for all $p \in [p_1, p_2]$ and $1 \leq m \leq M$. In addition, there exists a constant $C_6 > 0$ such that

$$\max\left\{\sup_{p_1 \leq p \leq p_2, x \in \mathcal{X}, y \in \mathcal{Y}} \left|1/\widehat{f}(P \mid X, Y)\right|, \|\widehat{Q}\|_\infty\right\} \leq C_6.$$

Assumption 5.4.1 requires that the inventory level Y is uniformly bounded, which is similar to Assumption 5.2.3(b), and they are reasonable in practice. Both Assumptions 5.2.3(b) and 5.4.1 can be relaxed by imposing a bounded condition on the second moments of D and Y respectively and using the truncation argument. In this case, it would be hard to derive the high probability bound that decays exponentially fast. Therefore, for simplicity, we consider these two relatively stronger assumptions. Assumption 5.4.2 basically states that the policy class Π_0 has finite Vapnik-Chervonenkis (VC) dimension (see Definition 2.1 of Chernozhukov et al. 2014). Assumption 5.4.3(a) imposes a high-level condition on the estimation of the surrogate outcome R in (5.8). As we discussed in the previous section, we estimate R by estimating the conditional survival function $H(\cdot)$ in (5.10). By standard nonparametric methods such as kernel Kaplan-Meier estimator, this assumption is satisfied as discussed in Dabrowska (1989) and Khardani and Semmar (2014). See Theorem 3.2 and the proof of Khardani and Semmar (2014) for more details. Assumption 5.4.3(b) imposes high-level conditions on the L_2-norm convergence rates of the estimated nuisance functions when the cross-fitting technique is applied. For our theoretical results below, we only require $\alpha + \beta > 1/2$, which is a mild assumption. For example, if linear models are imposed in estimating Q function, then $\alpha = \min(\frac{1}{2}, \delta)$ can be obtained by the least squares method as long as the linear model is correct. When non-parametric models such as linear sieve/neural networks are used to approximate Q, under some regularity conditions, one can show that $\alpha = \min(\frac{\omega}{q+1+2\omega}, \delta)$, where ω is the smoothness coefficient of the true Q. Assuming $2\omega > (q+1)$, we have $\alpha > \min(1/4, \delta)$. See Chen (2007) and Schmidt-Hieber (2020) for more details to attain these rates. Similar analysis can be performed for f. For example, if we model f by a Gaussian distribution, then a maximum likelihood estimation can be used to estimate $\mu_P(X,Y)$ and $\mathbf{Var}(P \mid X, Y)$ by a parametric model. If such a model is correct, then we have $\beta = 1/2$, which is a typical convergence rate in a parametric model. Therefore as long as we can estimate R reasonably well, we can guarantee that $\alpha + \beta > 1/2$ for a wide range of parametric and non-parametric models.

Let $\pi_h^* \in \arg\max_{\pi \in \Pi_0} V_h(\pi)$, and we let the approximation error of π_h^* be

$$\Lambda(\lambda_n) = V_h(\pi_h^*) - \sup_{\pi \in \Pi_0} \left\{ V_h(\pi) - \lambda_n J(\pi) \right\}.$$

The finite sample regret bound for $\widehat{\pi}_n$ is given by the following theorem.

Theorem 5.4.1 *Suppose that Assumptions 5.2.1–5.4.3 hold. If* $\lambda_n \leq 1$ *and* $\alpha + \beta > 1/2$, *then for any* $x > 0$, $\varepsilon \in (0,1)$ *with probability at least* $1 - \exp(-x) - \varepsilon$, *Algorithm 10 admits the following regret upper bound*

$$\begin{aligned} \mathbf{Regret}(\widehat{\pi}_n) \lesssim \Lambda(\lambda_n) + 2C_3 h + \max\{1, x\}\sqrt{v}\lambda_n^{-\frac{1}{2}} n^{-\frac{1}{2}}/h^2 \\ + C_5(\varepsilon)\max\{1, x\}\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\mathbb{P}(\Delta = 0), \end{aligned} \tag{5.20}$$

where the regret is defined in (5.19).

The proof of Theorem 5.4.1 is given in Appendix D.2. The first term $\Lambda(\lambda_n)$ in (5.20) can be interpreted as the approximation error of using Π_0. Under some mild conditions, if a universal kernel (e.g., Micchelli et al. (2006)) or a properly chosen neural network model (e.g., Barron (1994)) is used for constructing Π_0, one can show that $\Lambda(\lambda_n) \to 0$ as $\lambda_n \to 0$. The second term $2h$ is the approximation error incurred by the use of kernel function in (5.5). The third term of (5.20) can be understood as the estimation error when there is no censored demand. Note that when the tuning parameters λ_n and h are constants, the resulting rate of this estimation error is $n^{-1/2}$, which matches the optimal parametric rate. The last term in (5.20) is caused by the potential censored demand, which is controlled by the chance of observing censored demand, i.e., $\mathbb{P}(\Delta = 0)$ and the estimation error for the surrogate profit R, i.e., $n^{-\delta}$. Lastly, if we additionally assume that there exists $\zeta \in (0, 1]$ such that, for every n,

$$\Lambda(\lambda_n) \lesssim \lambda_n^{\zeta}, \tag{5.21}$$

then we have an explicit error bound for the regret of our estimated pricing strategy $\widehat{\pi}_n$ given by the following corollary. Note that (5.21) is a typical assumption in machine learning (See Steinwart and Christmann (2008) for more details).

Corollary 5.4.1 *Assume all conditions in Theorem 5.4.1 hold and* (5.21) *is satisfied, by choosing* $h = n^{-\frac{(12\zeta+1)\min(\frac{1}{2},\delta)}{6(6\zeta+1)}}$ *and* $\lambda_n = n^{-\frac{\min(\frac{1}{2},\delta)}{6\zeta+1}}$, *with probability at least* $1 - \exp(-x) - \varepsilon$, *Algorithm 10 admits the following regret upper bound*

$$\mathbf{Regret}(\widehat{\pi}_n) \lesssim \max\{1, x\} n^{-\frac{\zeta \min(\frac{1}{2},\delta)}{6\zeta+1}}. \tag{5.22}$$

Corollary 5.4.1 is obtained by plugging (5.21) into (5.20) and then optimizing the upper bound with respect to λ_n and h. We omit the proof for brevity. As we can see from (5.22), the regret bound of our estimated pricing strategy $\widehat{\pi}_n$ decreases as β increases (and the approximation error $\Lambda(\lambda_n)$ decreases). When $\zeta = 1$, we obtain the convergence rate $n^{-\min(\frac{1}{2},\delta)/7}$.

Note that with the probability bound in Theorem 5.4.1 and Corollary 5.4.1, one can also derive an upper bound for the expected regret. For example, suppose that with probability $1-\varepsilon$, $\text{regret}(\widehat{\pi}) \le \omega(n, \varepsilon)$ for some generic rate $\omega(n, \varepsilon)$. Then since $\text{regret}(\pi)$ is non-negative and uniformly bounded by some generic constant C for every π due to Assumptions 5.2.3(b) and 5.4.1, we have

$$\begin{aligned}\mathbb{E}[\mathbf{Regret}(\widehat{\pi}_n)] &\le \mathbb{E}[\mathbf{Regret}(\widehat{\pi}_n)\mathbb{1}(\mathbf{Regret}(\widehat{\pi}_n) \le \omega(n,\varepsilon))] + \mathbb{E}[\mathbf{Regret}(\widehat{\pi}_n)\mathbb{1}(\mathbf{Regret}(\widehat{\pi}_n) > \omega(n,\varepsilon))]\\ &\le \omega(n,\varepsilon) + C\varepsilon,\end{aligned}$$

For some concrete $\omega(n, \varepsilon)$ such as the one in Corollary 5.4.1, one can minimize the right-hand-side of the above inequality over ε to get the bound for the expected regret. We omit the details here.

In the literature considering a discrete action space with a fixed policy class and without the censoring issue, the minimax lower bound is of order $\sqrt{\text{VC}(\Pi_0)/n}$, where $\text{VC}(\Pi_0)$ is referred to as VC dimension of the policy class Π_0. Due to the complication of continuous action space and censored outcome, it still remains an open question for the optimal bound of the regret. Since this is beyond the scope of our chapter, we decide to leave it for future work.

5.5 Numerical Experiments

We carry out extensive numerical studies to demonstrate the efficacy of our proposed offline learning algorithm. We follow the implementation details of Algorithm 10, given in Section 5.3.2.

5.5.1 Experimental Setup

We present a simple but nontrivial numerical example. Even though this example is relatively simple, it gives insights into how the proposed causal inference based approach resolves the issue of demand censoring and achieves near-optimal feature-based pricing.

Data Generation. Consider a setting with only two product features X_1 and X_2. We generate each record of (X_1, X_2, P, D, Y, S) as follows.

(i) The first feature $X_1 \sim \text{Uniform}[0, 1]$.

(ii) The second feature $X_2 \sim \text{Uniform}[0, 1]$.

(iii) The price distribution $P \sim \text{Normal}(0.5, 0.5^2)$ truncated at $[0, 1]$.

(iv) The underlying demand distribution $D \sim \text{Poisson}(\lambda_1) + 1$ with rate $\lambda_1 = 5 + X_1^2 + X_2^2 - 5P$.

(v) The inventory Y is set as follows.

- For the dataset used in Section 5.5.2 without demand censoring, we let $Y = \infty$ for all records, so that the sales S is exactly demand D.
- For the dataset used in Section 5.5.3 where any demand exceeding inventory Y is censored, we let $Y = \lfloor \mathcal{N}(6, 2) \rfloor$.

(vi) $S = \min\{Y, D\}$ denotes the sales data.

Parameters. We set the sample size as $n = 2000$. The stock-out cost per unit is set to be $c = 0.1$. The neural network for estimating $\widehat{f}^{(-m(i))}(P_i \mid X_i, Y_i)$ has 3 inner layers, each with 84 nodes. The neural network for finding optimal prices has 4 inner layers, each with

12 nodes. We also discuss the sensitivity of the result with respect to the neural network parameters in Appendix D.4.2, and the running time is reported in Appendix D.4.3.

For the choice of h, we use $h = 0.01$ in our numerical implementation, which performs well empirically. Theoretically, we should choose h as small as possible since the kernel approximates the true value when $h \to 0$. However, there is always a trade-off between the estimator's bias and variance. In practice, given the kernel function, one can select the h value based on the size and quality of the sample data. For example, we may apply Silverman's rule of thumb, which is $h = 0.9 \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) n^{-\frac{1}{5}}$ where interquartile range (IQR) is the difference between the 75th and 25th percentiles of the data. Or, we can tune h by cross-validation.

For λ_n, we use $\lambda_n = 10^{-4}$ for ridge regularization term in our case. The selection of regularization parameters depends on the sample size, network structure, and the regularization type. In practice, it can be tuned through cross-validation techniques.

Kernel Approximations. We use kernel approximation in the value function estimation. Also, we use the finite sum to approximate any integrals.

(a) When calculating (5.17), the first term involves an integral, which is approximated by a finite sum with bandwidth 0.01. More precisely, the estimated Q values are given by

$$\begin{aligned} \widehat{Q}^{(-m(i))}(X_i, Y_i, \pi(X,Y)) &= \sum_{i=1}^{n} \int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(X_i, Y_i, p) K\left(\frac{\pi(X_i, Y_i) - p}{h}\right) dp \quad (5.23) \\ &\approx \sum_{i=1}^{n} \sum_{j=1}^{100} 0.01 \times \widehat{Q}^{(-m(i))}(X_i, Y_i, p_j) K\left(\frac{\pi(X_i, Y_i) - p_j}{h}\right), \end{aligned}$$

where $p_j = 0.01 \times j$ and $h = 0.01$.

(b) We use a Gaussian kernel function, $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$ with bandwidth $h = 0.01$, to approximate the indicator function in (5.4).

To see how our approximations perform, we shall compare the true and estimated Q values under different prices $\pi(X_i, Y_i) \in [0, 1]$. The true $Q(X_i, Y_i, \pi(X_i, Y_i))$ value is given by

$$Q(X, Y, P) = \mathbb{E}[R|X, Y, P] = (6 + X_1^2 + X_2^2 - 5P)P. \quad (5.24)$$

Note here that $Y = \infty$, ignoring the inventory impact. The estimated Q values are given by (5.23).

The numerical results show that the finite sum of the Gaussian kernel approximates the true Q values very well. Figure 5.3 shows an example where $X = [0.74, 0.22], Y = \infty$. As we can see, the estimated Q values based on the dataset are very close to the theoretical

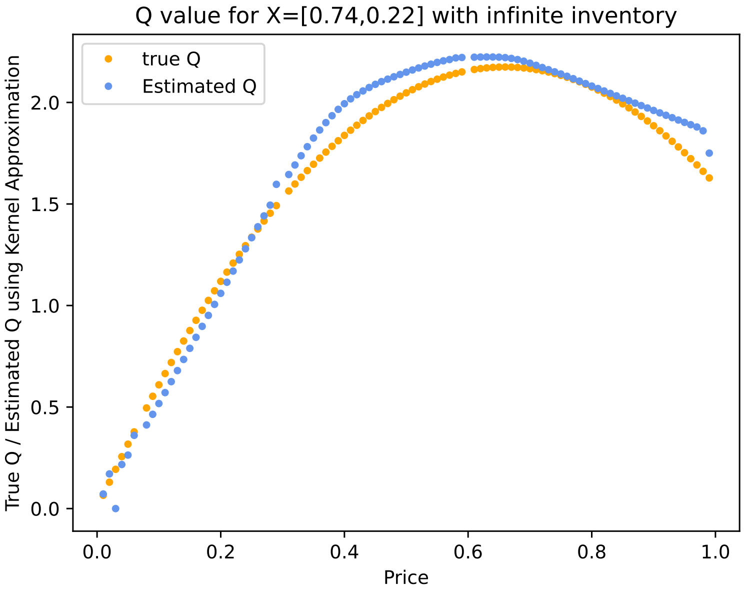


Figure 5.3: Comparison between the kernel approximated Q and the true Q

true Q values. Moreover, the estimated expected reward can capture the peak reward when iterating over different pricing choices, ensuring that the maximization of (5.17) would yield the pricing strategy with maximum profit.

Performance Measure. We randomly generate $L = 100$ pairs of training and testing datasets each with size $n = 2000$ according to our generative model. For each $\ell = 1, \dots, L$, we train our model independently using the ℓ^{th} training dataset and then test the performance on the ℓ^{th} testing dataset. Let $\mathbf{Eval}(\pi)$ be the average reward gained under any given pricing strategy π calculated on L testing datasets. Then we define the performance ratio as

$$\kappa(\pi) = \mathbf{Eval}(\pi)/\mathbf{Eval}(\pi^*), \tag{5.25}$$

where π^* is the theoretical optimal pricing policy. In our numerical experiment, we are particularly interested in $\kappa(\widehat{\pi}_n)$ and $\kappa(\tilde{\pi}_n)$. Here $\widehat{\pi}_n$ is our offline learning algorithm with demand censoring. Meanwhile, $\tilde{\pi}_n$ implements our algorithm by hypothetically treating the observed sales quantities as the true uncensored demand quantities.

5.5.2 The Case with Unlimited Inventory $Y = \infty$

Before implementing our algorithm with demand censoring, we first consider applying the algorithm to the same dataset but with unlimited inventory. More precisely, we hypothetically reset inventory $Y = \infty$ in the dataset (so that all demands are satisfied).

For our example in (5.24), we can calculate the theoretical optimal prices in closed-form as

$$\pi^*(X, Y) = \frac{(6 + X_1^2 + X_2^2)}{10} \in [0, 1],$$

which is used to benchmark against our numerical result (with unlimited inventory).

Estimation of Q Values. To check whether the estimates $\widehat{Q}^{(-m(i))}(X_i, Y_i, P_i)$ are accurate, we plot the estimated values and the true values of Q on one particular dataset in Figure 5.4.

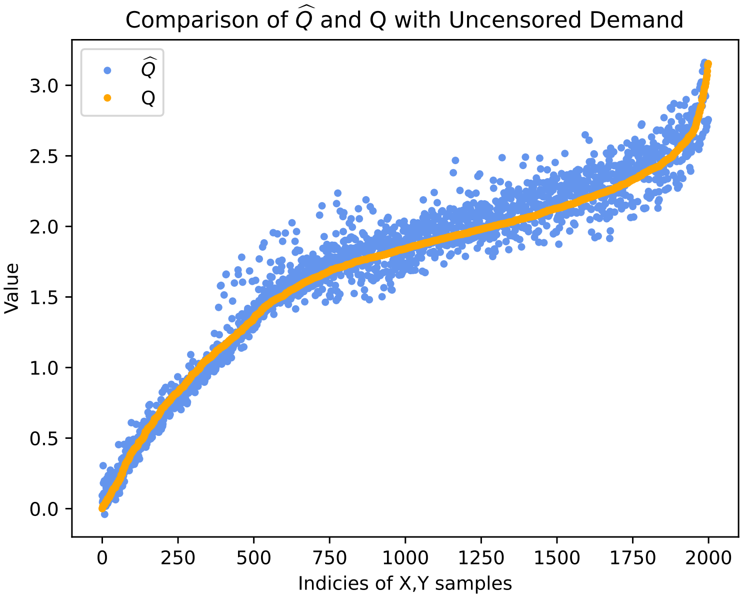


Figure 5.4: Accuracy of the estimated $\widehat{Q}$ values in the case of unlimited inventory

In Figure 5.4, the x-axis denotes the indices of samples in the dataset, which are ranked according to the true Q values from low to high. The true Q values (represented using orange dots) are calculated by (5.24) while our estimated $\widehat{Q}$ values (represented using blue dots) are given by MLP regressor. We can see that the estimated $\widehat{Q}$ values approximate the true Q values very well.

Doubly Robust Pricing Strategy. Now we apply Algorithm 10 according to (5.17). As the inventory is unlimited, there is no censoring in demand. Hence, we can obtain the true reward R_i for each record i using

$$R_i = P_i \times S_i - c \times (D_i - S_i)^+, \ \forall i \in [n].$$

Then upon computing the estimated generalized propensity scores $\widehat{f}^{(-m(i))}(P_i|X_i,Y_i)$ and the estimated Q values $\widehat{Q}^{(-m(i))}(X_i,Y_i,p)$, we compute the optimal pricing strategy via building a neural network for the pricing policy π with **"tensorflow"** package (version 2.6.0). With the loss function of this neural network being the negative of the right-hand side of (5.17), the regularization term is implemented by using the built-in "kernel_regularizer" parameter of the neural network layers. Then we apply the "Adam" optimization algorithm again to find the solution. So the optimal pricing strategy is obtained by solving the following problem:

$$\begin{aligned}\widehat{\pi}_n \in \underset{\pi\in\Pi_0}{\arg\max}\Bigg\{ & \frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(X_i,Y_i,p)K\left(\frac{p-\pi(X_i,Y_i)}{h}\right)dp \\ & +\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\widehat{f}^{(-m(i))}(P_i|X_i,Y_i)}K\left(\frac{P_i-\pi(X_i,Y_i)}{h}\right)(R_i-\widehat{Q}^{(-m(i))}(X_i,Y_i,P_i))-\lambda_n J(\pi)\Bigg\}.\end{aligned}$$

As a hypothetical benchmark, if we were to use true Q and f values, we would use

$$\begin{aligned}\breve{\pi}_n \in \underset{\pi\in\Pi_0}{\arg\max}\Bigg\{ & \frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2}Q(X_i,Y_i,p)K\left(\frac{p-\pi(X_i,Y_i)}{h}\right)dp \\ & +\frac{1}{n}\sum_{i=1}^{n}\frac{1}{hf(P_i|X_i,Y_i)}K\left(\frac{P_i-\pi(X_i,Y_i)}{h}\right)(R_i-Q(X_i,Y_i,P_i))-\lambda_n J(\pi)\Bigg\}.\end{aligned}$$

Figure 5.5 shows the performance of pricing strategies $\widehat{\pi}_n$ and $\breve{\pi}_n$, relative to the theoretical optimal prices π^*. The x-axis denotes the indices of samples in the dataset, ranked according to the theoretical optimal prices from low to high. We use translucent dots to represent the prices given by $\widehat{\pi}_n$ and $\breve{\pi}_n$. The fitted lines are generated by doing a polynomial fit of degree 4 for these dots. Our doubly robust pricing strategy $\widehat{\pi}_n$ is represented using the (fitted) blue line; the hypothetical benchmark $\breve{\pi}_n$ (with access to true Q and f) is represented using the (fitted) pink line; and the theoretical optimal price strategy is represented using the orange line. We can see that the fitted lines of the pricing strategies $\breve{\pi}_n$ and $\widehat{\pi}_n$ are close to each other, indicating excellent estimations of $\widehat{f}^{(-m(i))}(P_i|X_i,Y_i)$ and $\widehat{Q}^{(-m(i))}(X_i,Y_i,p)$. Both are consistent with the theoretical optimal prices but have a wider range.

In terms of profit performance via the notion of performance ratio defined in (5.25),

$$\kappa(\widehat{\pi}_n)=97.03\%\ (2.29\mathrm{e}^{-3}) \qquad \text{and} \qquad \kappa(\breve{\pi}_n)=96.78\%\ (3.04\mathrm{e}^{-3})$$

where the value in the parenthesis is the standard error of the average performance ratio.

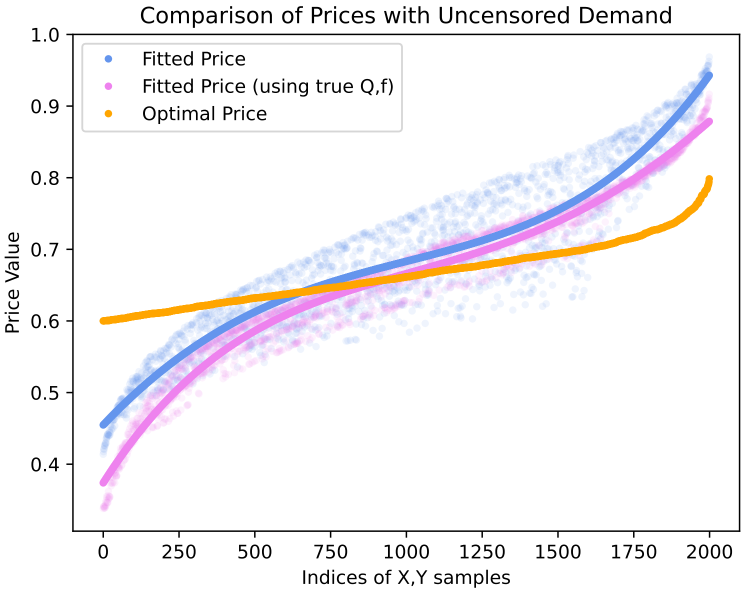


Figure 5.5: Prices given by doubly robust pricing strategy in the case of unlimited inventory

5.5.3 The Case with Limited Inventory (Demand Censoring)

Now we recover the finite values of Y in the dataset and consider the case with demand censoring.

Estimation of Q Values. With demand censoring, we no longer have access to the true profit, since any demand exceeding the limited inventory Y is lost and censored. Hence, we use the surrogate profit $\widehat{R}(X, P, S, \Delta)$. Specifically, as discussed in Ishwaran et al. (2008), we first draw B bootstrap samples from the whole dataset and grow a survival tree for each bootstrap sample using the extremely randomized tree in Geurts et al. (2006). Note that here B is a tuning parameter. After some experiments, we find that the output is robust with respect to the choice of B. In our use of R package **"ranger"** downloaded from R-CRAN, we adopt the default choice of B which is 500. Second, we estimate the conditional survival functions $H(t \,|\, X, P) = \mathbb{P}(D > t \,|\, X, P)$ as stated in Lemma 5.3.1 for all terminal nodes of the estimated survival trees using the Nelson–Aalen estimator (Aalen 1978, Nelson 1972) to obtain the cumulative hazard function and aggregate over all B trees. We remark that the terminal nodes are the unique sales values S in the data records. Then to estimate the conditional expected demand, the integral in (5.26) is approximated by Riemann sum, where the partition is decided by the terminal nodes, i.e. the observed sales values in the data. The survival function values are estimated by the "last observation carried forward" approach using the estimated conditional survival functions of the terminal nodes, which means for the t where we do not have the estimated survival function $\widehat{H}(t \mid X, P)$, we approximate it by using the estimated survival function of the closest sales value smaller than t. This step generates imputations for the censored samples. Then, by plugging the estimator of $H(t \mid X, P)$, we estimate the conditional expectation of demand using (5.10) that

$$\begin{aligned}\mathbb{E}[D \mid X, P, S, \Delta = 0] &= S + \int_S^{D_{\max}} \frac{H(t \mid X, P)}{H(S \mid X, P)} dt \\ &\approx S + \sum_{i=k}^{l-1} \frac{\widehat{H}(S_{[i]} \mid X, P)(S_{[i+1]} - S_{[i]})}{\widehat{H}(S \mid X, P)},\end{aligned} \tag{5.26}$$

where $S_{[k]} = S$ and $S_{[i]}$ denotes the ith order statistic of the unique sales values in the data and ℓ is the number of unique sales values. Based on the survival analysis above, $\widehat{R}(X, P, S, \Delta)$ is estimated via (5.18). In the neural network for estimating $Q(X, Y, P)$, we apply the multi-layer perceptron (MLP) regressor (Pedregosa et al. 2011).

Note that the true Q values (for our simple example) can be computed as follows.

$$\begin{aligned}Q(X, Y, P) &= \mathbb{E}[R|X, Y, P] = \mathbb{E}[R(X, P, S, \Delta)|X, Y, P] \\ &= \sum_{D=1}^{\infty} f(D) R(X, P, \min(D, Y), \mathbb{1}(D \le Y)) \\ &= \sum_{D=1}^{Y} f(D) PD + \sum_{D=Y+1}^{\infty} f(D) \left(PY + c(Y - \mathbb{E}[D|X, P, Y, \Delta = 0])\right) \\ &= \sum_{D=1}^{Y} f(D) PD + (1 - F(Y)) \left(PY + c\left(Y - \frac{\sum_{D=Y+1}^{\infty} D f(D)}{1 - F(Y)}\right)\right) \\ &= \sum_{D=1}^{Y} \frac{\lambda_1^{D-1} e^{-\lambda_1}}{(D-1)!} PD - \sum_{D=Y+1}^{\infty} \frac{\lambda_1^{D-1} e^{-\lambda_1}}{(D-1)!} cD + (1 - F(Y))(P + c)Y.\end{aligned} \tag{5.27}$$

Figure 5.6 shows the comparison of true Q values and the estimated $\widehat{Q}^{(-m(i))}(X_i, Y_i, P_i)$ using MLP. The x-axis denotes the indices of the sample in the data, ranked according to the true Q values from low to high. The true Q values (represented using orange dots) are calculated by (5.27) while our estimated $\widehat{Q}$ values (represented using blue dots) are given by MLP regressor. We use the "Adam" optimizer which is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments by Kingma and Ba (2017). We can see that the estimated $\widehat{Q}$ values approximate the true Q values very well.

Note that there is a slight overestimate of the true Q values, caused by the error from the estimation of $R(X, P, S, \Delta)$. Specifically, when demand is censored, the way we obtained

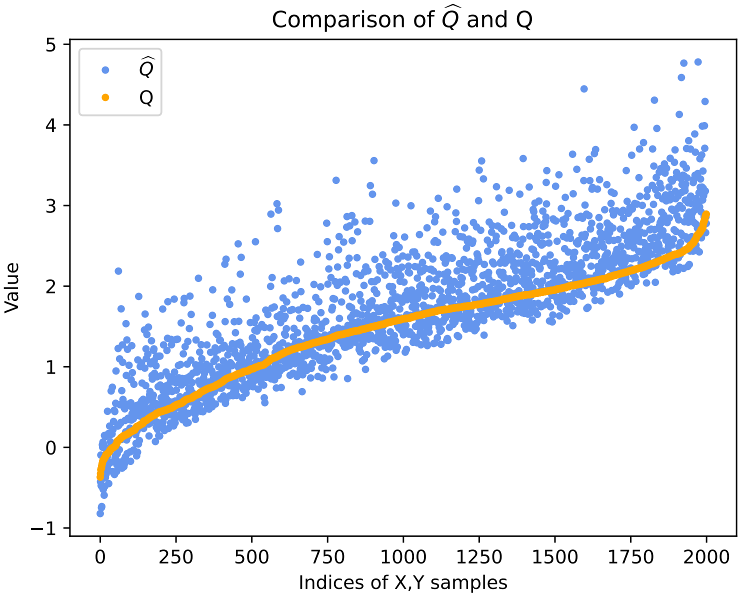


Figure 5.6: Accuracy of the estimated $\widehat{Q}$ values in the case of limited inventory

$\widehat{Q}(X,Y,P) = \mathbb{E}[\widehat{R}\,|\,X,Y,P]$ is by using neural networks, with feeding data $\widehat{R}(X,P,S,\Delta)$ achieved from the following approximation in implementation that

$$\widehat{R}(X,P,S,\Delta) \approx P \times S + c \times \mathbb{1}(\Delta = 0)\sum_{i=k}^{l-1} \frac{\widehat{H}(S_{[i]}\,|\,X,P)(S_{[i+1]} - S_{[i]})}{\widehat{H}(S\,|\,X,P)}, \tag{5.28}$$

where $S_{[k]} = S$ and $S_{[i]}$ denotes the ith order statistic of the unique sales values in the data and ℓ is the number of unique sales values. Note that (5.28) might result in under-estimating $\widehat{\mathbb{E}}\,[D|X,P,S,\Delta = 0]$ due to we let $H(t\,|\,X,P) \approx 0, t \geq S_{[\ell]}$, leading to the over-estimation of $\widehat{R}(X,P,S,\Delta)$ for each record. Therefore, $\widehat{Q}(X,Y,P)$ may be over-estimated given the over-estimation of the training data.

Computing Theoretical Optimal Prices. Unlike the unlimited inventory case, the theoretical optimal prices in the limited inventory case do not enjoy a closed-form expression, due to demand censoring. Thus, we “simulate” the theoretical optimal prices as follows.

(a) First, we generate another independent training dataset with size $N = 50000$ with the true reward $R_j^{\mathbf{TRUE}}(X_j, Y_j, D_j, P_j)$ for each record $j \in [N]$:

$$R_j^{\mathbf{TRUE}}(X_j, Y_j, D_j, P_j) = P_j \times \min\{Y_j, D_j\} - c \times (D_j - Y_j)^+.$$

(b) Second, we use deep neural networks and MLP regressor to obtain $\widehat{Q}\,(X_i, Y_i, p)$ with

$p = [0, 0.001, 0.002, \ldots, 1]$ for each record $i \in [n]$ in the original dataset. Note that the reward used in the MLP regressor for $\widehat{Q}^{(-k(j))}(X_j, Y_j, p)$ is the true reward $R_j^{\mathbf{TRUE}}(X_j, Y_j, D_j, P_j), \forall j \in [N]$ instead of $\widehat{R}_j(X_j, P_j, S_j, \Delta_j)$.

(c) Third, for each record $i \in [n]$, find

$$P_i^* = \underset{p \in \{0, 0.001, 0.002, \ldots, 1\}}{\arg\max} \widehat{Q}(X_i, Y_i, p), \forall i \in [n].$$

Use P_i^* as our estimated theoretical optimal prices for the original dataset, which are presented by the orange dots in Figure 5.7.

Doubly Robust Pricing Strategy. Now we apply Algorithm 10 according to (5.17). Recall that upon computing the estimated potential profit $\widehat{R}_i$, the estimated generalized propensity scores $\widehat{f}^{(-m(i))}(P_i|X_i, Y_i)$, and the estimated Q values $\widehat{Q}^{(-m(i))}(X_i, Y_i, p)$, we compute the optimal pricing strategy via

$$\widehat{\pi}_n \in \underset{\pi \in \Pi_0}{\arg\max} \left\{ \frac{1}{nh} \sum_{i=1}^{n} \int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(X_i, Y_i, p) K\left(\frac{p - \pi(X_i, Y_i)}{h}\right) dp \right.$$
$$\left. + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\widehat{f}^{(-m(i))}(P_i|X_i, Y_i)} K\left(\frac{P_i - \pi(X_i, Y_i)}{h}\right) (\widehat{R}_i - \widehat{Q}^{(-m(i))}(X_i, Y_i, P_i)) - \lambda_n J(\pi) \right\}.$$

Figure 5.7 shows the performance of pricing strategy $\widehat{\pi}_n$, relative to the theoretical optimal prices π^*. The x-axis denotes the indices of samples in the dataset, ranked according to the theoretical optimal prices from low to high. We use translucent dots to represent the prices given by $\widehat{\pi}_n$. The (blue) fitted lines are generated by doing a polynomial fit of degree 4 for these dots. Our doubly robust pricing strategy $\widehat{\pi}_n$ is represented using the (fitted) blue line, and the theoretical optimal price strategy is represented using the orange line. We can see that $\widehat{\pi}_n$ is consistent with the theoretical optimal prices but have a higher variance due to the finite sample data.

Value of Considering Demand Censoring. Figure 5.7 above also depicts the value of factoring in demand censoring. The reasoning is as follows. The variant $\tilde{\pi}_n$ implements our algorithm $\widehat{\pi}_n$ except for the fact that $\tilde{\pi}_n$ hypothetically treated the observed sales quantities as the true demand quantities. In other words, all the procedures are the same except that instead of using $\widehat{R}_i(X_i, P_i, S_i, \Delta_i)$ obtained from survival analysis, we now only use the observed sales profit

$$R_i^{\mathbf{OBS}}(X_i, P_i, S_i, \Delta_i) := P_i \times S_i, \ \forall i \in [n]$$

as the expected reward to estimate $Q(X, Y, P)$.

Figure 5.7 demonstrates that the estimated optimal prices without factoring in demand

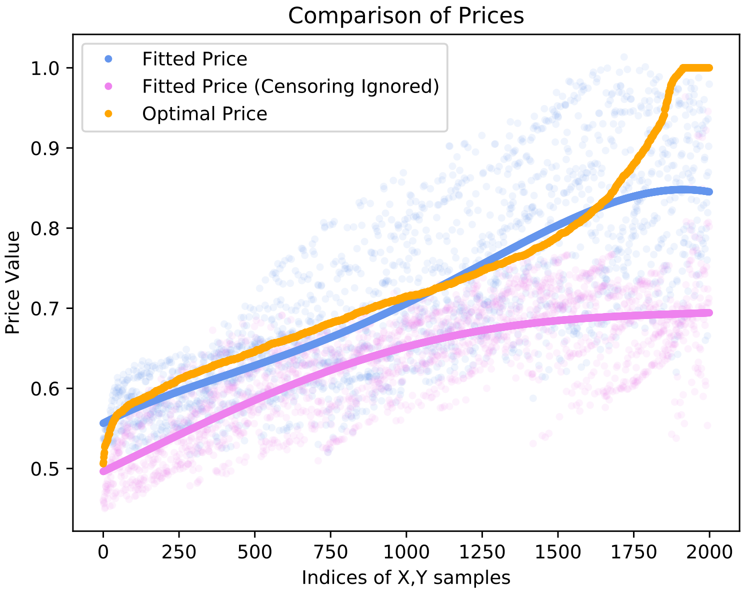


Figure 5.7: Prices given by doubly robust pricing strategy in the case of limited inventory

censoring (represented using the pink line) are significantly below the theoretical optimal prices and also our prices. To make sense of this, imagine a scenario where $Y = 10$ and $D = 100$. In this case, if the manager uses sales $\min(Y, D) = 10$ as true uncensored demand, she may post some moderate price, due to seemingly one-to-one matching between supply and demand. But in fact, the true demand is 100, which is much higher, competing for the same 10 units of inventory, then the manager should significantly increase the prices to gain more potential profit.

In terms of profit performance via the notion of performance ratio defined in (5.25),

$$\kappa(\widehat{\pi}_n) = 96.18\% \ (4.93\text{e}^{-3}) \qquad \text{and} \qquad \kappa(\tilde{\pi}_n) = 92.46\% \ (1.84\text{e}^{-2})$$

where the value in the parenthesis is the standard error of the average performance ratio.

We can see that our algorithm is near-optimal while ignoring the issue of demand censoring could significantly undermine the potential profit, underperforming by more than 4%. Also, the performance is also more stable than that of ignoring the demand censoring, with the standard error being about one-third of the latter one.

Robustness Check. We also test different variants as shown in Table 5.1 based on this model. Recall that n is the number of samples in the dataset, h is the bandwidth in the kernel approximation, and the batch size is the parameter in the neural network of optimizing pricing strategy.

Instance	Sample Size n	Bandwidth h	Batch Size
1	2,000	0.010	10
2	3,000	0.010	10
3	2,000	0.005	10
4	2,000	0.010	20

Table 5.1: Robustness Experiments

Instance 1 is the basic instance whereas we change the sample size in Instance 2, the kernel bandwidth in Instance 3, and the batch size in the neural network of maximizing (5.17) in Instance 4. By changing these parameters, we find that our algorithm is robust and that the results in these variants are consistent and robust. The results of applying our algorithms in these variants are shown in Figures D.1 – D.4 in Appendix D.4.1.

Comparison with Existing Methods. While there are few studies on offline learning under censored demand, Bu et al. (2023) proposed an algorithm named "D2ACD" for a single-product pricing problem based on inventory level under a linear demand model with zero penalty cost for lost demand. We compare the performance of our non-parametric algorithm with theirs under both linear and non-linear demand settings.

Specifically, we run $\widehat{\pi}_n$ and D2ACD (the algorithm proposed by Bu et al. (2023)) on $L = 50$ datasets and apply the pricing strategies obtained on 50 testing datasets respectively. For each training dataset, we have $K = 200$ groups of (y, p) pairs. For each group $i \in [K]$, there are $N_i = 10$ records and $\gamma_i = \mathbb{P}[\xi < y_i + bp_i] > 0$ as assumed in Bu et al. (2023). There are 2000 records of inventory levels in each testing dataset. Following the noise distribution in Bu et al. (2023), we set η as a centered geometric random variable with parameter 1/30. Table 5.2 is the comparison of the performance of the two algorithms.

Demand Function	$\kappa(\widehat{\pi}_n)$	$\kappa(D2ACD)$
$D = 120 - P + \eta$	98.40% (2.63e^{-3})	97.41% (3.16e^{-3})
$D = 120 - \frac{1}{10}P^2 + \eta$	99.34% (0.62e^{-3})	76.63% (8.3e^{-3})

Table 5.2: Comparison of $\widehat{\pi}_n$ with D2ACD

We can see that our algorithm outperforms D2ACD, especially in the nonlinear case. As the D2ACD assumes a linear relationship between demand and price, the results make sense. Our method outperforms Bu et al. (2023) in linear cases due to two potential reasons. First, the application of our method requires a large number of samples with random prices, which differs from the experimental settings in Bu et al. (2023) where the number of groups is relatively small (e.g., $K = 2$) and there are ample samples within each group (e.g., $N_i =$

$100, i \in [K]$). In our setting, the presence of a large number of groups and a small number of records within each group may undermine the advantage of D2ACD for linear cases. Second, while our method relies on the performance of supervised learning techniques for nonparametric estimation of variables, the relatively simple structure of the revenue function, which does not penalize lost sales, combined with multiple samples for the same price and inventory pair, makes it easier for the supervised learning techniques to learn the demand function and achieve favorable performance, even when compared to parametric approaches. The second reason also explains why the performance of our algorithm is better in this setting than the one studied before. Compared with $c = 0.1$ (the penalty for the lost demand) in our previous simulation study, here we do not have the penalty for the lost demand ($c = 0$), leading to the access to the true values of

$$R(X, P, S, \Delta) = P \times S + c \times \mathbb{1}(\Delta = 0) \left(S - \mathbb{E}\left[D | X, P, S, \Delta = 0\right]\right).$$

In addition, the underlying demand functions are also simpler than the previous setting, without covariates. Thus, the estimator $\widehat{Q}(X, Y, P)$ is more accurate and the performance of our algorithm is better than when our algorithm also suffers from the estimation error from the survival analysis for $R(X, P, S, \Delta)$.

5.6 Conclusions

We studied a feature-based pricing problem with demand censoring in an offline data-driven setting, and through the lens of causal inference, we proposed a novel data-driven algorithm that is based on survival analysis and doubly robust estimation. We derived a finite sample regret bound of our offline learning algorithm and also demonstrated the computational efficacy through simulation experiments. We also discussed the value of factoring in demand censoring, by demonstrating numerically that the resulting prices will be significantly lower than the theoretical optimal prices if one simply treats the sales data as the uncensored demand data.

To close this chapter, we would like to point out several promising future research avenues. First, it is interesting to extend our method to the longitudinal or multi-period setting. In practice, we may have longitudinal data over some time periods. In this case, the prices of certain products fluctuate, and we observe sales within each time period. It is promising to develop new models to learn the causal relationship between price and demand given such data. Second, it is worthwhile to consider the competitive setting where other suppliers are offering the same or similar products. In such cases, one may incorporate the competitors' features into the causal inference model to learn the optimal price given other competitors' information. Third, beyond pricing problems, one possible direction is to consider the offline learning problem in inventory control (or newsvendor) settings. One could potentially employ a similar causal inference model to prescribe target inventory levels with feature information. Finally, we believe that investigating the joint pricing and inventory control problem holds promise for future research. This problem introduces additional complexity as demand is influenced by price (and is no longer exogenous). Studying this problem through the lens of causal inference presents another intriguing challenge to investigate.

CHAPTER 6

Conclusion

This dissertation proposes online and offline learning algorithms in OM through four specific application settings. Chapter 2 is focused on designing an online learning algorithm with perturbed gradients to solve multi-product inventory control problems with product upgrading. Chapter 3 proposes an online learning algorithm, integrating bandit control and SAA, to solve dual sourcing systems. Chapter 4 is devoted to designing learning algorithms for two-sided markets. Finally, Chapter 5 develops an offline learning algorithm for feature-based pricing problems with demand censoring.

There are numerous future research directions for online and offline learning in OM.

(a) **Integrating causal inference into the optimization process.** While the literature on causal inference focuses on improving estimation performance, the role of causal inference in efficiently solving optimization problems remains open to study. Integrating causal inference into optimization algorithms for OM problems in a synergistic manner can help better understand the underlying mechanism and thus better decision-making.

(b) **Incorporating fairness constraints into resource allocation and pricing problems.** For instance, dynamic pricing may impact the fairness of customer transactions, and this can be addressed by introducing constraints on price trends. Similar considerations apply to resource allocation challenges, such as labor assignments, where the well-being of the workforce needs to be factored into the model.

(c) **Applications of generative AI in decision making.** While generative AI especially Large Language Models (LLMs) displayed huge potential in text/image generation, the impact of generative AI on decision sciences remains an open field to investigate. Just as the impact of machine learning on decision sciences in the past decade, massive datasets and the leap in computational power will provide brand new interpretations to models and algorithms. Integration of state-of-the-art techniques or the underlying ideas into algorithms in OM is an exciting and potentially fruitful direction.

APPENDIX A

Appendix For Chapter 2

A.1 Summary of Major Notation

Table A.1: Summary of Major Notation in the Problem Formulation

n	the number of demand or supply types considered
T	the number of periods in consideration
r_i	the revenue of satisfying one unit of demand $i, \forall i \in [n]$
c_j	the cost of ordering one unit of supply $j, \forall j \in [n]$
h_j	the cost of holding one unit of supply j for one period, $\forall j \in [n]$
D_i^t	random variable, the quantity of demand i in period $t, \forall i \in [n], t \in [T]$
$\bar{y}_i$	constant, the upper bound for $D_i^t, \forall i \in [n], t \in [T]$
0	constant, the lower bound for $D_i^t, \forall i \in [n], t \in [T]$
d_i^t	the realization of quantity of demand i in period $t, \forall i \in [n], t \in [T]$
x_j^t	the starting inventory level of supply j in period t, $\forall j \in [n], t \in [T]$
y_j^t	the inventory level of supply j after replenishment in period t, $\forall j \in [n], t \in [T]$
u_{ij}^t	the allocation of supply j to demand i with $i \geq j, i, j \in [n], t \in [T]$
y_j^*	the base-stock level vector for supply j in each period in the optimal policy π^*
$\mathbf{Regret}(\pi, T)$	the T-period expected cumulative regret of the policy π
$\mathbf{R}\left(\mathbf{y}\right)$	the expected single period profit with after-replenishment level $\mathbf{y}$ following the allocation policy π^{*A}
$\mathbf{R}\left(\mathbf{y}; \mathbf{d}\right)$	the sample-path single period profit given after-replenishment inventory level $\mathbf{y}$ and demand realization $\mathbf{d}$ following the allocation policy π^{*A}
$\hat{y}_j^t$	the implemented base-stock level for supply j in period $t, \forall j \in [n], t \in [T]$
$J^t\left(\mathbf{x}^t\right)$	the expected optimal profit from period t till the end of horizon with starting inventory level $\mathbf{x}^t$

A.2 Proof of Propositions and Theorems

A.2.1 Proof of Proposition 2.3.1 and Theorem 4.2.1

Before proving Theorem 4.2.1 by induction, we first establish the optimal policy for a single-period allocation problem. Then we study the optimal policy in period T and then analyze

Table A.2: Summary of Major Notation in the SGD-PG Algorithm

$\mathcal{L}$	the ordered set of all demand i and supply j pairs in decreasing order of $r_i - c_j$
l_i^t	the real-time imbalance between type i supply and demand in the second round allocation, $\forall i \in [n], t \in [T]$
$\bar{\mathbf{y}}$	upper bound for the optimal base-stock levels $\mathbf{y}^*$ and the base-stock levels in SGD-PG algorithm, set to be $\bar{y}_i = \bar{y}_i$
ϵ^t	algorithm parameters, $\forall t \in [T]$
γ	constant, a parameter in SGD-PG algorithm
$\mathbf{G}^t(\hat{\mathbf{y}}^t)$	the gradient estimator, a random variable in the algorithm
e_i^t	index of supply(demand) which depletes demand(supply) i in period t
a	temporal index of supply(demand) where a chain ends

period t by induction. Finally, the optimal policy is attained by the assumption that the system starts with zero inventory, as well as the fact that the allocation quantity is non-negative.

We first propose two lemmas below. The first one is the preservation of concavity to facilitate the proof and the second one is the optimal allocation policy in a single-period problem.

Lemma A.2.1 *Consider the following optimization problem with parameter vector* $\mathbf{b} \in \mathbb{R}^m$*, matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and decision variables* $\mathbf{x}^n$*. The objective function* $f(\mathbf{x}, \mathbf{b})$ *is concave and constraints are linear.*

$$\begin{aligned} g(\mathbf{b}) := \max_{\mathbf{x}} \ & f(\mathbf{x}, \mathbf{b}) \\ s.t. \ & \mathbf{A}\mathbf{x} \preceq \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0}, \end{aligned} \qquad (\bar{P}(\mathbf{b}))$$

Then $g(\mathbf{b})$ *is concave in* $\mathbf{b}$.

Lemma A.2.2 *For the following allocation problem* $P^A(\mathbf{y}, \mathbf{d})$ *with inventory vector* $\mathbf{y} \in \mathbb{R}^n$*, demand realization* $\mathbf{d} \in \mathbb{R}^n$*, and the decision variables* $u_{ij}, \forall 1 \leq j \leq i \leq n$.

$$\begin{aligned} \max_{u_{ij}, 1 \leq j \leq i \leq n} \ & \sum_{1 \leq j \leq i \leq n} (r_i - c_j + h_j) u_{ij} \\ s.t. \ & \sum_{i=j}^{n} u_{ij} \leq y_j, \forall j \in [n] \\ & \sum_{j=1}^{i} u_{ij} \leq d_i, \forall i \in [n] \\ & u_{ij} \geq 0, \forall 1 \leq j \leq i \leq n. \end{aligned} \qquad (P^A)$$

Then it is optimal to follow the allocation rule π^{*A} *specified in Theorem* 4.2.1.

Proof of Lemma A.2.1. For any given vectors $\mathbf{b}^1, \mathbf{b}^2$. For $k = 1, 2$, denote $\mathbf{x}^{k*}$ as the optimal solution to the problem, $\bar{P}\left(\mathbf{b}^k\right)$. So $g\left(\mathbf{b}^k\right) = f\left(\mathbf{x}^{k*}\right)$. For any $\lambda \in [0, 1]$, we have $\mathbf{x}' := \lambda \mathbf{x}^{1*} + (1 - \lambda)\mathbf{x}^{2*}$ is a feasible solution to the problem $\bar{P}\left(\lambda\left(\mathbf{b}^1\right) + (1 - \lambda)\left(\mathbf{b}^2\right)\right)$ since the constraints are all linear. By definition of $g(\cdot)$,

$$\begin{aligned} g\left(\lambda\left(\mathbf{b}^1\right) + (1-\lambda)\left(\mathbf{b}^2\right)\right) \geq & f\left(\mathbf{x}', \lambda\left(\mathbf{b}^1\right) + (1-\lambda)\left(\mathbf{b}^2\right)\right) \\ = & f\left(\lambda\mathbf{x}^{1*} + (1-\lambda)\mathbf{x}^{2*}, \lambda\left(\mathbf{b}^1\right) + (1-\lambda)\left(\mathbf{b}^2\right)\right) \\ \geq & \lambda f\left(\mathbf{x}^{1*}, \mathbf{b}^1\right) + (1-\lambda) f\left(\mathbf{x}^{2*}, \mathbf{b}^2\right) \\ = & \lambda g\left(\mathbf{b}^1\right) + (1-\lambda) g\left(\mathbf{b}^2\right), \end{aligned}$$

where the second inequality is due to concavity. Therefore, Lemma A.2.1 holds. **Q.E.D.**

Proof of Lemma A.2.2. The problem can be seen as a multi-knapsack problem with special structures. There are three steps of the proof to show it is optimal to

(I) not allocate supply j to demand i with $r_i - c_j + h_j < 0$;

(II) greedily match pair (i, j) with $r_i - c_j + h_j \geq 0$ in non-decreasing order of the layer index;

(III) determine the order of pairs in the same layer arbitrarily.

Step I: We first show that $u^*_{ij} = 0$ for any pair (i, j) with $r_i - c_j + h_j < 0$ by contradiction. Suppose $u^*_{ij} > 0$, then we can always decrease u^*_{ij} to 0 maintaining feasibility and strictly increasing the profit, contradicting the assumption that it is the optimal solution.

Step II: Then we prove the optimality of the greedy matching by induction of layers 1 to n. For the pairs in layer 1, we prove the optimal allocation policy by showing that if there exists k such that $r_k - c_k + h_k \geq 0$ and $u^*_{kk} < \min(y_k, d_k)$, then we can always increase it by a positive amount without harming the profit. Specifically, when $u^*_{kk} < \min(y_k, d_k)$, there are four cases.

1. $u^*_{ik} = u^*_{kj} = 0$ for all pairs $(i, k), (k, j)$ in layers 2 to n, which means no demand 2 is satisfied or supply 2 is used for the any pair in layers 2 to n. Then the remaining supply k is $y_k - \sum_{i=k}^n u^*_{ik} > 0$ and the unmet demand k is $d_k - \sum_{j=1}^k u^*_{kj} > 0$. So there exists δ such that

$$0 < \delta < \min\left(y_k - \sum_{i=k}^n u^*_{ik}, d_k - \sum_{j=1}^k u^*_{kj}\right).$$

Consider a new feasible solution with $u'_{kk} = u^*_{kk} + \delta$ and others remain the same. The new solution generates no less profit than U^*, since $r_k - c_k + h_k \geq 0$.

2. There exists a supply $j < k$ such that $u^*_{kj} > 0$ but $u^*_{ik} = 0$ for all pairs (i, k) in layer 2 to n, which means demand k is satisfied by some other supply $j < k$ for some pair (k, j) in layers 2 to n, and supply k is not used by any pair in layers 2 to n. So the remaining supply k is $y_k - \sum_{i=k}^{n} u^*_{ik} > 0$ and there exists δ such that

$$0 < \delta < \min\left(y_k - \sum_{i=k}^{n} u^*_{ik}, u^*_{kj}\right).$$

   Consider a new feasible solution with $u'_{kk} = u^*_{kk} + \delta, u'_{kj} = u^*_{kj} - \delta$ and others remain the same. The new solution generates no less profit than U^*, since $r_k - c_k + h_k \geq r_k - c_j + h_j$.

3. There exists a demand $i > k$ such that $u^*_{ik} > 0$ but $u^*_{kj} = 0$ for all pairs (k, j) in layer 2 to n, which means supply k is used to satisfy by some other demand $i > k$ for some pair (i, k) in layers 2 to n, and demand k is not satisfied by any pair in layers 2 to n. So the unmet demand k is $d_k - \sum_{j=1}^{k} u^*_{kj} > 0$ and there exists δ such that

$$0 < \delta < \min\left(d_k - \sum_{j=1}^{k} u^*_{kj}, u^*_{ki}\right).$$

   Consider a new feasible solution with $u'_{kk} = u^*_{kk} + \delta, u'_{ik} = u^*_{ik} - \delta$ and others remain the same. The new solution generates no less profit than U^*, since $r_k - c_k + h_k \geq r_i - c_k + h_k$.

4. There exist a supply $j < k$ such that $u^*_{kj} > 0$ and a demand $i > k$ such that $u^*_{ik} > 0$, which means demand k and supply k are both utilized by some pairs in layers 2 to n. So there exists δ such that

$$0 < \delta < \min\left(u^*_{ik}, u^*_{kj}\right).$$

   (a) If $r_i - c_j + h_j \geq 0$, then consider a new feasible solution with

$$u'_{kk} = u^*_{kk} + \delta,\ u'_{ij} = u^*_{ij} + \delta,\ u'_{kj} = u^*_{kj} - \delta,\ u'_{ik} = u^*_{ik} - \delta.$$

   and others remain the same, i.e., the new solution takes δ pairs (i, k) and (k, j) to form δ pairs (k, k) and (i, j). The new solution generates the same profit than U^*, since $r_k - c_k + h_k + r_i - c_j + h_j = r_i - c_k + h_k + r_k - c_j + h_j$.

(b) If $r_i - c_j + h_j < 0$, then consider a new feasible solution with

$$u'_{kk} = u^*_{kk} + \delta,\ u'_{kj} = u^*_{kj} - \delta,\ u'_{ik} = u^*_{ik} - \delta.$$

and others remain the same, i.e., the new solution takes δ pairs (i,k) and (k,j) to form δ pairs (k,k). The new solution generates more profit than U^*, since $r_k - c_k + h_k > r_i - c_k + h_k + r_k - c_j + h_j$.

In sum, if there exists k such that $u^*_{kk} < \min(y_k, d_k)$, then we can always increase it by a positive amount without harming the objective profit. In this way, it is optimal to set $u^*_{kk} = \min(y_k, d_k)$ for $k \in [n]$. So the greedy matching of pairs (i,j) with $r_i - c_j + h_j$ in layer 1 is optimal.

Now suppose the greedy matching of (i,j) with $r_i - c_j + h_j$ in layer 1 to $m-1$ is optimal. We next show for layer m containing the pairs $\{(m,1), \ldots, (n, n-m+1)\}$, if the matching between (k,l) with $r_k - c_j + h_j \geq 0$ is not greedy, then there are four cases.

1. $u^*_{kj} = u^*_{il} = 0$ for all pairs $(k,j), (i,l)$ in layer $m+1$ to n, which means demand k and supply l is not used by any pair in layers 2 to n. Then the remaining supply l is $y_l - \sum_{i=l}^{n} u^*_{il} > 0$ and the unmet demand k is $d_k - \sum_{j=1}^{k} u^*_{kj} > 0$. So there exists δ such that

$$0 < \delta < \min\left(y_l - \sum_{i=l}^{n} u^*_{il}, d_k - \sum_{j=1}^{k} u^*_{kj}\right).$$

Consider a new feasible solution with $u'_{kl} = u^*_{kl} + \delta$ and others remain the same. The new solution generates no less profit than U^*, since $r_k - c_l + h_l \geq 0$.

2. There exists a pair (k,j) in layers $m+1$ to n such that $u^*_{kj} > 0$ but $u^*_{il} = 0$ for all pairs (i,l) in layer $m+1$ to n, which means demand k is also satisfied by some pair in layers $m+1$ to n and supply l is only satisfied by pairs in layers 1 to m. So the remaining supply l is $y_l - \sum_{i=l}^{n} u^*_{il} > 0$ and there exists δ such that

$$0 < \delta < \min\left(y_l - \sum_{i=l}^{n} u^*_{il}, u^*_{kj}\right).$$

Consider a new feasible solution with $u'_{kl} = u^*_{kl} + \delta, u'_{kj} = u^*_{kj} - \delta$ and others remain the same. The new solution generates no less profit than U^*, since $r_k - c_l + h_l \geq r_k - c_j + h_j$.

3. There exists a pair (i,l) in lays $m+1$ to n such that $u^*_{il} > 0$ but $u^*_{kj} = 0$ for all pairs (k,j) in layer $m+1$ to n, which means supply l is used by some pair in layers 2 to

n and demand k is only satisfied using pairs in layers 1 to m. So the unmet demand $d_k - \sum_{j=1}^{k} u^*_{kj} > 0$ and there exists δ such that

$$0 < \delta < \min\left(d_k - \sum_{j=1}^{k} u^*_{kj}, u^*_{ki}\right).$$

Consider a new feasible solution with $u'_{kl} = u^*_{kl} + \delta, u'_{il} = u^*_{il} - \delta$ and others remain the same. The new solution generates no less profit than U^*, since $r_k - c_l + h_l \geq r_i - c_k + h_k$.

4. There exist pairs (k, j) and (i, l) such that $u^*_{kj} > 0$ and $u^*_{il} > 0$, which means supply l and demand k are both used by some pairs in layers $m + 1$ to n. So there exists δ such that

$$0 < \delta < \min\left(u^*_{il}, u^*_{kj}\right).$$

   (a) If $r_i - c_j + h_j \geq 0$, then consider a new feasible solution with

$$u'_{kl} = u^*_{kl} + \delta,\ u'_{ij} = u^*_{ij} + \delta,\ u'_{kj} = u^*_{kj} - \delta,\ u'_{il} = u^*_{il} - \delta.$$

   and others remain the same, i.e., the new solution takes δ pairs (i, l) and (k, j) to form δ pairs (l, k) and (i, j). The new solution generates the same profit than U^*, since $r_k - c_l + h_l + r_i - c_j + h_j = r_i - c_k + h_k + r_k - c_j + h_j$.

   (b) If $r_i - c_j + h_j < 0$, then consider a new feasible solution with

$$u'_{kl} = u^*_{kl} + \delta,\ u'_{kj} = u^*_{kj} - \delta,\ u'_{il} = u^*_{il} - \delta.$$

   and others remain the same, i.e., the new solution takes δ pairs (i, k) and (k, j) to form δ pairs (k, l). The new solution generates more profit than U^*, since $r_k - c_l + h_l > r_i - c_k + h_k + r_k - c_j + h_j$.

In sum, if the matching between (k, l) with $r_k - c_j + h_j \geq 0$ is not greedy, then we can always increase the matching of pair (k, l) by a positive amount without harming the profit. In this way, it is optimal to greedily match pair (i, j) with $r_i - c_j + h_j \geq 0$ in layer m.

Step III: Finally, the allocation order in layer 1 obviously does not matter. For any layer $m \geq 2$, there are at most two pairs in the same layer with supply or demand being k for any $k \in [n]$. Note that following the greedy allocation policy in layer 1, supply and demand k cannot simultaneously have a positive quantity when it comes to layer $m \geq 2$. So we only need to consider one of the two pairs in the same layer and the order of allocation within the same layer also does not matter.

Therefore, the allocation policy is π^{*A} as described in Theorem 4.2.1 is optimal for this optimization problem. **Q.E.D.**

Now we are ready to prove Theorem 4.2.1. We have the following lemma.

Lemma A.2.3 *For any period $t \in [T]$,*

(A) $J^t(\mathbf{x}^t)$ is concave in $\mathbf{x}^t$ and $V^t(\mathbf{y}^t)$ is concave in $\mathbf{y}^t$;

(B) when $\mathbf{x}^t \preceq \mathbf{y}^*$, it is optimal to order up to $\mathbf{y}^*$ and follow the allocation rule π^{*A} in Theorem 4.2.1.

Proof of Lemma A.2.3. We prove the results by induction. First, when $t = T$,

$$\begin{aligned}
J^T(\mathbf{x}^T) &= \max_{\mathbf{y}^T \succeq \mathbf{x}^T} V^T\left(\mathbf{y}^T\right) + \sum_{i=1}^n c_i x_i^T, \\
V^T\left(\mathbf{y}^T\right) &= -\sum_{j=1}^n c_j y_j^T + \mathbb{E}_{\mathbf{D}^T} \max_{\substack{\sum_{i=j}^n u_{ij}^T \le y_j^T, \forall j \in [n] \\ \sum_{j=1}^i u_{ij}^T \le D_i^T, \forall i \in [n] \\ u_{ij}^T \ge 0, 1 \le j \le i \le n}} \left[\sum_{j=1}^n \left(\sum_{i=j}^n r_i u_{ij}^T + (c_j - h_j)\left(y_j^T - \sum_{i=j}^n u_{ij}^T\right)\right)\right] \\
&= -\sum_{j=1}^n h_j y_j^T + \mathbb{E}_{\mathbf{D}^T} \max_{\substack{\sum_{i=j}^n u_{ij}^T \le y_j^T, \forall j \in [n] \\ \sum_{j=1}^i u_{ij}^T \le D_i^T, \forall i \in [n] \\ u_{ij}^T \ge 0, 1 \le j \le i \le n}} \sum_{1 \le j \le i \le n} (r_i - c_j + h_j)\, u_{ij}^T \qquad \text{(Eq } V^T\text{)} \\
&:= -\sum_{j=1}^n h_j y_j^T + \mathbb{E}_{\mathbf{D}^T} \mathbf{Q}\left(\mathbf{y}^T, \mathbf{D}^T\right).
\end{aligned}$$

The objective function of the optimization problem in (Eq V^T) is linear and thus concave. By Lemma A.2.1, we have $\mathbf{Q}\left(\mathbf{y}^T, \mathbf{d}^T\right)$ is concave in $\mathbf{y}^T$ for any realized $\mathbf{d}^T$. Then after taking expectation, $\mathbb{E}_{\mathbf{D}^T}\mathbf{Q}\left(\mathbf{y}^T, \mathbf{D}^T\right)$ is concave in $\mathbf{y}^T$. Therefore, $V^T\left(\mathbf{y}^T\right)$ is concave in $\mathbf{y}^T$. Since $\{\mathbf{y}^T : \mathbf{y}^T \succeq \mathbf{x}^T\}$ is a convex set, the maximization preserves concavity and $J^T(\mathbf{x}^T)$ is concave in $\mathbf{x}^T$. So Lemma A.2.3A holds for $t = T$.

Let's consider allocation decisions, which is exactly the problem $P^A\left(\mathbf{y}^T, \mathbf{d}^T\right)$ specified in Lemma A.2.2. So it is optimal to follow π^{*A} for the allocation decisions. Now

$$\begin{aligned}
V^T\left(\mathbf{y}^T\right) =& \mathbb{E}_{D^T} \sum_{j=1}^n \left(-c_j y_j + \sum_{i=j}^n r_i u_{ij}^*\left(\mathbf{y}^T, \mathbf{D}^T\right) + (c_j - h_j)\left(y_j - \sum_{i=j}^n u_{ij}^*\left(\mathbf{y}^T, \mathbf{D}^T\right)\right)\right) \\
=& \mathbf{R}\left(\mathbf{y}^T\right).
\end{aligned}$$

So $V^T\left(\mathbf{y}^T\right)$ is maximized at $\mathbf{y}^*$ and when $\mathbf{x}^T \preceq \mathbf{y}^T$, it is optimal to order up to $\mathbf{y}^*$. So

Lemma A.2.3B, i.e., the optimality of π^{*A} and π^{*R} holds for period T.

Now suppose Lemma A.2.3 holds for period $t+1$.Then we consider period t.

$$\begin{aligned}
J^t(\mathbf{x}^t) &= \max_{\mathbf{y}^t \succeq \mathbf{x}^t} V^t\left(\mathbf{y}^t\right) + \sum_{i=1}^n c_i x_i^t, \\
V^t\left(\mathbf{y}^t\right) &= -\sum_{j=1}^n c_j y_j^t + \mathbb{E}_{\mathbf{D}^t} \max_{\substack{\sum_{i=j}^n u_{ij}^t \le y_j^t, \forall j \in [n] \\ \sum_{j=1}^i u_{ij}^t \le D_i^t, \forall i \in [n] \\ u_{ij}^t \ge 0, 1 \le j \le i \le n}} \left[\sum_{j=1}^n \left(\sum_{i=j}^n r_i u_{ij}^t - h_j \left(y_j^t - \sum_{i=j}^n u_{ij}^t \right) \right) + J^{t+1}\left(\mathbf{y}^t - U^{t\intercal}\mathbf{1}\right) \right] \\
&:= -\sum_{j=1}^n c_j y_j^t + \mathbb{E}_{\mathbf{D}^t} \mathbf{H}^t\left(\mathbf{y}^t, \mathbf{D}^t\right). \qquad \text{(Eq } V^t\text{)}
\end{aligned}$$

By induction hypothesis, $J^{t+1}\left(\mathbf{x}^{t+1}\right)$ is concave in $\mathbf{x}^{t+1}$. Define $\mathbf{u}^t := \left[u_{11}^t, u_{21}^t, \ldots, u_{n(n-1)}^t\right]$, i.e., the elements of $\mathbf{u}^t$ is $u_{ij}^t, 1 \le j \le i \le n$ in non-decreasing order of j and in non-decreasing order of i within the same j index. So $\mathbf{x}^{t+1} = \mathbf{y}^t - U^{t\intercal}\mathbf{1}$ is an affine transformation of vector $(\mathbf{y}^t, \mathbf{u}^t)$, which preserves concavity. Therefore, for any realized $\mathbf{d}^t$, the objective of the optimization problem in (Eq V^t) is concave. By Lemma A.2.1, we have $\mathbf{H}^t\left(\mathbf{y}^t, \mathbf{D}^t\right)$ is concave in $\left(\mathbf{y}^t, \mathbf{D}^t\right)$. After taking expectation with $\mathbf{D}^t$, we have $V^t\left(\mathbf{y}^t\right)$ is concave in $\mathbf{y}^t$. Since $\{\mathbf{y}^t : \mathbf{y}^t \succeq \mathbf{x}^t\}$ is a convex set, the maximization preserves concavity and $J^t(\mathbf{x}^t)$ is concave in $\mathbf{x}^t$. So Lemma A.2.3A holds for period t.

If $\mathbf{x}^t \preceq \mathbf{y}^*$, $\mathbf{x}^{t+1} \preceq \mathbf{y}^*$ since non-negative quantity of supply will be allocated to the demand in period t. By induction hypothesis, it is optimal to order up to $\mathbf{y}^*$ in period $t+1$. Therefore, we have

$$J^{t+1}(\mathbf{x}^{t+1}) = V^{t+1}\left(\mathbf{y}^*\right) + \sum_{i=1}^n c_i x_i^{t+1}.$$

Thus for period t,

$$\begin{aligned}
V^t\left(\mathbf{y}^t\right) &= -\sum_{j=1}^n c_j y_j^t + \mathbb{E}_{\mathbf{D}^t} \max_{\substack{\sum_{i=j}^n u_{ij}^t \le y_j^t, \forall j \in [n] \\ \sum_{j=1}^i u_{ij}^t \le D_i^t, \forall i \in [n] \\ u_{ij}^t \ge 0, 1 \le j \le i \le n}} \left[\sum_{j=1}^n \left(\sum_{i=j}^n r_i u_{ij}^t + (c_j - h_j) \left(y_j^t - \sum_{i=j}^n u_{ij}^t \right) \right) \right] + V^{t+1}\left(\mathbf{y}^*\right) \\
&= -\sum_{j=1}^n h_j y_j^t + \mathbb{E}_{\mathbf{D}^t} \max_{\substack{\sum_{i=j}^n u_{ij}^t \le y_j^t, \forall j \in [n] \\ \sum_{j=1}^i u_{ij}^t \le D_i^t, \forall i \in [n] \\ u_{ij}^t \ge 0, 1 \le j \le i \le n}} \sum_{1 \le j \le i \le n} \left(r_i - c_j + h_j\right) u_{ij}^t + V^{t+1}\left(\mathbf{y}^*\right).
\end{aligned}$$

For any realized demand $\mathbf{d}^t$, the allocation problem is exactly $P^A(\mathbf{y}^t, \mathbf{d}^t)$ specified in Lemma

A.2.2. So it is optimal to follow π^{*A} for the allocation decisions. Now

$$\begin{aligned}V^t\left(\mathbf{y}^t\right) =&\mathbb{E}_{D^t}\sum_{j=1}^{n}\left(-c_j y_j+\sum_{i=j}^{n} r_i u^*_{ij}\left(\mathbf{y}^t,\mathbf{D}^t\right)+\left(c_j-h_j\right)\left(y_j-\sum_{i=j}^{n} u^*_{ij}\left(\mathbf{y}^t,\mathbf{D}^t\right)\right)\right)+V^{t+1}\left(\mathbf{y}^*\right)\\ =&\mathbf{R}\left(\mathbf{y}^t\right)+V^{t+1}\left(\mathbf{y}^*\right).\end{aligned}$$

So $V^t\left(\mathbf{y}^t\right)$ is maximized at $\mathbf{y}^*$ and when $\mathbf{x}^t \preceq \mathbf{y}^t$, it is optimal to order up to $\mathbf{y}^*$. So Lemma A.2.3B, i.e., the optimality of π^{*A} and π^{*R} holds for period t. **Q.E.D.**

If the system starts with zero inventory, then $\mathbf{x}^1 \preceq \mathbf{y}^*$. Then following the optimal policy indicated in Lemma A.2.3, $\mathbf{x}^t \preceq \mathbf{y}^*$ for any period t and Theorem 4.2.1 is proven.

A.2.2 Proof of Proposition 2.4.1

The main idea is to compute the profit change due to perturbing the inventory level of supply i by an infinitesimal amount δ. The n-dimensional vector $\mathbf{e}^t$ keeps track of the terminal state of each product $i \in [n]$. This vector is initialized to be value M. At the end of period t, there are three possible index values for each e_i^t: If $e_i^t = M$, it means that there is either unmet demand i or excess supply i. Otherwise, $e_i^t = j$ when unmet demand i is depleted by supply j or unmet supply i is depleted by demand j. The while loop identifies where the chain ends.

Let l_i^t denote the real-time imbalance between the supply and demand of product i. At the end of period t, $(l_i^t)^+$ denotes excess supply i and $(l_i^t)^-$ denotes unmet demand i.

1. If $l_i^t = 0$, there are two possible cases after perfect matching:

   (a) There is excess supply i after perfect matching (e.g., supply 3 in Figure 2.6).

   (b) There is unmet demand i after perfect matching (e.g., demand 4 in Figure 2.6).

   For either case, the influence of increasing the inventory level by δ is the same, as long as the terminal node of the chain is the same. Because all the supply and demand nodes in the middle of the chain will be depleted, the only difference takes place at the terminal node a.

   (a) If $l_a^t > 0$, which means the chain ends at excess supply a, then there will be δ more supply a carried over to the next period. Thus, the profit change is $\delta(-c_i+c_a-h_a)$.

   (b) If $l_a^t < 0$, which means the chain ends at unmet demand a, then there will be δ less unmet demand a. Thus, the profit change is $\delta(-c_i+r_a)$. (This is the case of $y_3^t \to y_3^t+\delta$ or the case of $y_4^t \to y_4^t+\delta$ in Figure 2.6.)

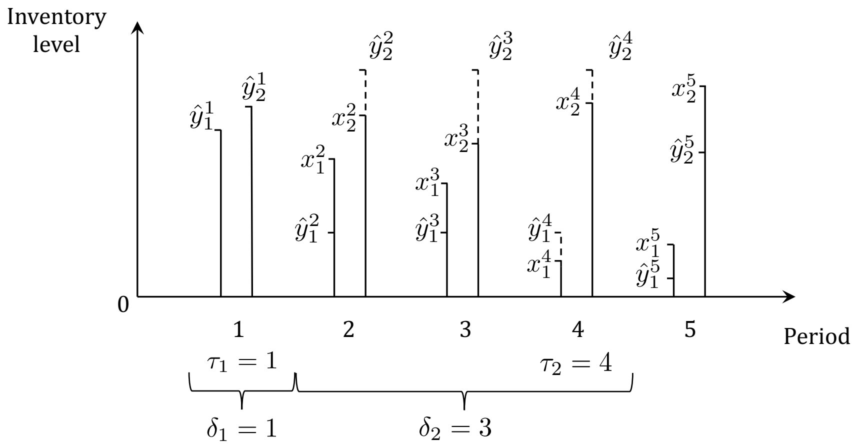


Figure A.1: One Sample Path with $n = 2$

2. If $l_i^t > 0$, there is still excess supply i at the end of period t. Hence, increasing supply i by δ will result in δ more excess supply i. Thus, the profit change is $\delta(-c_i + c_i - h_i) = \delta(-h_i)$. (This is the case of $y_6^t \to y_6^t + \delta$ in Figure 2.6.)

3. If $l_i^t < 0$, there is still unmet demand i at the end of period t. Hence, increasing supply i by δ will result in δ more demand i met. Thus, the profit change is $\delta(-c_i + r_i)$. (This is the case of $y_1^t \to y_1^t + \delta$ in Figure 2.6).

This exhausts all possible cases, thereby proving the clhuh2009nonparametric.

A.2.3 Proof of Theorem 2.4.1

Proof of Lemma 2.4.1.

$$\begin{aligned}\mathbf{R}(\mathbf{y}) &= -\sum_{j=1}^{n} c_j y_j + \mathbb{E}_{\mathbf{D}} \max_{\substack{\sum_{i=j}^{n} u_{ij} \le y_j, \forall j \in [n] \\ \sum_{j=1}^{i} u_{ij} \le d_i, \forall i \in [n] \\ u_{ij} \ge 0, 1 \le j \le i \le n}} \left[\sum_{j=1}^{n} \left(\sum_{i=j}^{n} r_i u_{ij} + (c_j - h_j) \left(y_j - \sum_{i=j}^{n} u_{ij} \right) \right) \right] \\ &= -\sum_{j=1}^{n} h_j y_j + \mathbb{E}_{\mathbf{D}} \mathbf{Q}(\mathbf{y}, \mathbf{d}).\end{aligned}$$

By Lemma A.2.1, we have $\mathbf{Q}(\mathbf{y}, \mathbf{D})$ is concave in $\mathbf{y}$ for any realized $\mathbf{d}$. Then after taking expectation, $\mathbb{E}_{\mathbf{D}} \mathbf{Q}(\mathbf{y}, \mathbf{D})$ is concave in $\mathbf{y}$. **Q.E.D.**

Before we prove the upper bound of $\Lambda_1(T)$ and $\Lambda_2(T)$, we first study the dynamics of the system. First, we establish a result that the length of an epoch is a sub-exponential random variable.

Lemma A.2.4 *Denote D as a non-negative random variable and $\mathbb{E}[D] = \alpha > 0$. Define $m := \max\{l : \sum_{i=0}^{l} D_i \leq M\}$. Then for any $M \in \mathbb{R}$, there exists non-negative numbers (ν, η) such that*

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\frac{\nu^2 \lambda^2}{2}}, \ \forall \, |\lambda| < \frac{1}{\eta}.$$

Proof of Lemma A.2.4. Fix $\bar{D} > 0$. Define another variable

$$D' := \begin{cases} D, & \text{if } D \leq \bar{D}, \\ \bar{D}, & \text{otherwise.} \end{cases}$$

Denote $\mathbb{E}[D'] = \alpha' > 0$. We have $D' \leq D$ almost surely, so

$$\mathbb{P}\left[\sum_{i=1}^{n} D_i \leq M\right] \leq \mathbb{P}\left[\sum_{i=1}^{n} D_i' \leq M\right].$$

If $M - n\alpha' < 0$, then by Hoeffding's inequality.

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^{n} D_i' \leq M\right] &= \mathbb{P}\left(\sum_{i=1}^{n} D_i' - n\alpha' \leq M - n\alpha'\right) \\ &\leq e^{-\frac{2(n\alpha' - M)^2}{n\bar{D}^2}} \\ &= e^{\frac{M^2 - 4\alpha' M}{\bar{D}^2}} \cdot e^{-\frac{2\alpha'}{\bar{D}^2} n}. \end{aligned}$$

If $M - n\alpha' \geq 0$, i.e., $n \leq \frac{M}{\alpha'}$

$$\mathbb{P}\left[\sum_{i=1}^{n} D_i' \leq M\right] \leq 1 \leq e^{\frac{2M\alpha'}{\bar{D}^2}} \cdot e^{-\frac{2\alpha'^2}{\bar{D}^2} n}.$$

Define $\xi_1 := \max\{e^{\frac{M^2 - 4\alpha' M}{\bar{D}^2}}, e^{\frac{2M\alpha'}{\bar{D}^2}}\}$ and $\xi_2 := \frac{2\alpha_i^2}{\bar{D}^2}$. So for any $s > 0$,

$$\begin{aligned} \mathbb{P}(m \geq s) =& \mathbb{P}\left[\sum_{t=\tau_{k-1}}^{\tau_{k-1} + \lceil s \rceil} D_i^t \leq \bar{y}_i\right] \leq \max\{e^{\frac{M^2 - 4\alpha' M}{\bar{D}^2}}, e^{\frac{2M\alpha'}{\bar{D}^2}}\} \cdot e^{-\frac{2\alpha_i^2}{\bar{D}^2}\lceil s \rceil} \\ \leq& \xi_1 + e^{-\xi_2 \lceil s \rceil} \leq \xi_1 + e^{-\xi_2 s}. \end{aligned}$$

By Theorem 2.13 in Wainwright (2019), m is sub-exponential. So there exists non-negative numbers (ν, η) such that

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\frac{\nu^2\lambda^2}{2}},\ \forall\, |\lambda| < \frac{1}{\eta}.$$

Q.E.D.

Then we denote the period when $x^t \leq \hat{y}^t$ as $\tau_1, \tau_2, \cdots, \tau_k$, which means starting inventory levels do not exceed the target levels and let $\tau_0 = 0, \tau_{k+1} = T$. Since the system starts with zero inventory, we have $\tau_1 = 1$. By definition, for any $\tau_l < t < \tau_{l+1}$, there exists some $j \in [n]$ such that $x_j^t > \hat{y}_j^t$. in addition, we denote the length between two such periods as $\delta_k := \tau_k - \tau_{k-1}, k = 1, \cdots, K+1$ as shown in Figure A.1. Recall that in Algorithm 1, the order-up-to levels will not be updated until all inventory levels drop below the target inventory levels. In Figure A.1, the order up to levels are updated in period 1 and 4 waiting for the inventory level of product 1 to drop below.

We also define an auxiliary variable Δ_k which is the length of epoch k is upper bounded by the length of time it takes for product i to drop $\bar{y}_i$ for all $i \in [n]$ *without upgrading.* Then $\delta_k \leq \Delta_k$ almost surely, since $x_i^{t+1} \leq y_i^t - d_i^t$ by the allocation rule that the supply i is first used to satisfy demand i as much as possible and then used to satisfy other demand if possible. We then clhuh2009nonparametric that the expected length of an epoch is of a constant order.

Lemma A.2.5 *For any given K, with $\delta_k = \tau_k - \tau_{k-1}, k = 1, \cdots, K+1$, we have*

$$\mathbb{E}\left[\delta_k - 1\right] \leq O\left(\log n\right),$$

where n is the number of products.

Proof of Lemma A.2.5. Now we consider $m_i^k := \max\left\{l : \sum_{t=\tau_{k-1}}^{\tau_{k-1}+l} d_i^t \leq \bar{y}_i\right\}$, which is the maximum number of periods whose cumulative demand does not exceed the upper bound of the order-up-to level of product i, i.e., $\bar{y}_i$. So $m_i^k + 1 = \min\left\{l : \sum_{t=\tau_{k-1}}^{\tau_{k-1}+l} d_i^t > \bar{y}_i\right\}$ which is the number of periods it takes for the inventory level of product i to drop $\bar{y}_i$ and $\Delta_k = \max_{i\in[n]} m_i^k + 1$.

$$\mathbb{E}\left[\delta_k - 1\right] \leq \mathbb{E}\left[\Delta_k - 1\right] = \mathbb{E}\left[\max_{i\in[n]} m_i^k\right]. \tag{A.1}$$

Note that m_i^k is a renewal process with inter-arrival times $D_i^{\tau_{k-1}}, D_i^{\tau_{k-1}+1}, \ldots$. Define $K_i := \frac{\bar{y}_i}{\alpha_i} + \frac{\sigma_i^2+\alpha_i^2}{\alpha_i^2}, \forall i \in [n]$ where $\alpha_i = \mathbb{E}\left[D_i\right]$ and $\sigma_i^2 = Var\left[D_i\right]$ defined in Assumption 2.2.1. We have

$$\mathbb{E}\left[m_i^k\right] \leq \frac{\bar{y}_i}{\mathbb{E}\left[D_i\right]} + \frac{\mathbb{E}\left[D_i^2\right]}{\left(\mathbb{E}\left[D_i\right]\right)^2} = K_i, \tag{A.2}$$

where the inequality is by Lorden's inequality in Asmussen (2003) Proposition 6.2.

In addition, by Lemma A.2.4, for any $i \in [n], k \in [K+1]$, there exists non-negative numbers (ν_i, η_i) such that

$$\mathbb{E}\left[e^{\lambda m_i^k}\right] \leq e^{\frac{\nu_i^2 \lambda^2}{2}}, \ \forall \left|\lambda\right| < \frac{1}{\eta_i}. \tag{A.3}$$

Let $\lambda = \min_{i\in[n]}\left(\frac{1}{2\eta_i}\right)$. We have

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[n]} m_i^k\right] =& \frac{1}{\lambda}\mathbb{E}\left[\log e^{\lambda \max_{i\in[n]} m_i^k}\right] \leq \frac{1}{\lambda}\log\mathbb{E}\left[e^{\lambda \max_{i\in(n]} m_i^k}\right] \\ \leq& \frac{1}{\lambda}\log\sum_{i\in[n]}\mathbb{E}\left[e^{\lambda m_i^k}\right] = \frac{1}{\lambda}\log\sum_{i\in[n]}\mathbb{E}\left[e^{\lambda\left(m_i^k - \mathbb{E}\left[m_i^k\right]\right)}\cdot e^{\mathbb{E}\left[m_i^k\right]}\right] \qquad \text{(A.4)} \\ \leq& \frac{1}{\lambda}\log\sum_{i\in[n]}\mathbb{E}\left[e^{\lambda\left(m_i^k - \mathbb{E}\left[m_i^k\right]\right)}\cdot e^{K_i}\right] = \frac{1}{\lambda}\log\sum_{i\in[n]} e^{K_i}\mathbb{E}\left[e^{\lambda\left(m_1^k - \mathbb{E}\left[m_i^k\right]\right)}\right] \\ \leq& \frac{1}{\lambda}\log\sum_{i\in[n]} e^{K_i + \frac{\nu_i^2\lambda^2}{2}} \leq O\left(\log n\right), \qquad \text{(A.5)}\end{aligned}$$

where the inequality in (A.4) is due to (A.2) and the inequality in (A.5) is due to (A.3). Combined with inequality (A.1), we have $\mathbb{E}\left[\delta_k - 1\right] \leq O\left(\log n\right)$ holds. **Q.E.D.**

Now we are ready to bound the two terms respectively.

A.2.4 Bound of $\Lambda_1(T)$.

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbf{R}\left(\mathbf{y}^*\right) - \mathbf{R}\left(\hat{\mathbf{y}}^t\right)\right)\right] =& \mathbb{E}\left[\sum_{k=0}^{K}\sum_{t=\tau_k+1}^{\tau_{k+1}}\left[\mathbf{R}\left(\mathbf{y}^*\right) - \mathbf{R}\left(\hat{\mathbf{y}}^t\right)\right]\right] \\ =& \mathbb{E}\left[\sum_{k=1}^{K+1}\delta_k\left[\mathbf{R}\left(\mathbf{y}^*\right) - \mathbf{R}\left(\hat{\mathbf{y}}^{\tau_k}\right)\right]\right] \\ \leq& \mathbb{E}_K\left[\sum_{k=1}^{K+1}\Delta_k\left(\mathbf{R}\left(\mathbf{y}^*\right) - \mathbf{R}\left(\mathbf{y}^{\tau_k}\right)\right) \mid K\right],\end{aligned}$$

where the inequality is due to $\delta_k \leq \Delta_k$ almost surely. Since Δ_k depends on demand after period τ_{k-1}, while $\hat{\mathbf{y}}^{\tau_k} = \hat{\mathbf{y}}^{\tau_{k-1}+1}$ depends on demand before (including) period τ_{k-1}, we have Δ_k and $\hat{\mathbf{y}}^{\tau_k}$ are independent. Thus

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbf{R}\left(\mathbf{y}^*\right)-\mathbf{R}\left(\hat{\mathbf{y}}^*\right)\right)\right] \leq & \mathbb{E}_K\left[\sum_{k=1}^{K+1}\mathbb{E}\left[\Delta_k\right]\mathbb{E}\left[\mathbf{R}\left(\mathbf{y}^*\right)-\mathbf{R}\left(\hat{\mathbf{y}}^{\tau_k}\right)\right] \mid K\right]\\ \leq & (C_0\log n+1)\mathbb{E}\left[\sum_{k=1}^{K+1}\left(\mathbf{R}\left(\mathbf{y}^*\right)-\mathbf{R}\left(\hat{\mathbf{y}}^{\tau_k}\right)\right)\right],\end{aligned}$$

for some constant C_0 by Lemma A.2.5. By Lemma 2.4.1, we have that for each period t, $\Phi\left(\mathbf{y}^t\right) := -\mathbf{R}\left(\mathbf{y}^t\right)$ is convex in $\mathbf{y}^t$. By Lemma 2.4.1, $\mathbf{G}^t\left(\hat{\mathbf{y}}^t\right)$ is a random vector where its realization is a valid gradient of $\mathbf{R}\left(\hat{\mathbf{y}}^t;\mathbf{d}^t\right)$ under the demand realization $\mathbf{d}^t$. Hence, taking the expectation over all sample paths of demand, $\mathbb{E}_D\left[-\mathbf{G}^t\left(\hat{\mathbf{y}}^t\right) \mid \hat{\mathbf{y}}^t\right] = \nabla\Phi\left(\hat{\mathbf{y}}^t\right)$.

The following lemma is a standard result in online convex optimization.

Lemma A.2.6 (Theorem 3.4 in Hazan et al. (2016)) *Let f be a convex function on a bounded convex and compact set $\mathcal{K}$ with an upper bound on the diameter D. We denote by $G > 0$ an upper bound on the norm of the subgradients of f over $\mathcal{K}$, i.e., $\|\nabla f(\mathbf{x})\| \leq G$ for all $\mathbf{x} \in \mathcal{K}$. For $t = 1,\ldots,T$, we update $\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t\tilde{\nabla}_t, \mathbf{x}_{t+1} = \prod_{\mathcal{K}}\left(\mathbf{y}_{t+1}\right)$ where $\mathbf{E}\left[\tilde{\nabla}_{\mathbf{x}}\right] = \nabla f(\mathbf{x})$. Then*

$$\mathbf{E}\left[\frac{1}{T}\sum_t f\left(\mathbf{x}_t\right)\right] \leq \min_{\mathbf{x}^\star\in\mathcal{K}} f\left(\mathbf{x}^\star\right) + \frac{3GD}{2\sqrt{T}}$$

with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$.

We can readily apply Lemma A.2.6 to $\Phi\left(\mathbf{y}^t\right)$ which is a convex function with $S = [0,\bar{y}]^n$ being a convex set. Let $\bar{\mathbf{y}} = \bar{y}\mathbf{1}$ and $\boldsymbol{\theta} = \max\{r_1 - c_n, c_1 - c_n + h\}\mathbf{1}$. Because $c_1 - r_n \leq h$ and $c_1 - c_n \geq 0$,

$$\begin{aligned}-\mathbf{G}_i^t(\hat{\mathbf{y}}^t) &\leq \max\{c_i - c_a + h, -r_a + c_i, h, -r_i + c_i\}\\ &\leq \max\{c_1 - c_n + h, -r_n + c_1, h, 0\} \leq c_1 - c_n + h.\end{aligned}$$

Also, by $r_i - c_i + h \geq 0$, we have

$$\begin{aligned}\mathbf{G}_i^t(\hat{\mathbf{y}}^t) &\leq \max\{-c_i + c_a - h, r_a - c_i, -h, r_i - c_i\}\\ &\leq \max\{-c_n + c_1 - h, r_1 - c_n, -h\} \leq r_1 - c_n.\end{aligned}$$

Thus, $\|-\mathbf{G}\left(\mathbf{y}\right)\| \leq \|\boldsymbol{\theta}\|$ for all $\mathbf{y}$ according to Algorithm 2. We have shown that

$\mathbb{E}\left[-\mathbf{G}_i\left(\mathbf{y}^t\right)|y_i\right] = \nabla_i \Phi\left(\mathbf{y}^t\right),\ \forall i \in [n]$. Thus, we have that for all $T \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbf{R}\left(y^*\right)-\mathbf{R}\left(\hat{y}^*\right)\right)\right] =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\Phi\left(\hat{\mathbf{y}}^t\right)-\Phi\left(\mathbf{y}^*\right)\right)\right] \\ &\leq \left(C_0 \log n + 1\right)\mathbb{E}\left[\sum_{k=1}^{K+1}\left(\Phi\left(\hat{\mathbf{y}}^{\tau_k}\right)-\Phi\left(\mathbf{y}^*\right)\right)\right] \leq O\left(\sqrt{T}\right). \qquad \text{(A.6)}\end{aligned}$$

A.2.5 Bound of $\Lambda_2(T)$.

We first show the difference between the expected profit can be upper bounded by differences in y_i^t and $\hat{y}_i^t$.

Lemma A.2.7 *For any $t \in [T]$, the difference in expected profit satisfies*

$$\mathbf{R}\left(\hat{\mathbf{y}}^t\right)-\mathbf{R}\left(\mathbf{y}^t\right) \leq \sum_{j=1}^{n} h_j\left(y_j^t - \hat{\mathbf{y}}_j^t\right).$$

Proof of Lemma A.2.7. By definition,

$$\begin{aligned}\mathbf{R}\left(\hat{\mathbf{y}}^t\right)-\mathbf{R}\left(\mathbf{y}^t\right) =& -\sum_{j=1}^{n} h_j \hat{y}_j^t + \mathbb{E}_{\mathbf{D}^t} \max_{\substack{\sum_{i=j}^{n} u_{ij}^t \leq \hat{y}_j^t, \forall j \in [n] \\ \sum_{j=1}^{i} u_{ij}^t \leq D_i^t, \forall i \in [n] \\ u_{ij}^t \geq 0, 1 \leq j \leq i \leq n}} \sum_{1 \leq j \leq i \leq n}\left(r_i - c_j + h_j\right) u_{ij}^t \\ &+ \sum_{j=1}^{n} h_j y_j^t - \mathbb{E}_{\mathbf{D}^t} \max_{\substack{\sum_{i=j}^{n} u_{ij}^t \leq y_j^t, \forall j \in [n] \\ \sum_{j=1}^{i} u_{ij}^t \leq D_i^t, \forall i \in [n] \\ u_{ij}^t \geq 0, 1 \leq j \leq i \leq n}} \sum_{1 \leq j \leq i \leq n}\left(r_i - c_j + h_j\right) u_{ij}^t \\ =& \sum_{j=1}^{n} h_j\left(y_j^t - \hat{y}_j^t\right) + \mathbb{E}_D\left[OPT\left(\hat{\mathbf{y}}^t, D\right) - OPT\left(\mathbf{y}^t, D\right)\right],\end{aligned}$$

where $OPT\left(\mathbf{y}, D\right)$ is the optimal value of the following problem $P(\mathbf{y}, D)$.

$$\begin{aligned}OPT\left(\mathbf{y}, \mathbf{d}\right) := \max & \sum_{1 \leq j \leq i \leq n}\left(r_i - c_j + h_j\right) u_{ij} \qquad \left(P_0\left(\mathbf{y}, \mathbf{d}\right)\right) \\ \text{s.t.} & \sum_{i=j}^{n} u_{ij} \leq \left(y_j - D_j\right)^+, \forall j \in [n] \\ & \sum_{j=1}^{i} u_{ij} \leq \left(D_i - y_i\right)^+, \forall i \in [n] \\ & u_{ij} \geq 0, 1 \leq j \leq i \leq n.\end{aligned}$$

Note that for any demand realization $\mathbf{d}$, solution $u^*_{ij}\left(\hat{\mathbf{y}}^t, \mathbf{d}\right)$ is feasible for the problem $P_0(\mathbf{y}^t, \mathbf{d})$ and thus $OPT(\mathbf{y}^t, \mathbf{d}) \geq OPT\left(\hat{\mathbf{y}}^t, \mathbf{d}\right)$. Thus for any $t \in [T]$,

$$\mathbf{R}\left(\hat{\mathbf{y}}^t\right) - \mathbf{R}\left(\mathbf{y}^t\right) \leq \sum_{j=1}^{n} h_j \left(y^t_j - \hat{y}^t_j\right).$$

Q.E.D.

By Lemma A.2.7, we have

$$\begin{aligned}
\sum_{t=1}^{T} \left[\mathbf{R}\left(\hat{\mathbf{y}}^t\right) - \mathbf{R}\left(\mathbf{y}^t\right)\right] \leq & \mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{n} h_i \left(y^t_i - \hat{y}^t_i\right)\right] \\
= & \mathbb{E}\left[\sum_{k:\tau_k+1\leq\tau_{k+1}-1}\ \sum_{t=\tau_k+1}^{\tau_{k+1}-1}\sum_{i=1}^{n} h_i \left(y^t_i - \hat{y}^t_i\right)\right].
\end{aligned}$$

Since for any period t such that $\tau_k + 1 \leq t \leq \tau_{k+1} - 1$, we have

$$\hat{y}^t_i = \hat{y}^{\tau_{k+1}}_i,\ \hat{y}^{\tau_{k+1}}_i < y^t_i \leq \hat{y}^{\tau_k}_i,$$

which means the inventory level after replenishment has not reached target $\hat{y}^{\tau_{k+1}}_i$ but does not exceed $\hat{y}^{\tau_k}_i$ due to the base-stock policy. Thus

$$\begin{aligned}
\sum_{t=1}^{T} \left[\mathbf{R}\left(\hat{\mathbf{y}}^t\right) - \mathbf{R}\left(\mathbf{y}^t\right)\right] \leq & \mathbb{E}\left[\sum_{k=2}^{K+1} (\delta_k - 1) \sum_{i=1}^{n} h_i \left|\hat{y}^{\tau_{k-1}}_i - \hat{y}^{\tau_k}_i\right|\right] \\
\leq & \mathbb{E}\left[\sum_{k=2}^{K+1}\left((\delta_k - 1)\left(\sum_{i=1}^{n} \frac{\|\bar{\mathbf{y}}\|\gamma}{\|\boldsymbol{\theta}\|\sqrt{\tau_{k-1}}}\right)\right)\right] \\
= & \mathbb{E}_K\left[\sum_{k=2}^{K+1} \mathbb{E}\left[(\delta_k - 1)\frac{\|\bar{\mathbf{y}}\| n\gamma}{\|\boldsymbol{\theta}\|\sqrt{\tau_{k-1}}}\right] \middle| K\right] \leq \mathbb{E}_K\left[\sum_{k=2}^{K+1} \mathbb{E}\left[(\delta_k - 1)\frac{\|\bar{\mathbf{y}}\| n\gamma}{\|\boldsymbol{\theta}\|\sqrt{k-1}}\right] \middle| K\right] \\
= & \mathbb{E}_K\left[\sum_{k=2}^{K+1} \mathbb{E}\left[(\delta_k - 1)\right]\frac{\|\bar{\mathbf{y}}\| n\gamma}{\|\boldsymbol{\theta}\|\sqrt{k-1}} \middle| K\right] \\
\leq & \mathbb{E}_k\left[\sum_{k=2}^{K+1} C_0 \log n \cdot \frac{\|\bar{\mathbf{y}}\| n\gamma}{\|\boldsymbol{\theta}\|\sqrt{k-1}} \middle| K\right] = C_0 \frac{\|\bar{\mathbf{y}}\|\gamma n \log n}{\|\boldsymbol{\theta}\|}\mathbb{E}\left[\sum_{k=2}^{K+1}\frac{1}{\sqrt{k-1}}\right] \leq O\left(\sqrt{T}\right),
\end{aligned} \tag{A.7}$$

where the last inequality is by Lemma A.2.5.

Finally, combining the bounds from (A.6) and (A.7) proves Theorem 2.4.1.

APPENDIX B

Appendix For Chapter 3

B.1 Summary of Major Notation

Table B.1: Summary of Major Notation for Model Formulation

T	the total number of periods
D^t	the demand in period t, random variable
D	time generic demand variable
d^t	the realized demand in period t
D_k^t	the cumulative demand from period t to $t+k$, $D_k^t = \sum_{i=0}^{k} D^{t+i}$, $\forall t \in [T], k \in \mathbb{Z}$
d_k^t	the realization of D_k^t
c_e	the unit ordering cost of inventory from expedited channel
c_r	the unit ordering cost of inventory from regular channel
l_e	the lead time of inventory from expedited channel
l_r	the lead time of inventory from regular channel
l	$l_r - l_e \geq 1$, the difference between two lead times
h	the unit holding cost for excess inventory
b	the unit penalty cost for unmet demand
IP_e^t	the expedited inventory position in period t
IP_r^t	the regular inventory position in period t
z_e	the order-up-to level for expedited inventory position
z_r	the order-up-to level for regular inventory position
Δ	$z_r - z_e$, the difference between the order-up-to levels
q_e^t	the order in period t from expedited channel
q_r^t	the order in period t from regular channel
I^t	the on-hand inventory in period t
O^t	the overshoot in period t, random variable
o^t	the realized overshoot in period t

Table B.2: Summary of Major Notation for Objective and Regret

$W^t(z_e, z_r)$	$(q_r^{t-1}, \ldots, q_r^{t-l+1}, IP_e^t + q_r^{t-l}) \in \mathbb{R}_+^{l-1} \times \mathbb{R}$ state variable under dual-index policy z_e, z_r
$C^t(z_e, z_r)$	the cost in period t under dual-index policy (z_e, z_r)
$\bar{W}^t(z_e, z_r)$	$(q_r^{t-1}, \ldots, q_r^{t-l+1}, 0, \ldots, 0, \max(z_e, IP_e^t + q_r^{t-l})) \in \mathbb{R}_+^l$
$O^\infty(\Delta)$	the random variable following the steady state distribution of the overshoot $O^t = (IP_e^t + q_r^{t-l} - z_e)^+$
$W^\infty(z_e, z_r)$	the steady state of process $W^t(z_e, z_r)$ under dual-index policy z_e, z_r

$C^{\infty}(z_e, z_r)$	the cost per period in the steady state $W^{\infty}(z_e, z_r)$
D_{l_e}	the sum of $l_e + 1$ random variable D
$\mathbf{ALG}$	an algorithm for the inventory replenishment in the dual-sourcing system
$C^t_{\mathbf{ALG}}$	the revenue in period t of an algorithm $\mathbf{ALG}$
$\mathcal{R}^{\mathbf{ALG}}_T$	the regret of an algorithm $\mathbf{ALG}$ in T periods
z_e^*	the expedited order-up-to level in the clairvoyant dual-index policy
z_r^*	the regular order-up-to level in the clairvoyant dual-index policy
Δ^*	the gap between the regular and the expedited order-up-to levels in the clairvoyant dual-index policy
$\gamma(z_e, z_r)$	$\mathbb{P}\left(D \leq \dfrac{z_r - z_e}{l_r + 1}\right)$
$\underline{\Delta}$	the lower limit of the gap between the regular and expedited order-up-to level Δ
λ	a known upper bound of $\mathbb{P}\left(D \leq \frac{\bar{Z}}{l_r+1}\right)$ with $\lambda < 1$

Table B.3: Summary of Major Notation for Online Learning Algorithm

$\bar{D}$	the upper bound of the demand variable
$\bar{Z}$	the upper limit of the regular order-up-to level z_r
μ	the expectation of the demand D
$\underline{\mu}$	the lower limit of the demand mean μ
B^n	the length of each epoch n
L^n	$L^n = \sum_{i=1}^n B^i$, $\forall n \in [N-1]$ with $L^0 = 0, L^N = T$
N	the total number of epochs, $N = \min\left\{n : \sum_{i=1}^n \lceil \frac{2^i}{\log T} \rceil \geq T\right\}$
J	the number of discretized values for Δ in the algorithm, $J = \lfloor \sqrt{T} \rfloor$
F_{Δ}	the CDF of the distribution of variable $D_{l_e} - O^{\infty}(\Delta)$
$z_e^*(\Delta)$	the optimal expedited order-up-to level given Δ
$\mathcal{A}^n$	the active set of choices for Δ in epoch n
j^n	the index in $\mathcal{A}^n$ of the Δ selected in epoch n
$\mathcal{D}^n$	the demand data set in epoch n
z_{ej}^n	the estimated optimal expedited order-up-to level given Δ_j in epoch n
$\hat{W}_j^t$	the simulated process for using Δ_j and z_{ej}^n where $t \in [L^n]$
$\hat{G}_j^n$	the estimated average period cost for the dual-index policy with Δ_j and z_{ej}^n in epoch n
$\mathcal{X}_j^n$	the data sample set of $X_j^t = D_{-l_e}^t - O^{t-l_e}(\Delta_j)$ for estimating the empirical quantile in epoch n
$\hat{F}_{\Delta_j}^n(\cdot)$	the empirical CDF of variable $D_{l_e} - O^{\infty}(\Delta_j)$ using $\mathcal{X}_j^n$
ε^n	the error bound to prune the active set for Δ
T_0	a constant defined in (3.9)

Table B.4: Summary of Major Notation for Regret Analysis

α^n, β^n	parameters in the algorithm
$\bar{C}$	an upper limit of the cost per period $(c_e + c_r + h)\bar{Z} + b((l_e + 1)\bar{D})$
K_0	constant, which is $e^{\frac{4(\bar{w}_j^1 \cdot \mathbf{1}^l - z_r)}{\bar{D}} + \frac{2\mu^2 l_r}{\bar{D}^2}}$
π	our proposed learning algorithm (Δ, z_e)
C_{π}^t	the cost in period t by running our algorithm $\pi = (\Delta, z_e)$
$\mathcal{R}_T^{\pi}$	the regret of our learning algorithm π in T periods
$C^t(\Delta, z_e)$	the cost in period t under dual-index policy $(z_e, z_e + \Delta)$
$C^{\infty}(\Delta, z_e)$	the steady-state per-period cost under dual-index policy $(z_e, z_e + \Delta)$
S	the event that the inventory position drops down below z_r after $\lceil \frac{5\bar{D}^2}{4\mu^2} \log T \rceil$ periods

τ	constant defined as $\tau = \lceil \frac{5\bar{D}^2}{4\mu^2} \log T \rceil + 2l_r \lceil 5 (\log T)^2 \rceil$
U	the event that the demand pattern (3.4) occurs during periods $\lceil \frac{5\bar{D}^2}{4\mu^2} \log T \rceil + 1$ to period $\lceil \frac{5\bar{D}^2}{4\mu^2} \log T \rceil + 2l_r \lceil 5 (\log T)^2 \rceil$
V^n	the event that two processes $W^t(z_e, z_r) = \tilde{W}^t(z_e, z_r)$ couple after τ periods in epoch n
M_j^n	the event that the estimated cost for arm j in epoch n is accurate enough
N_0	constant defined as $N_0 = \log_2 \log T + \log_2 \left(10l_r(\log T)^2 + \frac{5\bar{D}^2}{4\mu^2} \log T + 2l_r + 1\right)$
A_j^n	the event that $A_j^n = \left\{\left\lvert F_{\Delta_j}\left(z_{e_j}^n\right) - \frac{b}{b+h}\right\rvert \leq \alpha^n\right\}$

B.2 Proof of Theorems

B.2.1 Proof of Theorem 3.3.1

We prove the ergodicity in two disjoint cases depending on the initial regular inventory position.

B.2.1.1 Case 1: initial regular inventory position is at most z_r

Lemma B.2.1 *If* $\gamma(z_e, z_r) = \mathbb{P}\left(D \leq \dfrac{z_r - z_e}{l_r + 1}\right) > 0$, *then the Markov chain* $\{W^t(z_e, z_r) : t \geq 1\}$ *is ergodic with a steady state random vector* $W^\infty(z_e, z_r)$*. Moreover, for any* $t \geq 2l_r + 1$*, any initial vector* $w^1 \in \mathbb{R}_+^{l-1} \times \mathbb{R}$ *satisfying* $\bar{w}^1 \cdot \mathbf{1}^l \leq z_r$,

$$\delta^{t+1}\left(z_e, z_r, w^1\right) \leq \left(1 - \gamma(z_e, z_r)^{2l_r}\right)^{t/2l_r},$$

where we define $\bar{W}^t(z_e, z_r) = (q_r^{t-1}, \ldots, q_r^{t-l+1}, 0, \ldots, 0, \max(z_e, IP_e^t + q_r^{t-l})) \in \mathbb{R}_+^{l-1} \times \mathbb{R}$.

Proof of Lemma B.2.1. We say a measurable set $\bar{U} \subseteq \mathbb{R}_+^{l-1} \times \mathbb{R}$ is a small set with respect to a nontrivial measure ν provided that there exists $t^* > 0$ such that for any $w^1 \in \bar{U}$ and any measurable set $\Omega \subseteq \mathbb{R}_+^{l-1} \times \mathbb{R}$,

$$\mathbb{P}\left(W^{t^*}(z_e, z_r) \in \Omega \mid W^1(z_e, z_r) = w^1\right) \geq \nu(\Omega).$$

The following result appears in Theorem 16.0.2 in Meyn and Tweedie (1993). If $\bar{U}$ is a small set with respect to ν, then there exists stationary random variable $W^\infty(z_e, z_r)$ such that for any $w^1 \in \bar{U}$ and $t \geq t^*$,

$$\delta^{t+1}\left(z_e, z_r, w^1\right) \leq \left(1 - \nu(\mathbb{R}_+^{l-1} \times \mathbb{R})\right)^{t/(t^*-1)}.$$

Recall that $\bar{W}^t(z_e, z_r) = (q_r^{t-1}, \ldots, q_r^{t-l+1}, 0, \ldots, 0, \max(z_e, IP_e^t + q_r^{t-l})) \in \mathbb{R}_+^{l-1} \times \mathbb{R}$. Now we let $\bar{U} = \left\{ w^1 \in \mathbb{R}_+^{l-1} \times \mathbb{R} | \bar{w}^1 \cdot \mathbf{1} \leq z_r \right\}$.

For any $0 \leq k \leq l_r - 1$, let $\Omega_k \subseteq \mathbb{R}_+$ be any measurable set and let

$$\Omega = \left\{ (q_r^{-1}, q_r^{-2}, \ldots, q_r^{-l+1}, IP_e + q_r^{-l}) \in \mathbb{R}_+^{l-1} \times \mathbb{R} \middle| q_r^{-k} \in \Omega_k, \ \forall 1 \leq k \leq l-1, z_r - IP_e - q_r^{-l} - \sum_{k=1}^{l-1} q_r^{-k} \in \Omega_0 \right\}.$$

Define measure

$$\nu(\Omega) = \gamma(z_e, z_r)^{l_r + l_e} \cdot \prod_{k=0}^{l-1} \mathbb{P}\left(D \in \Omega_k \cap \left[0, \frac{z_r - z_e}{l_r + 1} \right] \right).$$

So we have $\nu(\mathbb{R}_+^{l-1} \times \mathbb{R}) = \gamma(z_e, z_r)^{2l_r} > 0$. So ν is a non-trivial measure.

To show $\bar{U}$ is a small set with respect to ν and $t^* = 2l_r + 1$, we define $\hat{\Omega}_k = \Omega_k \cap \left[0, \frac{z_r - z_e}{l_r + 1} \right]$, $\forall 1 \leq k \leq l - 1$. So we have

$$\mathbb{P}\left(W^{2l_r+1}(z_e, z_r) \in \Omega | W^1(z_e, z_r) = w^1 \right) \geq \mathbb{P}\left(W^{2l_r+1}(z_e, z_r) \in \hat{\Omega} | W^1(z_e, z_r) = w^1 \right),$$

and $\nu(\Omega) = \nu(\hat{\Omega})$. So if we can prove

$$\mathbb{P}\left(W^{2l_r+1}(z_e, z_r) \in \hat{\Omega} | W^1(z_e, z_r) = w^1 \right) \geq \nu(\hat{\Omega}),$$

we can show

$$\mathbb{P}\left(W^{2l_r+1}(z_e, z_r) \in \Omega | W^1(z_e, z_r) = w^1 \right) \geq \nu(\Omega).$$

Consider the following demand pattern of length $2l_r$.

$$\begin{aligned} D^t &\leq \frac{z_r - z_e}{l_r + 1}, \ \forall 1 \leq t \leq l_r + l_e, \\ D^{2l_r - k} &\in \hat{\Omega}_k, \ \forall k = l-1, \ldots, 0. \end{aligned}$$

This demand pattern happens with probability $\gamma(z_e, z_r)^{l_r + l_e} \cdot \prod_{k=0}^{l-1} \mathbb{P}\left(D \in \hat{\Omega}_k \right)$.

As $\bar{w}^1 \cdot \mathbf{1} \leq z_r$, we have

$$\begin{aligned} q_e^1 &= (z_e - IP_e^1 - q_r^{1-l})^+, \\ q_r^1 &= z_r - (IP_e^1 + q_r^{1-l} + \ldots + q_r^0 + q_e^1), \end{aligned}$$

for the first period. Also, from Veeraraghavan and Scheller-Wolf (2008) Eq. (4), we know

$q_e^{t+1} + q_r^{t+1} = d^t, \ \forall t \geq 1$. So we have $q_r^{t+1} \leq d^t, \ \forall t \geq 1$.

Thus, we have for $l_r + 1 \leq t \leq 2l_r$,

$$\begin{aligned} d^t \geq q_r^{t+1} &= z_r - (IP_r^{t+1} + q_e^{t+1}) \\ &= z_r - IP_r^{t+1} - (z_e - IP_e^{t+1} - q^{t+1-l})^+ \\ &= z_r - IP_e^{t+1} - q_r^{t-l+2} - \ldots - q_r^t - (z_e - IP_e^{t+1} - q^{t+1-l})^+ \\ &= z_r - \max(z_e, IP_e^{t+1} + q_r^{t+1-l}) - (q_r^{t-l+2} + \ldots + q_r^t). \end{aligned}$$

Therefore

$$\begin{aligned} \max(z_e, IP_e^{t+1} + q_r^{t+1-l}) &\geq z_r - d^t - (q_r^{t-l+2} + \ldots + q_r^t) \\ &\geq z_r - \frac{l(z_r - z_e)}{l_r + 1} > z_e, \end{aligned}$$

because $\dfrac{l}{l_r + 1} < 1$.

So $\max(z_e, IP_e^{t+1} + q_r^{t+1-l}) = IP_e^{t+1} + q_r^{t+1-l}$. Thus, $q_e^{t+1} = 0$ and $q_r^{t+1} = d^t, \ \forall l_r + 1 \leq t \leq 2l_r$.

Thus, we have

$$(q_r^{2l_r}, q_r^{2l_r-1}, \ldots, q_r^{l_r+2}) = (d^{2l_r-1}, d^{2l_r-2}, \ldots, d^{l_r+1}). \tag{B.1}$$

Also,

$$q_r^{2l_r+1} = z_r - IP_e^{2l_r+1} - q_r^{2l_r} - \ldots - q_r^{2l_r+1-l},$$

and therefore,

$$z_r - (IP_e^{2l_r+1} + q_r^{2l_r+1-l}) - \sum_{k=2l_r+2-l}^{2l_r} q_r^k = d^{2l_r}.$$

Thus, after the demand pattern, $W^{2l_r+1}(z_e, z_r) = (q_r^{2l_r}, q_r^{2l_r-1}, \ldots, q_r^{l_r+l_e+2}, IP_e^{2l_r+1} + q_r^{2l_r+1-l}) \in \hat{\Omega}$ as $q_r^{2l_r} = d^{2l_r-1} \in \hat{\Omega}_1, \ldots, q_r^{l_r+l_e+2} = d^{l_r+l_e+1} \in \hat{\Omega}_{l-1}, z_r - (IP_e^{2l_r+1} + q_r^{2l_r+1-l}) - \sum_{k=2l_r+2-l}^{2l_r} q_r^k = d^{2l_r} \in \hat{\Omega}_0$. So

$$\mathbb{P}\left(W^{2l_r+1}(z_e, z_r) \in \hat{\Omega} | W^1(z_e, z_r) = w^1\right) \geq \gamma(z_e, z_r)^{l_r+l_e} \cdot \prod_{k=0}^{l-1} \mathbb{P}\left(D \in \hat{\Omega}_k\right) = \nu(\hat{\Omega}).$$

And therefore $\bar{U}$ is a small set with respect to ν and $t^* = 2l_r + 1$. Hence, for any $t \geq 2l_r + 1$, any initial vector $w^1 \in \mathbb{R}_+^{l-1} \times \mathbb{R}$ satisfying $\bar{w}^1 \cdot \mathbf{1}^l \leq z_r$,

$$\delta^{t+1}\left(z_e, z_r, w^1\right) \leq (1-\gamma(z_e, z_r)^{2l_r})^{t/2l_r}.$$

Q.E.D.

B.2.1.2 Case 2: Initial regular inventory position exceeds z_r

The proof for this section is similar to the proof of Theorem 3 in Huh et al. (2009) Case 2.

Denote $F(\cdot)$ as the distribution function of D and $\mu = \mathbb{E}[D]$. Below is the Lemma 5 in Huh et al. (2009):

Lemma B.2.2 (Lemma 5 in Huh et al. (2009)) *For any $\eta \in \mathbb{R}$ and $t \geq 1$,*

$$\mathbb{P}\left(\sum_{\ell=1}^{t} D^\ell \leq \eta\right) \leq \begin{cases} F(\eta)^t, & \textit{if } D \textit{ has an infinite support,} \\ e^{4\eta/\bar{D}} \cdot e^{-2t\mu^2/\bar{D}^2}, & \textit{if } D \leq \bar{D} \textit{ with probability one.} \end{cases}$$

Also, similar to Lemma 6 in Huh et al. (2009), we have the following lemma:

Lemma B.2.3 *Consider dual-index policy with base stock levels (z_e, z_r). For any regular starting inventory position $w^1 \in \mathbb{R}_+^{l-1} \times \mathbb{R}$ and $t \geq 2l_r$, we have:*

$$\begin{aligned} &\mathbb{P}\left(\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l > z_r | \bar{W}^1(z_e, z_r) = \bar{w}^1\right) \\ &\leq \begin{cases} F(\bar{w}^1 \cdot \mathbf{1}^l - z_r)^{t-l_r}, & \textit{if } D \textit{ has infinite support,} \\ e^{4(\bar{w}^1 \cdot \mathbf{1}^l - z_r)/\bar{D}} \cdot e^{-2\mu^2(t-l_r)/\bar{D}^2}, & \textit{if } D \leq \bar{D} \textit{ with probability one.} \end{cases} \end{aligned}$$

Proof of Lemma B.2.3. According to the base-stock policy in the dual-index policy for the regular inventory position, we have

$$\max\left\{\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l, z_r\right\} \leq \max\left\{\bar{W}^1(z_e, z_r) \cdot \mathbf{1}^l, z_r\right\}, \ \forall t \geq 1.$$

So

$$\begin{aligned} &\mathbb{P}\left(D^1 + D^2 + \ldots + D^{t-l_r} < \bar{w}^1 \cdot \mathbf{1}^l - z_r\right) \\ =&\mathbb{P}\left(D^{l_r} + D^{l_r+1} + \ldots + D^{t-1} < \bar{w}^1 \cdot \mathbf{1}^l - z_r\right) \\ \geq&\mathbb{P}\left(D^{l_r} + D^{l_r+1} + \ldots + D^{t-1} < \bar{w}^{l_r} \cdot \mathbf{1}^l - z_r\right). \end{aligned}$$

So $\forall t \geq 2l_r$, we claim that $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l > z_r$ if and only if $\bar{W}^{l_r}(z_e, z_r) \cdot \mathbf{1}^l - (D^{l_r} + D^{l_r+1} + \ldots + D^{t-1}) > z_r$.

If $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l > z_r$, then according to the dual-index policy, we have $q_r^k = 0,\ \forall k \leq t$. So we have $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l = \max(z_e, IP_e^t)$.

Because $z_e \leq z_r$, we have $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l = IP_e^t > z_r \geq z_e$. Also, $\bar{W}^{l_r}(z_e, z_r) \cdot \mathbf{1}^l = IP_e^{l_r} > z_r \geq z_e$. So $q_e^k = 0,\ \forall l_r \leq k \leq t$.

So

$$\begin{aligned}
&\bar{W}^{l_r}(z_e, z_r) \cdot \mathbf{1}^l - (D^{l_r} + D^{l_r+1} + \ldots + D^{t-1}) \\
=&IP_e^{l_r} - (D^{l_r} + D^{l_r+1} + \ldots + D^{t-1}) \\
=&IP_e^t > z_r.
\end{aligned}$$

If $\bar{W}^{l_r}(z_e, z_r) \cdot \mathbf{1}^l - (D^{l_r} + D^{l_r+1} + \ldots + D^{t-1}) > z_r$, then $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l > z_r$.

Hence, $\bar{W}^t(z_e, z_r) \cdot \mathbf{1}^l > z_r$ if and only if $\bar{W}^{l_r}(z_e, z_r) \cdot \mathbf{1}^l - (D^{l_r} + D^{l_r+1} + \ldots + D^{t-1}) > z_r$.

Thus, by Lemma B.2.2, the results of Lemma B.2.3 follow. **Q.E.D.**

Next, we only prove the result when D has infinite support. The proof for the situation when D is bounded is similar.

Let $\mathbb{P}_{\bar{w}^1}$ and $\mathbb{E}_{\bar{w}^1}$ be the probability and expectation conditioned on the event that $\bar{W}^1 \cdot \mathbf{1}^l = \bar{w}^1$. Then for any measurable set $\Omega \subseteq \mathbb{R}_+^{l-1} \times \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}_{\bar{w}^1}[W^{t+1}(z_e, z_r) \in \Omega] \\
=&\mathbb{E}_{\bar{w}^1}\left[\mathbb{P}_{\bar{w}^1}[W^{t+1}(z_e, z_r) \in \Omega \mid W^{\lceil \frac{t}{2} \rceil}(z_e, z_r)]\right] \\
=&\mathbb{E}_{\bar{w}^1}\left[\mathbb{1}(W^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r) \cdot \mathbb{P}\left(W^{t+1}(z_e, z_r) \in \Omega | W^{\lceil \frac{t}{2} \rceil}(z_e, z_r)\right)\right] \\
&+ \mathbb{E}_{\bar{w}^1}\Big[\mathbb{1}(\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l > z_r) \cdot \mathbb{P}\Big(W^{t+1}(z_e, z_r) \in \Omega | \bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r)\Big)\Big)],
\end{aligned}$$

where the last equality is from Markov property. Then we have

$$\begin{aligned}
A :=&\mathbb{P}_{\bar{w}^1}[W^{t+1}(z_e, z_r) \in \Omega] - \mathbb{P}\left(W^{\infty}(z_e, z_r) \in \Omega\right) \\
=&\mathbb{E}_{\bar{w}^1}\left[\mathbb{1}(\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r) \cdot \phi(\Omega)\right] + \mathbb{E}_{\bar{w}^1}\left[\mathbb{1}(\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l > z_r) \cdot \phi(\Omega)\right],
\end{aligned}$$

where $\phi(\Omega) = \mathbb{P}\left(W^{t+1}(z_e, z_r) \in \Omega | W^{\lceil \frac{t}{2} \rceil}(z_e, z_r)\right) - \mathbb{P}\left(W^{\infty}(z_e, z_r) \in \Omega\right)$.

Also, we have almost surely

$$|\phi(\Omega)| \leq \delta_{t - \lceil \frac{t}{2} \rceil + 2}(z_e, z_r, W^{\lceil \frac{t}{2} \rceil}(z_e, z_r)),$$

and $\phi(\Omega) \leq 1$, so

$$\begin{aligned}|A| \leq &\mathbb{E}_{\bar{w}^1}\left[\mathbb{1}(\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r) \cdot \delta_{t-\lceil \frac{t}{2} \rceil+2}(z_e, z_r, W^{\lceil \frac{t}{2} \rceil}(z_e, z_r))\right] \\ &\quad + \mathbb{P}_{\bar{w}^1}[\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l > z_r].\end{aligned}$$

By Lemma B.2.1, the first term

$$\begin{aligned}&\mathbb{E}_{\bar{w}^1}\left[\mathbb{1}(\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r) \cdot \delta_{t-\lceil \frac{t}{2} \rceil+2}(z_e, z_r, W^{\lceil \frac{t}{2} \rceil}(z_e, z_r))\right] \\ \leq &\mathbb{P}_{\bar{w}^1}[\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l \leq z_r] \cdot \left(1 - \gamma(z_e, z_r)^{2l_r}\right)^{(t-\lceil \frac{t}{2} \rceil+1)/2l_r} \\ \leq &\left(1 - \gamma(z_e, z_r)^{2l_r}\right)^{(t-\lceil \frac{t}{2} \rceil+1)/2l_r} \\ \leq &\left(1 - \gamma(z_e, z_r)^{2l_r}\right)^{t/4l_r}.\end{aligned}$$

By Lemma B.2.3, the second term

$$\begin{aligned}\mathbb{P}_{\bar{w}^1}[\bar{W}^{\lceil \frac{t}{2} \rceil}(z_e, z_r) \cdot \mathbf{1}^l > z_r] &\leq F(\bar{w}^1 \cdot \mathbf{1}^l - z_r)^{\lceil \frac{t}{2} \rceil - l_r} \\ &\leq F(\bar{w}^1 \cdot \mathbf{1}^l - z_r)^{\frac{t}{2} - l_r}.\end{aligned}$$

Therefore, we obtain the bound for $\delta^{t+1}\left(z_e, z_r, w^1\right)$ as stated in Theorem 3.3.1.

B.3 Proof of Lemmas

B.3.1 Proof of Lemma 3.5.7

Because

$$L^{n-1} = \sum_{i=1}^{n-1} \left\lceil \frac{2^i}{\log T} \right\rceil \geq \sum_{i=1}^{n-1} \frac{2^i}{\log T} = \frac{2^n - 2}{\log T},$$

we have $\alpha^n \leq \frac{3}{2}\sqrt{3T_0}\dfrac{\log T}{\sqrt{2^n - 2}}$. Therefore,

$$\begin{aligned}\sum_{n=1}^{N} B^n \alpha^n \leq &\sum_{n=1}^{N} \left\lceil \frac{2^n}{\log T} \right\rceil \frac{3}{2}\sqrt{3T_0}\frac{\log T}{\sqrt{2^n - 2}} \\ \leq &\frac{3}{2}\sqrt{3T_0}\sum_{n=1}^{N}\left(\frac{2^n}{\sqrt{2^n - 2}} + \sqrt{\frac{\log T}{2^n - 2}}\right) \\ = &O(\sqrt{T \log T}),\end{aligned}$$

where the last line is because $N \leq \log_2(T\log T + 2) - 1$.

B.3.2 Proof of Lemma 3.5.11

For any epoch $n \in [N]$ and any arm $j \in \mathcal{A}^n$, as $\{W_j^t\}_{t=1}^{L^n}$ is the Markov chain of the states of the system following the dual-index policy $\left(\Delta_j, z_{ej}^n\right)$, let

$$\begin{aligned} g^n\left(\hat{W}_j^1, \ldots, \hat{W}_j^{L^n}\right) &= \frac{1}{L^n}\sum_{t=1}^{L^n} \hat{C}^t\left(\Delta_j, z_{e_j}^n\right) \\ &= \frac{1}{L^n}\sum_{t=1}^{L^n} c_e\hat{q}_{ej}^t + c_r\hat{q}_{r_j}^t + h\left(\hat{I}_j^{t+1}\right)^+ + b\left(\hat{I}_j^{t+1}\right)^- \end{aligned}$$

Also, notice that condition (3.8) holds with $\iota_i = \frac{\bar{C}}{L^n}$, $\forall i \in [L^n]$ as $C^t(\Delta, z_e) \leq \bar{C}$ with probability 1 for any Δ and z_e. Then Lemma 3.5.11 holds according to Lemma 3.5.5 with Markov chain being $\{W_j^t\}_{t=1}^{L^n}$ and the function f being g^n for any $n \in [N]$ and $j \in [J]$.

B.3.3 Proof of Lemma 3.5.12

For notational simplicity, let $\mathbb{E}\left[C^*(\Delta)\right] := \mathbb{E}\left[C^\infty(\Delta, z_e^*(\Delta))\right]$. For Δ_1 and Δ_2, without loss of generality, we assume $\mathbb{E}\left[C^*(\Delta_1)\right] \geq \mathbb{E}\left[C^*(\Delta_2)\right]$. Let $z_1 = \arg\min_z \mathbb{E}\left[C^\infty(\Delta_1, z_e)\right]$ and $z_2 = \arg\min_z \mathbb{E}\left[C^\infty(\Delta_2, z_e)\right]$.

$$\begin{aligned} &\mathbb{E}\left[C^*(\Delta_1)\right] - \mathbb{E}\left[C^*(\Delta_2)\right] \\ =&\mathbb{E}\left[C^\infty(\Delta_1, z_1)\right] - \mathbb{E}\left[C^\infty(\Delta_2, z_2)\right] \\ \leq&\mathbb{E}\left[C^\infty(\Delta_1, z_2)\right] - \mathbb{E}\left[C^\infty(\Delta_2, z_2)\right] \\ =&(c_e + c_r)\mathbb{E}\left[O^\infty(\Delta_1) - O^\infty(\Delta_2)\right] + h\mathbb{E}\left[(z_2 + O^\infty(\Delta_1) - D_{l_e})^+ - (z_2 + O^\infty(\Delta_2) - D_{l_e})^+\right] \\ &+ b\mathbb{E}\left[(z_2 + O^\infty(\Delta_1) - D_{l_e})^- - (z_2 + O^\infty(\Delta_2) - D_{l_e})^-\right]. \end{aligned}$$

We would like to offer an upper bound for the term $\mathbb{E}\left|O^\infty(\Delta_1) - O^\infty(\Delta_2)\right|$. First, we show that $\mathbb{E}\left[O^\infty(\Delta)\right]$ is non-decreasing in Δ by contradiction.

Suppose that $\Delta_1 \geq \Delta_2$ with $\mathbb{E}\left[O^\infty(\Delta_1)\right] < \mathbb{E}\left[O^\infty(\Delta_2)\right]$. From Equation (8) in Veeraraghavan and Scheller-Wolf (2008) which is $O^\infty(\Delta) = \Delta - (q_r^t + q_r^{t-1} + \ldots + q_r^{t-l+1})$, we have $\mathbb{E}\left[O^t(\Delta)\right] = \Delta - l\mathbb{E}\left[q_r^\infty\right]$. Consider the following two cases:

1. If $\mathbb{E}\left[O^\infty(\Delta_1)\right] = 0$, which means $\mathbb{E}\left[q_e^\infty\right] \geq 0$ and thus $\mathbb{E}\left[q_r^\infty\right] \leq \mathbb{E}\left[D\right]$.

$$\mathbb{E}\left[O^\infty(\Delta_1)\right] - \mathbb{E}\left[O^\infty(\Delta_2)\right]$$

$$=\Delta_1 - l\mathbb{E}\left[q_r^\infty(\Delta_1)\right] - \Delta_2 + l\mathbb{E}\left[D\right].$$

As $\mathbb{E}\left[q_r^\infty(\Delta_1)\right] \leq l\mathbb{E}\left[D\right]$ and $\Delta_1 \geq \Delta_2$, we have

$$\mathbb{E}\left[O^\infty(\Delta_1)\right] - \mathbb{E}\left[O^\infty(\Delta_2)\right] \geq 0,$$

which is a contradiction.

2. If $\mathbb{E}\left[O^\infty(\Delta_1)\right] > 0$, which means $\mathbb{E}\left[q_r^\infty\right] = \mathbb{E}\left[D\right]$. Then we have

$$\begin{aligned}&\mathbb{E}\left[O^\infty(\Delta_1)\right] - \mathbb{E}\left[O^\infty(\Delta_2)\right]\\ =&\Delta_1 - l\mathbb{E}\left[D\right] - \Delta_2 + l\mathbb{E}\left[D\right] \geq 0,\end{aligned}$$

which is a contradiction.

Therefore, we have $\mathbb{E}\left[O^\infty(\Delta)\right]$ is non-decreasing in Δ. So consider $\Delta_1 \geq \Delta_2$ without loss of generality. Then we have $\mathbb{E}\left[O^\infty(\Delta_1)\right] \geq \mathbb{E}\left[O^\infty(\Delta_2)\right]$.

1. If $\mathbb{E}\left[O^\infty(\Delta_1)\right] > 0$ and $\mathbb{E}\left[O^\infty(\Delta_2)\right] > 0$, then

$$\mathbb{E}\left[O^\infty(\Delta_1)\right] - \mathbb{E}\left[O^\infty(\Delta_2)\right] = \Delta_1 - \Delta_2.$$

2. If $\mathbb{E}\left[O^\infty(\Delta_1)\right] > 0$ and $\mathbb{E}\left[O^\infty(\Delta_2)\right] = 0$, then

$$\mathbb{E}\left[O^\infty(\Delta_1)\right] - \mathbb{E}\left[O^\infty(\Delta_2)\right] = \Delta_1 - \Delta_2 - l\mathbb{E}\left[D\right] + l\mathbb{E}\left[q_r^\infty(\Delta_2)\right] \leq \Delta_1 - \Delta_2.$$

3. If $\mathbb{E}\left[O^\infty(\Delta_1)\right] = 0$ and $\mathbb{E}\left[O^\infty(\Delta_2)\right] = 0$, then

$$\mathbb{E}\left[O^\infty(\Delta_1)\right] - \mathbb{E}\left[O^\infty(\Delta_2)\right] = 0.$$

In sum, $\mathbb{E}\left[|O^\infty(\Delta_1) - O^\infty(\Delta_2)|\right] \leq |\Delta_1 - \Delta_2|$.

Then, because $(x-a)^+ - (y-a)^+ \leq |x-y|$ and $(x-a)^- - (y-a)^- \leq |x-y|$, we have

$$\mathbb{E}\left[C^*(\Delta_1)\right] - \mathbb{E}\left[C^*(\Delta_2)\right] \leq (c_e + c_r + h + b)\mathbb{E}\left[|O^\infty(\Delta_1) - O^\infty(\Delta_2)|\right] \leq (c_e + c_r + h + b)\left|\Delta_1 - \Delta_2\right|,$$

which suggests that $\mathbb{E}\left[C^*(\Delta)\right]$ is Lipschitz in Δ.

B.4 Settings with Nonstationary Demand

The i.i.d. demand assumption is predominant in the dual sourcing literature (Allon and Van Mieghem 2010, Sheopuri et al. 2010). The dual-index policy was, therefore, developed for stationary demand (Veeraraghavan and Scheller-Wolf 2008), and its performance has never been explored under *non-stationary* demand.

Restart Learning Algorithm. When demand is non-i.i.d., the performance guarantee of our (Δ, z_e) algorithm may not hold (see Figure B.3) if our learning algorithm is naively implemented in this nonstationary environment. Thus, we need to come up with an alternative strategy. Borrowing the "restart" idea from Besbes et al. (2015), we re-design our algorithm by restarting the procedure for every τ_T periods. The details of our modified algorithm are in Algorithm 11. Regarding the choice of τ_T, in Besbes et al. (2015), $\tau_T = \left\lceil (T/V_T)^{2/3} \right\rceil$ where V_T is a known variation budget. In our problem, V_T is defined as the upper bound of $\sum_{t=2}^{T} \left\| C_t^{\infty} - C_{t-1}^{\infty} \right\| = \sum_{t=2}^{T} \sup_{(z_e,z_r)\in[0,\bar{Z}]\times[0,\bar{Z}]} |C_t^{\infty}(z_e, z_r) - C_{t-1}^{\infty}(z_e, z_r)|$ where $C_t^{\infty}(\cdot)$ is the stationary per-period cost when demand follows the same distribution as D^t. Note that our algorithm is different from the OGD studied in Besbes et al. (2015) and the f function is not necessarily convex, the tuning of the restarting interval will vary. For experimental purposes, we follow the choice of $\tau_T = \left\lceil (T/V_T)^{2/3} \right\rceil$ based on the intuition that the larger the variation budget is, the smaller the restart interval should be. Since the performance measure is the relative regret, the cost parameters can be scaled and so is the variation budget. We assume that the firm knows the variation budget $V_T = 1$ after rescaling the cost parameters.

Convergence Rate. Here we briefly analyze the performance guarantee of the proposed Algorithm 11 denoted as π'. Since the optimal inventory replenishment policy is complex and state-dependent even under stationary demand, we still choose the full-information optimal dual-index policies as the benchmark under non-stationary demand. Specifically, for any algorithm **ALG**, we define the performance metric under nonstationary demand as

$$\mathcal{R}_T^{\mathbf{ALG}} = \mathbb{E}\left[\sum_{t=1}^{T} C_{\mathbf{ALG}}^t - \sum_{t=1}^{T} C_t^{\infty}\left(z_e^{t*}, z_r^{t*}\right)\right],$$

where $C_{\mathbf{ALG}}^t$ is the cost in period t by running algorithm **ALG**. We define $C_t^{\infty}(z_e, z_r)$ as the stationary cost variable under the dual-index policy with dual indices (z_e, z_r) under the demand with the same distribution as D^t and $(z_e^{t*}, z_r^{t*}) := \arg\max_{(z_e,z_r)} C_t^{\infty}(z_e, z_r)$.

We conjecture that $\mathcal{R}_T^{\pi'} = \tilde{O}\left(T^{\frac{2}{3}} V_T^{\frac{1}{3}}\right)$, but we leave the rigorous proof to future work. Here, we only provide some intuitions and technical results, which would help build the foundation of a rigorous proof.

It is noteworthy that the establishment of Lemma 3.3.1 and Lemma 3.5.1 does not rely on the stationarity of the demand distribution, and both lemmas still hold (with an additional assumption restricting the cumulative demand from being excessively small for Lemma 3.5.1). In particular, $(W^t(z_e, z_r), t \geq 1)$ in the dual-sourcing system following a dual-index policy with parameters (z_e, z_r) under non-stationary demand still forms a (not necessarily homogeneous) Markov chain. Moreover, two processes driven by the same demand will couple after $O(\log T)^2$ periods with high probability.

As for another key result Lemma 3.5.5, which guarantees the performance of the empirical estimation framework, we here define the mixing time for Markov chains without assuming time homogeneity. We let $\mathcal{L}(X_{i+t} \mid X_i = x)$ be the conditional distribution of X_{i+t} given $X_i = x$.

$$\begin{aligned}
\bar{d}(t) &:= \max_{1 \leq i \leq N-t} \sup_{x,y \in \Omega_i} d_{\mathrm{TV}}\left(\mathcal{L}\left(X_{i+t} \mid X_i = x\right), \mathcal{L}\left(X_{i+t} \mid X_i = y\right)\right), \\
\tau(\epsilon) &:= \min\{t \in \mathbb{N} : \bar{d}(t) \leq \epsilon\}.
\end{aligned}$$

The following generalized result of Lemma 3.5.5 exists for not necessarily homogeneous Markov chains.

Lemma B.4.1 (Corollary 2.10 in Paulin (2015)) *Let $X := (X_1, \ldots, X_N)$ be a (not necessarily time-homogeneous) Markov chain, taking values in a Polish state space $\Lambda = \Lambda_1 \times \ldots \times \Lambda_N$, with mixing time $\tau(\epsilon)$(for $0 \leq \epsilon \leq 1$). Let $\tau_{\min} := \inf_{0 \leq \epsilon < 1} \tau(\epsilon) \cdot \left(\frac{2-\epsilon}{1-\epsilon}\right)^2$. Suppose that $f : \Lambda \to \mathbb{R}$ satisfies $f(x) - f(y) \leq \sum_{i=1}^n c_i \mathbb{1}\left[x_i \neq y_i\right]$ for every $x, y \in \Lambda$. Then for any $t \geq 0$,*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\left(\frac{-2t^2}{\|c\|^2 \tau_{\min}}\right).$$

Thus, we have a concentration inequality established for the sample average up to time t of some function with respect to the Markov chains with non-stationary transition kernels, compared with the true mean up to time t of the function under this nonstationary environment. With proper assumptions on the degree of changes in the underlying demand distribution, $\tau_{\min}$ would be of constant order. Consequently, following similar proof as in the case of a stationary environment, we can offer an upper bound for the difference between the cost incurred by the original (Δ, z_e) Algorithm (denoted as π) and the *static* optimal dual-index policy, i.e., $\mathscr{R}_{\tau_T}^{\pi} := \mathbb{E}\left[\sum_{t=1}^{\tau_T} C_\pi^t - \tau_T C^\infty\left(z_e^*, z_r^*\right)\right]$ where $\left(z_e^*, z_r^*\right) := \arg\max_{(z_e, z_r)} \mathbb{E}\left[\sum_{t=1}^{\tau_T} C_t^\infty\left(z_e, z_r\right)\right]$. Since all key results hold in the same order as in stationary cases, we speculate that $\mathscr{R}_{\tau_T}^{\pi} = \tilde{O}(\sqrt{\tau_T})$.

Combined with the following proposition, we can provide the upper bound for the dynamic regret.

Proposition B.4.1 (Proposition 2 in Besbes et al. (2015)) *Let π' be the policy de-*

fined by the restarting procedure that uses π as a subroutine with batch size τ_T. Then, for any $T \geqslant 1$,

$$\mathcal{R}_T^{\pi'} \leqslant \left\lceil \frac{T}{\tau_T} \right\rceil \cdot \mathscr{R}_{\tau_T}^{\pi} + 2\tau_T V_T.$$

Because the establishment of Proposition B.4.1 does not require the convexity property of the objective function, adopting the restarting procedure with batch size $\tau_T = \left\lceil (T/V_T)^{2/3} \right\rceil$ will incur a total regret $\mathcal{R}_T^{\pi'} = \tilde{O}\left(T^{\frac{2}{3}} V_T^{\frac{1}{3}}\right)$, matching the information-theoretic lower bound established in Besbes et al. (2015) up to logarithmic factors. If we consider the instance below where $V_T = O(1)$ and $\tau_T = \left\lceil (T/V_T)^{2/3} \right\rceil = 150$, the total regret is of order $O(T^{\frac{2}{3}} \log T)$ as shown in Figure B.1.

Non-IID Demand Instance. Consider the following test instance where the time horizon is five consecutive years with $T = 1825$. The lead times are set to be $l_e = 2, l_r = 4$. The demand in period t is set to follow the truncated normal distribution $\mathcal{N}(\mu_n, \sigma_n^2)$ where $n = \lceil t/365 \rceil$ as shown in Table B.5.

Table B.5: Demand Settings

Year	1	2	3	4	5
Distribution	$\mathcal{N}(30, 6)$	$\mathcal{N}(40, 7)$	$\mathcal{N}(50, 10)$	$\mathcal{N}(60, 12)$	$\mathcal{N}(70, 14)$
Truncated at	$[0, 60]$	$[10, 70]$	$[20, 80]$	$[30, 90]$	$[40, 100]$

Computational Performance. We emphasize that the firm does not know the evolution of the demand distribution nor the distributions themselves when running the learning algorithm. The relative regret is defined as

$$\textbf{Relative Regret} := \frac{\sum_{t=1}^{T} C^t(z_e^{t\pi}, z_r^{t\pi}) - \sum_{t=1}^{T} C^t(z_e^{t*}, z_r^{t*})}{\sum_{t=1}^{T} C^t(z_e^{t*}, z_r^{t*})},$$

where (z_e^{t*}, z_r^{t*}) are the optimal order-up-to levels for the dual-sourcing system with demand following $\mathcal{N}(\mu_{\lceil t/365 \rceil}, \sigma^2_{\lceil t/365 \rceil})$. Figure B.1 shows the relative regret averaged over the instances with (b, h) pairs taking values $(\{5, 10\} \times \{1, 4\})$ as in Table 3.1. For each instance, we run 1000 times and take the average of the relative regret. Also, Figure B.2 shows the performance of the restart (Δ, z_e) algorithm when the restart point is tuned to be the change point. For comparison, Figure B.3 shows the performance of the (Δ, z_e) algorithm without restart. It is evident that the restarting procedure is necessary under non-stationary demand, and the modified restart algorithm works well when the restarting point is close to the change point of demand.

Unknown Variation Budget. Algorithm 11 requires prior knowledge of the total

Algorithm 11 The "restart" (Δ, z_e) learning algorithm for the dual-index policy

```
for $k = 1, \ldots, \lceil T/\tau_T \rceil$:  do                                        ▷ Restart
    Let $T_0 = (\min\{\tau_T, T-(k-1)\tau_T\})$ and $N = \min\left\{n : \sum_{i=1}^{n} \lceil \frac{2^i}{\log T_0} \rceil \geq T_0\right\}$ the number of epochs.
    Let $J = \#$ discrete $\Delta$'s and $B^i = \lceil \frac{2^i}{\log T_0} \rceil$ be the $i$-th epoch length.
    Let $L^n = \sum_{i=1}^{n} B^i,\ \forall n \in [N-1]$ with $L^0 = 0, L^N = T_0$.    ▷ Parameters
    Initialize the active set $\mathcal{A}^1 = \{1, \ldots, J\}$, $\mathcal{D}^0 = \varnothing$.    ▷ Initialization
    For $j \in \mathcal{A}^1$, define $\Delta_j = \underline{\Delta} + \frac{j}{J}\left|\bar{Z} - \underline{\Delta}\right|$ and assign $z_{ej}^1 \in [0, \bar{Z} - \Delta_j]$ arbitrarily.
    for $n = 1, 2, \ldots, N$ do                                                     ▷ Outer Loop
        Randomly select $j^n \in \mathcal{A}^n$. Let demand set $\mathcal{D}^n = \mathcal{D}^{n-1}$.
        for $t = (k-1)\tau_T + L^{n-1} + 1, \ldots, (k-1)\tau_T + L^n$:  do
            Apply the dual-index policy $(z_e^t, z_r^t) = (z_{ej^n}^n, z_{ej^n}^n + \Delta_{j^n})$.
            Append the realized demand $d^t$ into $\mathcal{D}^n$.
                $q_e^t = (z_{ej^n}^n - IP_e^t - q_r^{t-l})^+, \quad q_r^t = (z_{ej^n}^n + \Delta_{j^n} - IP_r^t - q_e^t)^+,$
                $IP_e^{t+1} = IP_e^t + q_e^t - d^t + q_r^{t-l}, \quad IP_r^{t+1} = IP_r^t + q_e^t + q_r^t - d^t,$
                $o^t = (IP_e^t + q_r^{t-l} - z_{ej^n}^n)^+, \quad I^{t+1} = I^t + q_e^{t-l_e} + q_r^{t-l_r} - d^t.$
        end for
        for $j \in \mathcal{A}^n$ do                                                   ▷ Inner Loop
            Simulate the policy $(z_{ej}^n, z_{ej}^n + \Delta_j)$ for $\min\{L^n, T_0\}$ periods using $\mathcal{D}^n$ and denote the state variables of this simulation by $\hat{W}_j^t := (\hat{q}_{rj}^{t-1}, \ldots, \hat{q}_{rj}^{t-l+1}, \widehat{IP}_{ej}^t + \hat{q}_{rj}^{t-l}) \in \mathbb{R}_+^{l-1} \times \mathbb{R},\ t = (k-1)\tau_T + 1, \ldots, (k-1)\tau_T + L^n$.
            Obtain the estimated average period cost:
                $\hat{G}_j^n = \frac{1}{L^n} \sum_{t \in [L^n]} c_e \hat{q}_{ej}^t + c_r \hat{q}_{rj}^t + h(\hat{I}_j^{t+1})^+ + b(\hat{I}_j^{t+1})^-.$
            Let $\mathcal{X}_j^n = \left\{d_{l_e}^t - \hat{o}_j^t, t \in [(k-1)\tau_T + 1, (k-1)\tau_T + L^n - l_e]\right\}$.
            Let $\hat{F}_{\Delta_j}^n(\cdot)$ be the empirical CDF of $X_j^t = D_{l_e}^t - \hat{O}_j^t(\Delta_j)$ with data sample $\mathcal{X}_j^n$.
            Update $z_{ej}^{n+1} = \hat{F}_{\Delta_j}^{n^{-1}}(\frac{b}{b+h})$.                ▷ Inner Layer Optimization
        end for
        Update and prune the active set                                              ▷ Outer Layer Optimization
                $\mathcal{A}^{n+1} = \left\{ j \in \mathcal{A}^n : \hat{G}_j^n - \min_{j' \in \mathcal{A}^n} \hat{G}_{j'}^n \leq \varepsilon^n \right\}.$
    end for
end for
```

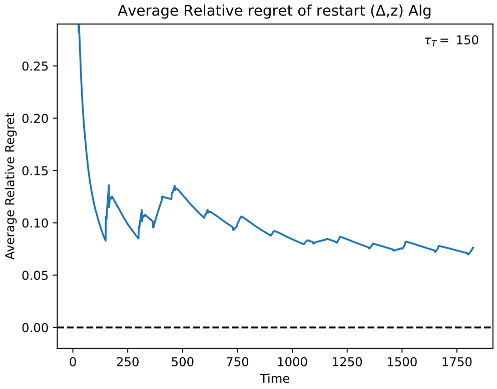


Figure B.1: Restart Interval $\tau_T = 150$

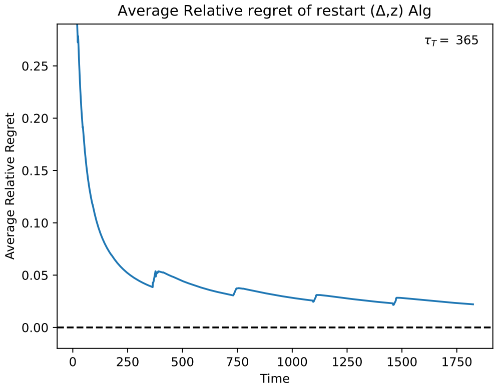


Figure B.2: Restart Interval $\tau_T = 365$

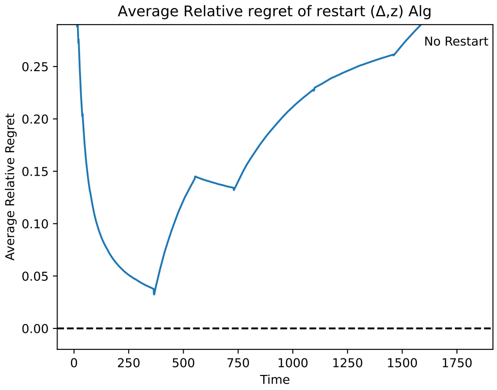


Figure B.3: (Δ, z_e) Algorithm without Restart

variation budget V_T to determine the length of restarting epoch τ_T. When there is no information on the degree of non-stationarity, one can schedule multiple instances of the base algorithm with different durations in a carefully-designed randomized scheme and restart based on the real-time detection result of the change of the environment as introduced in Wei and Luo (2021). The design and analysis of this framework for the dual-index policy in dual-sourcing systems are left to future work.

APPENDIX C

Appendix For Chapter 4

C.1 Summary of Major Notation

Table C.1: Summary of Major Notation for Model Formulation

p	the unit price set for customers
w	the unit remuneration for suppliers
$\lambda(\cdot), \mu(\cdot)$	unknown functions
ε, δ	zero-mean random variables with unknown distributions
$D(p)$	random demand depending on price p, specifically $D(p) = \lambda(p) + \varepsilon$
$S(w)$	random supply depending on remuneration w, specifically $S(w) = \mu(w) + \delta$
$R(p, w)$	expected revenue under price p and remuneration w, specifically $R(p, w) = (p - w)\mathbb{E}[\min(\lambda(p) + \varepsilon, \mu(w) + \delta)]$
$\bar{P}, \underline{P}$	known constants, the upper and lower bound of price p
$\bar{S}$	known constant, the upper bound for random demand and supply variables for any $0 \leq w \leq p \leq \bar{P}$
K_1	constant, $\lambda(p)$ is K_1-Lipschitz in p
K_2	constant, upper bound of $\mu'(w)$
K_3	constant, given p the optimal expected revenue $R\left(p, w^*(p)\right)$ is K_3-Lipschitz in p, specifically $K_3 = \max\{K_1\bar{P}, \max\{K_2, \bar{S}\}\bar{P} + \bar{S}\}$
p^*	clairvoyant optimal price
w^*	clairvoyant optimal remuneration
ALG	any algorithm
$\breve{R}^s_{\mathbf{ALG}}$	the realized revenue in period s by running algorithm **ALG**
$\mathbf{Regret}^T_{\mathbf{ALG}}$	the regret of any algorithm **ALG** for a T period finite horizon problem

Table C.2: Summary of Major Notation for Online Learning Algorithm

J	the number of discrete arms of price p
t	time indices, one time consists of 3 consecutive periods for brevity in analysis
p_j	the price corresponding to j-th arm, $p_j = \underline{P} + \frac{j}{J} \left\lvert \bar{P} - \underline{P} \right\rvert$
γ_i	the multiplier for various levels of confidence in bisection search, $\gamma_i = \frac{1}{2^i}, i \geq 1$
$[l_j, r_j]$	real-time updated interval of searching for $w^*(p_j)$
k_j	real-time updated index of level of confidence when querying the points in the current interval
m_j^t	the number of times of selecting price p_j in the current epoch up to the beginning of time t
n_j^t	the total number of times of selecting price p_j up to the beginning of time t
α_j^t	the sum of realized revenue using price p_j up to the beginning of time t
$\hat{R}_{j,x}^t, x \in \{l, c, r\}$	the cumulative sum of realized revenue using price p_j up to the beginning of time t in the current epoch
$\mathbf{Rad}_j^t$	the confidence radius of estimation of $R\left(p_j, w^*(p_j)\right)$ at the beginning of time t
j^t	the index of the arm pulled in time t
u	a temporary variable representing the length of the interval in time t, specifically $u = r_{j^t} - l_{j^t}$
$w_x^t, x \in \{l, c, r\}$	the three choices of remuneration in time t, corresponding to quartiles of the interval of arm j^t in time t
$\tilde{R}_x^t, x \in \{l, c, r\}$	the realized revenue using implementing (p_{j^t}, w_x^t) in time t
$\mathrm{UB}_{k_{j^t}}\left(w_x^t\right), x \in \{l, c, r\}$	upper bound of the confidence interval for estimation of $R\left(p_{j^t}, w_x^t\right)$ in time t
$\mathrm{LB}_{k_{j^t}}\left(w_x^t\right), x \in \{l, c, r\}$	lower bound of the confidence interval for estimation of $R\left(p_{j^t}, w_x^t\right)$ in time t

Table C.3: Summary of Major Notation for Regret Analysis

τ	index of an epoch for any arm
$L_{j,\tau}$	the set of time indices contained in price p_j's epoch τ
H_j	the set of time indices when the price is p_j
τ_j^t	the index of the epoch of price p_j in time t
$\left[l_j^\tau, r_j^\tau\right]$	the interval for estimation of $w^*(p_j)$ during epoch τ of price p_j
$\mathcal{E}_x^t, x \in \{l, c, r\}$	the event that the estimation of $R\left(p_{j^t}, w_x^t\right)$ is accurate enough, specifically, $\mathcal{E}_x^t = \left\{ \left\lvert \frac{\hat{R}_{j^t,x}^{t+1}}{m_{j^t}^{t+1}} - R\left(p_{j^t}, w_x^t\right) \right\rvert \leq \bar{P}\bar{S}\sqrt{\frac{\log T}{m_{j^t}^{t+1}}} \right\}$
$\mathcal{E}$	intersection of events $\left\{\mathcal{E}_x^t, x \in \{l, c, r\}, t \in [T]\right\}$
$\gamma_j^{t\,\min}$	the lower bound of the multiplier of price p_j used up to the end of time t, specifically, $\gamma_j^{t\,\min} = 2\bar{P}\bar{S}\sqrt{\frac{\log T}{n_j^{t+1}}}$
IN_j^t	the interval contained in the interval of price p_j in time t
$\breve{R}_{\mathbf{BBS}}^s$	the realized revenue in period s by running algorithm **BBS**
$\mathbf{Regret}_{\mathbf{BBS}}^{3T}$	the regret of any algorithm **BBS** for a $3T$ period finite horizon problem

C.2 Proof of Theorems and Lemmas

C.2.1 Proof of Theorem 4.2.1

Proof of Theorem 4.2.1a. We begin by proving the first part of Theorem 4.2.1, which essentially involves examining the second-order partial derivative.

$$
\begin{aligned}
\frac{\partial R(p,w)}{\partial w} &= -\mathbb{E}[\min(\lambda(p)+\varepsilon, \mu(w)+\delta)] + (p-w)\mu'(w)(1-F(\mu(w)-\lambda(p))), \\
\frac{\partial^2 R(p,w)}{\partial w^2} &= -2\mu'(w)(1-F(\mu(w)-\lambda(p))) + (p-w)\mu''(w)(1-F(\mu(w)-\lambda(p)))
\end{aligned}
$$

$$- (p-w)\left(\mu'(w)\right)^2 f(\mu(w)-\lambda(p)),$$

where $F(\cdot)$ and $f(\cdot)$ represent the CDF and PDF of the difference term $\varepsilon - \delta$. When $p > w$, Assumption 4.2.1d ensures $\frac{\partial^2 R}{\partial w^2} \leq 0$, leading to the concavity result.

Without loss of generality, suppose that $R\left(p, w_1\right) \geq R\left(p, w_2\right)$. Then we have

$$\begin{aligned}
&R\left(p, w_1\right) - R\left(p, w_2\right) \leq \left.\frac{\partial R(p,w)}{\partial w}\right|_{w=w_2} \left(w_1 - w_2\right) \\
&= \left(-\mathbb{E}\left[\min\left(\lambda\left(p\right)+\varepsilon, \mu\left(w_2\right)+\delta\right)\right] + \left(p-w_2\right)\mu'\left(w_2\right)\left(1-F\left(\mu\left(w_2\right)-\lambda\left(p\right)\right)\right)\right)\left(w_1-w_2\right).
\end{aligned}$$

Note that

$$-\bar{P}\bar{S} \leq -\mathbb{E}\left[\min\left(\lambda\left(p\right)+\varepsilon, \mu\left(w_2\right)+\delta\right)\right] + \left(p-w_2\right)\mu'\left(w_2\right)\left(1-F\left(\mu\left(w_2\right)-\lambda\left(p\right)\right)\right) \leq K_2\bar{P},$$

where the first inequality is from Assumption 4.2.1a and Assumption 4.2.1d, and the second inequality is from Assumption 4.2.1c. Consequently, $R\left(p, w_1\right) - R\left(p, w_2\right) \leq \max\{K_2, \bar{S}\}\bar{P}\left|w_1 - w_2\right|$. **Q.E.D.**

Proof of Theorem 4.2.1b. We prove the second part of Theorem 4.2.1. For ease of notation, we simply write $w_1^* = w^*(p_1)$ and $w_2^* = w^*(p_2)$. Without loss of generality, suppose $R\left(p_1, w_1^*\right) \geq R\left(p_2, w_2^*\right)$.

$$R\left(p_1, w_1^*\right) - R\left(p_2, w_2^*\right) \leq R\left(p_1, w_1^*\right) - R\left(p_2, \min\left(p_2, w_1^*\right)\right).$$

Then we have the following cases.

(1) If $w_1^* \leq p_2$, we have

$$\begin{aligned}
&R\left(p_1, w_1^*\right) - R\left(p_2, w_2^*\right) \\
&= \left(p_1 - w_1^*\right)\mathbb{E}\left[\min\left(\lambda\left(p_1\right)+\varepsilon, \mu\left(w_1^*\right)+\delta\right)\right] - \left(p_2 - w_1^*\right)\mathbb{E}\left[\min\left(\lambda\left(p_2\right)+\varepsilon, \mu\left(w_1^*\right)+\delta\right)\right].
\end{aligned} \tag{C.1}$$

(a) If $p_1 \geq p_2$, then $\lambda\left(p_1\right) \leq \lambda\left(p_2\right)$.

$$\text{(C.1)} \leq \left(p_1 - p_2\right)\mathbb{E}\left[\min\left(\lambda\left(p_2\right)+\varepsilon, \mu\left(w_1^*\right)+\delta\right)\right] \leq \bar{S}\left(p_1 - p_2\right).$$

(b) If $p_1 < p_2$, then $\lambda\left(p_1\right) > \lambda\left(p_2\right)$.

$$\text{(C.1)} \leq \left(p_2 - w_1^*\right)\left(\mathbb{E}\left[\min\left(\lambda\left(p_1\right)+\varepsilon, \mu\left(w_1^*\right)+\delta\right)\right] - \mathbb{E}\left[\min\left(x\left(p_2\right)+\varepsilon_1, \mu\left(w_1^*\right)+\delta\right)\right]\right)$$

$$
\begin{aligned}
&= (p_2 - w_1^*)\, \mathbb{E}\left[\min\left(\lambda\left(p_1\right) + \varepsilon, \mu\left(w_1^*\right) + \delta\right) - \min\left(\lambda\left(p_2\right) + \varepsilon, \mu\left(w_1^*\right) + \delta\right)\right] \\
&= (p_2 - w_1^*) \left(\int_{-\infty}^{\mu\left(w_1^*\right) - \lambda(p_1)} \left(\lambda\left(p_1\right) - \lambda\left(p_2\right)\right) f(\varepsilon - \delta) d(\varepsilon - \delta) \right. \\
&\qquad \left. + \int_{\mu\left(w_1^*\right) - \lambda(p_1)}^{\mu\left(w_1^*\right) - \lambda(p_2)} \left(\mu\left(w_1^*\right) + \delta - \lambda\left(p_2\right) - \varepsilon\right) f(\varepsilon - \delta) d(\varepsilon - \delta) \right).
\end{aligned}
$$

Because when $\varepsilon - \delta \in \left[\mu\left(w_1^*\right) - \lambda\left(p_1\right), \mu\left(w_1^*\right) - \lambda\left(p_2\right)\right]$, we have $\mu\left(w_1^*\right) + \delta \le \lambda\left(p_1\right) + \varepsilon$. Therefore,

$$
\begin{aligned}
(\text{C.1}) &\le (p_2 - w_1^*) \int_{-\infty}^{\mu\left(w_1^*\right) - \lambda(p_2)} \left(\lambda\left(p_1\right) - \lambda\left(p_2\right)\right) f(\varepsilon - \delta) d(\varepsilon - \delta) \\
&\le \bar{P}\left(\lambda\left(p_1\right) - \lambda\left(p_2\right)\right) \le \bar{P} K_1 \left|p_1 - p_2\right|,
\end{aligned}
$$

where the last inequality holds by Assumption 4.2.1b.

(2) If $w_1^* > p_2$, as $p_1 \ge w_1^*$, we have $p_1 > p_2$. Then

$$
\begin{aligned}
& R\left(p_1, w_1^*\right) - R\left(p_2, p_2\right) \\
=& R\left(p_1, w_1^*\right) - R\left(p_1, p_2\right) + R\left(p_1, p_2\right) - R\left(p_2, p_2\right) \\
\le& \left.\frac{\partial R\left(p_1, w\right)}{\partial w}\right|_{w = p_2} \left(w_1^* - p_2\right) + \left(p_1 - p_2\right) \bar{S} \\
\le& \left(\left.\frac{\partial R\left(p_1, w\right)}{\partial w}\right|_{w = p_2} + \bar{S} \right) \left(p_1 - p_2\right) \\
=& \left(-\mathbb{E}\left[\min\left(\lambda\left(p_1\right) + \varepsilon, \mu\left(p_2\right) + \delta\right)\right] + \left(p_1 - p_2\right) \mu'\left(p_2\right)\left(1 - F\left(\mu\left(p_2\right) - \lambda\left(p_1\right)\right)\right) + \bar{S}\right)\left(p_1 - p_2\right) \\
\le& \left(\max\{K_2, \bar{S}\} \bar{P} + \bar{S}\right)\left(p_1 - p_2\right),
\end{aligned}
$$

where the last inequality holds by Assumption 4.2.1c.

Combining the above cases, we obtain the desired Lipschitz continuity in p. **Q.E.D.**

C.2.2 Proof of Theorem 4.6.1

Proof of Theorem 4.6.1a. Define a new random variable $z := c_1 - c_2 + \varepsilon - \delta$ and $F(\cdot)$ and $f(\cdot)$ are the CDF and PDF of the random variable z. Since

$$
\begin{aligned}
R(\Delta, w) &= \Delta\left[c_1 - a_1(\Delta + w)\right] - \Delta \mathbb{E}\left[\left(z - a_1(\Delta + w) - a_2 w\right)^+\right], \\
\frac{\partial R(\Delta, w)}{\partial w} &= \Delta\left[a_2 - \left(a_1 + a_2\right) F\left(a_1 \Delta + a_1 w + a_2 w\right)\right],
\end{aligned}
$$

$$\frac{\partial^2 R(\Delta, w)}{\partial w^2} = -\left(a_1 + a_2\right)^2 \Delta f\left(a_1 p + a_2 w\right) < 0,$$

we have $\left|\frac{\partial R(\Delta,w)}{\partial w}\right| \leq \bar{P}\bar{a}$ where $\bar{a} = \max\{a_1, a_2\}$. Therefore Theorem 4.6.1a holds. **Q.E.D.**

Proof of Theorem 4.6.1b. Denote $w_1^*(\Delta) := \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right) - a_1\Delta}{a_1+a_2}$. Based on Theorem 4.6.1a, we can solve for the optimal w given any Δ by

$$w^*(\Delta) = \begin{cases} w_1^*(\Delta), & \text{if } \Delta < \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}, \\ 0, & \text{if } \Delta \geq \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}, \end{cases} \tag{C.2}$$

since we require $w \geq 0$.

Because $w^*(\Delta)$ is a piece-wise linear function of Δ, we first we consider when $\Delta < \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}$. By calculation, we have

$$\frac{d^2 R\left(\Delta, w_1^*(\Delta)\right)}{d\Delta^2} = -\frac{2a_1 w_2}{a_1 + a_2} < 0,$$

which means $(\Delta, w_1^*(\Delta))$ is concave in $\Delta \in \mathbb{R}$. So we can obtain the closed form of the optimizer for $(\Delta, w_1^*(\Delta))$ denoted as Δ_1^*:

$$\Delta_1^* := \frac{a_1 + a_2}{2a_1 a_2}\left(c_1 - \int_{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}^{\infty} z f(z) dz\right).$$

Then we consider when $\Delta \geq \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}$ where $w^*(\Delta) = 0$. We have

$$\frac{d^2 R(\Delta, 0)}{d\Delta^2} = -2a_1 F(a_1\Delta) - a_1^2 \Delta f(a_1\Delta) < 0,$$

which means $R(\Delta, 0)$ is concave in $\Delta \in \mathbb{R}$. Denote the optimizer of $R(\Delta, 0)$ as Δ_2^*, we have

$$\left.\frac{dR(\Delta, 0)}{d\Delta}\right|_{\Delta=\Delta_2^*} = c_1 - a_1\Delta_2^*[1 + F(a_1\Delta_2^*)] - \mathbb{E}\left[(z - a_1\Delta_2^*)^+\right] = 0.$$

We next combine these two pieces of functions and discuss the whole structure of

$R\left(\Delta, w^*(\Delta)\right)$, which is

$$R\left(\Delta, w^*(\Delta)\right) = \begin{cases} R\left(\Delta, w_1^*(\Delta)\right), \text{ if } \Delta < \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}, \\ R(\Delta, 0), \text{ if } \Delta \geq \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}. \end{cases}$$

Consider the sign of the value of the derivative of $R\left(\Delta, w_1^*(\Delta)\right)$ at the breakpoint $\Delta = \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}$:

$$\left.\frac{dR(\Delta,0)}{d\Delta}\right|_{\Delta=\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}} = c_1 - \frac{2a_2}{a_1+a_2}F^{-1}\left(\frac{a_2}{a_1+a_2}\right) - \int_{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}^{\infty} zf(z)dz. \tag{C.3}$$

1. If (C.3) ≤ 0, i.e. $\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1} \geq \Delta_2^*$, which means the breakpoint is no smaller than the optimal point of $R(\Delta, 0)$. Then $R(\Delta, 0)$ decreases in $\left[\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}, +\infty\right)$.

   Also, as (C.3) ≤ 0, we have

   $$c_1 - \int_{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}^{\infty} zf(z)dz \leq \frac{2a_2}{a_1+a_2}F^{-1}\left(\frac{a_2}{a_1+a_2}\right),$$

   which means $\Delta_1^* = \frac{a_1+a_2}{2a_1a_2}\left(c_1 - \int_{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}^{\infty} zf(z)dz\right) \leq \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}$ and Δ_1^* can thus be attained.

   So when (C.3) ≤ 0, $R\left(\Delta, w^*(\Delta)\right)$ increases in $(-\infty, \Delta_1^*)$ and decreases in $(\Delta_1^*, +\infty)$.

2. If (C.3) > 0, i.e. $\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1} < \Delta_2^*$, which means the breakpoint is to the left of the optimal point of $R(\Delta, 0)$. Then $R(\Delta, 0)$ first increases in $\left[\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}, \Delta_2^*\right)$ and then decreases in $[\Delta_2^*, \infty)$.

   Also, as (C.3) > 0, which means $c_1 - \int_{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}^{\infty} zf(z)dz > \frac{2a_2}{a_1+a_2}F^{-1}\left(\frac{a_2}{a_1+a_2}\right)$, i.e. $\Delta_1^* > \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}$ which means $R\left(\Delta, w_1^*(\Delta)\right)$ increases in $\left(-\infty, \frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}\right)$.

   So when (C.3) > 0, $R\left(\Delta, w^*(\Delta)\right)$ increases in $(-\infty, \Delta_2^*)$ and decreases in $[\Delta_2^*, +\infty)$.

In sum, $R\left(\Delta, w^*(\Delta)\right)$ consists of two concave functions $R\left(\Delta, w_1^*(\Delta)\right)$ and $R\left(\Delta, 0\right)$. Moreover, $R\left(\Delta, w^*(\Delta)\right)$ first increases and then decreases with the maximum point depending on

(C.3). Also, we have

$$\begin{aligned}
&\left.\frac{dR\left(\Delta, w_1^*(\Delta)\right)}{d\Delta}\right|_{\Delta=\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}}\\
=&c_1-\frac{2a_2}{a_1+a_2}F^{-1}\left(\frac{a_2}{a_1+a_2}\right)-\int_{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}^{\infty} zf(z)dz\\
=&\left.\frac{dR(\Delta,0)}{d\Delta}\right|_{\Delta=\frac{F^{-1}\left(\frac{a_2}{a_1+a_2}\right)}{a_1}},
\end{aligned}$$

which means the derivatives of the two sides of the breakpoint are the same. Also, as the two functions are both concave, we have $\mathbb{E}\left[R\left(\Delta, w^*(\Delta)\right)\right]$ is concave in Δ.

For the Lipschitz continuity, note that

$$\begin{aligned}
\frac{\partial R\left(\Delta, w_1^*(\Delta)\right)}{\partial \Delta}&=-\frac{2a_1a_2}{a_1+a_2}\Delta+c_1-\frac{a_1}{a_1+a_2}F^{-1}\left(\frac{a_2}{a_1+c_2}\right)-\mathbb{E}\left[\left(z-F^{-1}\left(\frac{a_2}{a_1+a_2}\right)\right)^+\right],\\
\frac{\partial R\left(\Delta, 0\right)}{\partial \Delta}&=-\mathbb{E}\left[\min\left\{c_1-a_1\Delta+\varepsilon, c_2+\delta\right\}\right]+\Delta F\left(a_1\Delta\right).
\end{aligned}$$

Therefore, by Assumption 4.2.1a, we have $-\left(\frac{2a_1a_2}{a_1+a_2}\bar{P}+\frac{2a_1+a_2}{a_1+a_2}\bar{S}\right)\leq\frac{\partial R\left(\Delta, w_1^*(\Delta)\right)}{\partial\Delta}\leq c_1$ and $0\leq\frac{\partial R(\Delta,0)}{\partial\Delta}\leq\bar{P}+\bar{S}$. Without loss of generality, we assume $R\left(\Delta_1, w^*(\Delta_1)\right)\leq R\left(\Delta_2, w^*(\Delta_2)\right)$. By concavity, we have

$$\begin{aligned}
R\left(\Delta_2, w^*(\Delta_2)\right)-R\left(\Delta_1, w^*(\Delta_1)\right)\leq&\max\left\{\left|\left.\frac{\partial R\left(\Delta, w_1^*(\Delta)\right)}{\partial\Delta}\right|_{\Delta=\Delta_2}\right|,\left|\left.\frac{\partial R\left(\Delta, 0\right)}{\partial\Delta}\right|_{\Delta=\Delta_2}\right|\right\}\left|\Delta_1-\Delta_2\right|\\
\leq&K_5\left|\Delta_1-\Delta_2\right|,
\end{aligned}$$

where $K_5=\max\left(c_1, \frac{2a_1a_2}{a_1+a_2}\bar{P}+\frac{2a_1+a_2}{a_1+a_2}\bar{S}, \bar{P}+\bar{S}\right)$. **Q.E.D.**

APPENDIX D

Appendix For Chapter 5

D.1 Summary of Major Notation

Table D.1: Summary of Major Notation for Problem Formulation

P	the price of a product
$\mathcal{P}$	$[p_1, p_2]$ the compact space of price values
Y	the amount of inventory available for sales
$\mathcal{Y}$	the set for inventory vales $\mathcal{Y} \subseteq [0, \infty)$
$D(p)$	the potential demand of a product if the price P is set to be a (deterministic) value p
q	number of dimensions of covariates associated with the product
X	observed q-dimensional covariates associated with the product
$\mathcal{X}$	some covariate space $\mathcal{X} \subset \mathbb{R}^q$
π	a pricing strategy being a measurable function: $(\mathcal{X}, \mathcal{Y}) \to \mathcal{P}$
Π	the class of all pricing strategies
$D(\pi)$	the potential outcome under a pricing strategy $\pi \in \Pi$
$V(\pi)$	the expected profit of a pricing strategy $\pi \in \Pi$
c	the stockout cost per unit
$f(P \mid X, Y)$	conditional probability density of the price, commonly referred to as the generalized propensity score
$f_{\min}$	an almost surely positive lower bound for the conditional probability density of the price $f(P \mid X, Y)$
$Q(X, Y, P)$	the expected profit of a product given the product covariates X, inventory amount Y and price P
π^*	the global optimal pricing strategy which maximizes the expected profit $V(\pi)$
$K(u)$	a kernel function $\mathbb{R} \to [0, \infty)$
h	the bandwidth used in kernel approximation
$V_h(\pi)$	approximated expected profit of pricing strategy π using kernel approximation with bandwidth h
C_1	a constant in Assumption 5.2.2(a)
C_2	a constant in Assumption 5.2.2(b)
C_3	a constant in Lemma 5.2.2
S	the observed sales quantity $S = \min\{D, Y\}$
Δ	censor indicator $\Delta = \mathbb{1}(D < Y)$
$R(X, P, S, \Delta)$	the surrogate profit given X, P, S, Δ of a product

Table D.2: Summary of Major Notation for Offline Feature-Based Pricing Strategy

$D_{\max}$	no-negative constant, upper bound of D in the assumption b
n	sample size

$\mathcal{D}_n$	n independent and identically distributed samples
$H(t\|X,P)$	the conditional survival function of the demand D, $H(t\|X,P)=\mathbb{P}(D>t\mid X,P)$
$\widehat{H}(t\|X,P)$	estimated conditional survival function using random forests method
ℓ	number of unique sales values in the dataset $\mathcal{D}_n$
$\widehat{\mathbb{E}}\left[D\|X,P,S,\Delta=0\right]$	estimated conditional expectation of demand using estimated $H(t\|X,P)$
$\widehat{R}(X,P,S,\Delta)$	estimated potential profit of a product
$\widehat{f}(P\mid X,Y)$	estimated conditional density function of the price
Π_0	some pre-specified class of pricing strategies
λ_n	a positive tuning parameter possibly depending on the sample size n
$\\|\cdot\\|_{\Pi_0}$	norm of the Hilbert space Π_0
$J(\pi)$	some regularization function on the policy π, set to be $\\|\pi\\|_{\Pi_0}^2$
β_0	constant term in the example of linear pricing strategies
β	parameters of the covariates in the example of linear pricing strategies, $\beta\in\mathbb{R}^{q+1}$
$\widehat{Q}(X,Y,P)$	estimated expected conditional potential profit $\widehat{Q}(X,Y,P)=\mathbb{E}[R\mid X,Y,P]$
$\widehat{V}_n^{DR}(\pi)$	a doubly robust estimator for estimating $V_h(\pi)$
$\widehat{\pi}$	the estimated global optimal pricing policy by solving (5.16)
$\widehat{\pi}_n$	the estimated global optimal pricing policy by solving (5.17)
B	number of survival trees in the random survival forests algorithm
M	number of folds in the cross-fitting technique
m	index of the fold in the cross-fitting technique, $m=1,\ldots,M$
$m(i)$	the fold containing the i-th observation
$\widehat{Q}^{(-m)}(X,Y,P)$	estimated expected conditional potential profit using data excluding fold m
$\widehat{f}^{(-m)}(P\mid X,Y)$	estimated conditional probability density of the price using data excluding fold m
$\mathcal{D}_n^{(-m)}$	the other $(M-1)$ folds data except k
ϕ_1	parameters of the neural network for the mean of the distribution of $(P\mid X,Y)$
ϕ_2	parameters of the neural network for the covariance of the distribution of $(P\mid X,Y)$
$\widehat{\phi}_1$	estimated ϕ_1 using MLE
$\widehat{\phi}_2$	estimated ϕ_2 using MLE
$\widehat{\mu}^{(-m(i))}(X_i,Y_i)$	estimated mean of the the multi-variate Gaussian distribution of $(P\mid X,Y)$ using $\widehat{\phi}_1$
$\widehat{\sigma}^{(-m(i))}(X_i,Y_i)$	estimated covariance matrix of the multivariate Gaussian distribution of the price $(P\mid X,Y)$ using $\widehat{\phi}_2$
ϕ_3	parameters of the neural network for the pricing policy
$\widehat{\phi}_3$	estimated ϕ_3 maximizing the right hand side of (5.17)

Table D.3: Summary of Major Notation for Regret Analysis and Double Robustness

Regret$(\widehat{\pi}_n)$	the regret of the pricing strategy $\widehat{\pi}_n$
C_4	a constant in Assumption 5.4.1
A	a constant in Assumption 5.4.2
v	a constant in Assumption 5.4.2
$\widetilde{Q}$	some probability measure on (X,Y)
$\\|\cdot\\|_{Q,2}$	the L_2-norm under $\widetilde{Q}$ on (X,Y)
F	the envelope function of Π_0
C_5	a constant in Assumption 5.4.3(a)
C_6	a constant in Assumption 5.4.3(b)
α,β	constants in Assumption 5.4.3(b)
ω	the smoothness coefficient of the true Q
π_h^*	the estimated optimal pricing policy by maximizing $V_h(\pi)$

D.2 Technical Proofs

Proof of Lemma 5.2.1. By definition, we have

$$\begin{aligned}
V(\pi) &= \mathbb{E}\left\{\pi(X,Y) \times \min\{D(\pi), Y\} - c \times (D(\pi) - Y)^+\right\}, \\
&= \mathbb{E}\left\{\mathbb{E}\left\{\pi(X,Y) \times \min\{D(\pi), Y\} - c \times (D(\pi) - Y)^+ \,|\, X, Y\right\}\right\}, \\
&= \mathbb{E}\left\{\mathbb{E}\left\{\pi(X,Y) \times \min\{D(\pi), Y\} - c \times (D(\pi) - Y)^+ \,|\, X, Y, P = \pi(X,Y)\right\}\right\}, \\
&= \mathbb{E}\left\{\int_{p_1}^{p_2} \mathbb{E}\left\{p \times \min\{D, Y\} - c \times (D - Y)^+ \,|\, X, Y, P = p\right\} \mathbb{1}(\pi(X,Y) = p) dp\right\}, \\
&= \mathbb{E}\left\{Q(X, Y, \pi(X,Y))\right\},
\end{aligned}$$

where the third equality is by Assumption 5.2.1(c) and the fourth equality is by Assumption 5.2.1(a). **Q.E.D.**

Proof of Lemma 5.2.2. We first consider the unbounded support of $\mathcal{P}$. As seen from Lemma 5.2.1, $V(\pi) = \mathbb{E}\left[Q(X,Y,\pi(X,Y))\right]$. By a similar derivation, we can show that

$$\begin{aligned}
V_h(\pi) &= \mathbb{E}\left\{\frac{Q(X,Y,P)K(\frac{P-\pi(X,Y)}{h})}{hf(P|X,Y)}\right\} \\
&= \mathbb{E}\left\{\int \frac{Q(X,Y,p)K(\frac{p-\pi(X,Y)}{h})}{h} dp\right\} \\
&= \mathbb{E}\left\{\int Q(X,Y,th + \pi(X,Y))K(t)dt\right\},
\end{aligned}$$

where the last equality is based on the change of variables. Then it can be seen that

$$\begin{aligned}
|V_h(\pi) - V(\pi)| &= \left|\mathbb{E}\left\{\int Q(X,Y,th + \pi(X,Y))K(t)dt\right\} - \mathbb{E}\left\{\int Q(X,Y,\pi(X,Y))\right\}\right| \\
&\leq \mathbb{E}\left\{\int \left|Q(X,Y,th + \pi(X,Y)) - Q(X,Y,\pi(X,Y))\right| K(t)dt\right\} \\
&\leq \mathbb{E}\left\{\sup_{p_1 \leq p < p' \leq p_2} \left|\frac{Q(X,Y,p) - Q(X,Y,p')}{p' - p}\right| \int |th|\, K(t)dt\right\} \\
&\leq hC_2 \int |t|\, K(t)dt \\
&\leq C_3 h,
\end{aligned}$$

where the second inequality is given by Assumption 5.2.2(b). When $\mathcal{P}$ has a bounded support, we need to normalize the kernel by

$$\widetilde{K}(\frac{p-\pi(X,Y)}{h}) = K(\frac{p-\pi(X,Y)}{h}) / \int_{p_1}^{p_2} K(\frac{p-\pi(X,Y)}{h}) dp.$$

Then by a similar proof as the unbounded case, we can show that $|V_h(\pi) - V(\pi)| \leq C_3 h$, which completes our proof. **Q.E.D.**

Proof of Lemma 5.2.3. Consider the following quantity:

$$\begin{aligned}
&\mathbb{E}\left[\left(P \times S - c \times (D-Y)^+\right) | X, P, S, Y\right] \\
&\quad = \mathbb{E}\left[\left(P \times S - c \times (D-Y)^+\right) | X, P, S, Y, \Delta = 1\right] \mathbb{1}(\Delta = 1) \\
&\qquad + \mathbb{E}\left[\left(P \times S - c \times (D-Y)^+\right) | X, P, S, Y, \Delta = 0\right] \mathbb{1}(\Delta = 0) \\
&\quad = \mathbb{1}(\Delta = 1) PS + \mathbb{1}(\Delta = 0) \mathbb{E}\left[\left(P \times S - c \times (D-Y)^+\right) | X, P, S, Y, \Delta = 0\right] \\
&\quad = \mathbb{1}(\Delta = 1) PS + \mathbb{1}(\Delta = 0) \mathbb{E}\left[\left(P \times S - c \times (D-Y)\right) | X, P, S, D > S, Y = S\right] \\
&\quad = \mathbb{1}(\Delta = 1) PS + (P+c) S \mathbb{1}(\Delta = 0) - c \mathbb{1}(\Delta = 0) \mathbb{E}\left[D | X, P, S, D > S\right] \\
&\quad = PS + cS\mathbb{1}(\Delta = 0) - c\mathbb{1}(\Delta = 0) \mathbb{E}\left[D | X, P, S, D > S\right] = R,
\end{aligned}$$

where the last but two equality is based on Assumption 5.2.3(a). In addition, we can rewrite $V(\pi)$ as

$$\begin{aligned}
V(\pi) &= \mathbb{E}\left\{\int_{p_1}^{p_2} \mathbb{E}\left[\left(P \times S - c \times (D-Y)^+\right) | X, Y, P = p\right] \mathbb{1}(\pi(X,Y) = p) dp\right\} \\
&= \mathbb{E}\left\{\int_{p_1}^{p_2} \mathbb{E}\left[R | X, Y, P = p\right] \mathbb{1}(\pi(X,Y) = p) dp\right\} \\
&= \mathbb{E}\left\{R | X, Y, P = \pi(X,Y)\right\},
\end{aligned}$$

which concludes our proof. **Q.E.D.**

Proof of Lemma 5.3.1. We show this lemma by interchanging the order of integration. Let $h(w \,|\, X, P)$ as the conditional probability density of survival function H. Note that

$$\begin{aligned}
&\int_S^{D_{\max}} \frac{H(t|X,P)}{H(S|X,P)} dt = \int_S^{D_{\max}} \int_t^{D_{\max}} \frac{h(w|X,P)}{H(S|X,P)} dw dt \\
&\quad = \int_S^{D_{\max}} \left\{\int_S^w \frac{h(w|X,P)}{H(S|X,P)} dt\right\} dw = \int_S^{D_{\max}} (w - S) \frac{h(w|X,P)}{H(S|X,P)} dw \\
&\quad = \int_S^{D_{\max}} w \frac{h(w|X,P)}{H(S|X,P)} dw - S = \mathbb{E}\left[D \,|\, X, P, S, \Delta = 0\right] - S,
\end{aligned}$$

which concludes the proof. **Q.E.D.**

Proof of Theorem 5.3.1. Since by the assumption our estimator $\widehat{R}$ and either $\widetilde{Q}$ or $\widetilde{f}$ are consistent in terms of sup-norm, without loss of generality, we assume $\widehat{R} = R$ and consider either $Q = \widetilde{Q}$ or $f = \widetilde{f}$. If $Q = \widetilde{Q}$, we have

$$\begin{aligned}\widehat{V}_n^{DR}(\pi) =& \frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2} Q(X_i, Y_i, p) K\left(\frac{p - \pi(X_i, Y_i)}{h}\right) dp \\ &+ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h\widetilde{f}(P_i|X_i, Y_i)} K\left(\frac{P_i - \pi(X_i, Y_i)}{h}\right)(R_i - Q(X_i, Y_i, P_i)).\end{aligned}$$

By the law of large numbers, we can show that $\widehat{V}_n^{DR}(\pi)$ converges in probability to

$$\frac{1}{h}\mathbb{E}\left[\int_{p_1}^{p_2} Q(X, Y, p) K\left(\frac{p - \pi(X, Y)}{h}\right) dp\right] + \mathbb{E}\left[\frac{1}{h\widetilde{f}(P\,|\,X, Y)} K\left(\frac{P - \pi(X, Y)}{h}\right)(R - Q(X, Y, P))\right] = V_h(\pi),$$

where the equation is given by $\mathbb{E}\left[R - Q(X, Y, P)\,|\,X, Y, P\right] = 0$. If $f = \widetilde{f}$, then by the law of large numbers again, we can show that $\widehat{V}_n^{DR}(\pi)$ converges in probability to

$$\begin{aligned}&\frac{1}{h}\mathbb{E}\left[\int_{p_1}^{p_2} \widetilde{Q}(X, Y, p) K\left(\frac{p - \pi(X, Y)}{h}\right) dp\right] + \mathbb{E}\left[\frac{1}{hf(P\,|\,X, Y)} K\left(\frac{P - \pi(X, Y)}{h}\right)(R - \widetilde{Q}(X, Y, P))\right] \\ &= V_h(\pi) + \frac{1}{h}\mathbb{E}\left[\int_{p_1}^{p_2} \widetilde{Q}(X, Y, p) K\left(\frac{p - \pi(X, Y)}{h}\right) dp\right] - \mathbb{E}\left[\frac{1}{hf(P\,|\,X, Y)} K\left(\frac{P - \pi(X, Y)}{h}\right)\widetilde{Q}(X, Y, P)\right] \\ &= V_h(\pi).\end{aligned}$$

The proof is complete by noticing that $|V_h(\pi) - V(\pi)| \leq C_3 h$ given by Lemma 5.2.2. **Q.E.D.**

Proof of Theorem 5.4.1. For notational simplicity, let $Z = (X, Y)$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We further let $U(\pi) = -V(\pi)$, $U_h(\pi) = -V_h(\pi)$. By Lemma 5.2.2, we can show that

$$\begin{aligned}V(\pi^*) - V(\widehat{\pi}_n) &= U(\widehat{\pi}_n) - U(\pi^*) \\ &\leq U_h(\widehat{\pi}_n) + C_3 h - U_h(\pi^*) + C_3 h + U_h(\pi_h^{\lambda_n}) + \lambda_n J(\pi_h^{\lambda_n}) + \lambda_n J(\widehat{\pi}_n) - \{U_h(\pi_h^{\lambda_n}) + \lambda_n J(\pi_h^{\lambda_n})\} \\ &\leq U_h(\pi_h^{\lambda_n}) + \lambda_n J(\pi_h^{\lambda_n}) - U_h(\pi_h^*) + U_h(\widehat{\pi}_n) + \lambda_n J(\widehat{\pi}_n) - \{U_h(\pi_h^{\lambda_n}) + \lambda_n J(\pi_h^{\lambda_n})\} + 2C_3 h \\ &= \Lambda(\lambda_n) + \underbrace{U_h(\widehat{\pi}_n) + \lambda_n J(\widehat{\pi}_n) - \{U_h(\pi_h^{\lambda_n}) + \lambda_n J(\pi_h^{\lambda_n})\}}_{(I)} + 2C_3 h,\end{aligned}$$

where $\pi_h^{\lambda_n} \in \arg\min_{\pi \in \Pi_0}\{U_h(\pi) + \lambda_n J(\pi)\}$. In the following, we apply the empirical process theory to bound Term (I) on the right hand side of the inequality above. Let

$$\begin{aligned}\mathcal{G}_\pi \triangleq \Bigg\{&\int_{p_1}^{p_2} Q(Z, p)\frac{K((p - \pi_h^{\lambda_n}(Z))/h)}{h} dp + \frac{1}{hf(P|Z)} K(\frac{P - \pi_h^{\lambda_n}(Z)}{h})(R - Q(Z, P)) + \lambda_n J(\pi) \\ &- \int_{p_1}^{p_2} Q(Z, p)\frac{K((p - \pi(Z))/h)}{h} dp - \frac{1}{hf(P|Z)} K(\frac{P - \pi(Z)}{h})(R - Q(Z, P)) - \lambda_n J(\pi_h^{\lambda_n}) \,|\, J(\pi) \lesssim \lambda_n^{-1}, \pi \in \Pi_0\Bigg\}.\end{aligned}$$

Based on the definition of $\mathcal{G}_\pi$, we use g_π to denote any generic element in $\mathcal{G}_\pi$. Recall that $J(\pi) = \|\pi\|_{\Pi_0}^2$. We consider a constraint class on π by the following argument. By Assumptions 5.4.1 and 5.4.3(b), all nuisance functions in (5.17) are bounded. Then according to the optimization property, we can show that

$$\begin{aligned}&\frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(Z_i,p)K(\frac{p-\widehat{\pi}_n(Z_i)}{h})dp+\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\widehat{f}^{(-m(i))}(P_i|Z_i)}K(\frac{P_i-\widehat{\pi}_n(Z_i)}{h})(\widehat{R}_i-\widehat{Q}(Z_i,P_i))+\lambda_n J(\widehat{\pi}_n)\\ &\leq\frac{1}{nh}\sum_{i=1}^{n}\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(Z_i,p)K(\frac{p}{h})dp+\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\widehat{f}^{(-m(i))}(P_i|Z_i)}K(\frac{P_i}{h})(\widehat{R}_i-\widehat{Q}(Z_i,P_i)),\end{aligned}$$

which implies that $\lambda_n J(\widehat{\pi}_n) \lesssim 1$. Based on this, we can further show that for any $g_\pi \in \mathcal{G}_\pi$,

$$\|g_\pi\|_\infty \lesssim 1/h + \|\pi\|_{\Pi_0}/h \lesssim \lambda_n^{-\frac{1}{2}}/h,$$

since K is Lipschitz with respect to $\|\bullet\|_{\Pi_0}$ and $\lambda_n \to 0$ with $\lambda_n \leq 1$. The remaining proof consists of two steps. In the first step, we show

$$\mathbb{E}_n(g_{\widehat{\pi}_n}) \leq \varepsilon_1,$$

for some $\varepsilon_1 > 0$ with a high probability. In the second step, we aim to show that

$$\sup_{g_\pi\in\mathcal{G}_\pi}|\mathbb{E}_n(g_\pi) - \mathbb{E}(g_\pi)| \leq \varepsilon_2,$$

with a high probability for some ε_2. Then combining two, we are able to show $(I) \leq \varepsilon_1 + \varepsilon_2$ with some high probability.

Step 1: We first notice that

$$\begin{aligned}&\mathbb{E}_n(g_{\widehat{\pi}_n})\\ &=\mathbb{E}_n\left\{\int_{p_1}^{p_2}Q(Z,p)\frac{K((p-\pi_h^{\lambda_n}(Z))/h)}{h}dp+\frac{1}{hf(P|Z)}K(\frac{P-\pi_h^{\lambda_n}(Z)}{h})(R-Q(Z,P))\right\}+\lambda_n J(\widehat{\pi}_n)\\ &\quad-\mathbb{E}_n\left\{\int_{p_1}^{p_2}Q(Z,p)\frac{K((p-\widehat{\pi}_n(Z))/h)}{h}dp+\frac{1}{hf(P|Z)}K(\frac{P-\widehat{\pi}_n(Z)}{h})(R-Q(Z,P))\right\}-\lambda_n J(\pi_h^{\lambda_n})\\ &=\mathbb{E}_n\left\{\int_{p_1}^{p_2}Q(Z,p)\frac{K((p-\pi_h^{\lambda_n}(Z))/h)}{h}dp+\frac{1}{hf(P|Z)}K(\frac{P-\pi_h^{\lambda_n}(Z)}{h})(R-Q(Z,P))\right\}\\ &\quad-\mathbb{E}_n\left\{\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(Z,p)\frac{K((p-\pi_h^{\lambda_n}(Z))/h)}{h}dp+\frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{P-\pi_h^{\lambda_n}(Z)}{h})(\widehat{R}-\widehat{Q}^{(-m(i))})(Z,P))\right\}\\ &\quad+\mathbb{E}_n\left\{\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(Z,p)\frac{K((p-\pi_h^{\lambda_n}(Z))/h)}{h}dp+\frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{P-\pi_h^{\lambda_n}(Z)}{h})(\widehat{R}-\widehat{Q}^{(-m(i))}(Z,P))\right\}-\lambda_n J(\pi_h^{\lambda_n})\\ &\quad-\mathbb{E}_n\left\{\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(Z,p)\frac{K((p-\widehat{\pi}_n(Z))/h)}{h}dp+\frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{P-\widehat{\pi}_n(Z)}{h})(\widehat{R}-\widehat{Q}^{(-m(i))}(Z,P))\right\}+\lambda_n J(\widehat{\pi}_n)\\ &\quad+\mathbb{E}_n\left\{\int_{p_1}^{p_2}\widehat{Q}^{(-m(i))}(Z,p)\frac{K((p-\widehat{\pi}_n(Z))/h)}{h}dp+\frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{P-\widehat{\pi}_n(Z)}{h})(\widehat{R}-\widehat{Q}^{(-m(i))}(Z,P))\right\}\\ &\quad-\mathbb{E}_n\left\{\int_{p_1}^{p_2}Q(Z,p)\frac{K((p-\widehat{\pi}_n(Z))/h)}{h}dp+\frac{1}{hf(P|Z)}K(\frac{P-\widehat{\pi}_n(Z)}{h})(R-Q(Z,P))\right\}\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}_n \left\{ \int_{p_1}^{p_2} Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(R-Q(Z,P)) \right\} \\
&\quad - \mathbb{E}_n \left\{ \int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(\widehat{R}-\widehat{Q}^{(-m(i))})(Z,P)) \right\} \\
&\quad + \mathbb{E}_n \left\{ \int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(Z,p)\frac{K((\widehat{\pi}_n(Z)-p)/h)}{h}dp + \frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{\widehat{\pi}_n(Z)-P}{h})(\widehat{R}-\widehat{Q}^{(-m(i))}(Z,P)) \right\} \\
&\quad - \mathbb{E}_n \left\{ \int_{p_1}^{p_2} Q(Z,p)\frac{K((\widehat{\pi}_n(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\widehat{\pi}_n(Z)-p}{h})(R-Q(Z,P)) \right\},
\end{aligned}
$$

where the last inequality is given by the optimization property in (5.17). In the following, we bound right hand side of the above inequality. It suffices to focus on the first two terms on the right hand side while the other two terms can be bounded similarly.

Specifically, we consider bounding the following term, defined as

$$
\begin{aligned}
E_1 \triangleq \mathbb{E}_n &\left\{ \int_{p_1}^{p_2} Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(R-Q(Z,P)) \right\} \\
&- \mathbb{E}_n \left\{ \int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(\widehat{R}-\widehat{Q}^{(-m(i))}(Z,P)) \right\}
\end{aligned}
$$

We remark that we can write

$$
\begin{aligned}
&\int_{p_1}^{p_2} Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(R-Q(Z,P)) \\
=&\int_{p_1}^{p_2} \underbrace{Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h} + \frac{\mathbb{1}(P=p)}{hf(p|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-p}{h})(R-Q(Z,p))}_{E_1(p)}\,dp,
\end{aligned}
$$

where $\mathbb{1}(P=p)$ is indeed a Dirac measure. For a fix p, it can be seen that

$$
\begin{aligned}
E_1(p) = &\frac{1}{nh}\sum_{i=1}^{n}(1-\frac{\mathbb{1}(P_i=p)}{f(p|Z_i)})(\widehat{Q}^{-m(i)}(Z_i,P_i)-Q(Z_i,p))K(\frac{\widehat{\pi}_n(Z_i)-p}{h}) \\
&+\frac{1}{nh}\sum_{i=1}^{n}(\frac{\mathbb{1}(P_i=p)}{\widehat{f}^{-m(i)}(p|Z_i)}-\frac{\mathbb{1}(P_i=p)}{f(p|Z_i)})(R_i-Q(Z_i,P_i))K(\frac{\widehat{\pi}_n(Z_i)-p}{h}) \\
&+\frac{1}{nh}\sum_{i=1}^{n}(\frac{\mathbb{1}(P_i=p)}{\widehat{f}^{-m(i)}(p|Z_i)}-\frac{\mathbb{1}(P_i=p)}{f(p|Z_i)})(\widehat{R}_i-\widehat{Q}^{-m(i)}(Z_i,P_i)-(R_i-Q(Z_i,P_i)))K(\frac{\widehat{\pi}_n(Z_i)-p}{h}) \\
&+\frac{1}{nh}\sum_{i=1}^{n}\frac{\mathbb{1}(P_i=p)}{f(p|Z_i)}(\widehat{R}_i-R_i)K(\frac{\widehat{\pi}_n(Z_i)-p}{h}) \\
&\triangleq E_2(p)+E_3(p)+E_4(p)+E_5(p).
\end{aligned}
$$

In the following, we bound each of the above four terms. For $E_3(p)$, consider

$$
\mathcal{G}_{1,\pi} \triangleq \left\{ \int_{p_1}^{p_2} (\frac{\mathbb{1}(P=p)}{\widehat{f}^{-(k)}(p|Z)}-\frac{\mathbb{1}(P=p)}{f(p|Z)})(R-Q(Z,P))K(\frac{\pi(Z)-P}{h})dp \,|\, J(\pi)\leq\lambda_n^{-1}, \pi\in\Pi_0 \right\}.
$$

By the sample splitting, we can show that $\mathbb{E}\left[R - Q(Z,P)|Z,P,f^{-(m(i))}(p|Z)\right] = 0$. Therefore we can observe that $\mathbb{E}[g_\pi] = 0$ for any $g_\pi \in \mathcal{G}_{1,\pi}$. In addition, the envelop function of $\mathcal{G}_1$, defined as G_1, is proportional to $\int_{p_1}^{p_2} |\frac{\mathbb{1}(P=p)}{\widehat{f}^{-(k)}(p|Z)} - \frac{\mathbb{1}(P=p)}{f(p|Z)}||R - Q(Z,P)|\lambda_n^{-\frac{1}{2}}/hdp$ by the Lipschitz boundness on K in Assumption 5.2.2(a). Therefore $\|G_1\|_{2,P} \lesssim n^{-\beta}\lambda_n^{-\frac{1}{2}}/h$ by the error bound condition on $\widehat{f}^{-(m)}(p|Z)$ given in Assumption 5.4.3(b). By the entropy condition in Assumption 5.4.2 and Lipschitz property of K in Assumption 5.2.2(a), we can further show that

$$\sup_{\widetilde{Q}} N(\mathcal{G}_{1,\pi}, \widetilde{Q}, \varepsilon\|G_1\|_{2,\widetilde{Q}}) \lesssim \left(\frac{1}{\varepsilon}\right)^v,$$

which implies that

$$J(1, \mathcal{G}_{1,\pi}, G_1) \triangleq \int_0^1 \sup_{\widetilde{Q}} \sqrt{\log N(\mathcal{G}_{1,\pi}, \widetilde{Q}, \varepsilon\|G_1\|_{2,\widetilde{Q}})}d\varepsilon \lesssim \sqrt{v}.$$

By leveraging the maximal inequality in the empirical process theory, we can show that

$$\mathbb{E} \sup_{g \in \mathcal{G}_{1,\pi}} |\mathbb{E}_n g| \lesssim \sqrt{v} n^{-\frac{1}{2}} n^{-\beta} \lambda_n^{-\frac{1}{2}}/h.$$

Then by Talagrand's inequality, we can show with probability $1 - e^{-x}$,

$$\begin{aligned}\int_{p_1}^{p_2} E_3(p)dp &\lesssim \frac{1}{h}\left\{\mathbb{E} \sup_{g \in \mathcal{G}_{1,\pi}} |\mathbb{E}_n g| + 2\sqrt{x}\sqrt{\frac{4\sqrt{v} n^{-\frac{1}{2}-\beta}\lambda_n^{-1}/h^2 + C_0 n^{-2\beta}\lambda_n^{-1}/h^2}{n}} + \frac{3x\lambda_n^{-\frac{1}{2}}}{nh}\right\} \\ &\lesssim \max\{1,x\}\sqrt{v} n^{-\frac{1}{2}} n^{-\beta} \lambda_n^{-\frac{1}{2}}/h^2.\end{aligned}$$

Similarly, we can show

$$\int_{p_1}^{p_2} E_2(p)dp \lesssim \max\{1,x\}\sqrt{v} n^{-\frac{1}{2}} n^{-\alpha} \lambda_n^{-\frac{1}{2}}/h^2,$$

with probability at least $1 - e^{-x}$. In addition, we can bound $\int_{p_1}^{p_2} E_4(p)dp$ term by Cauchy-Schwarz inequality, i.e., with probability at least $1 - 2e^{-x}$,

$$\begin{aligned}\int_{p_1}^{p_2} E_4(p)dp &\leq 1/h^2\left(\mathbb{E}_n\left[\frac{1}{\widehat{f}^{-(m)}(P|Z)} - \frac{1}{f(P|Z)}\right]^2\right)^{\frac{1}{2}} \times \left(\mathbb{E}_n\left[\widehat{R} - \widehat{Q}^{-m(i)}(Z,P) - (R - Q(Z,P))\right]^2\right)^{\frac{1}{2}} \lambda_n^{-\frac{1}{2}} \\ &\leq 1/h^2\left(\mathbb{E}_n\left[\frac{1}{\widehat{f}^{-(m)}(P|Z)} - \frac{1}{f(P|Z)}\right]^2\right)^{\frac{1}{2}} \times \left\{\left(\mathbb{E}_n\left[\widehat{Q}^{-m(i)}(Z,P) - Q(Z,P)\right]^2\right)^{\frac{1}{2}} + n^{-\delta}\right\}\lambda_n^{-\frac{1}{2}} \\ &\lesssim \max\{1,x\}\left(n^{-(\alpha+\beta)} + n^{-\beta}n^{-\delta}\right)\lambda_n^{-\frac{1}{2}}/h^2.\end{aligned}$$

The last inequality is due to Bernstein's inequality, i.e.,

$$\mathbb{E}_n\left[\frac{1}{\widehat{f}^{-(k)}(p|Z)}-\frac{1}{f(p|Z)}\right]^2 \lesssim n^{-2\beta}+n^{-\frac{1}{2}-\beta}\sqrt{2x}+\frac{x}{3n},$$

and

$$\mathbb{E}_n\left[\widehat{Q}^{-m(i)}(Z,P)-Q(Z,P)\right]^2 \lesssim n^{-2\beta}+n^{-\frac{1}{2}-\beta}\sqrt{2x}+\frac{x}{3n},$$

by the uniformly bounded assumption in Assumptions 5.4.1 and 5.4.3(b) and the error bound condition on nuisance function estimation in Assumption 5.4.3(b).

For the last term $\int_{p_1}^{p_2} E_5(p)dp$, we can show that with probability at least $1-e^x-\varepsilon$,

$$\begin{aligned}\int_{p_1}^{p_2} E_5(p)dp &= \frac{1}{nh}\sum_{i=1}^n \frac{1}{f(P_i|Z_i)}(\widehat{R}_i-R_i)K(\frac{\widehat{\pi}_n(Z_i)-p}{h})\\ &\lesssim C_5(\varepsilon)\frac{1}{nh^2}\sum_{i=1}^n \mathbb{1}(\Delta_i=0)n^{-\delta}\lambda_n^{-1/2}\\ &\lesssim C_5(\varepsilon)\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\left(\mathbb{P}(\Delta=0)+\sqrt{\frac{x}{n}}\right).\end{aligned}$$

Combining the results above together, we can show that with probability at least $1-5e^{-x}-\varepsilon$,

$$\begin{aligned}\int_{p_1}^{p_2} E_1(p)dp \lesssim{}& \max\{1,x\}\sqrt{v}n^{-\frac{1}{2}}n^{-\min(\beta,\alpha)}\lambda_n^{-\frac{1}{2}}/h^2\\ &+\max\{1,x\}n^{-(\alpha+\beta)}\lambda_n^{-\frac{1}{2}}/h^2+C_5(\varepsilon)\max\{1,x\}\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\mathbb{P}(\Delta=0).\end{aligned}$$

Similar results can be obtained if we replace $\widehat{\pi}_n$ by $\pi_h^{\lambda_n}$ in E_1. Then we have

$$\begin{aligned}\mathbb{E}_n(g_{\widehat{\pi}_n}) \lesssim{}& \max\{1,x\}\sqrt{v}n^{-\frac{1}{2}}n^{-\min(\beta,\alpha)}\lambda_n^{-\frac{1}{2}}/h^2\\ &+\max\{1,x\}n^{-(\alpha+\beta)}\lambda_n^{-\frac{1}{2}}/h^2+C_5(\varepsilon)\max\{1,x\}\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\mathbb{P}(\Delta=0),\end{aligned}$$

with probability $1-10\exp(-x)-2\varepsilon$.

Step 2: Again by applying Talagrand's inequality and maximal inequality, we can similarly show that with probability at least $1-e^{-x}$,

$$\sup_{g_\pi\in\mathcal{G}_\pi}|\mathbb{E}_n(g_\pi)-\mathbb{E}(g_\pi)| \lesssim \max\{1,x\}\sqrt{v}\lambda_n^{-\frac{1}{2}}n^{-\frac{1}{2}}/h^2.$$

Summarizing Steps 1 and 2, we can show that with probability $1-e^{-x}-\varepsilon$,

$$
\begin{aligned}
\mathbf{Regret}(\widehat{\pi}_n) =& V(\pi^*) - V(\widehat{\pi}_n) \\
&\lesssim \Lambda(\lambda_n) + 2C_3 h + \max\{1,x\}\sqrt{v}\lambda_n^{-\frac{1}{2}} n^{-\frac{1}{2}}/h^2 \\
&\quad + \max\{1,x\}\sqrt{v} n^{-\frac{1}{2}} n^{-\min(\beta,\alpha)} \lambda_n^{-\frac{1}{2}}/h^2 \\
&\quad + \max\{1,x\} n^{-(\alpha+\beta)} \lambda_n^{-\frac{1}{2}}/h^2 + C_5(\varepsilon)\max\{1,x\}\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\mathbb{P}(\Delta=0).
\end{aligned}
$$

which concludes our proof. **Q.E.D.**

D.3 Dependent Data Scenario and Its Analysis

In this section, we consider the scenario where $\{(X_i, Y_i, P_i, S_i, \Delta_i)\}_{1\leq i\leq n}$ are not i.i.d. copies. Instead, we assume our data come from M centers, where data collected at each center are dependent and cross the center are independent. Specifically, for center $1 \leq m \leq M$, our offline data consist of $\left\{(X_t^{(m)}, Y_t^{(m)}, P_t^{(m)}, S_t^{(m)}, \Delta_t^{(m)})\right\}_{1\leq t\leq n}$. Since data across different centers are independent, we can apply the previously proposed method (5.17) to learn the optimal pricing strategy that

$$
\begin{aligned}
\widehat{\pi}_n \in \underset{\pi\in\Pi_0}{\arg\max} \Bigg\{ & \frac{1}{nMh}\sum_{i=1}^{nM}\int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(X_i, Y_i, p) K\left(\frac{p-\pi(X_i,Y_i)}{h}\right) dp \\
& + \frac{1}{nM}\sum_{i=1}^{nM} \frac{1}{h\widehat{f}^{(-m(i))}(P_i|X_i,Y_i)} K\left(\frac{P_i-\pi(X_i,Y_i)}{h}\right)(\widehat{R}_i - \widehat{Q}^{(-m(i))}(X_i,Y_i,P_i)) - \lambda_n J(\pi)\Bigg\},
\end{aligned}
\tag{D.1}
$$

where $m(i)$ denotes the center containing the i-th observation.

In the following, we provide a theoretical guarantee for our approach under the non-i.i.d. case. For any two σ-fields $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{F}$, we define $\beta(\mathcal{A},\mathcal{B}) := \sup\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left|P\left(A_i\cap B_j\right) - P\left(A_i\right)P\left(B_j\right)\right|$ where the supremum is taken over all pairs of (finite) partitions $\{A_1,\ldots,A_I\}$ and $\{B_1,\ldots,B_J\}$ of Ω such that $A_i \in \mathcal{A}$ for each i and $B_j \in \mathcal{B}$ for each j. We also define the σ-field for the sequence of random variables $X := (X_k, k\in\mathbb{Z})$. For each $n \geq 1$, the dependence coefficient (β-mixing coefficient) is defined as $\beta(n) := \sup_{j\in\mathbb{Z}}\beta\left(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{\infty}\right)$. Then the random sequence X is said to be β-mixing if $\beta(n)\to 0$ as $n\to\infty$. We now make the following assumption to characterize the dependency among observations in each center.

Assumption D.3.1 *For each $1 \leq m \leq M$, the stochastic process $\left\{(X_t^{(m)}, Y_t^{(m)}, P_t^{(m)}, S_t^{(m)}, \Delta_t^{(m)})\right\}_{t\geq 1}$ is a stationary and exponential β-mixing process with β-mixing coefficient $\beta(t) \leq \beta_0\exp(-\beta_1 t)$ for some $\beta_0 \geq 0$ and $\beta_1 > 0$.*

Assumption D.3.1 characterizes the dependence among observations in each center. An exponential β-mixing process is a type of stochastic process that satisfies certain conditions regarding the dependence structure of its random variables. Specifically, a process is β-mixing if the correlation between two variables decreases exponentially fast as the time separation between them increases. The β-mixing coefficient is a measure of the rate at which the correlation between two variables decays as the time separation between them increases. A smaller β-mixing coefficient indicates a faster decay of correlation and hence a weaker dependence structure. In particular, the upper bound on the β-mixing coefficient at the time lag t basically means that the dependence decays to 0 at the slowest possible exponential rate with respect to t. See Bradley (2005) for more details. Without loss of generality, we assume $M = 2$. Then we have the following regret guarantee for our estimated optimal pricing strategy given in (D.1).

Theorem D.3.1 *Suppose that Assumptions 5.2.1–D.3.1 hold. If $\lambda_n \leq 1$ and $\alpha + \beta > 1/2$, then for any $\varepsilon \in (0, 1)$ with probability at least $1 - 1/n - \varepsilon$, Algorithm 10 admits the following regret upper bound*

$$\begin{aligned}\textbf{Regret}(\widehat{\pi}_n) \lesssim \Lambda(\lambda_n) + 2C_3 h + \log(n)\sqrt{v}\lambda_n^{-\frac{1}{2}} n^{-\frac{1}{2}}/h^2 \\ + C_5(\varepsilon)\log(n)\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\mathbb{P}(\Delta = 0),\end{aligned} \tag{D.2}$$

where the regret is defined in (5.19)*. Furthermore, if* (5.21) *holds, then with probability at least* $1 - 1/n - \varepsilon$

$$\textbf{Regret}(\widehat{\pi}_n) \lesssim \log(n) n^{-\frac{\zeta \min(\frac{1}{2}, \delta)}{6\zeta + 1}}. \tag{D.3}$$

Under Assumption D.3.1, one can obtain a similar regret result as that in Theorem 5.4.1.

Lemma D.3.1 *Suppose that a stochastic process $\{Z_t\}_{t\geq 1}$ is a stationary and exponential β-mixing process with β-mixing coefficient $\beta(q) \leq \beta_0 \exp(-\beta_1 q)$ for some $\beta_0 \geq 0$ and $\beta_1 > 0$. Let $\mathcal{G}$ be a class of measurable functions that take Z_t as input. For any $g \in \mathcal{G}$, assume $\mathbb{E}[g(Z_t)] = 0$ for any $t \geq 0$. Suppose that the envelop function of $\mathcal{G}$ is uniformly bounded by some constant $C > 0$. In addition, if $\mathcal{G}$ belongs to the class of VC-typed functions such that $\sup_{\widetilde{Q}} \mathcal{N}(\mathcal{G}, \widetilde{Q}, \epsilon \| \bullet \|_{\widetilde{Q},2}) \lesssim (1/\epsilon)^\alpha$ for a constant $\alpha > 0$. Then with a probability at least $1 - 1/T$,*

$$\sup_{g\in\mathcal{G}} \left| \frac{1}{T}\sum_{t=1}^{T} g(Z_{i,t}) \right| \lesssim \log(T)\sqrt{\frac{\alpha}{T}}.$$

If $\sigma^2 = \sup_{g\in\mathcal{G}} \mathbb{E}[g^2(Z_t)]$ for $1 \le t \le T$, then with probability at least $1 - 1/T$,

$$\sup_{g\in\mathcal{G}} \left| \frac{1}{T} \sum_{t=1}^{T} g(Z_{i,t}) \right| \lesssim \frac{\log(T)(\sqrt{\log(T)\alpha}\|G\|_2 + \sqrt{T\sigma^2} + 1)}{T}.$$

Proof of Lemma D.3.1. To prove the case (i) of the lemma, we focus on bounding $\sup_{g\in\mathcal{G}} \left|\sum_{t=1}^{T} g(Z_t)\right|$. Specifically, we apply Berbee's coupling lemma (Berbee 1979) and follow the remark below Lemma 4.1 of Dedecker and Louhichi (2002). Let q be some positive integer. One can always construct a sequence $\{\widetilde{Z}_t\}_{t\ge1}$ such that with probability at least $1 - (T\beta(q))/q$,

$$\sup_{g\in\mathcal{G}} \left| \sum_{t=1}^{T} g(Z_t) \right| = \sup_{g\in\mathcal{G}} \left| \sum_{t=1}^{T} g(\widetilde{Z}_t) \right|.$$

In the same time, the block sequence $\widetilde{X}_k(g) = \{g(\widetilde{Z}_{(k-1)q+j})\}_{1\le j\le q}$ are identically distributed for $k \ge 1$. In addition, the sequence $\{\widetilde{X}_k(g) \mid k = 2\omega, \omega \ge 1\}$ are independent and so are the sequence $\{\widetilde{X}_k(g) \mid k = 2\omega+1, \omega \ge 0\}$. Let $I_r = \{\lfloor T/q\rfloor q+1, \cdots, T\}$ with $\mathrm{Card}(I_r) < q$. Then we can show that with probability at least $1 - (T\beta(q))/q$,

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} \left| \sum_{t=1}^{T} g(Z_t) \right| \\
\le &\sup_{g\in\mathcal{G}} \left| \sum_{t=1}^{q\lfloor T/q\rfloor} g(\widetilde{Z}_t) \right| + \sup_{g\in\mathcal{G}} \left| \sum_{t\in I_r} g(Z_t) \right|.
\end{aligned}$$

In the following, we always assume that the above inequality holds. Then it is sufficient to bound each of the above two terms separately. First of all, without loss of generality, we assume $\lfloor T/q\rfloor$ is an even number. Then for the first term, we have

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} \left| \sum_{t=1}^{q\lfloor T/q\rfloor} g(\widetilde{Z}_t) \right| \\
\le &\sum_{j=1}^{2q} \sup_{g\in\mathcal{G}} \left| \sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i) \right|.
\end{aligned}$$

By the previous construction, $\sup_{g\in\mathcal{G}} \left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right|$ is a suprema empirical process of i.i.d. sequences. Then by conditions in Lemma D.3.1 and Mcdiarmid's inequality, we have with probability at least $1-\varepsilon$,

$$\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right| \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right|\right] + \sqrt{\frac{T\log(1/\varepsilon)}{q}}.$$

Given the condition that

$$\sup_{\widetilde{Q}}\mathcal{N}(\mathcal{G},\widetilde{Q},\epsilon\|\bullet\|_{\widetilde{Q},2}) \lesssim (1/\epsilon)^{\alpha}.$$

By a standard maximal inequality using uniform entropy integral (e.g., Van Der Vaart and Wellner 2011), we can show that with probability at least $1-\varepsilon$,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right|\right] \lesssim \sqrt{\frac{\alpha T}{q}}.$$

By letting $\varepsilon = 1/T$, we can show that with probability at least $1-1/T$,

$$\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right| \lesssim \sqrt{\frac{\alpha T}{q}} + \sqrt{\frac{T\log(T)}{q}}.$$

Next, we can bound $\sup_{g\in\mathcal{G}}\left|\sum_{t\in I_r} g(Z_t)\right|$ by Cq. By letting $q \asymp \log(T)$, we can show that with probability at least $1-1/T$,

$$\begin{aligned}&\sup_{g\in\mathcal{G}}\left|\sum_{t=1}^{T} g(Z_t)\right| \\ \lesssim&\log(T)\sqrt{\frac{\alpha T}{\log(T)}} + \log(T)\sqrt{\frac{T\log(T)}{\log(T)}} + \log(T) \\ \lesssim&\log(T)\sqrt{T\alpha}.\end{aligned}$$

This concludes our proof of case (i) by dividing both sides by T.

In the second part of our proof, we have $\sigma^2 = \sup_{g\in\mathcal{G}}\mathbb{E}[g^2(Z_t)]$ for $1 \le t \le T$. Then by conditions in Lemma D.3.1 and Talagrand's inequality, we have with probability at least $1-\varepsilon$,

$$\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right| \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right|\right] + \sqrt{2\eta_n\log(1/\varepsilon)} + C\log(1/\varepsilon),$$

where $\eta_n = 2T/q\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right|\right] + T/q\sigma^2$. We can deploy another maximal inequality to show that

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right|\right] \lesssim J(1,\mathcal{G},G)\|G\|_2 \lesssim \sqrt{\alpha}\|G\|_2.$$

by letting $\varepsilon = 1/T$, we can show that with probability at least $1-1/T$,

$$\sup_{g\in\mathcal{G}}\left|\sum_{k=1}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_i)\right| \lesssim \sqrt{\alpha}\|G\|_2 + \sqrt{\left(\sqrt{\alpha}\|G\|_2 + \frac{T}{q}\sigma^2\right)\log(T)} + C\log(T).$$

By letting $q \asymp \log(T)$, we can show that with probability at least $1-1/T$,

$$\begin{aligned}&\sup_{g\in\mathcal{G}}\left|\sum_{t=1}^{T} g(Z_t)\right| \\ &\lesssim \log(T)\sqrt{\alpha}\|G\|_2 + \log(T)\sqrt{\left(\sqrt{\alpha}\|G\|_2 + \frac{T}{\log(T)}\sigma^2\right)\log(T)} + \log(T) \\ &\lesssim \log(T)(\sqrt{\log(T)\alpha}\|G\|_2 + \sqrt{T\sigma^2} + 1)\end{aligned}$$

The result follows by dividing both sides by T. **Q.E.D.**

Proof of Theorem D.3.1. For notational simplicity, let $Z_t = (X_t, Y_t)$ for $1 \le t \le n$ and $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$. We further let $U(\pi) = -V(\pi)$, $U_h(\pi) = -V_h(\pi)$. By similar derivation, we can show that

$$V(\pi^*) - V(\widehat{\pi}_n) \le \Lambda(\lambda_n) + \underbrace{U_h(\widehat{\pi}) + \lambda_n J(\widehat{\pi}) - \{U_h(\pi_h^{\lambda_n}) + \lambda_n J(\pi_h^{\lambda_n})\}}_{(I)} + 2C_3 h,$$

where $\pi_h^{\lambda_n} \in \arg\min_{\pi\in\Pi_0}\{U_h(\pi) + \lambda_n J(\pi)\}$. In the following, we apply Lemma D.3.1 to bound Term (I) on the right hand side of the inequality above. Let

$$\begin{aligned}\mathcal{G}_\pi \triangleq \Bigg\{&\int_{p_1}^{p_2} Q(Z,p)\frac{K((p-\pi_h^{\lambda_n}(Z))/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{P-\pi_h^{\lambda_n}(Z)}{h})(R-Q(Z,P)) + \lambda_n J(\pi) \\ &-\int_{p_1}^{p_2} Q(Z,p)\frac{K((p-\pi(Z))/h)}{h}dp - \frac{1}{hf(P|Z)}K(\frac{P-\pi(Z)}{h})(R-Q(Z,P)) - \lambda_n J(\pi_h^{\lambda_n}) \,|\, J(\pi) \lesssim \lambda_n^{-1}, \pi\in\Pi_0\Bigg\}.\end{aligned}$$

Recall that $J(\pi) = \|\pi\|_{\Pi_0}^2$. We consider a constraint class on π based on the same argument in Theorem 5.4.1. In the first step, we show

$$\mathbb{E}_n(g_{\widehat{\pi}_n}) \le \varepsilon_1,$$

for some $\varepsilon_1 > 0$ with a high probability. In the second step, we aim to show that

$$\sup_{g_\pi \in \mathcal{G}_\pi} |\mathbb{E}_n(g_\pi) - \mathbb{E}(g_\pi)| \leq \varepsilon_2,$$

with a high probability for some ε_2. Then combining two, we are able to show $(I) \leq \varepsilon_1 + \varepsilon_2$ with some high probability.

Step 1: We can similarly derive that

$$\begin{aligned}
&\mathbb{E}_n(g_{\widehat{\pi}_n}) \\
&\leq \mathbb{E}_n\left\{\int_{p_1}^{p_2} Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(R-Q(Z,P))\right\} \\
&\quad - \mathbb{E}_n\left\{\int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(\widehat{R}-\widehat{Q}^{(-m(i))})(Z,P))\right\} \\
&\quad + \mathbb{E}_n\left\{\int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(Z,p)\frac{K((\widehat{\pi}_n(Z)-p)/h)}{h}dp + \frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{\widehat{\pi}_n(Z)-P}{h})(\widehat{R}-\widehat{Q}^{(-m(i))}(Z,P))\right\} \\
&\quad - \mathbb{E}_n\left\{\int_{p_1}^{p_2} Q(Z,p)\frac{K((\widehat{\pi}_n(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\widehat{\pi}_n(Z)-p}{h})(R-Q(Z,P))\right\},
\end{aligned}$$

In the following, we bound the right-hand side of the above inequality. It suffices to focus on the first two terms on the right-hand side while the other two terms can be bounded similarly.

Specifically, we consider bounding the following term, defined as

$$\begin{aligned}
E_1 \triangleq\ &\mathbb{E}_n\left\{\int_{p_1}^{p_2} Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(R-Q(Z,P))\right\} \\
&- \mathbb{E}_n\left\{\int_{p_1}^{p_2} \widehat{Q}^{(-m(i))}(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{h\widehat{f}^{(-m(i))}(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(\widehat{R}-\widehat{Q}^{(-m(i))})(Z,P)\right\}
\end{aligned}$$

We again notice that

$$\begin{aligned}
&\int_{p_1}^{p_2} Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h}dp + \frac{1}{hf(P|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-P}{h})(R-Q(Z,P)) \\
&= \int_{p_1}^{p_2} \underbrace{Q(Z,p)\frac{K((\pi_h^{\lambda_n}(Z)-p)/h)}{h} + \frac{\mathbb{1}(P=p)}{hf(p|Z)}K(\frac{\pi_h^{\lambda_n}(Z)-p}{h})(R-Q(Z,p))}_{E_1(p)}\,dp,
\end{aligned}$$

where $\mathbb{1}(P=p)$ is indeed a Dirac measure. For a fix p, it can be seen that

$$\begin{aligned}
E_1(p) = &\frac{1}{nh}\sum_{i=1}^{n}(1-\frac{\mathbb{1}(P_i=p)}{f(p|Z_i)})(\widehat{Q}^{-m(i)}(Z_i,P_i)-Q(Z_i,p))K(\frac{\widehat{\pi}_n(Z_i)-p}{h}) \\
&+ \frac{1}{nh}\sum_{i=1}^{n}(\frac{\mathbb{1}(P_i=p)}{\widehat{f}^{-m(i)}(p|Z_i)} - \frac{\mathbb{1}(P_i=p)}{f(p|Z_i)})(R_i-Q(Z_i,P_i))K(\frac{\widehat{\pi}_n(Z_i)-p}{h}) \\
&+ \frac{1}{nh}\sum_{i=1}^{n}(\frac{\mathbb{1}(P_i=p)}{\widehat{f}^{-m(i)}(p|Z_i)} - \frac{\mathbb{1}(P_i=p)}{f(p|Z_i)})(\widehat{R}_i-\widehat{Q}^{-m(i)}(Z_i,P_i)-(R_i-Q(Z_i,P_i)))K(\frac{\widehat{\pi}_n(Z_i)-p}{h})
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{nh} \sum_{i=1}^{n} \frac{\mathbb{1}(P_i = p)}{f(p|Z_i)} (\widehat{R}_i - R_i) K(\frac{\widehat{\pi}_n(Z_i) - p}{h}) \\
&\triangleq E_2(p) + E_3(p) + E_4(p) + E_5(p).
\end{aligned}
$$

In the following, we bound each of the above four terms. For $E_3(p)$,consider

$$
\mathcal{G}_{1,\pi} \triangleq \left\{ \int_{p_1}^{p_2} (\frac{\mathbb{1}(P = p)}{\widehat{f}^{-(k)}(p|Z)} - \frac{\mathbb{1}(P = p)}{f(p|Z)})(R - Q(Z, P))K(\frac{\pi(Z) - P}{h})dp \,|\, J(\pi) \leq \lambda_n^{-1}, \pi \in \Pi_0 \right\}.
$$

By the problem setting and the independence across centers, we can show that $\mathbb{E}\left[R - Q(Z, P)|Z, P, f^{-(m(i))}(p|Z)\right] = 0$. Therefore we can observe that $\mathbb{E}[g_\pi] = 0$ for any $g_\pi \in \mathcal{G}_{1,\pi}$. In addition, the envelop function of $\mathcal{G}_1$, defined as G_1, is proportional to $\int_{p_1}^{p_2} |\frac{\mathbb{1}(P=p)}{\widehat{f}^{-(k)}(p|Z)} - \frac{\mathbb{1}(P=p)}{f(p|Z)}||R - Q(Z, P)|\lambda_n^{-\frac{1}{2}}/hdp$ by the Lipschitz boundness on K in Assumption5.2.2(a). Therefore $\|G_1\|_{2,P} \lesssim n^{-\beta}\lambda_n^{-\frac{1}{2}}/h$ by the error bound condition on $\widehat{f}^{-(m)}(p|Z)$ given in Assumption 5.4.3(b). By the entropy condition in Assumption 5.4.2 and Lipschitz property of K in Assumption 5.2.2(a), we can further show that

$$
\sup_{\widetilde{Q}} N(\mathcal{G}_{1,\pi}, \widetilde{Q}, \varepsilon\|G_1\|_{2,\widetilde{Q}}) \lesssim \left(\frac{1}{\varepsilon}\right)^{v},
$$

which implies that

$$
J(1, \mathcal{G}_{1,\pi}, G_1) \triangleq \int_0^1 \sup_{\widetilde{Q}} \sqrt{\log N(\mathcal{G}_{1,\pi}, \widetilde{Q}, \varepsilon\|G_1\|_{2,\widetilde{Q}})} d\varepsilon \lesssim \sqrt{v}.
$$

By leveraging the result in Lemma D.3.1, we can show that with probability at least $1 - 1/n$,

$$
\int_{p_1}^{p_2} E_3(p)dp \lesssim \log(n)\sqrt{v}n^{-\frac{1}{2}}n^{-\beta}\lambda_n^{-\frac{1}{2}}/h^2.
$$

Similarly, we can show

$$
\int_{p_1}^{p_2} E_2(p)dp \lesssim \log(n)\sqrt{v}n^{-\frac{1}{2}}n^{-\alpha}\lambda_n^{-\frac{1}{2}}/h^2,
$$

with probability at least $1 - 1/n$. In addition, we can bound $\int_{p_1}^{p_2} E_4(p)dp$ term by Cauchy-Schwarz inequality, i.e., with probability at least $1 - 1/n$,

$$
\int_{p_1}^{p_2} E_4(p)dp \leq 1/h^2 \left(\mathbb{E}_n \left[\frac{1}{\widehat{f}^{-(m)}(P|Z)} - \frac{1}{f(P|Z)}\right]^2\right)^{\frac{1}{2}} \times \left(\mathbb{E}_n \left[\widehat{R} - \widehat{Q}^{-m(i)}(Z, P) - (R - Q(Z, P))\right]^2\right)^{\frac{1}{2}} \lambda_n^{-\frac{1}{2}}
$$

$$
\begin{aligned}
&\leq 1/h^2 \left(\mathbb{E}_n \left[\frac{1}{\widehat{f}^{-(m)}(P|Z)} - \frac{1}{f(P|Z)} \right]^2 \right)^{\frac{1}{2}} \times \left\{ \left(\mathbb{E}_n \left[\widehat{Q}^{-m(i)}(Z,P) - Q(Z,P) \right]^2 \right)^{\frac{1}{2}} + n^{-\delta} \right\} \lambda_n^{-\frac{1}{2}} \\
&\lesssim \log(n) \left(n^{-(\alpha+\beta)} + n^{-\beta} n^{-\delta} \right) \lambda_n^{-\frac{1}{2}} / h^2.
\end{aligned}
$$

The last inequality is due to Bernstein's inequality in the dependent case using the uniformly bounded assumption in Assumptions 5.4.1 and 5.4.3(b) and the error bound condition on nuisance function estimation in Assumption 5.4.3(b). See theorem 8 of Fu et al. (2022) for more details.

For the last term $\int_{p_1}^{p_2} E_5(p)dp$, we can show that with probability at least $1-\varepsilon$,

$$
\begin{aligned}
\int_{p_1}^{p_2} E_5(p)dp &= \frac{1}{nh} \sum_{i=1}^n \frac{1}{f(P_i|Z_i)} (\widehat{R}_i - R_i) K(\frac{\widehat{\pi}_n(Z_i) - p}{h}) \\
&\lesssim C_5(\varepsilon) \frac{1}{nh^2} \sum_{i=1}^n \mathbb{1}(\Delta_i = 0) n^{-\delta} \lambda_n^{-1/2} \\
&\lesssim C_5(\varepsilon) \frac{n^{-\delta} \lambda_n^{-1/2}}{h^2} \left(\mathbb{P}(\Delta = 0) + \sqrt{\frac{x}{n}} \right).
\end{aligned}
$$

Combining the results above together, we can show that with probability at least $1 - 5/n - \varepsilon$,

$$
\begin{aligned}
\int_{p_1}^{p_2} E_1(p)dp \lesssim & \log(n) \sqrt{v} n^{-\frac{1}{2}} n^{-\min(\beta,\alpha)} \lambda_n^{-\frac{1}{2}} / h^2 \\
& + \log(n) n^{-(\alpha+\beta)} \lambda_n^{-\frac{1}{2}} / h^2 + C_5(\varepsilon) \log(n) \frac{n^{-\delta} \lambda_n^{-1/2}}{h^2} \mathbb{P}(\Delta = 0).
\end{aligned}
$$

Similar results can be obtained if we replace $\widehat{\pi}_n$ by $\pi_h^{\lambda_n}$ in E_1. Then we have

$$
\begin{aligned}
\mathbb{E}_n(g_{\widehat{\pi}_n}) \lesssim & \log(n) \sqrt{v} n^{-\frac{1}{2}} n^{-\min(\beta,\alpha)} \lambda_n^{-\frac{1}{2}} / h^2 \\
& + \log(n) n^{-(\alpha+\beta)} \lambda_n^{-\frac{1}{2}} / h^2 + C_5(\varepsilon) \log(n) \frac{n^{-\delta} \lambda_n^{-1/2}}{h^2} \mathbb{P}(\Delta = 0),
\end{aligned}
$$

with probability $1 - 10/n - 2\varepsilon$.

Step 2: Again by applying Lemma D.3.1, we can similarly show that with probability at least $1 - 1/n$,

$$
\sup_{g_\pi \in \mathcal{G}_\pi} |\mathbb{E}_n(g_\pi) - \mathbb{E}(g_\pi)| \lesssim \log(n) \sqrt{v} \lambda_n^{-\frac{1}{2}} n^{-\frac{1}{2}} / h^2.
$$

Summarizing Steps 1 and 2, we can show that with probability $1 - 1/n - \varepsilon$,

$$
\begin{aligned}
\mathbf{Regret}(\widehat{\pi}_n) =& V(\pi^*) - V(\widehat{\pi}_n) \\
& \lesssim \Lambda(\lambda_n) + 2C_3 h + \log(n)\sqrt{v}\lambda_n^{-\frac{1}{2}} n^{-\frac{1}{2}}/h^2 \\
& \quad + \log(n)\sqrt{v} n^{-\frac{1}{2}} n^{-\min(\beta,\alpha)} \lambda_n^{-\frac{1}{2}}/h^2 \\
& \quad + \log(n) n^{-(\alpha+\beta)} \lambda_n^{-\frac{1}{2}}/h^2 + C_5(\varepsilon)\log(n)\frac{n^{-\delta}\lambda_n^{-1/2}}{h^2}\mathbb{P}(\Delta = 0).
\end{aligned}
$$

which concludes our proof for the first statement. The second statement holds the same argument as that in Corollary 5.4.1. **Q.E.D.**

D.4 Numerical Experiment Supplementary

D.4.1 Robustness Experiments

Figures D.1 – D.4 are the results of the experiments in Table 5.1.

D.4.2 Sensitivity Analysis

The neural network parameters' sensitivity will be instance-dependent. For the instance we use, after experiments with various neural network parameters for the neural networks generating $\widehat{Q}(X, Y, P)$, $\hat{f}(P \mid X, Y)$, and $\widehat{\pi}_n$, we find the outputs are not too sensitive to the parameters. For example, Figures D.5 – D.8 are the results of the same instance obtained using four sets of parameters, where the numbers in the parentheses denote the number of neurons in each hidden layer and the number of scalars denotes the number of hidden layers used for the neural network for estimating each variable:

Set	$\widehat{Q}(X,Y,P)$	$\hat{f}(P\|X,Y)$	$\widehat{\pi}_n$
1	(100,100)	(48)	(12)
2	(100,100)	(48)	(24)
3	(100,100)	(24)	(12)
4	(200,100)	(48)	(12)

Table D.4: Neural Network Parameters

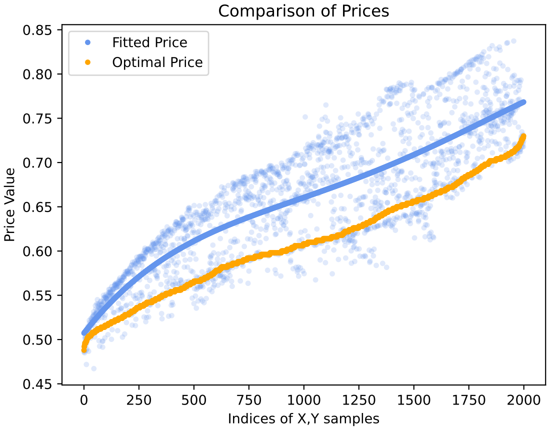


Figure D.1: Instance 1

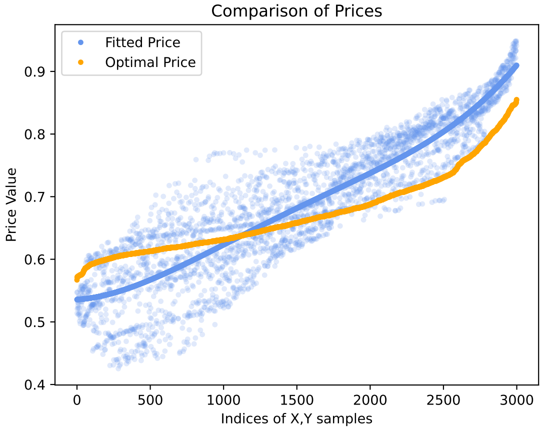


Figure D.2: Instance 2

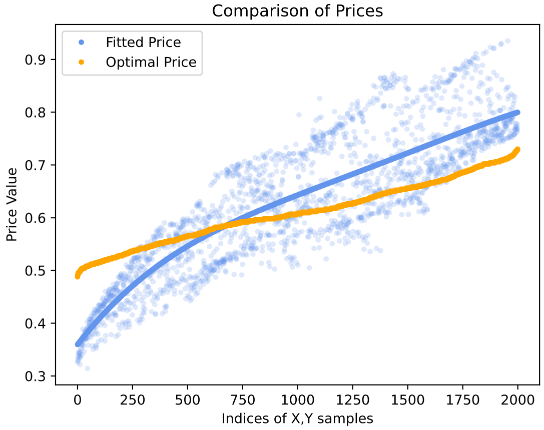


Figure D.3: Instance 3

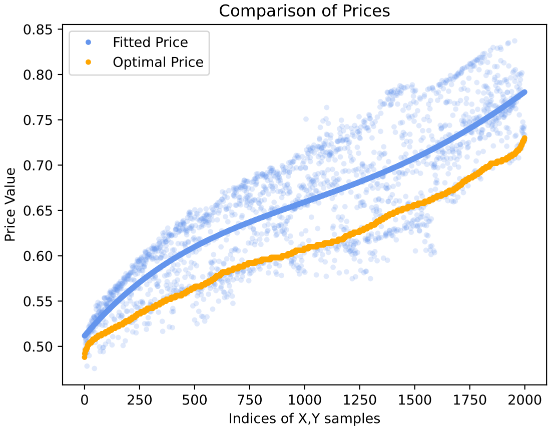


Figure D.4: Instance 4

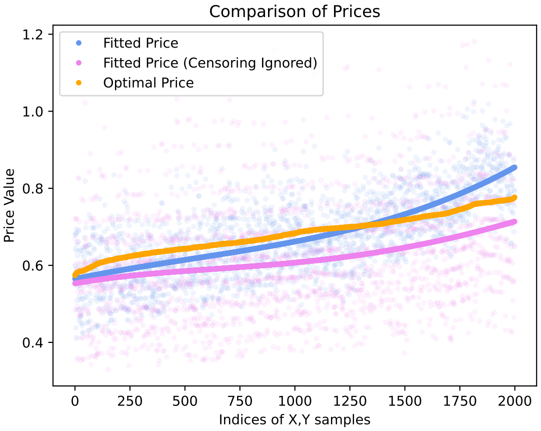


Figure D.5: Set 1

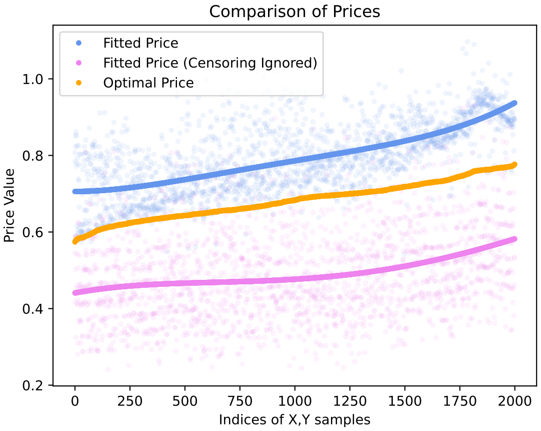


Figure D.6: Set 2

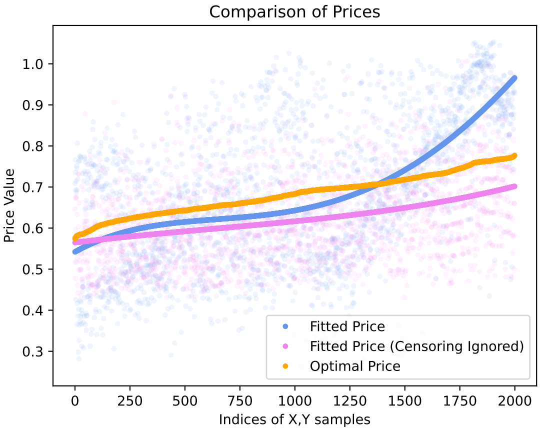


Figure D.7: Set 3

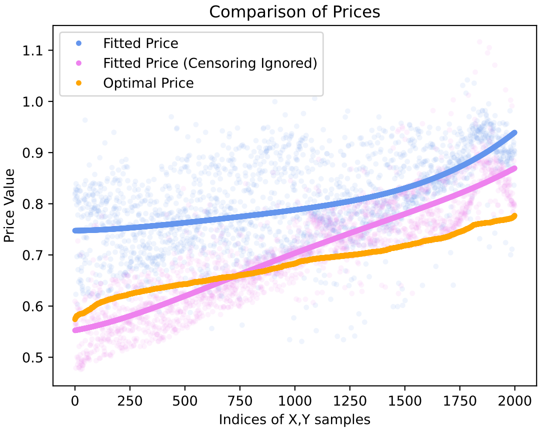


Figure D.8: Set 4

D.4.3 Running Time

The main time-consuming part of the algorithm lies in 1) in each fold, the training of the neural networks used to estimate $\widehat{Q}^{(-m(i))}(X_i, Y_i, p)$, $\widehat{Q}^{(-m(i))}(X_i, Y_i, P_i)$ and $\widehat{f}^{(-m(i))}(P_i|X_i, Y_i)$; 2) the training of the neural network used to estimate $\widehat{\pi}_n$. Note that the number of folds used in cross-validation is usually not too large. In our case, we choose $K = 3$. Below we report the running time to train the neural network for each statistic, averaged over 100 instances of sample size 2000 and also 3 folds if in the inner loop, with the value in the parenthesis being the standard deviation. So in general, it takes around 457.28 seconds to run our algorithm for an instance with a sample size of 2000 which is reasonable.

Neural Network	$\widehat{Q}^{(-m(i))}(X_i, Y_i, p)$	$\widehat{Q}^{(-m(i))}(X_i, Y_i, P_i)$	$\widehat{f}^{(-m(i))}(P_i\|X_i, Y_i)$	$\widehat{\pi}_n$
Average Running Time (s)	33.95(13.00)	30.26(12.35)	22.34(1.96)	197.73(16.29)

Table D.5: Running Time

D.4.4 Random Survival Forests Description

We briefly describe the random survival forests method introduced by (Ishwaran et al. 2008). Define the censoring indicator to be 0 if the data is right-censored and otherwise 1. Given a data set with each record comprising the individual's survival time and the $0-1$ censoring indicator, the random survival forests algorithm consists of the following steps:

1. First, we draw B bootstrap samples from the original data where B is a given parameter. Note that each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data).

2. For each bootstrap sample, grow a survival tree, where p candidate variables randomly selected are used at each node. Then the node is split using the candidate variable that maximizes survival difference between daughter nodes to separate the dissimilar cases, by searching over all possible x variables and split values c, and choosing that x^* and c^* that maximizes survival difference. The tree is grown to a full size such that a terminal node should have at least $d_0 > 0$ unique deaths, where d_0 is also a specified parameter.

3. Calculate a cumulative hazard function (CHF) for each tree. Average to obtain the ensemble CHF.

4. Calculate prediction error for the ensemble CHF using OOB data.

Using the non-parametric random survival forests method above, we can obtain the estimated CHF $h(t\,|\,X,P)$ for each record (X,P) and thus the survival function $H(t\,|\,X,P)$ in (5.10) via $H(t\,|\,X,P)=e^{-h(t\,|\,X,P)}$.

BIBLIOGRAPHY

Aalen O (1978) Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6(4):701–726. 139

Agarwal A, Foster DP, Hsu DJ, Kakade SM, Rakhlin A (2011) Stochastic convex optimization with bandit feedback. Shawe-Taylor J, ed., *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 1035–1043, NIPS'11 (Red Hook, NY, USA: Curran Associates Inc.). 8, 76, 77, 80, 88, 89, 93, 94, 98, 104, 105, 106

Agrawal R (1995) The continuum-armed bandit problem. *SIAM Journal on Control and Optimization* 33(6):1926–1951. 80

Agrawal S, Jia R (2022) Learning in structured MDPs with convex cost functions: Improved regret bounds for inventory management. *Operations Research* 70(3):1646–1664. 16, 43, 116

Allon G, Van Mieghem JA (2010) Global dual sourcing: Tailored base-surge allocation to near-and offshore production. *Management Science* 56(1):110–124. 43, 70, 174

Armstrong M (2006) Competition in two-sided markets. *The RAND Journal of Economics* 37(3):668–691. 79

Asmussen S (2003) *Applied probability and queues*, volume 2 (Springer). 160

Auer P, Cesa-Bianchi N, Fischer P (2002a) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47:235–256. 2

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002b) The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32(1):48–77. 2

Auer P, Jaksch T, Ortner R (2008) Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems* 21. 1

Azizzadenesheli K, Lazaric A, Anandkumar A (2016) Reinforcement learning of pomdps using spectral methods. *Conference on Learning Theory*, 193–256 (PMLR). 55

Baccara M, Lee S, Yariv L (2020) Optimal dynamic matching. *Theoretical Economics* 15(3):1221–1278. 14

Bai J, So KC, Tang CS, Chen X, Wang H (2019) Coordinating supply and demand on an on-demand service platform with impatient customers. *Manufacturing & Service Operations Management* 21(3):556–570. 78

Ban GY (2020) Confidence intervals for data-driven inventory policies with demand censoring. *Operations Research* 68(2):309–326. 16, 115

Ban GY, Keskin NB (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science* 67(9):5549–5568. 113, 116

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108. 4, 16, 115

Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973. 8, 114, 127

Barron AR (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning* 14(1):115–133. 132

Bassok Y, Anupindi R, Akella R (1999) Single-period multiproduct inventory models with substitution. *Operations Research* 47(4):632–642. 14

Benjaafar S, Hu M (2020) Operations management in the age of the sharing economy: What is old and what is new? *Manufacturing & Service Operations Management* 22(1):93–101. 75

Berbee HC (1979) *Random walks with stationary increments and renewal theory* (Mathematisch Centrum, Amsterdam, Netherlands). 197

Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Operations Research* 63(5):1227–1244. 70, 174, 175, 176

Besbes O, Muharremoglu A (2013) On implications of demand censoring in the newsvendor problem. *Management Science* 59(6):1407–1424. 116, 123

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* 57(6):1407–1425. 79, 82

Besbes O, Zeevi A (2012) Blind network revenue management. *Operations Research* 60(6):1537–1550. 79, 82

Bickel PJ (1982) On adaptive estimation. *Annals of Statistics* 10(3):647–671. 128

Boucheron S, Lugosi G, Massart P (2013) *Concentration inequalities: A nonasymptotic theory of independence* (Oxford University Press, Oxford, UK). 2, 103

Bradley RC (2005) Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys* 2(1):107–144. 196

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32. 2

Bu J, Gong X, Yao D (2020) Constant-order policies for lost-sales inventory models with random supply functions: Asymptotics and heuristic. *Operations Research* 68(4):1063–1073. 78

Bu J, Simchi-Levi D, Wang L (2023) Offline pricing and demand learning with censored data. *Management Science* 69(2):885–903. 4, 112, 115, 120, 144

Bu J, Simchi-Levi D, Xu Y (2022) Online pricing with offline data: Phase transition and inverse square law. *Management Science* 68(12):8568–8588. 115

Bulinskaya EV (1964) Some results concerning optimum inventory policies. *Theory of Probability & Its Applications* 9(3):389–403. 41, 43

Cai H, Shi C, Song R, Lu W (2021) Jump interval-learning for individualized decision making. Technical report, North Carolina State University, Raleigh, NC. Available at arXiv:2111.08885. 117

Chen B, Chao X (2020a) Dynamic inventory control with stockout substitution and demand learning. *Management Science* 66(11):5108–5127. 16

Chen B, Chao X (2020b) Dynamic inventory control with stockout substitution and demand learning. *Management Science* 66(11):5108–5127. 116

Chen B, Chao X, Ahn HS (2019a) Coordinating pricing and inventory replenishment with nonparametric demand learning. *Operations Research* 67(4):1035–1052. 8, 16, 44, 73, 77, 88, 116

Chen B, Chao X, Shi C (2021a) Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. *Mathematics of Operations Research* 46(2):726–756. 16, 44, 73, 116

Chen B, Chao X, Wang Y (2020a) Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Operations Research* 68(5):1445–1456. 44, 116

Chen B, Shi C (2020) Tailored base-surge policies in dual-sourcing inventory systems with demand learning. Technical report, University of Michigan, Ann Arbor, MI. Available at SSRN 3456834. 4, 8, 16, 44, 73, 77, 80, 88

Chen B, Simchi-Levi D, Wang Y, Zhou Y (2022a) Dynamic pricing and inventory control with fixed ordering cost and incomplete demand information. *Management Science* 68(8):5684–5703. 4, 16, 44, 73, 116

Chen B, Wang Y, Zhou Y (2020b) Optimal policies for dynamic pricing and inventory control with nonparametric censored demands. Technical report, University of Illinois at Chicago, Chicago, IL. Available at SSRN 3750413. To appear in *Management Science.* 16, 78

Chen G, Zeng D, Kosorok MR (2016) Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association* 111(516):1509–1521. 117

Chen J (1997) *Substitution and inspection models in production-inventory systems.* Ph.D. thesis, Columbia University, New York, NY. 14, 21

Chen L, Plambeck EL (2008) Dynamic inventory management with learning about the demand distribution and substitution probability. *Manufacturing & Service Operations Management* 10(2):236–256. 15

Chen N, Gallego G (2021) Nonparametric pricing analytics with customer covariates. *Operations Research* 69(3):974–984. 113, 116

Chen N, Gallego G (2022) A primal–dual learning algorithm for personalized dynamic pricing with an inventory constraint. *Mathematics of Operations Research* 47(4):2585–2613. 79

Chen Q, Jasin S, Duenyas I (2019b) Nonparametric self-adjusting control for joint learning and optimization of multiproduct pricing with finite resource capacity. *Mathematics of Operations Research* 44(2):601–631. 79

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. 2

Chen W, Shi C, Duenyas I (2020c) Optimal learning algorithms for stochastic inventory systems with random capacities. *Production and Operations Management* 29(7):1624–1649. 16, 43

Chen X (2007) Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* 6(76):5549–5632. 132

Chen X, Gao X, Pang Z (2018) Preservation of structural properties in optimization with decisions truncated by random variables and its applications. *Operations Research* 66(2):340–357. 78

Chen X, Jasin S, Shi C (2022b) *The Elements of Joint Learning and Optimization in Operations Management* (Springer, New York, NY). 79

Chen X, Miao S, Wang Y (2021b) Differential privacy in personalized pricing with nonparametric demand models. Technical report, New York University, New York, NY. Available at SSRN 3919807. To appear in Operations Research. 117

Chen X, Simchi-Levi D (2004a) Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case. *Operations Research* 52(6):887–896. 76, 109

Chen X, Simchi-Levi D (2004b) Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The infinite horizon case. *Mathematics of Operations Research* 29(3):698–723. 76

Chen X, Simchi-Levi D (2012) *Pricing and Inventory Management* (The Oxford Handbook of Pricing Management, Oxford University Press, Oxford, UK). 78

Chen X, Simchi-Levi D, Wang Y (2022c) Privacy-preserving dynamic personalized pricing with demand learning. *Management Science* 68(7):4878–4898. 117

Chen X, Zhang X, Zhou Y (2021c) Fairness-aware online price discrimination with nonparametric demand models. Technical report, New York University, New York, NY. Available at arXiv:2111.08221. 8, 77, 88, 117

Chen Y, Shi C (2023) Network revenue management with online inverse batch gradient descent method. *Production and Operations Management* 32(7):2123–2137, URL http://dx.doi.org/https://doi.org/10.1111/poms.13960. 4, 79

Chen YJ, Dai T, Korpeoglu CG, Körpeoğlu E, Sahin O, Tang CS, Xiao S (2020d) OM Forum—innovative online platforms: Research opportunities. *Manufacturing & Service Operations Management* 22(3):430–445. 73

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1):1–68. 128

Chernozhukov V, Chetverikov D, Kato K, et al. (2014) Gaussian approximation of suprema of empirical processes. *Annals of Statistics* 42(4):1564–1597. 131

Chernozhukov V, Demirer M, Lewis G, Syrgkanis V (2019) Semi-parametric efficient policy learning with continuous actions. *Advances in Neural Information Processing Systems* 32(1):1–9. 117

Cheung WC, Ma W, Simchi-Levi D, Wang X (2022) Inventory balancing with online learning. *Management Science* 68(3):1776–1807. 43

Cheung WC, Simchi-Levi D (2019) Sampling-based approximation schemes for capacitated stochastic inventory control models. *Mathematics of Operations Research* 44(2):668–692. 4, 79, 114

Chou MC, Sim CK, Teo CP, Zheng H (2012) Newsvendor pricing problem in a two-sided market. *Production and Operations Management* 21(1):204–208. 73, 79

Ciarallo FW, Akella R, Morton TE (1994) A periodic review, production planning model with uncertain capacity and uncertain demand—optimality of extended myopic policies. *Management Science* 40(3):320–332. 78

Cohen MC, Elmachtoub AN, Lei X (2022) Price discrimination with fairness constraints. *Management Science* 68(12):8536–8552. 117

Cohen MC, Lobel I, Paes Leme R (2020) Feature-based dynamic pricing. *Management Science* 66(11):4921–4943. 113, 116

Cohen MC, Miao S, Wang Y (2021) Dynamic pricing with fairness constraints. Technical report, McGill University, Montreal, Canada. Available at SSRN 3930622. 117

Columbus L (2020) 10 ways AI improves pricing and revenue management. URL https://www.forbes.com/sites/louiscolumbus/2020/09/07/10-ways-ai-improves-pricing-and-revenue-management/?sh=6b05cf972f33, online; posted 7-September-2020. 114

Cope EW (2009) Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control* 54(6):1243–1253. 80

Cortes C, Vapnik V (1995) Support-vector networks. *Machine learning* 20(3):273–297. 2

Cui Y, Pang JS, Sen B (2018) Composite difference-max programs for modern statistical estimation problems. *SIAM Journal on Optimization* 28(4):3344–3374. 130

Cui Y, Zhu R, Kosorok M (2017) Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic journal of statistics* 11(2):3927. 124

Dabrowska DM (1989) Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics* 17(3):1157–1167. 132

Dedecker J, Louhichi S (2002) *Maximal inequalities and empirical central limit theorems* (Springer, New York, NY). 197

Den Boer AV (2015) Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science* 20(1):1–18. 116

Diggle P, Kenward MG (1994) Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics* 43(1):49–73. 123

Duenyas I, Hopp WJ, Bassok Y (1997) Production quotas as bounds on interplant JIT contracts. *Management Science* 43(10):1372–1386. 78

Duenyas I, Tsai C (2000) Control of a manufacturing system with random product yield and downward substitutability. *IIE Transactions* 32(9):785–795. 11, 14, 21

Elmachtoub AN, Grigas P (2022) Smart "predict, then optimize". *Management Science* 68(1):9–26. 4

Elmachtoub AN, Gupta V, Hamilton ML (2021) The value of personalized pricing. *Management Science* 67(10):6055–6070. 111, 116

Elmachtoub AN, Yao D, Zhou Y (2019) The value of flexibility from opaque selling. Working Paper, Columbia University, New York, NY. Available at SSRN 3483872. 14

EuroCommission E (2022) Online platforms. URL https://digital-strategy.ec.europa.eu/en/policies/online-platforms, last accessed on May 03, 2013. 73

Even-Dar E, Mannor S, Mansour Y (2002) PAC bounds for multi-armed bandit and Markov decision processes. Kivinen J, Sloan RH, eds., *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, volume 2375 of *Lecture Notes in Computer Science*, 255–270 (Berlin / Heidelberg, Germany: Springer), ISBN 3-540-43836-X, URL http://www.ece.mcgill.ca/~smanno1//public/banCOLTcamera.pdf. 2

Fan J, Guo Y, Yu M (2022) Policy optimization using semiparametric models for dynamic pricing. *Journal of the American Statistical Association* just-in-press:1–37. 116

Federgruen A, Heching A (1999) Combined pricing and inventory control under uncertainty. *Operations Research* 47(3):454–475. 76

Federgruen A, Liu Z, Lu L (2020) Synthesis and generalization of structural results in inventory management: A generalized convexity property. *Mathematics of Operations Research* 45(2):547–575. 71

Federgruen A, Liu Z, Lu L (2022) Dual sourcing: Creating and utilizing flexible capacities with a second supply source. *Production and Operations Management* 31(7):2789–2805. 71

Federgruen A, Yang N (2011) Procurement strategies with unreliable suppliers. *Operations Research* 59(4):1033–1039. 78

Feng Q (2010) Integrating dynamic pricing and replenishment decisions under supply capacity uncertainty. *Management Science* 56(12):2154–2172. 78

Feng Q, Gallego G, Sethi SP, Yan H, Zhang H (2005) Periodic-review inventory model with three consecutive delivery modes and forecast updates. *Journal of Optimization Theory and Applications* 124(1):137–155. 70

Feng Q, Shanthikumar JG (2018a) How research in production and operations management may evolve in the era of big data. *Production and Operations Management* 27(9):1670–1684. 1, 115

Feng Q, Shanthikumar JG (2018b) Supply and demand functions in inventory models. *Operations Research* 66(1):77–91. 78

Flaxman A, Kalai AT, McMahan HB (2005) Online convex optimization in the bandit setting: gradient descent without a gradient. Buchsbaum A, ed., *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 385–394, SODA '05 (USA: Society for Industrial and Applied Mathematics). 80

Foster DJ, Krishnamurthy A, Simchi-Levi D, Xu Y (2021) Offline reinforcement learning: Fundamental barriers for value function approximation. Technical report, MIT, Cambridge, MA. Available at arXiv:2111.10919. 115

Fu Z, Qi Z, Wang Z, Yang Z, Xu Y, Kosorok MR (2022) Offline reinforcement learning with instrumental variables in confounded markov decision processes. Technical report, Northwestern University, Evanston, IL. Available at arXiv:2209.08666. 202

Fujimoto S, Meger D, Precup D (2019) Off-policy deep reinforcement learning without exploration. *International Conference on Machine Learning*, 2052–2062. 2

Fukuda Y (1964) Optimal policies for the inventory problem with negotiable leadtime. *Management Science* 10(4):690–708. 41, 43

Gallego G, Katircioglu K, Ramachandran B (2006) Semiconductor inventory management with multiple grade parts and downgrading. *Production Planning & Control* 17(7):689–700. 14

Gallego G, Van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* 40(8):999–1020. 79

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning* 63(1):3–42. 139

Godfrey GA, Powell WB (2001) An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science* 47(8):1101–1112. 15

Gong XY, Simchi-Levi D (2023) Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Management Science* . 4, 44

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems* 27(1):1–9. 126

Güllü R (1998) Base stock policies for production/inventory problems with uncertain capacity levels. *European Journal of Operational Research* 105(1):43–51. 78

Hazan E, et al. (2016) Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2(3-4):157–325. 161

Henig M, Gerchak Y (1990) The structure of periodic review policies in the presence of random yield. *Operations Research* 38(4):634–643. 78

Hirano K, Imbens GW (2004) The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* 1(7):73–84. 117, 119

Hu M, Liu Y (2023) Precommitments in two-sided market competition. *Manufacturing & Service Operations Management* 25(2):704–718. 78

Hu M, Zhou Y (2020) Price, wage, and fixed commission in on-demand matching. Technical report, University of Toronto, Toronto, Ontario, Canada. Available at SSRN 2949513. 78, 82

Hu M, Zhou Y (2021) Dynamic type matching. To appear in *Manufacturing & Service Operations Management.* 11, 14

Hu X, Duenyas I, Kapuscinski R (2008) Optimal joint inventory and transshipment control under uncertain capacity. *Operations Research* 56(4):881–897. 11, 15, 21

Hua Z, Yu Y, Zhang W, Xu X (2015) Structural properties of the optimal policy for dual-sourcing systems with general lead times. *IIE Transactions* 47(8):841–850. 41, 43

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009) An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research* 34(2):397–416. 4, 16, 37, 43, 49, 116, 169

Huh WT, Levi R, Rusmevichientong P, Orlin JB (2011) Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research* 59(4):929–941. 116

Huh WT, Nagarajan M (2010) Linear inflation rules for the random yield problem: Analysis and computations. *Operations Research* 58(1):244–251. 75, 78

Huh WT, Rusmevichientong P (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research* 34(1):103–123. 4, 7, 112, 116

Iglehart DL (1964) The dynamic inventory problem with unknown demand distribution. *Management Science* 10(3):429–440. 15

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, Cambridge, UK). 120

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *The Annals of Applied Statistics* 2(3):841–860. 126, 129, 139, 206

Jain A, Moinzadeh K, Dumrongsiri A (2015) Priority allocation in a rental model with decreasing demand. *Manufacturing & Service Operations Management* 17(2):236–248. 13

Janakiraman G, Seshadri S (2017) Dual sourcing inventory systems: On optimal policies and the value of costless returns. *Production and Operations Management* 26(2):203–210. 43

Janakiraman G, Seshadri S, Sheopuri A (2015) Analysis of tailored base-surge policies in dual sourcing inventory systems. *Management Science* 61(7):1547–1561. 43

Javanmard A, Nazerzadeh H (2019) Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research* 20(1):315–363. 116

Kallus N, Zhou A (2018) Policy evaluation and optimization with continuous treatments. *International Conference on Artificial Intelligence and Statistics* 84(1):1243–1251. 117, 121

Kazaz B, Webster S (2015) Price-setting newsvendor problems with uncertain supply and risk aversion. *Operations Research* 63(4):807–811. 75

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 3146–3154. 2

Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* 62(5):1142–1167. 78, 82, 99, 103

Khardani S, Semmar S (2014) Nonparametric conditional density estimation for censored data based on a recursive kernel. *Electronic Journal of Statistics* 8(2):2541–2556. 132

Kingma DP, Ba J (2017) Adam: A method for stochastic optimization. *International Conference on Learning Representations* 13(5):1–9. 140

Kleinbaum DG, Klein M (2010) *Survival analysis* (Springer, New York, NY). 124, 126

Kleinberg R (2004) Nearly tight bounds for the continuum-armed bandit problem. Lawrence K Saul LB Y Weiss, ed., *Proceedings of the 17th International Conference on Neural Information Processing Systems*, 697–704, NIPS'04 (Cambridge, MA, USA: MIT Press). 80

Kleinberg R, Slivkins A, Upfal E (2008) Multi-armed bandits in metric spaces. Dwork C, ed., *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, 681–690, STOC '08 (New York, NY, USA: Association for Computing Machinery). 2, 80

Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12(2):479–502. 15, 43, 79

Konda V, Tsitsiklis J (1999) Actor-critic algorithms. *Advances in neural information processing systems* 12. 1

Kosorok MR, Laber EB (2019) Precision medicine. *Annual Review of Statistics and Its Application* 6(1):263–286. 117

Kouvelis P, Xiao G, Yang N (2018) On the properties of yield distributions in random yield problems: Conditions, class of distributions and relevant applications. *Production and Operations Management* 27(7):1291–1302. 75

Kumar A, Fu J, Soh M, Tucker G, Levine S (2019) Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 11784–11794. 2

Kumar A, Zhou A, Tucker G, Levine S (2020) Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, volume 33, 1179–1191. 2

Le Guen T (2008) *Data-driven pricing.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA. 82

Lei YM, Jasin S, Sinha A (2014) Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. Technical report, Queen's University at Kingston, Kingston, Ontario, Canada. Available at SSRN 2509425. 8, 77, 79, 80, 82, 88

Levi R, Perakis G, Uichanco J (2015) The data-driven newsvendor problem: new bounds and insights. *Operations Research* 63(6):1294–1306. 4, 15, 43, 79, 114

Levi R, Roundy RO, Shmoys DB (2007a) Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research* 32(4):821–839. 4, 43, 61, 79, 114

Levi R, Roundy RO, Shmoys DB (2007b) Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research* 32(4):821–839. 15

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*, 661–670. 2

Li Q, Yu P (2014) Multimodularity and its applications in three stochastic dynamic inventory problems. *Manufacturing & Service Operations Management* 16(3):455–463. 41, 43

Li Q, Zheng S (2006) Joint inventory replenishment and pricing control for systems with uncertain yield and demand. *Operations Research* 54(4):696–705. 75

Lin M, Huh WT, Krishnan H, Uichanco J (2022) Data-driven newsvendor problem: Performance of the sample average approximation. *Operations Research* 70(4):1996–2012. 4, 79

Lu X, Song JS, Zhu K (2005) On "the censored newsvendor and the optimal acquisition of information". *Operations Research* 53(6):1024–1026. 15

Lu X, Song JS, Zhu K (2008) Analysis of perishable-inventory systems with censored demand data. *Operations Research* 56(4):1034–1038. 15

Lugosi G, Markakis MG, Neu G (2017) On the hardness of inventory management with censored demand data. Technical report, Pompeu Fabra University, Barcelona, Spain. Available at arXiv:1710.05739. 116

Luo Y, Sun WW, et al. (2021) Distribution-free contextual dynamic pricing. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC. Avaiable at arXiv:2109.07340. To appear in Mathematics of Operations Research. 117

Mahajan S, van Ryzin G (2001) Stocking retail assortments under dynamic consumer substitution. *Operations Research* 49(3):334–351. 13

Mao W, Zhang K, Zhu R, Simchi-Levi D, Bacsar T (2020) Model-free non-stationary RL: Near-optimal regret and applications in multi-agent RL and inventory control. Working Paper, MIT, Cambridge, MA. 16

Martens J (2010) Deep learning via hessian-free optimization. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 735–742 (Omnipress). 2

Meyn S, Tweedie R (1993) *Markov Chains and Stochastic Stability*, volume 92 (Cambridge University Press, Cambridge, UK). 166

Miao R, Qi Z, Shi C, Lin L (2023) Personalized pricing with invalid instrumental variables: Identification, estimation, and policy learning. Technical report, George Washington University, Washington, DC. 117

Miao S, Chen X, Chao X, Liu J, Zhang Y (2022) Context-based dynamic pricing with online clustering. *Production and Operations Management* 31(9):3559–3575. 113, 116

Miao S, Wang Y (2021) Network revenue management with nonparametric demand learning:\sqrt{T}-regret and polynomial dimension dependency. Technical report, McGill University, Montreal, Quebec, Canada. Available at SSRN 3948140. 79

Micchelli CA, Xu Y, Zhang H (2006) Universal kernels. *Journal of Machine Learning Research* 7(12):2651–2667. 132

Mills ES (1959) Uncertainty and price theory. *The Quarterly Journal of Economics* 73(1):116–130. 78, 99

Mockus J (1994) Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization* 4:347–365. 2

Murray GR, Silver EA (1966) A bayesian analysis of the style goods inventory problem. *Management Science* 12(11):785–797. 15

Nagarajan M, Rajagopalan S (2008) Inventory models for substitutable products: Optimal policies and heuristics. *Management Science* 54(8):1453–1466. 13

Nambiar M, Simchi-Levi D, Wang H (2019) Dynamic learning and pricing with model misspecification. *Management Science* 65(11):4980–5000. 117

Neal RM (2012) *Bayesian learning for neural networks*, volume 118 (Springer Science & Business Media). 2

Nelson W (1972) Theory and applications of hazard plotting for censored failure data. *Technometrics* 14(4):945–966. 139

Ortner R (2020) Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research* 67:115–128. 7, 42, 61

Parker GG, Van Alstyne MW (2005) Two-sided network effects: A theory of information product design. *Management Science* 51(10):1494–1504. 79

Parker R, Olsen T (2010) Dynamic inventory competition with stockout-based substitution. *Proceedings of the Behavioral and Quantitative Game Theory: Conference on Future Directions*, 1–1. 13

Parzen E (1962) On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* 33(3):1065–1076. 122

Pasternack BA, Drezner Z (1991) Optimal inventory policies for substitutable commodities with stochastic demand. *Naval Research Logistics (NRL)* 38(2):221–240. 13

Paulin D (2015) Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability* 20(none):1 – 32. 7, 42, 175

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research* 12(1):2825–2830. 140

Petruzzi NC, Dada M (1999) Pricing and the newsvendor problem: A review with extensions. *Operations Research* 47(2):183–194. 76, 78, 99

Powell W, Ruszczyński A, Topaloglu H (2004) Learning algorithms for separable approximations of discrete stochastic optimization problems. *Mathematics of Operations Research* 29(4):814–836. 15

Qiang S, Bayati M (2016) Dynamic pricing with demand covariates. Technical report, Stanford University, Stanford, CA. Available at SSRN 2765257. 116

Qin H, Simchi-Levi D, Wang L (2022) Data-driven approximation schemes for joint pricing and inventory control models. *Management Science* 68(9):6591–6609. 4, 79, 115

Rao U, Swaminathan J, Zhang J (2004) Multi-product inventory planning with downward substitution, stochastic demand and setup costs. *IIE Transactions* 36:59–71. 14

Rasmussen CE, Williams CKI (2003) Gaussian processes in machine learning. *Summer School on Machine Learning* 63–71. 2

Robbins H (1952) Some aspects of the sequential design of experiments . 2

Robbins H, Monro S (1951) A stochastic approximation method. *The annals of mathematical statistics* 400–407. 2

Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7(9-12):1393–1512. 119, 121

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55. 8, 114

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688. 8, 113, 118, 119

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *nature* 323(6088):533–536. 2

Rummery GA, Niranjan M (1994) On-line q-learning using connectionist systems. volume 37 (University of Cambridge, Department of Engineering Cambridge, UK). 1

Scarf H (1959) Bayes solutions of the statistical inventory problem. *The Annals of Mathematical Statistics* 30(2):490–508. 15

Schlapp J, Fleischmann M (2018) Multiproduct inventory management under customer substitution and capacity restrictions. *Operations Research* 66. 14

Schmidt-Hieber J (2020) Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics* 48(4):1875–1897. 132

Schulz J, Moodie EE (2021) Doubly robust estimation of optimal dosing strategies. *Journal of the American Statistical Association* 116(533):256–268. 117

Sen B (2018) A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University* 11:28–29. 2

Shamir O (2013) On the complexity of bandit and derivative-free stochastic convex optimization. Shalev-Shwartz S, Steinwart I, eds., *COLT 2013 - The 26th Annual Conference on Learning Theory*, 3–24, JMLR Workshop and Conference Proceedings (New York, NY, USA: JMLR.org). 80

Shanthikumar JG, Yao DD, Zijm WHM (2003) *Stochastic modeling and optimization of manufacturing systems and supply chains*, volume 63 (Springer Science & Business Media, New York, NY). 21

Sheopuri A, Janakiraman G, Seshadri S (2010) New policies for the stochastic inventory control problem with two supply sources. *Operations Research* 58(3):734–745. 41, 43, 70, 174

Shi C, Chen W, Duenyas I (2016) Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research* 64(2):362–370. 4, 16, 43

Shumsky RA, Zhang F (2009) Dynamic capacity management with substitution. *Operations Research* 57(3):671–684. 15, 21

Simchi-Levi D, Xu Y (2022) Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research* 47(3):1904–1931. 116

Snyder LV, Shen ZJM (2019) *Fundamentals of supply chain theory* (John Wiley & Sons, Hoboken, NJ). 82

Steinwart I, Christmann A (2008) *Support vector machines* (Springer, New York, NY). 133

Sun J, Van Mieghem JA (2019) Robust dual sourcing inventory management: Optimality of capped dual index policies and smoothing. *Manufacturing & Service Operations Management* 21(4):912–931. 41, 43

Svoboda J, Minner S, Yao M (2021) Typology and literature review on multiple supplier inventory control models. *European Journal of Operational Research* 293(1):1–23. 41, 70

Talluri KT, Van Ryzin G, Van Ryzin G (2004) *The theory and practice of revenue management* (Springer, New York, NY). 82

Taylor JB (1974) Asymptotic properties of multiperiod control rules in the linear regression model. *International Economic Review* 472–484. 78, 99

Taylor TA (2018) On-demand service platforms. *Manufacturing & Service Operations Management* 20(4):704–720. 78

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3-4):285–294. 2

Van Der Vaart A, Wellner JA (2011) A local maximal inequality under uniform entropy. *Electronic Journal of Statistics* 5(1):192–203. 198

Vapnik V (2013) *The nature of statistical learning theory* (Springer science & business media). 2

Veeraraghavan S, Scheller-Wolf A (2008) Now or later: A simple policy for effective dual sourcing in capacitated systems. *Operations Research* 56(4):850–864. 41, 43, 46, 50, 70, 167, 172, 174

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press). 2, 158

Wang CH, Wang Z, Sun WW, Cheng G (2020) Online regularization for high-dimensional dynamic pricing algorithms. Technical report, Purdue University, West Lafayette, IN. Available at arXiv:2007.02470. 116

Wang H, Talluri K, Li X (2021a) On dynamic pricing with covariates. Technical report, Imperial College London, London, UK. Avaiable at arXiv:2112.13254. 117

Wang L, Tchetgen ET (2018) Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 80(3):531–550. 120

Wang Y (2014) Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research* 62(6):1244–1258. 79, 82

Wang Y (2023) Estimation of high-dimensional contextual pricing models with nonparametric price confounders. Technical report, University of Texas at Dallas, Richardson, TX. 117

Wang Y, Chen B, Simchi-Levi D (2021b) Multimodal dynamic pricing. *Management Science* 67(10):6136–6152. 4, 80, 98, 99

Wang Y, Gerchak Y (1996) Periodic review production models with variable capacity, random yield, and uncertain demand. *Management Science* 42(1):130–137. 78

Wang Z, Mersereau AJ (2017) Bayesian inventory management with potential change-points in demand. *Production and Operations Management* 26(2):341–359. 15

Watkins CJCH (1989) *Learning from delayed rewards.* Ph.D. thesis, University of Cambridge England. 1

Wei CY, Luo H (2021) Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. Belkin M, Kpotufe S, eds., *Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, 4300–4354, PMLR (PMLR). 178

Wellner J, et al. (2013) *Weak convergence and empirical processes: with applications to statistics* (Springer Science & Business Media). 2

Whitin TM (1955) Inventory control and price theory. *Management Science* 2(1):61–68. 73, 78

Whittemore AS, Saunders SC (1977) Optimal inventory under stochastic demand with two supply options. *SIAM Journal on Applied Mathematics* 32(2):293–305. 41, 43

Wu Y, Tucker G, Nachum O (2019) Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361* . 2

Xin L, Goldberg DA (2018) Asymptotic optimality of tailored base-surge policies in dual-sourcing inventory systems. *Management Science* 64(1):437–452. 43

Xin L, Van Mieghem JA (2021) Dual-sourcing, dual-mode dynamic stochastic inventory models: A review. Technical report, Chicago University, Chicago, IL. Available at SSRN 3885147. 40

Xu H, Yao DD, Zheng S (2011) Optimal control of replenishment and substitution in an inventory system with nonstationary batch demand. *Production and Operations Management* 20(5):727–736. 14, 21

Xu J, Wang YX (2021) Logarithmic regret in feature-based dynamic pricing. *Advances in Neural Information Processing Systems* 34(1):13898–13910. 116

Xu Y, Lu B, Ghose A, Dai H, Zhou W (2023) The interplay of earnings, ratings, and penalties on sharing platforms: An empirical investigation. Technical report, University of North Carolina at Chapel Hill, NC. Available at SSRN 3609132. Forthcoming in *Management Science*. 74

Yu Y, Chen X, Zhang F (2015) Dynamic capacity management with general upgrading. *Operations Research* 63(6):1372–1389. 10, 11, 15, 21

Yu Y, Shou B, Ni Y, Chen L (2017) Optimal production, pricing, and substitution policies in continuous review production-inventory systems. *European Journal of Operational Research* 260(2):631–649. 15, 21

Yuan H, Luo Q, Shi C (2021) Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science* 67(10):6089–6115. 4, 16, 44, 73, 116

Zhang H, Chao X, Shi C (2018) Perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand. *Operations Research* 66(5):1276–1286. 4, 16, 43, 116

Zhang H, Chao X, Shi C (2020) Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Science* 66(5):1962–1980. 4, 16, 28, 43, 53, 116