University of California

Santa Barbara

Dynamic Pricing as an Online Decision-Making Problem

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy

in

Computer Science

by

Jianyu Xu

Committee in charge:

Professor Yu-Xiang Wang, Chair
Professor Ambuj Singh
Professor Daniel Lokshtanov
Professor Erik Eyster

September 2024

The Dissertation of Jianyu Xu is approved.

Professor Ambuj Singh

Professor Daniel Lokshtanov

Professor Erik Eyster

Professor Yu-Xiang Wang, Committee Chair

August 2024

Dynamic Pricing as an Online Decision-Making Problem

by

Jianyu Xu

To my family and friends.

Acknowledgements

First and foremost, I am immensely grateful to my advisor, Prof. Yu-Xiang Wang, for his unwavering support and encouragement throughout my PhD journey. Yu-Xiang has been an invaluable mentor, imparting extensive knowledge and skills in machine learning, optimization, statistics, and computing and information sciences. His generous support has not only sustained my academic pursuits but also significantly aided my job search. Yu-Xiang exemplifies the ideals of professionalism, modesty, and morality, setting a high standard for what it means to be a truly great advisor.

I would also like to extend my heartfelt thanks to Professors Zheng Zhang, Yuan Xie, and Guoqi Li for their invaluable advice during my studies. Their guidance was crucial to my achievements and has deeply influenced my academic development. Additionally, I am grateful to Professors Xi Chen, Yining Wang, Lei Deng, and Chong Liu for their support in my research projects and for their gracious endorsements in the job market.

For those significant contributions to our projects, I am deeply appreciative of my collaborators Dheeraj, Dan, Ming, Wenhu, Ling, Xing, and Jiayue at UC Santa Barbara, as well as external scholars Xuan Wang, Jiashuo Jiang, Pau Pereira, and Chengyi Lyu, for their expertise and dedication. Their involvement and insights have been crucial in advancing our research and achieving our collective goals.

I extend my sincere thanks to my labmates in the S^2ML group: Yuqing, Xuandong, Rachel, Kaiqi, Esha, Erchi, Peng, Mengye, Sunil, Momin, Yingjia, and Shuai. Your help and enthusiasm have been vital to my growth and progress. I am also grateful to Professors Ambuj Singh, Daniel Lokshitanov, and Erik Eyster for their service on my PhD committee. I deeply appreciate their commitment and constructive feedback at each milestone of my PhD journey, which have been instrumental in shaping my research.

During my study at UCSB, I had the privilege of taking a variety of excellent courses that were delivered with both passion and deep insight. I particularly enjoyed Daniel's random algorithms, Omer's combinatorics, Kenneth's information theory, Trinabh's cryptography, and Jeffrey's Linguistics 7 for teaching assistantship. Additionally, I benefited greatly from online courses such as Larry Wasserman's statistical machine learning, Hung-yi Lee's deep learning, and Tim Roughgarden's algorithmic game theory.

I would like to thank Tong Zhang, Wei He, Qingpei Hu, Maxime Cohen, Yiyun Luo, Will Wei Sun, Elynn Chen, Yaqi Duan, Renato Paes Leme, Yifeng Teng, Akshay Krishnamuthy, Wenpeng Zhang and Lijun Zhang for their graciously-offered guidance across various dimensions. Special thanks are due to Aarti Singh and Bryan Wilder for extending a postdoctoral offer to me at CMU MLD, which marks a significant milestone in my career. Additionally, my thanks go to those providing me with valuable career tips. This includes Jing Lei, Zhiyi Huang, Lirong Xia, Zizhuo Wang, Zhuoyu Long, Jinzhi Bu, Ali Aouad, Chara Podimata, Zili Meng, Shiji Zhou, and another Jianyu Xu (professor at XJ-Liverpool).

I extend my heartfelt thanks to my friends at UCSB, Tsinghua, AntGroup, Amazon, HKUST and beyond. Especially, I would like to thank my dear friend-couple, Zhaodong Chen and Carina Quan, for their unwavering support, constant encouragement, and steadfast companionship during my difficult times over the past years. Besides, I sincerely appreciate my friend Bingxu Chen for his sharp-eyed intuition and insightful suggestions on my decisions ever since we met as undergraduate classmates.

Finally, I would like to express my deepest gratitude to my parents, whose unconditional love and support have been my foundation and strength. I am eternally grateful for everything they have done.

Curriculum Vitæ

Jianyu Xu

Education

2024	Ph.D. in Computer Science (Expected), University of California, Santa Barbara.
2019	B.S. in Measurement, Control & Instrumentation, Tsinghua University.

Publication

	Xu, Jianyu, Yining Wang, Xi Chen, and Yu-Xiang Wang. "Online Dynamic Pricing with Inventory-Censored Demands." *In Submission.*
ICML 2024	**Xu, Jianyu**, and Yu-Xiang Wang. "Pricing with Contextual Elasticity and Heteroscedastic Valuation." *Forty-first International Conference on Machine Learning.* 2024.
AIStats 2023	**Xu, Jianyu**, Dan Qiao, and Yu-Xiang Wang. "Doubly-Fair Dynamic Pricing." *International Conference on Artificial Intelligence and Statistics*, 9941-9975, 2023
TMLR	Baby, Dheeraj*, **Jianyu Xu***, and Yu-Xiang Wang. "Non-Stationary Contextual Pricing with Safety Constraints." *Transactions on Machine Learning Research*, 2023.
AIStats 2022	**Xu, Jianyu**, and Yu-Xiang Wang. "Towards agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise." *International Conference on Artificial Intelligence and Statistics*, 9643-9662, 2022.
NeurIPS 2021	**Xu, Jianyu**, and Yu-Xiang Wang. "Logarithmic regret in feature-based dynamic pricing." *Advances in Neural Information Processing Systems*, 34, 13898-13910, 2021.
EMNLP 2023	Chen, Wenhu, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, **Jianyu Xu**, Xinyi Wang, and Tony Xia. "Theoremqa: A theorem-driven question answering dataset." *Empirical Methods in Natural Language Processing*, 7889-7901. 2023.
IEEE-JSTSP	Liang, Ling, **Jianyu Xu**, et al. "Fast search of the optimal contraction sequence in tensor networks." *IEEE Journal of Selected Topics in Signal Processing*, 15.3 (2021): 574-586.
Phy. Rev. E	**Xu, Jianyu**, Ling Liang, Lei Deng, Changyun Wen, Yuan Xie, and Guoqi Li. "Towards a polynomial algorithm for optimal contraction sequence of tensor networks from trees." *Physical Review E* 100.4 (2019): 043309.

Abstract

Dynamic Pricing as an Online Decision-Making Problem

by

Jianyu Xu

The intersection of pricing and machine learning has gained considerable attention in recent years, positioning pricing strategy as a decision-making problem for the sellers who are tasked with setting prices in real-time and learning optimal prices through observed demands. This thesis explores dynamic pricing within the framework of online decision-making, where sequential decisions are informed by continuously evolving observations.

We contribute novel technical approaches to dynamic pricing through the study of two main aspects:

I **Feature-Based Dynamic Pricing.** In Part I, we address the challenge of pricing highly differentiated products, each characterized by specific features.Assuming linear and noisy customer valuations with binary decision outcomes, we explore settings with known, unknown, and heteroscedastic noise distributions. We propose algorithms for each scenario, providing rigorous analysis of their regret guarantees. Our findings illustrate that the difficulty of solving feature-based dynamic pricing is contingent on the seller's knowledge of noise distributions.

II **Dynamic Pricing under Constraints.** In Part II, we examine constraints that affect pricing strategies in modern markets, focusing on fairness and inventory limits.

We firstly introduce two fairness notions and develop a randomized pricing mechanism that accommodates multiple fairness constraints simultaneously, achieving optimal regret and fairness outcomes. In the other project, we tackle pricing under inventory constraints, addressing challenges posed by censored demands to achieve optimal regrets.

Our algorithmic solutions are rigorously evaluated through the metric of information-theoretic regret bounds. The practical relevance of our methodologies is further validated by comprehensive empirical studies using simulated data. The combination of theoretical and practical justifications demonstrates the robustness and applicability of our approaches across various dynamic pricing scenarios.

Contents

Chapter 1

Introduction

Dynamic pricing is a strategy where prices are flexibly adjusted in response to market demand and other variables. In the ever-evolving landscape of E-commerce and online markets, it plays a pivotal role in enhancing the profitability and market responsiveness of firms. Having been studied since Cournot [1897], dynamic pricing is formulated as a complex decision-making problem that intertwines elements of machine learning, operations research, and computer science.

In this thesis, we investigate dynamic pricing as an online decision-making problem, exploring both its theoretical underpinnings and practical applications to develop strategies that optimize pricing decisions in real-time, maximizing profit while accommodating the inherent characteristics and constraints of online markets.

1.1 Dynamic Pricing Problems

Dynamic pricing involves a scenario where a seller proposes prices and observes the realized demands from customers over time. The primary aim is to estimate the *demand curve* effectively – understanding how customer demand varies with changes in price. Correspondingly, an accurate estimation of the demand curve allows the seller to adjust prices wisely to maximize profitability. This setting induces continual interactions between price proposals and demand observations, necessitating dynamic decision-making strategies on the seller's side based on immediate customer responses. Therefore, the challenge lies in determining the most effective pricing strategy that not only approaches the optimal price asymptotically but also minimizes the costs associated with the adaptation process.

1.2 Online Decision-Making Problems

The framework of online decision-making is particularly apt for addressing problems where decisions are made sequentially and where the feedback of these decisions inform future actions. In this context, a player takes an action at each time and receives a corresponding action-dependent reward from the environment. This iterative process involves estimating the potential rewards of various actions and selecting the optimal action (or the optimal action-taking policy) based on these estimates. The complexity of online decision-making is further accentuated by the diversity of problem models it encompasses, including but not limited to multi-armed bandits [MAB, Lai and Robbins, 1985], contextual bandits [CB, Langford and Zhang, 2007], online convex optimization [OCO, Hazan, 2016], Markov decision processes [MDP, Puterman, 1990], and Reinforcement learning [RL, Kaelbling et al., 1996], each presenting unique challenges and solutions.

1.3 Bridging Dynamic Pricing with Online Decision-Making

This thesis posits that dynamic pricing can be effectively modeled and analyzed within the framework of online decision-making. The synthesis of these domains is beneficial for several reasons:

1. **Natural alignment in problem setting.** The inherent structure of dynamic pricing problems can be formulated as online decision-making, which involves sequential actions and real-time feedback.

2. **Applicability of performance metric.** The concept of *regret*, a performance metric prevalent in online decision-making, is readily applicable to dynamic pricing, as it provides a measure for evaluating the efficacy of pricing strategies by comparing the cumulative rewards against a benchmark of optimal actions/policies.

3. **Algorithmic and analytic solutions.** Online decision-making has spawned a wealth of algorithms that can potentially be adapted for dynamic pricing. Existing literature also offers a rich toolkit for analyzing the quantitative performance of decision-making algorithms from theoretical and practical aspects.

However, despite these synergies, existing methods from the domain of online decision-making are not directly transferrable to dynamic pricing due to unique characteristics such as *continuous decision spaces*, *partial observations*, and *constraints*. These distinctions necessitate the development of specialized algorithms that can handle the specificities of dynamic pricing while striving to minimize regret.

This dissertation aims to bridge this gap by developing novel algorithms that incorporate

advanced methodologies in online decision-making studies with the nuanced structures of dynamic pricing. We endeavor to contribute to the field by constructing a theoretical foundation for dynamic pricing, developing algorithms that address its unique challenges, and demonstrating their information-theoretic performance through quantitative analysis. It not only enhances the understandings of dynamic pricing within statistical and computational frameworks, but also provides concrete algorithms that can be implemented in real-world scenarios.

1.4 Organization of Dissertation

This thesis will be organized into two parts, each containing a few chapters and presenting a sequence of concrete research works. Each chapter represents a previously published/submitted paper and has self-concordant introduction, conclusion, and analysis.

1.4.1 Feature-based Dynamic Pricing

Consider the following pricing problem: At each time $t = 1, 2, \ldots, T$, a *feature* vector x_t describing the current sales session is revealed by the nature. After observing x_t, the seller proposes a *price* p_t, while the buyer generates a *valuation* y_t but keeps it in secret. The *demand* is a binary decision of purchase, i.e. $\mathbb{1}_t := \mathbb{1}[p_t \leq y_t]$, based on the comparison between price and valuation.

In Part I, we study this problem under a variety of modeling assumptions. We firstly consider a *linear valuation model* where y_t is a linear noisy mapping from x_t. Chapter 2 considers the case when the noise distribution is *known* to the seller, and Chapter 3

focuses on the problem when the noise distribution is *agnostic*. Moreover, Chapter 4 generalizes the linear valuation model by allowing a feature-dependent noise variance. In chapters listed above, we propose algorithms that achieve sub-linear regrets. We also prove information-theoretic regret lower bounds as indicators of optimality. At the end of Part I, we will conclude our results and discuss on the contemporary and future development of feature-based dynamic pricing studies.

1.4.2 Dynamic Pricing under Constraints

In real-world scenarios, the proposed prices and realized demands are always restricted under certain constraints. One example lies in customers' concerns on pricing fairness, i.e. how different are the prices for me versus for other customers. This requires the prices across different customer groups are *similar* under certain metrics. Another example originates from a limited inventory quantity: What if the potential demand exceeds the inventory we currently have in storage?

In Part II, we study dynamic pricing problem under these two types of constraints. On the one hand, Chapter 5 introduces the concepts of *procedural* and *substantive fairness* for pricing, and proposes an algorithm that achieves optimal regret and unfairness rates simultaneously. On the other hand, Chapter 6 introduces a pricing problem with inventory-censored demand and proposes an optimal algorithm.

Part I

Feature-Based Dynamic Pricing

Overview

In this part, we study the problem of *feature-based dynamic pricing*, which describes a scenario where the seller is selling a variety of items on the fly. At each time period, the nature may introduce a new product, unfamiliar to the seller, yet similar in *features* to previously sold items. Leveraging historical sales data, the seller progressively refines their pricing strategy based on these features.

We generally define our problem setting as follows: For each time period $t = 1, 2, \ldots, T$, a feature vector $x_t \in \mathbb{R}^d$, which characterizes an item and its associated sales context, is disclosed to both the customer and the seller. Upon reviewing x_t, the customer internally assesses a *secret valuation* y_t, while the seller simultaneously proposes a *price* p_t. The customer then decides whether to purchase the item based on a comparison of p_t and y_t. The seller, not having access to y_t, only observes the customer's binary purchase decision, represented as $\mathbb{1}_t := \mathbb{1}[p_t \leq y_t]$.

This model was initially proposed and studied in Cohen et al. [2020], which posited a linear valuation framework with negligible or no noise interference. Later Javanmard and Nazerzadeh [2019] investigated a more general *linear noisy valuation* model and addressed the sparsity of feature vectors. A stream of subsequent works (including ours) have further developed this topic from various perspectives, which we will discuss in more

detail throughout Part I.

In the following chapters, we will introduce three of our works that contributes to this problem employing a (generalized) linear noise valuation model. Specifically, we investigate the following three aspects.

1. In Chapter 2, we focus on the linear noisy valuation model $y_t = x_t^\top \theta^* + N_t$ where θ^* is an unknown parameter and N_t is an i.i.d. noise with a *known distribution*. We propose two algorithms, EMLP and ONSP, that achieve the $O(d \log T)$ optimal regret for stochastic and adversarial $\{x_t\}$ sequences respectively.

2. In Chapter 3, we continue with the same valuation model but assume N_t is drawn from an *unknown* noise distribution. We propose a D2-EXP4 algorithm, which achieves a regret of $O(T^{3/4})$ without any specific assumptions about the noise distribution beyond boundedness. On the other hand, we show the problem hardness by proving a $\tilde{\Omega}(T^{2/3})$ regret lower bound.

3. In Chapter 4, we expand the valuation model to account for heteroscedasticity, acknowledging that the variances of customers' valuations towards different items can vary significantly and are not necessarily proportional to the valuations themselves. Therefore, we model this effect by incorporating a contextual multiplier on the original linear model, such that $y_t = \frac{1}{x_t^\top \eta^*} \cdot (x_t^\top \theta^* + N_t)$. Under this model, we propose a PwP algorithm that achieves $\tilde{O}(\sqrt{dT})$ optimal regret.

A overview of regret rates under all comparable assumptions (i.e. the linear noisy valuation model and those closely related) is summarized and illustrated in Figure 1.1. Each chapter will provide comprehensive details, including motivations, model formulations, assumptions, algorithmic strategies, analyses, experiments, and discussions.

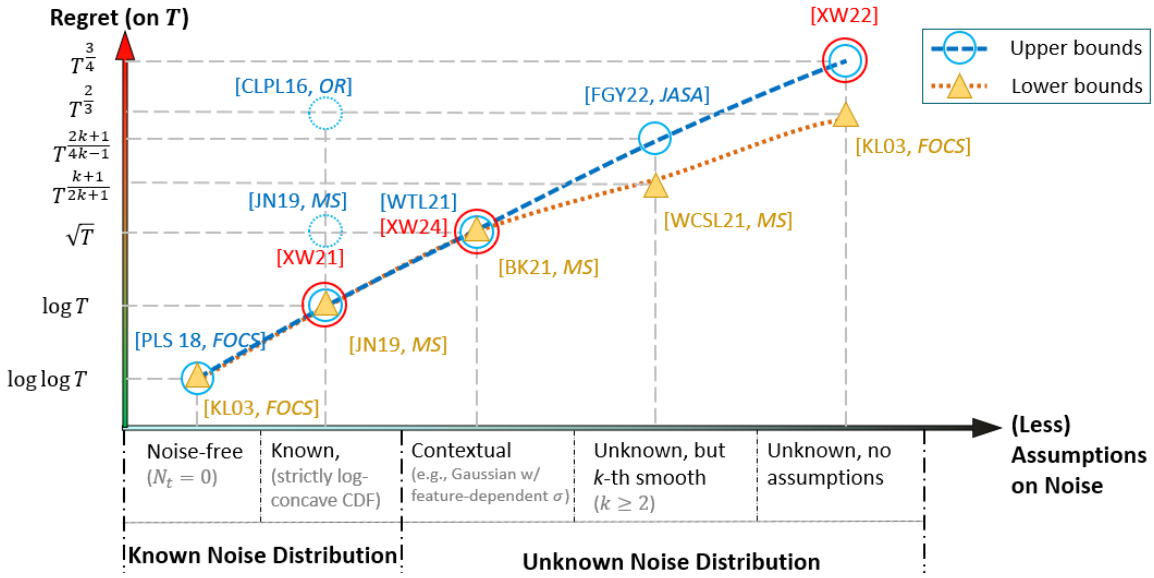


Figure 1.1: The regret rates in existing works (until July 2024) on the feature-based dynamic pricing problem. The x-axis represents the assumptions on the valuation noise distribution, and the y-axis represents the regret rates. Blue circles and yellow triangles represent the algorithmic upper bound and information-theoretic lower bound, respectively. A matching circle-and-triangle means a closed gap, indicating both of them are optimal.

Chapter 2

Logarithmic Regret: When Noise Distribution is Known

As outlined in the overview, we study a feature-based dynamic pricing where each customer's valuation is a linear and noisy mapping from the revealed features, and the customer's demand is binary that depends on the comparison between price and evaluation. In this chapter, we focus on scenarios where the seller possesses precise knowledge of the noise distribution affecting customers' valuations. We provide two algorithms (EMLP and ONSP) for stochastic and adversarial feature settings, respectively, and prove the optimal $O(d \log T)$ regret bounds for both. In comparison, the best existing results are $O\left(\min\left\{\frac{1}{\lambda_{\min}^2} \log T, \sqrt{T}\right\}\right)$ and $O(T^{2/3})$ respectively, with $\lambda_{\min}$ being the smallest eigenvalue of $\mathbb{E}[xx^T]$ that could be arbitrarily close to 0. We also prove an $\Omega(\sqrt{T})$ information-theoretic lower bound for a slightly more general setting, which demonstrates that "knowing-the-demand-curve" leads to an exponential improvement in feature-based dynamic pricing.

2.1 Introduction

The problem of pricing — to find a high-and-acceptable price — has been studied since Cournot [1897]. In order to locate the optimal price that maximizes the revenue, a firm may adjust their prices of products frequently, which inspires the *dynamic pricing* problem. Existing works [Kleinberg and Leighton, 2003, Broder and Rusmevichientong, 2012, Chen and Farias, 2013, Besbes and Zeevi, 2015] primarily focus on pricing a single product, which usually will not work well in another setting when thousands of new products are being listed every day with no prior experience in selling them. Therefore, we seek methods that approach an acceptable-and-profitable price with only observations on this single product and some historical selling records of other products.

In this chapter, we consider a "feature-based dynamic pricing" problem, which was studied by Amin et al. [2014], Cohen et al. [2020], Javanmard and Nazerzadeh [2019]. In this problem setting, a sales session (product, customer and other environmental variables) is described by a feature vector, and the customer's *expected* valuation is modeled as a linear function of this feature vector.

Feature-based dynamic pricing. For $t = 1, 2, ..., T$:

1. A feature vector $x_t \in \mathbb{R}^d$ is revealed that describes a sales session (product, customer and context).
2. The customer valuates the product as $w_t = x_t^\top \theta^* + N_t$.
3. The seller proposes a price $p_t > 0$ concurrently (according to x_t and historical sales records).
4. The transaction is successful if $p_t \leq w_t$, i.e., the seller gets a reward (payment) of $r_t = p_t \cdot \mathbb{1}(p_t \leq w_t)$.

Here T is unknown to the seller (and thus can go to infinity), x_t's can be either stochastic

(e.g., each sales session is drawn i.i.d.) or adversarial (e.g., the sessions arrive in a strategic sequence), $\theta^* \in \mathbb{R}^d$ is a fixed parameter for all time periods, N_t is a zero-mean noise, and $\mathbb{1}_t = \mathbb{1}(p_t \leq w_t)$ is an indicator that equals 1 if $p_t \leq w_t$ and 0 otherwise. In this online-fashioned setting, we only see and sell one product at each time. Also, the feedback is *Boolean Censored*, which means we can only observe $\mathbb{1}_t$ instead of knowing w_t directly. The best pricing policy for this problem is the one that maximizes the *expected* reward, and the *regret* of a pricing policy is accordingly defined as the difference of expected rewards between this selected policy and the best policy.

Summary of Results. Our contributions are threefold.

1. When x_t's are independently and identically distributed (i.i.d.) from an unknown distribution, we propose an "Epoch-based Max-Likelihood Pricing (EMLP)" algorithm that guarantees a regret bound at $O(d \log T)$. The design of EMLP is similar to that of the RMLP algorithm in Javanmard and Nazerzadeh [2019], but our new analysis improves their regret bound at $O(\sqrt{T})$ when $\mathbb{E}[xx^\top]$ is near singular.

2. When x_t's are adversarial, we propose an "Online-Newton-Step Pricing (ONSP)" algorithm that achieves $O(d \log T)$ regret on constant-level noises for the first time, which exponentially improves the best existing result of $O(T^{2/3})$ [Cohen et al., 2020].[1]

3. Our methods that achieve logarithmic regret require knowing the exact distribution of N_t in advance, as is also assumed in Javanmard and Nazerzadeh [2019]. We prove an $\Omega(\sqrt{T})$ lower bound on the regret if $N_t \sim \mathcal{N}(0, \sigma^2)$ where σ is *unknown*, even with θ^* given and x_t fixed for all t.

[1] Previous works [Cohen et al., 2020, Krishnamurthy et al., 2021] did achieve polylog regrets, but only for negligible noise with $\sigma = O(\frac{1}{T \log T})$.

The $O(\log T)$ regret of EMLP and ONSP meets the information-theoretical lower bound [Theorem 5, Javanmard and Nazerzadeh, 2019]. In fact, the bound is optimal even when w_t is revealed to the learner [Mourtada, 2019]. From the perspective of characterizing the hardness of dynamic pricing problems, we generalize the classical results on "The Value of Knowing a Demand Curve" [Kleinberg and Leighton, 2003] by further dividing the random-valuation class with an exponential separation of: (1) $O(\log T)$ regret for knowing the *demand curve* exactly (even with adversarial features), and (2) $\Omega(\sqrt{T})$ regret for *almost* knowing the *demand curves* (up to a one-parameter parametric family).

2.2 Related Works

In this section, we discuss our results relative to existing works on feature-based dynamic pricing, and highlight the connections and differences to the related settings of contextual bandits and contextual search.

Feature-based Dynamic Pricing. There is a growing body of work on dynamic pricing with linear features [Amin et al., 2014, Qiang and Bayati, 2016, Cohen et al., 2020, Javanmard and Nazerzadeh, 2019]. Table 2.1 summarizes the differences in the settings and results[2]. Among these work, our paper directly builds upon [Cohen et al., 2020] and [Javanmard and Nazerzadeh, 2019], as we share the same setting of online feature vectors, linear and noisy valuations and Boolean-censored feedback. Relative to the results in [Javanmard and Nazerzadeh, 2019], we obtain $O(d \log T)$ regret under weaker assumptions on the sequence of input features — in both distribution-free stochastic feature setting and the adversarial feature setting. It is to be noted that [Javanmard and Nazerzadeh, 2019] also covers the sparse high-dimensional setting, and handles a

[2]We only concern the dependence on T since there are various different assumptions on d.

Table 2.1: Related Works and Regret Bounds w.r.t. T

Algorithm	Work	Regret (upper) bound	Feature	Noise
LEAP	[Amin et al., 2014]	$\tilde{O}(T^{\frac{2}{3}})$	i.i.d.	Noise-free
EllipsoidPricing	[Cohen et al., 2020]	$O(\log T)$	adversarial	Noise-free
EllipsoidEXP4	[Cohen et al., 2020]	$\tilde{O}(T^{\frac{2}{3}})$	adversarial	Sub-Gaussian
PricingSearch	[Leme and Schneider, 2018]	$O(\log\log(T))$	adversarial	Noise-free
RMLP	[Javanmard and Nazerzadeh, 2019]	$O(\log T/C_{\min}^2)^\dagger$ $O(\sqrt{T})$	i.i.d.	Log-concave, distribution-known
RMLP-2	[Javanmard and Nazerzadeh, 2019]	$O(\sqrt{T})$	i.i.d.	Known parametric family of log-concave.
ShallowPricing	[Cohen et al., 2020]	$O(poly(\log T))$	adversarial	Sub-Gaussian, known $\sigma = O(\frac{1}{T\log T})$
CorPV	[Krishnamurthy et al., 2021]			
Algorithm 2 (MSPP)	[Liu et al., 2021]	$O(\log\log(T))$	adversarial	Noise-free
EMLP	This paper	$O(\log T)$	i.i.d.	Strictly log-concave, distribution-known
ONSP	This paper	$O(\log T)$	adversarial	Strictly log-concave, distribution-known

† $C_{\min}$ is the restricted eigenvalue condition. It reduces to the smallest eigenvalue of $\mathbb{E}[xx^\top]$ in the low-dimensional case we consider.

slightly broader class of demand curves. Relative to [Cohen et al., 2020], in which the adversarial feature-based dynamic pricing was first studied, our algorithm ONSP enjoys the optimal $O(d\log T)$ regret when the noise-level is a constant. In comparison, Cohen et al. [2020] reduces the problem to contextual bandits and applies the (computationally inefficient) "EXP-4" algorithm [Auer et al., 2002b] to achieve a $\tilde{O}(T^{2/3})$ regret. The "bisection" style-algorithm in both Cohen et al. [2020] and Krishnamurthy et al. [2021] could achieve $\tilde{O}(poly(d)poly\log(T))$ regrets but requires a small-variance subgaussian noise satisfying $\sigma = O(\frac{1}{T\log T})$.

Lower Bounds. Most existing works focus on the lower regret bounds of non-feature-based models. Kleinberg and Leighton [2003] divides the problem setting as fixed, random, and adversarial valuations, and then proves each a $\Theta(\log\log T)$, $\Theta(\sqrt{T})$, and $\Theta(T^{2/3})$

regret, respectively. Broder and Rusmevichientong [2012] further proves a $\Theta(\sqrt{T})$ regret in general parametric valuation models. In comparison, we generalize the methods of Broder and Rusmevichientong [2012] to our feature-based setting and further narrow it down to a linear-feature Gaussian-noisy model. As a complement to Kleinberg and Leighton [2003], we further separate the exponential regret gap between: (1) $O(\log T)$ of the hardest (adversarial feature) totally-parametric model, and (2) $\Omega(\sqrt{T})$ of the simplest (fixed known expectation) unknown-σ Gaussian model.

Contextual Bandits. For readers familiar with the online learning literature, our problem can be reduced to a contextual bandits problem [Langford and Zhang, 2007, Agarwal et al., 2014] by discretizing the prices. But this reduction only results in $O(T^{2/3})$ regret, as it does not capture the special structure of the feedback: *an accepted price indicates the acceptance of all lower prices*, and vise versa. Moreover, when comparing to linear bandits [Chu et al., 2011], it is the valuation instead of the expected reward that we assume to be linear.

Contextual Search. Feature-based dynamic pricing is also related to the contextual search problem [Lobel et al., 2018, Leme and Schneider, 2018, Liu et al., 2021, Krishnamurthy et al., 2021], which often involves learning from Boolean feedbacks, sometimes with a "pricing loss" and "noisy" feedback. These shared jargons make this problem *appearing* very similar to our problem. However, except for the noiseless cases [Lobel et al., 2018, Leme and Schneider, 2018], contextual search algorithms, even with "pricing losses" and "Noisy Boolean feedback" [e.g., Liu et al., 2021], do *not* imply meaningful regret bounds in our problem setup due to several subtle but important differences in the problem settings. Specifically, the noisy-boolean feedback model of [Liu et al., 2021] is about randomly toggling the "purchase decision" determined by the *noiseless* valuation $x^\top \theta^*$ with probability $0.5 - \epsilon$. This is incompatible to our problem setting where the

purchasing decision is determined by a noisy valuation $x^\top\theta^* + \text{Noise}$. Ultimately, in the setting of [Liu et al., 2021], the optimal policy alway plays $x^\top\theta^*$, but our problem is harder in that we need to exploit the noise and the optimal price could be very different from $x^\top\theta^*$. [3] Krishnamurthy et al. [2021] also discussed this issue explicitly and considered the more natural noisy Boolean feedback model studied in this paper. Their result, similar to Cohen et al. [2020], only achieves a logarithmic regret when the noise on the valuation is vanishing in an $\tilde{O}(1/T)$ rate.

2.3 Problem Setup

Symbols and Notations. Now we introduce the mathematical symbols and notations involved in the following pages. The game consists of T rounds. $x_t \in \mathbb{R}^d$, $p_t \in \mathbb{R}_+$ and $N_t \in \mathbb{R}$ denote the feature vector, the proposed price and the noise respectively at round $t = 1, 2, ..., T$.[4] We denote the product $u_t := x_t^\top\theta^*$ as an *expected valuation*. At each round, we receive a payoff (reward) $r_t = p_t \cdot \mathbb{1}_t$, where the binary variable $\mathbb{1}_t$ indicates whether the price is accepted or not, i.e., $\mathbb{1}_t = \mathbf{1}(p_t \leq w_t)$. As we may estimate θ^* in our algorithms, we denote $\hat{\theta}_t$ as an estimator of θ^*, which we will formally define in the algorithms. Furthermore, we denote some functions that are related to noise distribution: $F(\omega)$ and $f(\omega)$ denote the cumulative distribution function (CDF) and probability density function (PDF) sequentially. We know that $F'(\omega) = f(\omega)$ if we assume differentiability. To concisely denote all data observed up to round τ (i.e., feature, price and payoff of all past rounds), we define $hist(\tau) = \{(x_t, p_t, \mathbb{1}_t) \text{ for } t = 1, 2, ..., \tau\}$. $hist(\tau)$ represents the *transcript* of all observed random variables before round $(\tau + 1)$.

[3] As an explicit example, suppose the valuation $x^\top\theta^* = 0$, then the optimal price must be > 0 in order to avoid zero return.

[4] In an epoch-design situation, a subscript (k, t) indicates round t of epoch k.

We define

$$l_t(\theta) := -\mathbb{1}_t \cdot \log\Big(1 - F(p_t - x_t^\top \theta)\Big) - (1 - \mathbb{1}_t) \log\Big(F(p_t - x_t^\top \theta)\Big) \tag{2.1}$$

as a negative log-likelihood function at round t. Also, we define an expected log-likelihood function $L_t(\theta)$:

$$L_t(\theta) := \mathbb{E}_{N_t}[l_t(\theta)|x_t] \tag{2.2}$$

Notice that we will later define an $\hat{L}_k(\theta)$ which is, however, not an expectation.

Definitions of Key Quantities. We firstly define an *expected reward* function $g(p, u)$.

$$g(p, u) := \mathbb{E}[r_t|p_t = p, x_t^\top \theta^* = u] = p \cdot P[p \leq x_t^\top \theta^* + N_t] = p \cdot (1 - F(p - u)). \tag{2.3}$$

This indicates that if the expected valuation is u and the proposed price is p, then the (conditionally) expected reward is $g(p, u)$. Now we formally define the *regret* of a policy (algorithm) $\mathcal{A}$ as is promised in Section 2.1.

Definition 2.3.1 (Regret)**.** Let $\mathcal{A} : \mathbb{R}^d \times \Big(\mathbb{R}^d, \mathbb{R}, \{0, 1\}\Big)^{t-1} \to \mathbb{R}$ be a policy of pricing, i.e. $\mathcal{A}(x_t, hist(t-1)) = p_t$. The regret of $\mathcal{A}$ is defined as follows.

$$Reg_{\mathcal{A}} = \sum_{t=1}^{T} \max_p g(p, x_t^\top \theta^*) - g(\mathcal{A}(x_t, hist(t-1)), x_t^\top \theta^*). \tag{2.4}$$

Here $hist(t-1)$ is the historical records until $(t-1)^{\text{th}}$ round.

Summary of Assumptions. We specify the problem settings by proposing three assumptions.

Assumption 2.3.2 (Known, bounded, strictly log-concave distribution)**.** The noise N_t is independently and identically sampled from a distribution whose CDF is F. Assume that $F \in \mathbb{C}^2$ is strictly increasing and that F and $(1 - F)$ are strictly log-concave. Also assume that f and f' are bounded, and denote $B_f := \sup_{\omega \in \mathbb{R}} f(\omega), B_{f'} := \sup_{\omega \in \mathbb{R}} |f'(\omega)|$ as two constants.

Assumption 2.3.3 (Bounded convex parameter space)**.** The true parameter $\theta^* \in \mathbb{H}$, where $\mathbb{H} \subseteq \{\theta : ||\theta||_2 \leq B_1\}$ is a bounded convex set and B_1 is a constant. Assume $\mathbb{H}$ is known to us (but θ^* is not).

Assumption 2.3.4 (Bounded feature space)**.** Assume $x_t \in D \subseteq \{x : ||x||_2 \leq B_2\}$, $\forall t = 1, 2, \ldots, T$. Also, $0 \leq x^\top \theta \leq B, \forall x \in D, \forall \theta \in \mathbb{H}$, where $B = B_1 \cdot B_2$ is a constant.

Assumption 2.3.3 and Assumption 2.3.4 are mild as we can choose B_1 and B_2 large enough. In Section 2.4.1, we may add further complement to Assumption 2.3.4 to form a stochastic setting. Assumption 2.3.2 is stronger since we might not know the exact CDF in practice, but it is still acceptable from an information-theoretic perspective. There are at least three reasons that lead to this assumption: Primarily, this is necessary if we hope to achieve an $O(\log(T))$ regret. We will prove in Section 2.5.3 that an $\Omega(\sqrt{T})$ is unavoidable if we cannot know one parameter exactly. Moreover, the pioneering work of Javanmard and Nazerzadeh [2019] also assumes a known noise distribution with log-concave CDF, and many common distributions are actually strictly log-concave, such as Gaussian and logistic.[5] Besides, although we did not present a method to precisely estimate σ in this chapter, it is a reasonable algorithm to replace with a plug-in estimator estimated using historical offline data. As we have shown, not knowing σ requires $O(\sqrt{T})$ regret in general, but the lower bound does not rule out the plug-in approach achieving a smaller regret for interesting subclasses of problems in practice.

Finally, we state a lemma and define an argmax function helpful for our algorithm design.

Lemma 2.3.5 (Uniqueness)**.** *For any $u \geq 0$, there exists a unique $p^* \geq 0$ such that*

[5]In fact, F and $(1 - F)$ are both log-concave if its PDF is log-concave, according to Prekopa's Inequality.

Algorithm 1 Epoch-based max-likelihood pricing (EMLP)

Input: Convex and bounded set $\mathbb{H}$

Observe x_1, randomly choose p_1 and get r_1.

Solve $\hat{\theta}_1 = \arg\min_{\theta\in\mathbb{H}} l_1(\theta)$;

for $k = 1$ **to** $\lfloor \log_2 T \rfloor + 1$ **do**

Set $\tau_k = 2^{k-1}$;

for $t = 1$ **to** τ_k **do**

Observe $x_{k,t}$;

Set price $p_{k,t} = J(x_{k,t}^\top \hat{\theta}_k)$;

Receive $r_{k,t} = p_{k_t} \cdot \mathbb{1}_t$;

end for

Solve: $\hat{\theta}_{k+1} = \arg\min_{\theta\in\mathbb{H}} \hat{L}_k(\theta)$, where $\hat{L}_k(\theta) = \frac{1}{\tau_k} \sum_{t=1}^{\tau_k} l_{k,t}(\theta)$.

end for

Algorithm 2 Online Newton Step Pricing (ONSP)

Input: Convex and bounded set $\mathbb{H}$, θ_1, parameter $\gamma, \epsilon > 0$

Set $A_0 = \epsilon \cdot I_d$;

for $t = 1$ **to** T **do**

Observe x_t;

Set price $p_t = J(x_t^\top \theta_t)$;

Receive $r_t = p_t \cdot \mathbb{1}_t$;

Set surrogate loss function $l_t(\theta)$;

Calculate $\nabla_t = \nabla l_t(\theta)$;

Rank-1 update: $A_t = A_{t-1} + \nabla_t \nabla_t^\top$;

Newton step: $\hat{\theta}_{t+1} = \theta_t - \frac{1}{\gamma} A_t^{-1} \nabla_t$;

Projection: $\theta_{t+1} = \prod_{\mathbb{H}}^{A_t} (\hat{\theta}_{t+1})$.

end for

$g(p^*, u) = \max_{p\in\mathbb{R}} g(p, u)$. *Thus, we can define a* greedily pricing *function that maximizes the expected reward:*

$$J(u) := \arg\max_{p} g(p, u) \tag{2.5}$$

Please see the proof of Lemma 2.3.5 in Section 2.9.1.

2.4 Algorithms

In this section, we propose two dynamic pricing algorithms: EMLP and ONSP, for stochastic and adversarial features respectively.

2.4.1 Pricing with Distribution-Free Stochastic Features

Assumption 2.4.1 (Stochastic features)**.** Assume $x_t \sim \mathbb{D} \subseteq D$ are independently identically distributed (i.i.d.) from an unknown distribution, for any $t = 1, 2, \ldots, T$.

The first algorithm, Epoch-based Max-Likelihood Pricing (EMLP) algorithm, is suitable for a stochastic setting defined by Assumption 2.4.1. EMLP proceeds in epochs with each stage doubling the length of the previous epoch. At the end of each epoch, we consolidate the observed data and solve a maximum likelihood estimation problem to learn θ. A max likelihood estimator (MLE) obtained by minimizing $\hat{L}_k(\theta) := \frac{1}{\tau_k} \sum_{t=1}^{\tau_k} l_{k,t}(\theta)$, which is then used in the next epoch as if it is the true parameter vector. In the equation, k, τ_k denotes the index and length of epoch k. The estimator is computed using data in $hist(k)$, which denotes the transcript for epoch $1 \sim k$. The pseudo-code of EMLP is summarized in Algorithm 1. In the remainder of this section, we discuss the computational efficiency and prove the upper regret bound of $O(d \log T)$.

Computational Efficiency. The calculations in EMLP are straightforward except for $\arg\min \hat{L}_k(\theta)$ and $J(u)$. As $g(p, u)$ is proved unimodal in Lemma 2.3.5, we may efficiently calculate $J(u)$ by binary search. We will prove that $l_{k,t}$ is exp-concave (and thus also convex). Therefore, we may apply any off-the-shelf tools for solving convex optimization.

MLE and Probit Regression. A closer inspection reveals that this log-likelihood function corresponds to a probit [Aldrich et al., 1984] or a logit model [Wright, 1995] for Gaussian or logistic noises. See Section 2.7.4.

Affine Invariance. Both optimization problems involved depend only on $x^\top \theta$, so if we add any affine transformation to x into $\tilde{x} = Ax$, the agent can instead learn a new parameter of $\tilde{\theta}^* = (A^\top)^{-1}\theta^*$ and achieve the same $u_t = x_t^\top \theta^*$. Also, the regret bound

is not affected as the upper bound B over $x^\top\theta$ does not change [6]. Therefore, it is only natural that the regret bound does not depend on the distribution x, nor the condition numbers of $\mathbb{E}[xx^\top]$ (i.e., the ratio of $\lambda_{\max}/\lambda_{\min}$).

2.4.2 Pricing with Adversarial Features

In this part, we propose an "Online Newton Step Pricing (ONSP)" algorithm that deals with adversarial $\{x_t\}$ series and guarantees $O(d\log T)$ regret. The pseudo-code of ONSP is shown as Algorithm 2. In each round, it uses the likelihood function as a surrogate loss and applies "Online Newton Step"(ONS) method to update $\hat{\theta}$. In the next round, it adopts the updated $\hat{\theta}$ and sets a price greedily. In the remainder of this section, we discuss some properties of ONSP and prove the regret bound.

The calculations of ONSP are straightforward. The time complexity of calculating the matrix inverse A_t^{-1} is $O(d^3)$, which is fair as d is small. In high-dimensional cases, we may use *Woodbury matrix identity*[7] to reduce it to $O(d^2)$ as we could get A^{-1} directly from the latest round.

2.5 Regret Analysis

In this section, we mainly prove the logarithmic regret bounds of EMLP and ONSP corresponding to stochastic and adversarial settings, respectively. Besides, we also prove an $\Omega(\sqrt{T})$ regret bound on fully parametric F with one parameter unknown.

[6] Here A is assumed invertible, otherwise the mapping from $\tilde{x}_t$ to u_t does not necessarily exist.
[7] $(A+xx^\top)^{-1} = A^{-1} - \frac{1}{1+x^\top A^{-1}x}A^{-1}x(A^{-1}x)^\top$.

2.5.1 $O(d \log T)$ Regret of EMLP

In this part, we present the regret analysis of Algorithm 1. First of all, we propose the following theorem as our main result on EMLP.

Theorem 2.5.1 (Overall regret)**.** *With Assumption 2.3.2, Assumption 2.3.3, Assumption 2.3.4 and Assumption 2.4.1, the expected regret of EMLP can be bounded by:*

$$\mathbb{E}[Reg_{EMLP}] \leq 2C_s d \log T, \tag{2.6}$$

where C_s is a constant that depends only on $F(\omega)$ and is independent to $\mathbb{D}$.

The proof of Theorem 2.5.1 is sophisticated. For the sake of clarity, we next present an inequality system as a roadmap toward the proof. After this, we formally illustrate each line of it with lemmas.

Since EMLP proposes $J(x_{k,t}^\top \hat{\theta}_k)$ in every round of epoch k, we may denote the per-round regret as $Reg_t(\hat{\theta}_k)$, where:

$$Reg_t(\theta) := g(J(x_t^\top \theta^*), x_t^\top \theta^*) - g(J(x_t^\top \theta), x_t^\top \theta^*). \tag{2.7}$$

Therefore, it is sufficient to prove the following Theorem:

Theorem 2.5.2 (Expected per-round regret)**.** *For the per-round regret defined in Eq.* (2.7)*, we have:*

$$\mathbb{E}[Reg_{k,t}(\hat{\theta}_k)] \leq C_s \cdot \frac{d}{\tau_k}.$$

The proof roadmap of Theorem 2.5.2 can be written as the following inequalities.

$$\begin{aligned} \mathbb{E}[Reg_{k,t}(\hat{\theta}_k)] \leq & C \cdot \mathbb{E}[(\hat{\theta}_k - \theta^*)^\top x_{k,t} x_{k,t}^\top (\hat{\theta}_k - \theta^*)] \leq \frac{2C}{C_{\text{down}}} \mathbb{E}[\hat{L}_k(\hat{\theta}_k) - \hat{L}_k(\theta^*)] \\ \leq & \frac{2C \cdot C_{\text{exp}}}{C_{\text{down}}^2} \frac{d}{\tau_k}. \end{aligned} \tag{2.8}$$

We explain Eq. (2.8) in details. The first inequality comes from the following Lemma 2.5.3.

Lemma 2.5.3 (Quadratic regret bound)**.** *We have:*

$$Reg_t(\theta) \leq C \cdot (\theta - \theta^*)^\top x_t x_t^\top (\theta - \theta^*), \forall \theta \in \mathbb{H}, \forall x_t \in \mathbb{D}. \tag{2.9}$$

Here $C = 2B_f + (B + J(0)) \cdot B_{f'}$.

The intuition is that function $g(J(u), u)$ is 2^{nd}-order-smooth at $(J(u^*), u^*)$. A detailed proof of Lemma 2.5.3 is in Section 2.9.2. Note that C is highly dependent on the distribution F. After this, we propose Lemma 2.5.4 that contributes to the second inequality of Eq. (2.8).

Lemma 2.5.4 (Quadratic likelihood bound)**.** *For the expected likelihood function* $L_t(\theta)$ *defined in Eq.* (2.2)*, we have:*

$$L_t(\theta) - L_t(\theta^*) \geq \frac{1}{2} C_{down} (\theta - \theta^*)^\top x_t x_t^\top (\theta - \theta^*), \forall \theta \in \mathbb{H}, \forall x \in \mathbb{D}, \tag{2.10}$$

$$\textit{where} \quad C_{down} := \inf_{\omega \in [-B, B+J(0)]} \min \left\{ \frac{d^2 \log(1 - F(\omega))}{d\omega^2}, \frac{d^2 \log(F(\omega))}{d\omega^2} \right\} > 0. \tag{2.11}$$

Proof. Since the true parameter always maximizes the expected likelihood function [Murphy, 2012], by Taylor Expansion we have $\nabla L(\theta^*) = 0$, and hence $L_t(\theta) - L_t(\theta^*) = \frac{1}{2}(\theta - \theta^*)^\top \nabla^2 L_t(\tilde{\theta})(\theta - \theta^*)$ for some $\tilde{\theta} = \alpha\theta^* + (1 - \alpha)\theta$. Therefore, we only need to prove the following lemma:

Lemma 2.5.5 (Strong convexity and Exponential Concavity)**.** *Suppose* $l_t(\theta)$ *is the negative log-likelihood function in epoch* k *at time* t*. For any* $\theta \in \mathbb{H}, x_t \sim \mathbb{D}$*, we have:*

$$\nabla^2 l_t(\theta) \succeq C_{down} x_t x_t^\top \succeq \frac{C_{down}}{C_{exp}} \nabla l_t(\theta) \nabla l_t(\theta)^\top \succeq 0, \tag{2.12}$$

$$\textit{where} \quad C_{exp} := \sup_{\omega \in [-B, B+J(0)]} \max \left\{ \frac{f(\omega)^2}{F(\omega)^2}, \frac{f(\omega)^2}{(1 - F(\omega))^2} \right\} < +\infty. \tag{2.13}$$

Proof of Lemma 2.5.5 is in Section 2.9.2. With this lemma, we see that Lemma 2.5.4 holds. ■

With Lemma 2.5.3 and Lemma 2.5.4, we can immediately get the following Lemma 2.5.6.

Lemma 2.5.6 (Surrogate Regret)**.** *The relationship between $Reg(\theta)$ and likelihood function can be shown as follows:*

$$Reg_t(\theta) \leq \frac{2 \cdot C}{C_{down}} \left(L_t(\theta) - L_t(\theta^*) \right), \tag{2.14}$$

$\forall \theta \in \mathbb{H}, \forall x \in \mathbb{D}$, where C and C_{down} are defined in Lemma 2.5.3 and Lemma 2.5.4 respectively.

Lemma 2.5.6 enables us to choose the negative log-likelihood function as a surrogate loss. This is not only an important insight of EMLP regret analysis, but also the foundation of ONSP design.

The last inequality of Eq. (2.8) comes from this lemma:

Lemma 2.5.7 (Per-epoch surrogate regret bound)**.** *Denoting $\hat{\theta}_k$ as the estimator coming from epoch $(k-1)$ and being used in epoch k, we have:*

$$\mathbb{E}_h[\hat{L}_k(\hat{\theta}_k) - \hat{L}_k(\theta^*)] \leq \frac{C_{exp}}{C_{down}} \cdot \frac{d}{\tau_k + 1}. \tag{2.15}$$

Here C_{exp} is defined in Eq. (2.13)*, and $\mathbb{E}_h[\cdot] = \mathbb{E}[\cdot | hist(k-1)]$.*

Proof of Lemma 2.5.7 is partly derived from the work Koren and Levy [2015], and here we give a proof sketch without specific derivations. A detailed proof lies in Section 2.9.2.

Proof sketch of Lemma 2.5.7. We list the four main points that contribute to the proof:

- Notice that $l_{k,t}(\theta)$ is strongly convex w.r.t. a seminorm $x_{k,t}x_{k,t}^\top$, we know $\hat{L}_k(\theta)$ is also strongly convex w.r.t. $\sum_{t=1}^{\tau_k} x_{k,t}x_{k,t}^\top$.

- For two strongly convex functions g_1 and g_2, we can upper bound the distance between their arg-minimals (scaled by some norm $||\cdot||$) with the dual norm of $\nabla(g_1 - g_2)$.

- Since a seminorm has no dual norm, we apply two methods to convert it into a norm: (1) separation of parameters and likelihood functions with a "leave-one-out" method (to separately take expectations), and (2) separation of the spinning space and the null space.

- As the dual data-dependent norm offsets the sum of $xx^\top$ to a constant, Lemma 2.5.7 holds.

We have so far proved Eq. (2.8) after proving Lemma 2.5.3, Lemma 2.5.4, Lemma 2.5.7. Therefore, Theorem 2.5.2 holds.

2.5.2 $O(d \log T)$ Regret of ONSP

Here we present the regret analysis of Algorithm 2 (ONSP). Firstly, we state the main theorem.

Theorem 2.5.8. *With Assumption 2.3.2, Assumption 2.3.3, Assumption 2.3.4, the regret of Algorithm 2 (ONSP) satisfies:*

$$Reg_{ONSP} \leq C_a \cdot d \log T, \tag{2.16}$$

where C_a is a function only dependent on F.

Proof. Proof of Theorem 2.5.8 here is more concise than Section 2.5.1, because the important Lemma 2.5.5 and Lemma 2.5.6 have been proved there. From Lemma 2.5.6, we have:

$$g(J(u_t^*), u_t^*) - g(J(u_t), u_t^*) \leq \frac{2 \cdot C}{C_{\text{down}}} \cdot \mathbb{E}_{N_t}[l_t(\theta_t) - l_t(\theta^*)]. \tag{2.17}$$

With Eq. (2.17), we may reduce the regret of likelihood functions as a surrogate regret of pricing. From Lemma 2.5.5 we see that the log-likelihood function is $\frac{C_{\text{down}}}{C_{\text{exp}}}$-exponentially concave[8]. This enables an application of Online Newton Step method to achieve a logarithmic regret. Therefore, by citing from the *Online Convex Optimization* [Hazan, 2016], we have the following Lemma.

Lemma 2.5.9 (Online Newton Step)**.** *With parameters* $\gamma = \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$ *and* $\epsilon = \frac{1}{\gamma^2 D^2}$, *and* $T > 4$ *guarantees:*

$$\sup_{\{x_t\}} \left\{ \sum_{t=1}^{T} l_t(\theta_t) - \min_{\theta \in \mathbb{H}} \sum_{t=1}^{T} l_t(\theta) \right\} \leq 5 \left(\frac{1}{\alpha} + GD \right) d \log T.$$

Here $\alpha = \frac{C_{down}}{C_{exp}}$, $D = 2 \cdot B_1$ *and* $G = \sqrt{C_{exp}} \cdot B_2$.

With Eq. (2.17) and Lemma 2.5.9, we have:

$$Reg = \sum_{t=1}^{T} \left(g(J(u_t^*), u_t^*) - \mathbb{E}_{N_1, N_2, \ldots, N_{t-1}}[g(J(u_t), u_t^*)] \right) \leq \frac{2 \cdot C}{C_{\text{down}}} \cdot 5 \left(\frac{1}{\alpha} + GD \right) d \log T. \tag{2.18}$$

Therefore, we have proved Theorem 2.5.8. ∎

2.5.3 Lower Bound for Unknown Distribution

In this part, we evaluate Assumption 2.3.2 and prove that an $\Omega(\sqrt{T})$ lower regret bound is unavoidable with even a slight relaxation: a Gaussian noise with unknown σ. Our proof

[8] A function $f(\mu)$ is α-exponentially concave iff $\nabla^2 f(\mu) \succeq \alpha \nabla f(\mu) \nabla f(\mu)^\top$.

is inspired by Broder and Rusmevichientong [2012] Theorem 3.1, while our lower bound relies on more specific assumptions (and thus applies to more general cases).

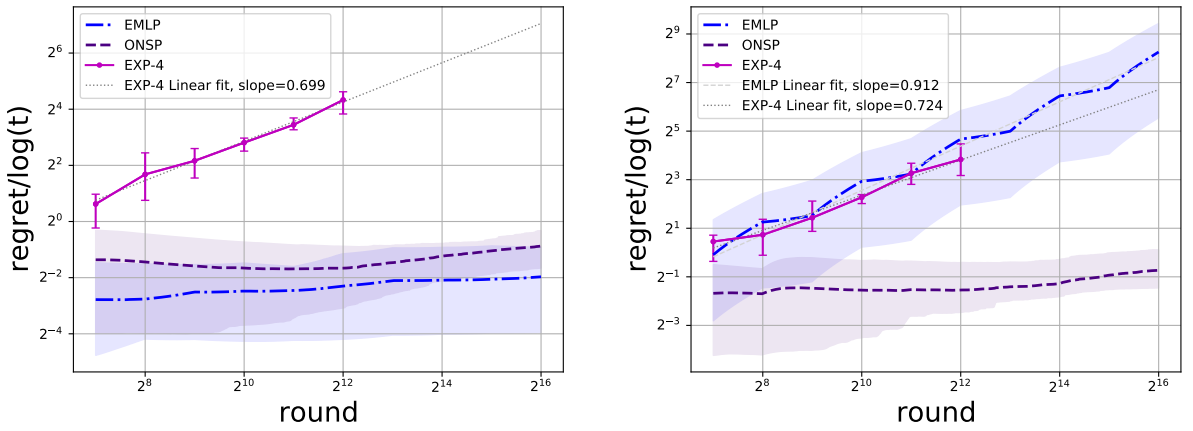
We firstly state Assumption 2.5.10 covering this part, and then state Theorem 2.5.11 as a lower bound:

Assumption 2.5.10. The noise $N_t \sim \mathcal{N}(0, \sigma^2)$ independently, where $0 < \sigma \leq 1$ is fixed and **unknown**.

Theorem 2.5.11 (Lower bound with unknown σ)**.** *Under Assumption 2.3.3, Assumption 2.3.4, Assumption 2.4.1 and Assumption 2.5.10, for any policy (algorithm)* $\Psi : \mathbb{R}^d \times \left(\mathbb{R}^d, \mathbb{R}, \{0,1\}\right)^{t-1} \to \mathbb{R}^+$ *and any* $T > 2$*, there exists a Gaussian parameter* $\sigma \in \mathbb{R}^+$*, a distribution* $\mathbb{D}$ *of features and a fixed parameter* θ^**, such that:* $Reg_\Psi \geq \frac{1}{24000} \cdot \sqrt{T}$.

Remark: Here we assume x_t to be i.i.d., which also implies the applicability on adversarial features. However, the minimax regret of the stochastic feature setting is $\Theta(\sqrt{T})$ [Javanmard and Nazerzadeh, 2019], while existing results have not yet closed the gap in adversarial feature settings.

Proof sketch of Theorem 2.5.11. Here we assume a fixed valuation, i.e. $u^* = x_t^\top \theta^*, \forall t = 1, 2, \ldots$. Equivalently, we assume a fixed feature. The main idea of proof is similar to that in Broder and Rusmevichientong [2012]: we assume $\sigma_1 = 1, \sigma_2 = 1 - T^{-\frac{1}{4}}$, and we prove that: (1) it is costly for an algorithm to perform well in both cases if the σ's are different by a lot, and (2) it is costly for an algorithm to distinguish the two cases if σ's are close enough to each other. We put the detailed proof in Section 2.9.3.



(a) Stochastic feature (b) Adversarial feature

Figure 2.1: The regret of EMLP, ONSP and EXP-4 on simulated examples (we only conduct EXP-4 up to $T = 2^{12}$ due to its exponential time consuming), with Figure a for stochastic features and Figure b for adversarial ones. The plots are in log-log scales with all regrets divided by a $\log(t)$ factor to show the convergence. For EXP-4, we discretize the parameter space with $T^{-\frac{1}{3}}$-size grids, which would incur an $\tilde{O}(T^{\frac{2}{3}})$ regret according to Cohen et al. [2020]. We also plot linear fits for some regret curves, where a slope-α line indicates an $O(T^{\alpha})$ regret. Besides, we draw error bars and bands with 0.95 coverage using Wald's test. The two diagrams reveal that (i) logarithmic regrets of EMLP and ONSP in the stochastic setting, (ii) a nearly-linear regret of EMLP in the adversarial setting, and (iii) $O(T^{\frac{2}{3}})$ regrets of EXP-4 in both settings.

2.6 Numerical Result

In this section, we conduct numerical experiments to validate EMLP and ONSP. In comparison with the existing work, we implement a discretized EXP-4 [Auer et al., 2002b] algorithm for pricing, as is introduced in Cohen et al. [2020] (in a slightly different setting). We will test these three algorithms in both stochastic and adversarial settings.

Basically, we assume $d = 2, B_1 = B_2 = B = 1$ and $N_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.25$. In both settings, we conduct EMLP and ONSP for $T = 2^{16}$ rounds. For ONSP, we empirically

select γ and ϵ that accelerates the convergence, instead of using the values specified in Lemma 2.5.9. Since EXP-4 consumes exponential time and requires the knowledge of T in advance to discretize the policy and valuation spaces, we execute EXP-4 for a series of $T = 2^k, k = 1, 2, \ldots, 12$. We repeat every experiment 5 times for each setting and then take an average.

Stochastic Setting. We implement and test EMLP, ONSP and EXP-4 with stochastic $\{x_t\}$'s. The numerical results are shown in Figure 2.1a on a log-log diagram, with the regrets divided by $\log(t)$. It shows $\log(t)$-convergences on EMLP and ONSP, while EXP-4 is in a t^α rate with $\alpha \approx 0.699$.

Adversarial Setting. We implement the three algorithms and test them with an adversarial $\{x_t\}$'s: for the k-th epoch, i.e. $t = 2^{k-1}, 2^{k-1}+1, \ldots, 2^k - 1$, we let $x_t = [1, 0]^\top$ if $k \equiv 1 (\mod 2)$ and $x_t = [0, 1]^\top$ if $k \equiv 0 (\mod 2)$. The numerical results are shown in Figure 2.1b on a log-log diagram, with the regrets divided by $\log(t)$. The log-log plots of ONSP and EXP-4 are almost the same as those in Figure 2.1a. However, EMLP shows an almost linear (t^α rate with $\alpha \approx 0.912$) regret in this adversarial setting. This is because the adversarial series only trains one dimension of θ in each epoch, while the other is arbitrarily initialized and does not necessarily converge. However, in the next epoch, the incorrect dimension is exploited. Therefore, a linear regret originates.

2.7 Discussion

Here we discuss more related works, regret dependence on other parameters, problem modeling, algorithm design, and potential extensions of future works.

2.7.1 More Related Works

Here we will briefly review the history and recent studies that are related to our work. For the historical introductions, we mainly refer to den Boer [2015] as a survey. For bandit approaches, we will review some works that apply bandit algorithms to settle pricing problems. For the structural models, we will introduce different modules based on the review in Chan et al. [2009]. Based on the existing works, we might have a better view of our problem setting and methodology.

History of Pricing It was the work of Cournot [1897] in 1897 that firstly applied mathematics to analyze the relationship between prices and demands. In that work, the price was denoted as p and the demand was defined as a *demand function* $F(p)$. Therefore, the revenue could be written as $pF(p)$. This was a straightforward interpretation of the general pricing problem, and the key to solving it was estimations of $F(p)$ regarding different products. Later in 1938, the work Schultz et al. [1938] proposed price-demand measurements on exclusive kinds of products. It is worth mentioning that these problems are "static pricing" ones, because F is totally determined by price p and we only need to insist on the optimal one to maximize our profits.

However, the static settings were qualified by the following two observations: on the one hand, a demand function may not only depends on the static value of p, but also be affected by the trend of p's changing [Evans, 1924, Mazumdar et al., 2005]; on the other hand, even if $F(p)$ is static, p itself might change over time according to other factors such as inventory level [Kincaid and Darling, 1963]. As a result, it is necessary to consider dynamics in both demand and price, which leads to a "dynamic pricing" problem setting.

Dynamic Pricing as Bandits As is said in Section 2.2, the pricing problem can be viewed as a stochastic contextual bandits problem [see, e.g., Langford and Zhang, 2007, Agarwal et al., 2014]. Even though we may not know the form of the demand function, we can definitely see feedback of demands, i.e. how many products are sold out, which enables us to learn a better decision-making policy. Therefore, it can be studied in a bandit module. If the demand function is totally agnostic, i.e. the evaluations (the highest prices that customers would accept) come at random or even at adversary over time, then it can be modeled as a Multi-arm bandit (MAB) problem [Whittle, 1980] exactly. In our paper, instead, we focus on selling different products with a great variety of features. This can be characterized as a Contextual bandit (CB) problem [Auer et al., 2002b, Langford and Zhang, 2007]. The work Cohen et al. [2020], which applies the "EXP-4" algorithm from Auer et al. [2002b], also mentions that "the arms represent prices and the payoffs from the different arms are correlated since the measures of demand evaluated at different price points are correlated random variables". A variety of existing works, including Kleinberg and Leighton [2003], Araman and Caldentey [2009], Chen and Farias [2013], Keskin and Zeevi [2014], Besbes and Zeevi [2015], has been approaching the demand function from a perspective of from either parameterized or non-parameterized bandits.

However, our problem setting is different from a contextual bandits setting in at least two perspectives: feedback and regret. The pricing problem has a specially structured feedback between full information and bandits setting. Specifically, $r_t > 0$ implies that all policies producing $p < p_t$ will end up receiving $r'_t = p$, and $r_t = 0$ implies that all policies producing $p > p_t$ will end up receiving $r'_t = 0$. However, the missing patterns are confounded with the rewards. Therefore it is non-trivial to leverage this structure to improve the importance sampling approach underlying the algorithm of Agarwal et al. [2014]. We instead consider the natural analog to the linear contextual bandits setting

[Chu et al., 2011][9] and demonstrate that in this case an exponential improvement in the regret is possible using the additional information from the censored feedback. As for regret, while in contextual bandits it refers to a comparison with the optimal policy, it is here referring to a comparison with the optimal *action*. In other words, though our approaches (both in EMLP and in ONSP) are finding the true parameter θ^*, the regret is defined as the "revenue gap" between the optimal price and our proposed prices. These are actually equivalent in our fully-parametric setting (where we assume a linear-valuation-known-noise model), but will differ a lot in partially parametric and totally agnostic settings.

2.7.2 Regret Dependence on Other Parameters

While a totally agnostic model guarantees the most generality, a structural model would help us better understand the mechanism behind the observation of prices and demands. The key to a structural pricing model is the *behavior* of agents in the market, including customers and/or firms. In other words, the behavior of each side can be described as a decision model. From the perspective of demand (customers), the work Kadiyali et al. [1996] adopts a linear model on laundry detergents market, Iyengar et al. [2007] and Lambrecht et al. [2007] study three-part-tariff pricing problems on wireless and internet services with mixed logit models. Besanko et al. assumed an aggregate logit model on customers in works Besanko et al. [1998] and Besanko et al. [2003] in order to study the competitive behavior of manufacturers in ketchup market. Meanwhile, the supply side is usually assumed to be more strategic, such as Bertrand-Nash behaviors [Kadiyali et al., 1996, Besanko et al., 1998, Draganska and Jain, 2006]. For more details, please see Chan et al. [2009].

[9] But do notice that our expected reward above is not linear, even if the valuation function is.

Coefficients of the regret bounds. The exact regret bounds of both EMLP and ONSP contain a constant $\frac{C_{\exp}}{C_{\text{down}}}$ that highly depends on the noise CDF F and could be large. A detailed analysis later in this subsection shows that $\frac{C_{\exp}}{C_{\text{down}}}$ is exponentially large w.r.t. $\frac{B}{\sigma}$ (see Eq. (2.20) and Lemma 2.7.1) for Gaussian noise $\mathcal{N}(0, \sigma^2)$, which implies that a smaller noise variance would lead to a (much) larger regret bound. This is very counter-intuitive as a larger noise usually leads to a more sophisticated situation, but similar phenomenons also occur in existing algorithms that are suitable for constant-variance noise, such as RMLP in Javanmard and Nazerzadeh [2019] and OORMLP in Wang et al. [2020]. In fact, it is because a (constantly) large noise would help explore the unknown parameter θ^* and smoothen the expected regret. In this work, this can be addressed by increasing T since we mainly concern the asymptotic regrets as $T \to \infty$ with fixed noise distributions. However, we admit that it is indeed a nontrivial issue for finite T and small σ situations. There exists a "ShallowPricing" method in Cohen et al. [2020] that can deal with a very-small-variance noise setting (when $\sigma = \tilde{O}(\frac{1}{T})$) and achieve a logarithmic regret. Specifically, its regret bound would decrease as the noise variance σ decreases (but would still not reach $O(\log \log T)$ as the noise vanishes). We might also apply this method as a preprocess to cut the parameter domain and decrease $\frac{B}{\sigma}$ within logarithmic trials (see Cohen et al. [2020] Thm. 3), but it is still open whether a $\log(T)$ regret is achievable when $\sigma = \Theta(T^{-\alpha})$ for $\alpha \in (0, 1)$.

Dependence on B and Noise Variance Here we use a concrete example to analyze the coefficients of regret bounds. Again, we assume that $N_t \sim \mathcal{N}(0, \sigma^2)$. Notice that both C_s and C_a have a component of $\frac{C_{\exp}}{C_{down}}$. In order to analyze $\frac{C_{\exp}}{C_{down}}$, we define a *hazard function* denoted as $\lambda(\omega)$ with $\omega \in \mathbb{R}$:

$$\lambda(\omega) := \frac{\varphi_1(\omega)}{1-\Phi_1(\omega)} = \frac{\varphi_1(-\omega)}{\Phi_1(-\omega)}, \tag{2.19}$$

where Φ_1 and φ_1 are the CDF and PDF of standard Gaussian distribution. The concept of hazard function comes from the area of *survival analysis*. From Eq. (2.11) and Eq. (2.13), we plug in Equation Eq. (2.19) and get:

$$\begin{aligned} C_{\text{down}} &\geq \inf_{\omega\in[-\frac{B}{\sigma},\frac{B}{\sigma}]} \left\{ \frac{1}{\sigma^2}\lambda(-\omega)^2 + \omega\cdot\lambda(-\omega) \right\} \\ C_{\text{exp}} &\leq \sup_{\omega\in[-\frac{B}{\sigma},\frac{B}{\sigma}]} \left\{ \frac{1}{\sigma^2}\lambda(-\omega)^2 \right\}. \end{aligned} \tag{2.20}$$

In Lemma 2.7.1, we will prove that $\lambda(\omega)$ is exponentially small as $\omega \to +\infty$, and is asymptotically close to $-\omega$ as $\omega \to -\infty$. Therefore, C_{down} is exponentially small and C_{exp} is quadratically large with respect to B/σ. Although we assume that B and σ are constant, we should be alert that the scale of B/σ can be very large as σ goes to zero, i.e. as the noise is "insignificant". In practice (especially when T is finite), this may cause extremely large regret at the beginning. A "Shallow Pricing" method introduced by Cohen et al. [2020] (as well as other domain-cutting methods in contextual searching) may serve as a good pre-process as it frequently conducts bisections to cut the feasible region of θ^* with high probability. According to Theorem 3 in Cohen et al. [2020], their Shallow Pricing algorithm will bisect the parameter set for at most logarithmic times to ensure that $\frac{B}{\sigma}$ has been small enough (i.e. upper-bounded by $O(poly\log(T))$). However, this does not necessarily means that we can use a $O(\log T)$-time pre-process to achieve the same effect, since they run the algorithm throughout the session while we only take it as a pre-process. Intuitively, at least under the adversarial feature assumption, we cannot totally rely on a few features occurring at the beginning (as they might be misleading) to cut the parameter set once and for all. A mixture approach of Shallow Pricing and

EMLP/ONSP might work, as the algorithm can detect whether current $\frac{B}{\sigma}$ is larger than a threshold of bisection. However, this requires new regret analysis as the operations parameter domain are changing over time. Therefore, we claim in Section 2.7 that the regret bound is still open if $\sigma = \Theta(T^{-\alpha})$ for $\alpha \in (0, 1)$.

Lemma 2.7.1 (Properties of $\lambda(\omega)$)**.** *For* $\lambda(\omega) := \frac{\varphi_1(\omega)}{1-\Phi_1(\omega)}$*, we have:*

$$
\begin{aligned}
&1, \qquad \frac{d}{d\omega}\lambda(\omega) > 0.\\
&2, \qquad \lim_{\omega\to-\infty} \omega^k \lambda(\omega) = 0, \ \forall k > 0.\\
&3, \qquad \lim_{\omega\to+\infty} \lambda(\omega) - \omega = 0.\\
&4, \qquad \lim_{\omega\to+\infty} \omega\left(\lambda(\omega) - \omega\right) = 1.
\end{aligned}
$$

The detailed proof of Lemma 2.7.1 is in Section 2.9.4.

2.7.3 Problem Modeling

Noise Distributions In this chapter, we have made four assumptions on the noise distribution: strict log-concavity, $2^{\text{nd}} -$ order smooth, known, and i.i.d.. Here we explain each of them specifically.

- The assumption of knowing the exact F is critical to the regret bound: If we have this knowledge, then we achieve $O(\log T)$ even with adversarial features; otherwise, an $\Omega(\sqrt{T})$ regret is unavoidable even with stochastic features.

- The strictly log-concave distribution family includes Gaussian and logistic distributions as two common noises. In comparison, Javanmard and Nazerzadeh [2019] assumes log-concavity that further covers Laplacian, exponential and uniform distributions. Javanmard and Nazerzadeh [2019] also considers the cases when (1)

the noise distribution is unknown but log-concave, and (2) the noise distribution is zero-mean and bounded by support of $[-\delta, \delta]$. For case (1), they propose an algorithm with regret $O(\sqrt{T})$ and meanwhile prove the same lower bound. For case (2), they propose an algorithm with linear regret.

- The assumption that F is $2^{\text{nd}}-$order smooth is also assumed by Javanmard and Nazerzadeh [2019] by taking derivatives $f'(p)$ and applying its upper bound in the proof. Therefore, we are still unaware of the regret bound if the noise distribution is discrete, where a lower bound of $\Omega(\sqrt{T})$ can be directly applied from Kleinberg and Leighton [2003].

- We even assume that the noise is identically distributed. However, the noise would vary among different people. The same problem happens on the parameter θ^*: can we assume different people sharing the same evaluation parameter? We may interpret it in the following two ways, but there are still flaws: (1) the "customer" can be the public, i.e. their performance is quite stable in general; or (2) the customer can be the same one over the whole time series. However, the former explanation cannot match the assumption that we just sell one product at each time, and the latter one would definitely undermine the independent assumption of the noise: people would do "human learning" and might gradually reduce their noise of making decisions. To this extent, it is closer to the fact if we assume noises as martingales. This assumption has been stated in Qiang and Bayati [2016].

Linear Valuations on Features There exist many products whose prices are not linearly dependent on features. One famous instance is a diamond: a kilogram of diamond powder is very cheap because it can be produced artificially, but a single 5-carat (or 1

gram) diamond might cost more than \$100,000. This is because of an intrinsic non-linear property of diamond: large ones are rare and cannot be (at least easily) compound from smaller ones. Another example lies in electricity pricing [Joskow and Wolfram, 2012], where the more you consume, the higher unit price you suffer. On the contrary, commodities tend to be cheaper than retail prices. These are both consequences of marginal costs: a large volume consuming of electricity may cause extra maintenance and increase the cost, and a large amount of purchasing would release the storage and thus reduce their costs. In a word, our problem setting might not be suitable for those large-enough features, and thus an upper bound of $x^\top\theta$ becomes a necessity.

2.7.4 Algorithm Design

Probit and Logistic Regressions A probit/logit model is described as follows: a Boolean random variable Y satisfies the following probabilistic distribution: $\mathbb{P}[Y = 1|X] = F(X^\top\beta)$, where $X \in \mathbb{R}$ is a random vector, $\beta \in \mathbb{R}$ is a parameter, and F is the cumulative distribution function (CDF) of a (standard) Gaussian/logistic distribution. In our problem, we may treat $\mathbb{1}_t$ as Y, $[{x_t}^\top, p_t]^\top$ as X and $[{\theta^*}^\top, -1]^\top$ as β, which exactly fits this model if we assume the noise as Gaussian or logistic. Therefore, $\hat{\theta}_k = \arg\min_\theta \hat{L}_k(\theta)$ can be solved via the highly efficient implementation of generalized linear models, e.g., GLMnet, rather than resorting to generic tools for convex programming. As a heuristic, we could leverage the vast body of statistical work on probit or logit models and adopt a fully Bayesian approach that jointly estimates θ and hyper-parameters of F. This would make the algorithm more practical by eliminating the need to choose the hyper-parameters when running this algorithm.

Advantages of EMLP over ONSP. For the stochastic setting, we specifically propose EMLP even though ONSP also works. This is because EMLP only "switch" the pricing policy $\hat{\theta}$ for $\log T$ times. This makes it appealing in many applications (especially for brick-and-mortar sales) where the number of policy updates is a bottleneck. In fact, the iterations within one epoch can be carried out entirely in parallel.

***Ex Ante* v.s. *Ex Post* Regrets** In this chapter, we considered the *ex ante* regret $Reg_{ea} = \sum_{t=1}^{T} \max_\theta \mathbb{E}[p_t^\theta \cdot \mathbb{1}(p_t^\theta \le w_t)] - \mathbb{E}[p_t \cdot \mathbb{1}(p_t \le w_t)]$, where $p_t^\theta = J(x_t^\top \theta)$ is the greedy price with parameter θ and $w_t = x_t^\top \theta^* + N_t$ is the realized random valuation. The *ex post* definition of the cumulative regret, i.e., $Reg_{ep} = \max_\theta \sum_{t=1}^{T} p_t^\theta \mathbb{1}(p_t^\theta \le w_t) - p_t \mathbb{1}(p_t \le w_t)$ makes sense, too. Note that we can decompose $\mathbb{E}[Reg_{ep}] = Reg_{ea} + \mathbb{E}[\max_\theta \sum_{t=1}^{T} p_t^\theta \mathbb{1}(p_t^\theta \le w_t) - \sum_{t=1}^{T} p_t^{\theta^*} \mathbb{1}(p_t^{\theta^*} \le w_t)]$. While it might be the case that the second term is $\Omega(\sqrt{dT})$ as the reviewer pointed out, it is a constant independent of the algorithm. For this reason, we believe using Reg_{ea} is without loss of generality, and it reveals more nuanced performance differences of different algorithms.

For an *ex post dynamic* regret, i.e., $Reg_d = \sum_{t=1}^{T} w_t - p_t \cdot \mathbb{1}(p_t \le w_t)$, it is argued in Cohen et al. [2020] that any policy must suffer an expected regret of $\Omega(T)$ (even if θ^* is known). We may also present a good example lies in $N_t \sim \mathcal{N}(0,1), x_t^\top \theta^* = \sqrt{\frac{\pi}{2}}$ where the optimal price is $\sqrt{\frac{\pi}{2}}$ as well but the probability of acceptance is only $1/2$, and this leads to a constant *per-step* regret of $\frac{1}{2}\sqrt{\frac{\pi}{2}}$.

2.7.5 Potential Extensions

Agnostic Dynamic Pricing: Explorations versus Exploitation At the moment, the proposed algorithm relies on the assumption of a linear valuation function. It will be

interesting to investigate the settings of model-misspecified cases and the full agnostic settings. The key would be to exploit the structural feedback in model-free policy-evaluation methods such as importance sampling. The main reason why we do not explore lies in the noisy model: essentially we are implicitly exploring a higher (permitted) price using the naturally occurring noise in the data. In comparison, there is another problem setting named "adversarial irrationality" where some of the customers will valuate the product adaptively and adversarially[10]. Existing work Krishnamurthy et al. [2021] adopts this setting and shows a linear regret dependence on the number of irrational customers, but they consider a different loss function (See Related Works Section).

Ethic Issues A field of study lies in "personalized dynamic pricing" [Aydin and Ziya, 2009, Chen and Gallego, 2021], where a firm makes use of information of individual customers and sets a unique price for each of them. This has been frequently applied in airline pricing [Krämer et al., 2018]. However, this causes first-order pricing discrimination. Even though this "discrimination" is not necessarily immoral, it must be embarrassing if we are witted proposing the same product with different prices towards different customers. For example, if we know the coming customer is rich enough and is not as sensitive towards a price (e.g., he/she has a variance larger than other customers), then we are probably raising the price without being too risky. Or if the customer is used to purchase goods from ours, then he or she might have a higher expectation on our products (e.g., he/she has a $\theta = a\theta^*, a > 1$), and we might take advantage and propose a higher price than others. These cases would not happen in an auction-based situation (such as a live sale), but might frequently happen in a more secret place such as a customized travel plan.

[10]An adaptive adversary may take actions adversarially in respond to the environmental changes. In comparison, what we allow for the "adversarial features" is actually chosen by an oblivious adversary before the interactions start.

2.8 Conclusion

In this chapter, we studied the problem of online feature-based dynamic pricing with a noisy linear valuation in both stochastic and adversarial settings. We proposed a max-likelihood-estimate-based algorithm (EMLP) for stochastic features and an online-Newton-step-based algorithm (ONSP) for adversarial features. Both of them enjoy a regret guarantee of $O(d \log T)$, which also attains the information-theoretic limit up to a constant factor. Compared with existing works, EMLP gets rid of strong assumptions on the distribution of the feature vectors in the stochastic setting, and ONSP improves the regret bound exponentially from $O(T^{2/3})$ to $O(\log T)$ in the adversarial setting. We also showed that knowing the noise distribution (or the demand curve) is required to obtain logarithmic regret, where we prove a lower bound of $\Omega(\sqrt{T})$ on the regret for the case when the noise is knowingly Gaussian but with an unknown σ. In addition, we conducted numerical experiments to empirically validate the scaling of our algorithms. Finally, we discussed the regret dependence on the noise variance, and proposed a subtle open problem for further study.

2.9 Proof Details

2.9.1 Proof of Lemma 2.3.5

Proof. Since $p^* = \operatorname{argmax} g(p, u)$, we have:

$$\begin{aligned}\frac{\partial g(p,u)}{\partial p}|_{p=p^*} = 0 \Leftrightarrow & 1 - F(p^* - u) - p^* \cdot f(p^* - u) = 0 \\ \Leftrightarrow & \frac{1 - F(p^* - u)}{f(p^* - u)} - (p^* - u) = u\end{aligned}$$

Define $\varphi(\omega) = \frac{1-F(\omega)}{f(\omega)} - \omega$, and we take derivatives:

$$\varphi'(\omega) = \frac{-f^2(\omega) - (1-F(\omega))f'(w)}{f^2(w)} - 1 = \frac{\mathrm{d}^2 \log(1-F(\omega))}{\mathrm{d}\omega^2} \cdot \frac{(1-F(\omega))^2}{(f(\omega))^2} - 1 < -1,$$

where the last equality comes from the strict log-concavity of $(1 - F(\omega))$. Therefore, $\varphi(\omega)$ is decreasing and $\varphi(+\infty) = -\infty$. Also, notice $\varphi(-\infty) = +\infty$, we know that for any $u \in \mathbb{R}$, there exists an ω such that $\varphi(\omega) = u$. For $u \geq 0$, we know that $g(p, u) \geq 0$ for $p \geq 0$ and $g(p, u) < 0$ for $p < 0$. Therefore, $p^* \geq 0$ if $u \geq 0$. ∎

2.9.2 Proofs in Section 2.5.1

Proof of Lemma 2.5.3

Proof. We denote $\varphi(\omega) = \frac{1-F(\omega)}{f(\omega)} - \omega$ as in Section 2.9.1. According to Eq. (2.5), we have:

$$\begin{aligned} \frac{\partial g(p,u)}{\partial p}|_{p=J(u)} = 0 \quad &\Rightarrow \varphi(J(u) - u) = u \\ \Rightarrow J(u) = u + \varphi^{-1}(u) \quad &\Rightarrow J'(u) = 1 + \frac{1}{\varphi'(\varphi^{-1}(u))}. \end{aligned} \tag{2.21}$$

The last line of Eq. (2.21) is due to the Implicit Function Derivatives Principle. From the result in Section 2.9.1, we know that $\varphi'(\omega) < -1, \forall \omega \in \mathbb{R}$. Therefore, we have $J'(u) \in (0, 1), u \in \mathbb{R}$, and hence $0 \geq J(u) < u + J(0)$ for $u \geq 0$. Since $u \in [0, B]$, we may assume that $p \in [0, B + J(0)]$ without losing generality. In the following part, we will frequently use this range.

Denote $u := x_t^\top \theta, u^* = x_t^\top \theta^*$. According to Eq. (2.7), we know that:

$$\begin{aligned}
Reg_t(\theta) &= g(J(u^*), u^*) - g(J(u), u^*) \\
&= -\frac{\partial g(p, u^*)}{\partial p}|_{p=J(u^*)}(J(u^* - J(u))) + \frac{1}{2}\left(-\frac{\partial^2 g(p, u^*)}{\partial p^2}|_{p=\tilde{p}}\right)(J(u^*) - J(u))^2 \\
&\leq 0 + \frac{1}{2}\max_{\tilde{p}\in[0, B+J(0)]}\left(-\frac{\partial^2 g(p, u^*)}{\partial p^2}|_{p=\tilde{p}}\right)\cdot(J(u^*) - J(u))^2 \\
&= \frac{1}{2}\max_{\tilde{p}\in[0, B+J(0)]}(2f(\tilde{p} - u^*) + \tilde{p}\cdot f'(\tilde{p} - u^*))\cdot(J(u^*) - J(u))^2 \\
&\leq \frac{1}{2}(2B_f + (B + J(0))\cdot B_{f'})(J(u^*) - J(u))^2 \\
&\leq \frac{1}{2}(2B_f + (B + J(0))\cdot B_{f'})(u^* - u)^2 \\
&= \frac{1}{2}(2B_f + (B + J(0))\cdot B_{f'})(\theta^* - \theta)^\top x_t x_t^\top (\theta^* - \theta).
\end{aligned}$$

Here the first line is from the definition of g and $Reg(\theta)$, the second line is due to Taylor's Expansion, the third line is from the fact that $J(u^*)$ maximizes $g(p, u^*)$ with respect to p, the fourth line is by calculus, the fifth line is from the assumption that $0 < f(\omega) \leq B_f, |f'(\omega)| \leq B_{f'}$ and $p \in [0, B + J(0)]$, the sixth line is because of $J'(u) \in (0, 1), \forall u \in \mathbb{R}$, and the seventh line is from the definition of u and u^*. ■

Proof of Lemma 2.5.5

Proof. We take derivatives of $l_t(\theta)$, and we get:

$$
\begin{aligned}
l_t(\theta) =& \mathbb{1}_t \left(-\log(1-F(p_t - x_t^\top \theta))\right) + (1-\mathbb{1}_t)\left(-\log(F(p_t - x_t^\top \theta))\right) \\
\nabla l_t(\theta) =& \mathbb{1}_t \left(-\frac{f(p_t - x_t^\top \theta)}{1-F(p_t - x_t^\top \theta)}\right) \cdot x_t + (1-\mathbb{1}_t)\left(\frac{f(p_t - x_t^\top \theta)}{F(p_t - x_t^\top \theta)}\right) \cdot x_t \\
\nabla^2 l_t(\theta) =& \mathbb{1}_t \cdot \frac{f(p_t - x_t^\top \theta)^2 + f'(p_t - x_t^\top \theta)\cdot(1-F(p_t - x_t^\top \theta))}{(1-F(p_t - x_t^\top \theta))^2} \cdot x_t x_t^\top \\
&+ (1-\mathbb{1}_t) \cdot \frac{f(p_t - x_t^\top \theta)^2 - f'(p_t - x_t^\top \theta) F(p_t - x_t^\top \theta)}{F(p_t - x_t^\top \theta)^2} \cdot x_t x_t^\top \\
=& \mathbb{1}_t \cdot \frac{-\mathrm{d}^2 \log(1-F(\omega))}{\mathrm{d}\omega^2}|_{\omega = p_t - x_t^\top \theta} \cdot x_t x_t^\top + (1-\mathbb{1}_t)\frac{-\mathrm{d}^2 \log(F(\omega))}{\mathrm{d}\omega^2}|_{\omega = p_t - x_t^\top \theta} \cdot x_t x_t^\top \\
\succeq & \inf_{\omega \in [-B, B+J(0)]} \min\left\{\frac{\mathrm{d}^2 \log(1-F(\omega))}{\mathrm{d}\omega^2}, \frac{\mathrm{d}^2 \log(F(\omega))}{\mathrm{d}\omega^2}\right\} \\
=& C_{\text{down}} x_t x_t^\top,
\end{aligned}
\tag{2.22}
$$

which directly proves the first inequality. For the second inequality, just notice that

$$
\begin{aligned}
\nabla l_t(\theta) \nabla l_t(\theta)^\top =& \mathbb{1}_t \left(\frac{f(p_t - x_t^\top \theta)}{1-F(p_t - x_t^\top \theta)}\right)^2 x_t x_t^\top + (1-\mathbb{1}_t)\left(\frac{f(p_t - x_t^\top \theta)}{F(p_t - x_t^\top \theta)}\right)^2 x_t x_t^\top \\
\preceq & \sup_{\omega \in [-B, B+J(0)]} \max\{\left(\frac{f(\omega)}{F(\omega)}\right)^2, \left(\frac{f(\omega)}{1-F(\omega)}\right)^2\} x_t x_t^\top \\
=& C_{\exp} x_t x_t^\top.
\end{aligned}
\tag{2.23}
$$

The only thing to point out is that $\frac{f(\omega)}{F(\omega)}$ and $\frac{f(\omega)}{1-F(\omega)}$ are all continuous for $\omega \in [-B, B+J(0)]$, as $F(\omega)$ is strictly increasing and thus $0 < F(\omega) < 1, \omega \in \mathbb{R}$. ∎

Proof of Lemma 2.5.7

Proof. In the following part, we consider a situation that an epoch of $n \geq 2$ rounds of pricing is conducted, generating $l_j(\theta)$ as negative likelihood functions, $j = 1, 2, \ldots, n$. Define a "**leave-one-out**"negative log-likelihood function

$$
\tilde{L}_i(\theta) = \frac{1}{n} \sum_{j=1, j\neq i}^{n} l_j(\theta),
$$

and let

$$\tilde{\theta}_i := \arg\min_{\theta} \tilde{L}_i(\theta).$$

Based on this definition, we know that $\tilde{\theta}_i$ is independent to $l_i(\theta)$ given historical data, and that $\tilde{\theta}_i$ are identically distributed for all $i = 1, 2, 3, \ldots, n$.

In the following part, we will firstly propose and proof the following inequality:

$$\frac{1}{n}\sum_{i=1}^{n}(l_i(\tilde{\theta}_i) - l_i(\hat{\theta})) \leq \frac{C_{\exp}}{C_{down}}\frac{d}{n} = O(\frac{d}{n}), \tag{2.24}$$

where $\hat{\theta}$ is the short-hand notation of $\hat{\theta}_k$ as we do not specify the epoch k in this part. We now cite a lemma from Koren and Levy [2015]:

Lemma 2.9.1. *Let g_1, g_2 be 2 convex function defined over a closed and convex domain $\mathcal{K} \subseteq \mathbb{R}^d$, and let $x_1 = \arg\min_{x\in\mathcal{K}} g_1(x)$ and $x_2 = \arg\min_{x\in\mathcal{K}} g_2(x)$. Assume g_2 is locally δ-strongly-convex at x_1 with respect to a norm $||\cdot||$. Then, for $h = g_2 - g_1$ we have*

$$||x_2 - x_1|| \leq \frac{2}{\delta}||\nabla h(x_1)||_*.$$

Here $||\cdot||_*$ denotes a dual norm.

The following is a proof of this lemma.

Proof. (of Lemma 2.9.1) According to convexity of g_2, we have:

$$g_2(x_1) \geq g_2(x_2) + \nabla g_2(x_2)^\top (x_1 - x_2). \tag{2.25}$$

According to strong convexity of g_2 at x_1, we have:

$$g_2(x_2) \geq g_2(x_1) + \nabla g_2(x_1)^\top (x_2 - x_1) + \frac{\delta}{2}||x_2 - x_1||^2. \tag{2.26}$$

Add Eq. (2.25) and Eq. (2.26), and we have:

$$\begin{aligned}
& g_2(x_1) + g_2(x_2) \geq g_2(x_2) + g_2(x_1) + (\nabla g_2(x_1) - \nabla g_2(x_2))^\top (x_2 - x_1) + \frac{\delta}{2}||x_2 - x_1||^2 \\
\Leftrightarrow \quad & (\nabla g_2(x_1) - \nabla g_2(x_2))^\top (x_1 - x_2) \geq \frac{\delta}{2}||x_1 - x_2||^2 \\
\Leftrightarrow \quad & (\nabla g_1(x_1) + \nabla h(x_1) - \nabla g_2(x_2))^\top (x_1 - x_2) \geq \frac{\delta}{2}||x_1 - x_2||^2 \\
\Leftrightarrow \quad & \nabla h(x_1)^\top (x_1 - x_2) \geq \frac{\delta}{2}||x_1 - x_2||^2 \\
\Rightarrow \quad & ||\nabla h(x_1)||_* ||x_1 - x_2|| \geq \frac{\delta}{2}||x_1 - x_2||^2 \\
\Rightarrow \quad & ||\nabla h(x_1)||_* \geq \frac{\delta}{2}||x_1 - x_2||.
\end{aligned} \tag{2.27}$$

The first step is trivial. The second step is a sequence of $g_2 = g_1 + h$. The third step is derived by the following 2 first-order optimality conditions: $\nabla g_1(x_1)^\top (x_1 - x_2) \leq 0$, and $\nabla g_2(x_2)^\top (x_2 - x_1) \leq 0$. The fourth step is derived from Holder's Inequality:

$$||\nabla h(x_1)||_* ||x_1 - x_2|| \geq \nabla h(x_1)^\top (x_1 - x_2).$$

Therefore, the lemma holds. ■

In the following part, we will set up a strongly convex function of g_2. Denote $H = \sum_{t=1}^n x_t x_t^\top$. From Lemma 2.5.5, we know that

$$\nabla^2 \hat{L}(\theta) \succeq C_{down} \frac{1}{n} H.$$

Here $\hat{L}(\theta)$ is the short-hand notation of $\hat{L}_k(\theta)$ as we do not specify k in this part. Since we do not know if H is invertible, i.e. if a norm can be induced by H, we cannot let $g_2(\theta) = \hat{L}(\theta)$. Instead, we change the variable as follows:

We first apply singular value decomposition to H, i.e. $H = U \Sigma U^\top$, where $U \in \mathbb{R}^{d \times r}, U^\top U = I_r, \Sigma = diag\{\lambda_1, \lambda_2, \ldots, \lambda_r\} \succ 0$. After that, we introduce a new variable $\eta := U^\top \theta$. Therefore, we have $\theta = U\eta + V\epsilon$, where $V \in \mathbb{R}^{d \times (d-r)}, V^\top V = I_{d-r}, V^\top U = 0$

is the standard orthogonal bases of the null space of U, and $\epsilon \in \mathbb{R}^{(d-r)}$. Similarly, we define $\tilde{\eta}_i = U^\top \tilde{\theta}_i$ and $\hat{\eta} = U^\top \hat{\theta}$. According to these, we define the following functions:

$$\begin{aligned} f_i(\eta) &:= l_i(\theta) = l_i(U\eta + V\epsilon) \\ \tilde{F}_i(\eta) &:= \tilde{L}_i(\theta) = \tilde{L}_i(U\eta + V\epsilon) \\ \hat{F}(\eta) &:= \hat{L}(\theta) = \hat{L}(U\eta + V\epsilon). \end{aligned} \tag{2.28}$$

Now we prove that $\hat{F}(\eta)$ is locally-strongly-convex. Similar to the proof of Lemma 2.5.5, we have:

$$\begin{aligned} \nabla^2 \hat{F}(\eta) =& \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(\eta) \\ =& \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial (x_i^\top \theta)^2} (\frac{\partial x_i^\top (U\eta + V\epsilon)}{\partial \eta})(\frac{\partial x_i^\top (U\eta + V\epsilon)}{\partial \eta})^\top \\ =& \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial (x_i^\top \theta)^2} (U^\top x_i)(U^\top x_i)^\top \\ \succeq& \frac{1}{n}\sum_{i=1}^{n} C_{down} U^\top x_i x_i^\top U \\ =& \frac{1}{n} C_{down} U^\top (\sum_{i=1}^{n} x_i x_i^\top) U^\top \\ =& \frac{1}{n} C_{down} U^\top H U = \frac{1}{n} C_{down} U^\top U \Sigma U^\top U = \frac{1}{n} C_{down} \Sigma \succ 0 \end{aligned} \tag{2.29}$$

That is to say, $\hat{F}(\eta)$ is locally $\frac{C_{down}}{n}$-strongly convex w.r.t Σ at η. Similarly, we can verify that $\tilde{F}_i(\eta)$ is convex (not necessarily strongly convex). Therefore, according to Lemma 2.9.1, let $g_1(\eta) = \tilde{F}_i(\eta), g_2(\eta) = \hat{F}(\eta)$, and then $x_1 = \tilde{\eta}_i = U^\top \tilde{\theta}_i$, $x_2 = \hat{\eta} = U^\top \hat{\theta}$. Therefore, we have:

$$||\hat{\eta} - \tilde{\eta}_i||_\Sigma \leq \frac{1}{C_{down}} ||\nabla f_i(\tilde{\eta}_i)||_\Sigma^*. \tag{2.30}$$

Now let us go back to the proof of the theorem:

$$\begin{aligned} l_i(\tilde{\theta}_i) - l_i(\hat{\theta}) =& f_i(\tilde{\eta}_i) - f_i(\hat{\eta}) \underset{\text{convexity}}{\uparrow\leq} \nabla f_i(\tilde{\eta}_i)^\top (\tilde{\eta}_i - \hat{\eta}) \underset{\text{Holder inequality}}{\uparrow\leq} ||\nabla f_i(\tilde{\eta}_i)||_\Sigma^* ||\tilde{\eta}_i - \hat{\eta}||_\Sigma \\ \underset{\text{Lemma 2.9.1}}{\uparrow\leq}& \frac{1}{C_{down}} (||\nabla f_i(\tilde{\eta}_i)||_\Sigma^*)^2. \end{aligned} \tag{2.31}$$

Given that, we have

$$\begin{aligned}\sum_{i=1}^{n} l_i(\tilde{\theta}_i) - l_i(\hat{\theta}) &\leq \frac{1}{C_{down}} \sum_{i=1}^{n} ||\nabla f_i(\tilde{\eta}_i)||_{\Sigma}^{*})^2 \\ &\leq \frac{1}{C_{down}} \sum_{i=1}^{n} (\frac{p}{\Phi})_{\max}^2 x_i^\top U \Sigma^{-1} U^\top x_i \\ &= \frac{C_{\exp}}{C_{down}} tr(U \Sigma^{-1} U^\top \sum_{i=1}^{n} x_i x_i^\top) \\ &= \frac{C_{\exp}}{C_{down}} tr(U \Sigma^{-1} U^\top H) \\ &= \frac{C_{\exp}}{C_{down}} tr(I_r) = \frac{C_{\exp}}{C_{down}} r \leq \frac{C_{\exp}}{C_{down}} d.\end{aligned} \tag{2.32}$$

Thus Eq. (2.24) is proved. After that, we have:

$$\begin{aligned}\mathbb{E}_h[L(\tilde{\theta}_n)] - L(\theta^*) =& \mathbb{E}_h[L(\tilde{\theta}_n)] - \mathbb{E}_h[\hat{L}(\theta^*)] \\ \leq& \mathbb{E}_h[L(\tilde{\theta}_n)] - \mathbb{E}_h[\hat{L}(\hat{\theta})] \\ =& \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_h[l_i(\tilde{\theta}_i) - l_i(\hat{\theta})] \leq \frac{C_{\exp}}{C_{down}} \cdot \frac{d}{n}\end{aligned}$$

Thus we has proved that $\mathbb{E}_h[L(\tilde{\theta}_n)] - L(\theta^*) \leq \frac{C_{\exp}}{C_{down}} \cdot \frac{d}{n}$. Notice that $\tilde{\theta}_n$ is generated by optimizing the leave-one-out likelihood function $\tilde{L}_n(\theta) = \sum_{j=1}^{n-1} l_j(\theta)$, which does not contain $l_n(\theta)$, and that the expected likelihood function $L(\theta)$ does not depend on any specific result occurring in this round. That is to say, every term of this inequality is not related to the last round $(x_n, p_n, \mathbb{1}_n)$ at all. In other words, this inequality is still valid if we only conduct this epoch from round 1 to $(n-1)$.

Now let $n = \tau + 1$, and then we know that $\tilde{\theta}_{\tau+1} = \hat{\theta}$. Therefore, the theorem holds. ∎

2.9.3 Proof of Lower bound in Section 2.5.3

Proof. We assume a fixed u^* such that $x^\top \theta^* = u^*, \forall x \in \mathbb{D}$. In other words, we are considering a non-context setting. Therefore, we can define a policy as $\Psi : \{0,1\}^t \rightarrow$

$\mathbb{R}^+, t = 1, 2, \ldots$ that does not observe x_t at all. Before the proof begins, we firstly define a few notations: We denote $\Phi_\sigma(\omega)$ and $\varphi_\sigma(\omega)$ as the CDF and PDF of Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and the corresponding $J_\sigma(u) = \arg\max_p p(1-\Phi_\sigma(p-u))$ as the pricing function.

Since we have proved that $J'(u) \in (0, 1)$ for $u \in \mathbb{R}$ in Section 2.9.2, we have the following lemma:

Lemma 2.9.2. *$u - J_\sigma(u)$ monotonically increases as $u \in (0, +\infty), \forall \sigma > 0$. Also, we know that $J_\sigma(0) > 0, \forall \sigma > 0$.*

Now consider the following cases: $\sigma_1 = 1, \sigma_2 = 1 - f(T)$, where $\lim_{T\to\infty} f(T) = 0, f'(T) < 0, 0 < f(T) < \frac{1}{2}$. We will later determine the explicit form of $f(T)$.

Suppose u^* satisfies $J_{\sigma_1}(u^*) = u^*$. Solve it and get $u^* = \sqrt{\frac{\pi}{2}}$. Therefore, we have $u \in (0, u^*) \Leftrightarrow J_1(u) > u$, and $u \in (u^*, +\infty) \Leftrightarrow J_1(u) < u$. As a result, we have the following lemma.

Lemma 2.9.3. *For any $\sigma \in (\frac{1}{2}, 1)$, we have $J_\sigma(u^*) \in (0, u^*)$.*

Proof. We have:

$$\begin{aligned} J_\sigma(u) =& \arg\max_p p\Phi_\sigma(u-p) = \arg\max_p p\Phi_1(\frac{u-p}{\sigma}) = \arg\max_{\omega=\frac{p}{\sigma}} \sigma\omega\Phi_1(\frac{u}{\sigma} - \omega) \\ =& \sigma \arg\max_\omega \Phi_1(\frac{u}{\sigma} - \omega) = \sigma J_1(\frac{u}{\sigma}). \end{aligned}$$

When $\sigma \in (\frac{1}{2}, 1)$, we know $\frac{u^*}{\sigma} > u^*$. Since $J_1(u^*) = u^*$ and that $u \in (u^*, +\infty) \Leftrightarrow J_1(u) < u$, we have $\frac{u^*}{\sigma} > J_1(\frac{u^*}{\sigma})$. Hence $u^* > \sigma J_1(\frac{u^*}{\sigma}) = J_\sigma(u^*)$.

■

Therefore, without losing generality, we assume that for the problem parameterized by σ_2, the price $p \in (0, u^*)$. To be specific, suppose $p^*(\sigma) = J_\sigma(u^*)$. Define $\Psi_{t+1} : [0,1]^t \to (0, u^*)$ as any policy that proposes a price at time $t+1$. Define $\Psi = \{\Psi_1, \Psi_2, \ldots, \Psi_{T-1}, \Psi_T\}$.

Define the sequence of price as $P = \{p_1, p_2, \ldots, p_{T-1}, p_T\}$, and the sequence of decisions as $\mathbb{1} = \{\mathbb{1}_1, \mathbb{1}_2, \ldots, \mathbb{1}_{T-1}, \mathbb{1}_T\}$. Denote $P^t = \{p_1, p_2, \ldots, p_t, \}$.

Define the probability (also the likelihood if we change u^* to other parameter u):

$$Q_T^{P,\sigma}(\mathbb{1}) = \prod_{t=1}^{T} \Phi_\sigma(u^* - p_t)^{\mathbb{1}_t} \Phi_\sigma(p_t - u^*)^{1-\mathbb{1}_t}. \tag{2.33}$$

Define a random variable $Y_t \in \{0,1\}^t, Y_t \sim Q_t^{P^t,\sigma}$ and one possible assignment $y_t = \{\mathbb{1}_1, \mathbb{1}_2, \ldots, \mathbb{1}_{t-1}, \mathbb{1}_t\}$. For any price p and any parameter σ, define the expected reward function as $r(p, \sigma) := p\Phi_\sigma(u^* - p)$. Based on this, we can further define the expected regret $\text{Regret}(\sigma, T, \Psi)$:

$$\text{Regret}(\sigma, T, \Psi) = \mathbb{E}[\sum_{t=1}^{T} r(J_\sigma(u^*), \sigma) - r(\Psi_t(y_{t-1}), \sigma)] \tag{2.34}$$

Now we have the following properties:

Lemma 2.9.4. *We have the following properties:*

1. $r(p^*(\sigma), \sigma) - r(p, \sigma) \geq \frac{1}{60}(p^*(\sigma) - p)^2$;
2. $|p^*(\sigma) - u^*| \geq \frac{2}{5}|1 - \sigma|$;
3. $|\Phi_\sigma(u^* - p) - \Phi_1(u^* - p)| \leq |u^* - p| \cdot |\sigma - 1|$.

Proof. 1. We have:

$$\frac{\partial r(p,\sigma)}{\partial p}|_{p=p^*(\sigma)} = 0 \frac{\partial^2 r(p,\sigma)}{\partial p^2} \quad = \frac{1}{\sigma^2}(p^2 - u^* p - 2\sigma^2)\varphi_\sigma(u^* - p)$$

Since $p \in (0, u^*)$, we have $(p^2 - u^* p - 2\sigma^2) < -2\sigma^2$. Also, since $\sigma \in (1/2, 1)$, we have $\varphi_\sigma(u^* - p) > \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(u^*)^2}{2\cdot(1/2)^2}} = \frac{1}{\sqrt{2\pi}e^\pi} > 0.017$. Therefore, we have

$$\frac{\partial^2 r(p,\sigma)}{\partial p^2} < -2 * 0.017 < -\frac{1}{30}$$

As a result, we have:

$$\begin{aligned} r(p^*(\sigma),\sigma) - r(p,\sigma) &= -(p^*(\sigma) - p)\frac{\partial r(p,\sigma)}{\partial p}|_{p=p^*(\sigma)} - \frac{1}{2}(p^*(\sigma) - p)^2 \frac{\partial^2 r(p,\sigma)}{\partial p^2}|_{p=\tilde{p}} \\ &= 0 - \frac{1}{2}(p^*(\sigma) - p)^2 \frac{\partial^2 r(p,\sigma)}{\partial p^2}|_{p=\tilde{p}} \\ &\geq \frac{1}{2} \cdot \frac{1}{30}(p^*(\sigma) - p)^2. \end{aligned} \tag{2.35}$$

2. According to the proof of Lemma 2.9.2, we know that:

$$p^*(\sigma) = \sigma J_1(\frac{u^*}{\sigma})$$

For $u \in (u^*, +\infty), J_1(u) < u$. According to Lemma 2.9.2, we have:

$$J_1'(u) = 1 + \frac{1}{J_1(u)(J_1(u) - u) - 2} > 1 + \frac{1}{0 - 2} = \frac{1}{2}.$$

Also, for $u \in (u^*, \frac{u^*}{\sigma})$, we have:

$$J_1'(u) = 1 - \frac{1}{2 + J_1(u)(u - J_1(u))} \underset{\substack{\uparrow \\ 0<J_1(u)<u}}{\leq} 1 - \frac{1}{2 + u(u-0)} \underset{\substack{\uparrow \\ u<\frac{u^*}{\sigma}<\frac{u^*}{2}}}{\leq} 1 - \frac{1}{2 + (\frac{u^*}{2})^2} < \frac{3}{5}.$$

Therefore, we have:

$$\begin{aligned} J_1(u^*) - J_\sigma(u^*) = J_1(u^*) - \sigma J_1(\frac{u^*}{\sigma}) &= J_1(u^*)(1-\sigma) - \sigma(J_1(\frac{u^*}{\sigma}) - J_1(u^*)) \\ &> u^*(1-\sigma) - \sigma \cdot \frac{3}{5}(\frac{u^*}{\sigma} - u^*) > \frac{2}{5}(1-\sigma). \end{aligned}$$

3. This is because:

$$\begin{aligned}|\Phi_\sigma(u^*-p)-\Phi_1(u^*-p)| &= |\Phi_\sigma(u^*-p)-\Phi_\sigma(\sigma(u^*-p))| \\ &\le \max|\varphi_\sigma|\cdot|(u^*-p)-\sigma(u^*-p)| \\ &\le (1-\sigma)|u^*-p|.\end{aligned}$$

■

In the following part, we will propose two theorems, which balance the cost of learning and that of uncertainty. This part is mostly similar to [BR12] Section 3, but we adopt a different family of demand curves here.

Theorem 2.9.5 (Learning is costly)**.** *Let $\sigma \in (1/2, 1)$ and $p_t \in (0, u^*)$, and we have:*

$$\mathcal{K}(Q^{P,1}; Q^{P,\sigma}) < 9900(1-\sigma)^2 \mathrm{Regret}(1, T, \Psi). \tag{2.36}$$

Here $p_t = \Psi(y_{t-1}), t = 1, 2, \ldots, T$.

Proof. First of all, we cite the following lemma that would facilitate the proof.

Lemma 2.9.6 (Corollary 3.1 in Taneja and Kumar, 2004)**.** *Suppose B_1 and B_2 are distributions of Bernoulli random variables with parameters q_1 and q_2, respectively, with $q_1, q_2 \in (0, 1)$. Then,*

$$\mathcal{K}(B_1; B_2) \le \frac{(q_1-q_2)^2}{q_2(1-q_2)}.$$

According to the definition of KL-divergence, we have:

$$\mathcal{K}(Q_T^{P,1}; Q_T^{P,\sigma}) = \sum_{s=1}^{T} \mathcal{K}(Q_s^{P^s,1}; Q_s^{P^s,\sigma} | Y_{s-1}).$$

For each term of the RHS, we have:

$$
\begin{aligned}
& \mathcal{K}(Q_s^{P^s,1}, Q_s^{P^s,\sigma}|Y_{s-1}) \\
= & \sum_{y_s\in\{0,1\}^s} Q_s^{P^s,1}(y_s)\log\left(\frac{Q_s^{P^s,1}(\mathbb{1}_s|y_{s-1})}{Q_s^{P^s,\sigma}(\mathbb{1}_s|y_{s-1})}\right) \\
\underset{\text{split } y_s \text{ as } y_{s-1} \text{ and } ind_s}{=} & \sum_{y_{s-1}\in\{0,1\}^{s-1}} Q_{s-1}^{P^{s-1},1}(y_{s-1})\cdot\sum_{\mathbb{1}_s\in\{0,1\}} Q_s^{P^s,1}(\mathbb{1}_s|y_{s-1})\log\left(\frac{Q_s^{P^s,1}(\mathbb{1}_s|y_{s-1})}{Q_s^{P^s,\sigma}(\mathbb{1}_s|y_{s-1})}\right) \\
= & \sum_{y_{s-1}\in\{0,1\}^{s-1}} Q_{s-1}^{P^{s-1},1}(y_{s-1})\mathcal{K}\left(Q_s^{P^s,1}(\cdot|y_{s-1}), Q_s^{P^s,\sigma}(\cdot|y_{s-1})\right) \\
\underset{\text{Lemma 2.9.6}}{\leq} & \sum_{y_{s-1}\in\{0,1\}^{s-1}} Q_{s-1}^{P^{s-1},1}(y_{s-1})\frac{(\Phi_1(u^*-p_s)-\Phi_\sigma(u^*-p_s))^2}{\Phi_\sigma(u^*-p_s)(1-\Phi_\sigma(u^*-p_s))} \\
= & \frac{1}{\Phi_\sigma(u^*-p_s)(1-\Phi_\sigma(u^*-p_s))}\sum_{y_{s-1}\in\{0,1\}^{s-1}} Q_{s-1}^{P^{s-1},1}(y_{s-1})(\Phi_1(u^*-p_s)-\Phi_\sigma(u^*-p_s))^2 \\
\underset{(**)}{\leq} & 165\cdot\sum_{y_{s-1}\in\{0,1\}^{s-1}} Q_{s-1}^{P^{s-1},1}(y_{s-1})(\Phi_1(u^*-p_s)-\Phi_\sigma(u^*-p_s))^2 \\
\underset{\text{Lemma 2.9.4 Property 3}}{\leq} & 165\cdot\sum_{y_{s-1}\in\{0,1\}^{s-1}} Q_{s-1}^{P^{s-1},1}(y_{s-1})(u^*-p_s)^2(1-\sigma)^2 \\
= & 165(1-\sigma)^2\mathbb{E}_{Y_{s-1}}[(u^*-p_s)^2].
\end{aligned}
$$

Here inequality (**) above is proved as follows: since $p_s \in (0, u^*)$ as is assumed, we have:

$$\frac{1}{2} < \Phi_\sigma(u^*-p_s) < \Phi_\sigma(u^*) = \sigma\cdot\Phi_1(\frac{u^*}{\sigma}) \leq 1\cdot\Phi_1(\frac{\sqrt{\frac{\pi}{2}}}{\frac{1}{2}}) \leq 0.9939\ .$$

As a result, we have $\frac{1}{\Phi_\sigma(u^*-p_s)(1-\Phi_\sigma(u^*-p_s))} \leq \frac{1}{0.9939\times 0.0061} = 164.7988 \leq 165$. Therefore, by summing up all s, we have:

$$
\begin{aligned}
\mathcal{K}(Q_T^{P,1}; Q_T^{P,\sigma}) &= \sum_{s=1}^T \mathcal{K}(Q_s^{P^s,1}; Q_s^{P^s,\sigma}|Y_{s-1}) \\
&\leq 165(1-\sigma)^2\sum_{s=1}^T \mathbb{E}_{Y_{s-1}}[(u^*-p_s)^2] \\
&\underset{\text{Lemma 2.9.4 Property 1}}{\leq} 165\times 60\cdot(1-\sigma)^2\sum_{s=1}^T (r(u^*,1)-r(p_s,1)) \\
&\underset{\text{definition of regret and } p_s=\Psi(y_{s-1}).}{=} 9900(1-\sigma)^2\mathrm{Regret}(1,T,\Psi),
\end{aligned}
$$

which concludes the proof. ■

Theorem 2.9.7 (Uncertainty is costly)**.** *Let $\sigma \leq 1 - T^{-\frac{1}{4}}$, and we have:*

$$\text{Regret}(1, T, \Psi) + \text{Regret}(\sigma, T, \Psi) \geq \frac{1}{24000} \cdot \sqrt{T} \cdot e^{-\mathcal{K}(Q^{P,1}; Q^{P,\sigma})}. \tag{2.37}$$

Here $p_t = \Psi(y_{t-1}), t = 1, 2, \ldots, T$.

Proof. First of all, we cite a lemma that would facilitate our proof:

Lemma 2.9.8. *Let Q_0 and Q_1 be two probability distributions on a finite space $\mathcal{Y}$; with $Q_0(y), Q_1(y) > 0, \forall y \in \mathcal{Y}$. Then for any function $F : \mathcal{Y} \to \{0, 1\}$,*

$$Q_0\{F = 1\} + Q_1\{J = 0\} \geq \frac{1}{2} e^{-\mathcal{K}(Q_0; Q_1)},$$

where $\mathcal{K}(Q_0; Q_1)$ denotes the KL-divergence of Q_0 and Q_1.

Define two intervals of prices:

$$C_1 = \{p : |u^*| \leq \frac{1}{10T^{\frac{1}{4}}}\} \quad \text{and} \quad C_2 = \{p : |J_\sigma(u^*) - p| \leq \frac{1}{10T^{\frac{1}{4}}}\}$$

Note that C_1 and C_2 are disjoint, since $|u^* - J_\sigma(u^*)| \geq \frac{2}{5}|1 - \sigma| = \frac{2}{5T^{1/2}}$ according to Lemma 2.9.4 Property 2. Also, for $p \in (0, u^*) \backslash C_2$, the regret is large according to Lemma 2.9.4 Property 1, because:

$$r(p^*(\sigma), \sigma) - r(p, \sigma) \geq \frac{1}{60}(p - p^*(\sigma))^2 \geq \frac{1}{6000T^{\frac{1}{2}}}.$$

Then, we have:

$$
\begin{aligned}
&\operatorname{Regret}(1,T,\Psi)+\operatorname{Regret}(\sigma,T,\Psi)\\
\geq&\sum_{t=1}^{T-1}\mathbb{E}_1[r(u^*,1)-r(p_{t+1},1)]+\mathbb{E}_\sigma[r(J_\sigma(u^*),\sigma)-r(p_{t+1},\sigma)]\\
\geq&\frac{1}{6000\sqrt{T}}\sum_{t=1}^{T-1}\mathbb{P}_1[p_{t+1}\notin C_1]+\mathbb{P}_\sigma[p_{t+1}\notin\{C_2\}]\\
\underset{\text{Suppose } F_{t+1}=\mathbb{1}[p_{t+1}\in C_2]}{\geq}&\frac{1}{6000\sqrt{T}}\sum_{t=1}^{T-1}\mathbb{P}_1[F_{t+1}=1]+\mathbb{P}_\sigma[F_{t+1}=0]\\
\underset{\text{Lemma 2.9.8}}{\geq}&\frac{1}{6000\sqrt{T}}\sum_{t=1}^{T-1}\frac{1}{2}e^{-\mathcal{K}(Q_t^{P^t,1};Q_t^{P^t,\sigma})}\\
\underset{\mathcal{K}(Q_t^{P^t,1};Q_t^{P^t,\sigma})\text{ not decreasing}}{\geq}&\frac{1}{6000\sqrt{T}}\frac{T-1}{2}e^{-\mathcal{K}(Q_T^{P,1};Q_T^{P,\sigma})}\\
\geq&\frac{1}{24000}\sqrt{T}e^{-\mathcal{K}(Q_T^{P,1};Q_T^{P,\sigma})}.
\end{aligned}
$$

■

According to Theorem 2.9.5 and Theorem 2.9.7, we can then prove Theorem 2.5.11. Let $\sigma=1-T^{-\frac{1}{4}}$

$$
\begin{aligned}
&2\left(\operatorname{Regret}(1,T,\Psi)+\operatorname{Regret}(\sigma,T,\Psi)\right)\\
\geq&\operatorname{Regret}(1,T,\Psi)+\left(\operatorname{Regret}(1,T,\Psi)+\operatorname{Regret}(\sigma,T,\Psi)\right)\\
\geq&\frac{1}{9900T^{-1/2}}\mathcal{K}(Q^{P,1};Q^{P,\sigma})+\frac{1}{24000}\cdot\sqrt{T}\cdot e^{-\mathcal{K}(Q^{P,1};Q^{P,\sigma})}\\
\geq&\frac{1}{24000}\sqrt{T}\left(\mathcal{K}(Q^{P,1};Q^{P,\sigma})+e^{-\mathcal{K}(Q^{P,1};Q^{P,\sigma})}\right)\\
\underset{\text{The fact } e^x\geq x+1,\forall x\in\mathbb{R}}{\geq}&\frac{1}{24000}\sqrt{T}.
\end{aligned}
$$

Thus Theorem 2.5.11 is proved valid. ■

2.9.4 Proof of Lemma 2.7.1

Proof. We prove Lemma 2.7.1 sequentially:

1. We have:

$$\begin{aligned}\lambda'(\omega) =& \frac{\varphi_1^2(-\omega) - p_1'(-\omega)\Phi_1(-\omega)}{\Phi_1(-\omega)^2} \\ =& \frac{\varphi_1^2(-\omega) - \omega\varphi_1(-\omega)\Phi_1(-\omega)}{\Phi_1(-\omega)^2} \\ =& \frac{\varphi_1(-\omega)\left(\varphi_1(-\omega) - \omega\Phi_1(-\omega)\right)}{\Phi_1(-\omega)^2}.\end{aligned} \tag{2.38}$$

Therefore, it is equivalent to prove that $\varphi_1(-\omega) - \omega\Phi_1(-\omega) > 0$.

Suppose $f(\omega) = \varphi_1(\omega) + \omega\Phi_1(\omega)$. We now take its derivatives as follows:

$$f'(\omega) = (-\omega)\varphi_1(\omega) + \Phi_1(\omega) + \omega \cdot \varphi_1(\omega) = \Phi_1(\omega) > 0 \tag{2.39}$$

Therefore, we know that $f(\omega)$ monotonically increases in $\mathbb{R}$. Additionally, since we have:

$$\begin{aligned}\lim_{\omega\to-\infty} f(\omega) =& \lim_{\omega\to-\infty} \varphi_1(\omega) + \lim_{\omega\to-\infty} \omega\Phi_1(\omega) = 0 + \lim_{\omega\to-\infty} \frac{1}{\sigma^2} \cdot \frac{\Phi_1(\omega)}{1/\omega} \\ =& \lim_{\omega\to-\infty} \cdot \frac{\varphi_1(\omega)}{-1/\omega^2} = \lim_{\omega\to-\infty} \cdot \left(-\frac{1}{\sqrt{2\pi}} \cdot \frac{\omega^2}{\exp\{\frac{\omega^2}{2}\}}\right) = 0\end{aligned} \tag{2.40}$$

Therefore, we know that $f(\omega) > 0$, $\forall\omega \in \mathbb{R}$, and as a result, $\lambda'(\omega) > 0$.

2. We have:

$$\lim_{\omega\to-\infty} \omega^k\lambda(\omega) = \lim_{\omega\to-\infty} \omega^k \frac{\varphi_1(-\omega)}{\Phi_1(-\omega)} = \frac{\lim_{\omega\to-\infty} \omega^k\varphi_1(-\omega)}{\lim_{\omega\to-\infty} \Phi_1(-\omega)} = \frac{0}{1} = 0. \tag{2.41}$$

3. It is sufficient to prove that

$$\lim_{\omega\to+\infty} \lambda(\omega) - \omega = 0.$$

Actually, we have:

$$\begin{aligned}\lim_{\omega\to+\infty} \lambda(\omega) - \omega =& \lim_{\omega\to+\infty} \frac{\varphi_1(-\omega) - \omega\Phi_1(-\omega)}{\Phi_1(-\omega)} = \lim_{\omega\to-\infty} \frac{\varphi_1(\omega) + \omega\Phi_1(\omega)}{\Phi_1(\omega)} \\ =& \underset{\uparrow \text{ L'Hospital's rule}}{\lim_{\omega\to-\infty}} \frac{(-\omega)\varphi_1(\omega) + \Phi_1(\omega) + \omega\varphi_1(\omega)}{\varphi_1(\omega)} = \lim_{\omega\to-\infty} \frac{\Phi_1(\omega)}{\varphi_1(\omega)} = 0\end{aligned} \tag{2.42}$$

4. This is because

$$\begin{aligned}\lim_{\omega\to+\infty}\omega(\lambda(\omega)-\omega) &= \lim_{\omega\to-\infty}\frac{-\omega\varphi_1(\omega)-\omega^2\Phi_1(\omega)}{\Phi_1(\omega)}\\ &\underset{\text{L'Hospital's rule}}{\uparrow}\lim_{\omega\to-\infty}\frac{-\varphi_1(\omega)-\omega(-\omega)\varphi_1(\omega)-\omega^2\varphi_1(\omega)-2\omega\cdot\Phi_1(\omega)}{\varphi_1(\omega)}\\ &= -1-2\lim_{\omega\to-\infty}\frac{\omega\Phi_1(\omega)}{\varphi_1(\omega)} = -1+2\lim_{\omega\to+\infty}\frac{1}{\frac{\lambda(\omega)}{\omega}} = -1+2=1.\end{aligned} \tag{2.43}$$

Thus the lemma holds.

■

Chapter 3

Towards Agnostic Feature-Based Dynamic Pricing

Prior research in feature-based dynamic pricing usually relies on assumptions of either *noiseless* linear valuation or *precisely-known* noise distribution, which limits the applicability of those algorithms in practice when these assumptions are hard to verify. In this chapter, we explore two more agnostic models to address this limitation:

a **Linear Policy** (LP) problem: We aim to compete with the best linear pricing policy while making no assumptions on the data. We show a $\tilde{\Theta}(d^{\frac{1}{3}}T^{\frac{2}{3}})$ minimax regret up to logarithmic factors.

b **Linear Valuation** (LV) problem: Customers' valuations are modeled as a linear function plus an unknown, assumption-free i.i.d. noise. We present an algorithm that achieves an $\tilde{O}(T^{\frac{3}{4}})$ regret. We also improve the existing lower bound from $\Omega(T^{\frac{3}{5}})$ to $\tilde{\Omega}(T^{\frac{2}{3}})$.

These results demonstrate that no-regret learning is possible for feature-based dynamic pricing under weak assumptions, but also reveal a disappointing fact that the seemingly richer *pricing feedback* is not significantly more useful than the *bandit feedback* in regret reduction.

3.1 Introduction

In a dynamic pricing process, a seller presents prices for the products and adjusts these prices according to customers' feedback (i.e., whether they decide to buy or not) to maximize the revenue. Existing works on the single-product pricing problem [Kleinberg and Leighton, 2003, Wang et al., 2021b] assume that customers make decisions only according to the comparisons between prices and their own (random) valuations, and the goal is to find out a best fixed price that maximizes the (expected) revenue. In general, the single-product pricing problem has been well studied under a variety of assumptions.

However, these methods are not applicable when there are thousands of highly differentiated products with no experience in selling them. This motivates the idea of "contextual pricing" [Cohen et al., 2020, Mao et al., 2018, Javanmard and Nazerzadeh, 2019, Liu et al., 2021], where each sale session is described by a context that also affects the valuation and pricing.

Contextual pricing. For $t = 1, 2, ..., T$:

1. A context $x_t \in \mathbb{R}^d$ is revealed that describes a sales session (product, customer and context).
2. The customer valuates the product as y_t using x_t.
3. The seller proposes a price $v_t > 0$ concurrently (according to x_t and historical sales records).
4. The transaction is successful if $v_t \leq y_t$, i.e., the seller gets a reward $r_t = v_t \cdot \mathbb{1}(v_t \leq y_t)$.

Here T is the time horizon known to the seller in advance[1], x_t's can be either stochastic (i.e., each x_t is independently and identically distributed) or adversarial (i.e., the sequence $\{x_t\}_{t=1}^T$ are arbitrarily chosen and fixed by nature before $t = 0$), and $\mathbb{1}_t := \mathbb{1}(v_t \le y_t)$ is an indicator that equals 1 if $v_t \le y_t$ and 0 otherwise. In this chapter, we consider two distinct problem setups that make use of the feature vector x_t.

(a) Linear Policy (**LP**): (x_t, y_t)'s are selected by nature (or an oblivious adversary) arbitrarily, and the learning goal is to compete with the optimal linear prices $v_t^* = x_t^\top \beta^*$ where β^* maximizes the cumulative reward in the hindsight.

(b) Linear Valuation (**LV**): assume *valuations* are linear + noise, i.e., $y_t = x_t^\top \theta^* + N_t$, where $\theta^* \in \mathbb{R}^d$ is a fixed vector and N_t is a market noise, drawn i.i.d. from a fixed *unknown* distribution $\mathbb{D}$. The learning goal is to compete with the *globally* optimal price $v_t^* = \text{argmax}_v \, v \cdot \Pr[v \le y_t | x_t]$ with no restrictions on the pricing policy.

These two problem setups — although quite similar at a glance — are intrinsically different. The LP problem makes no assumptions on the $x_t \to y_t$ mapping, i.e., agnostic learning. Customers' valuations are not necessarily linear (and can be deterministic/noisy/stochastic/adversarial), but the seller competes with the optimal policy in a constrained family. In contrast, the LV problem makes mild modeling assumptions about the distribution of y_t given x_t while keeping the policy class unrestricted. In other words, LP is modeling our strategy while LV is modeling the nature. We adopt *regret* as a metric of algorithmic performance: For the LP problem, we compare its (expected) reward with that of the optimal fixed β^* in hindsight (i.e., an *ex post* regret); For the LV problem, we compare its (expected) reward with the largest expected reward condition on θ^* and $\mathbb{D}$ (i.e., an *ex ante* regret). We will clarify the difference between LP and LV in Section 3.7

[1]Here we assume T known for simplicity of notations. In fact, if T is unknown, then we may apply a "doubling epoch" trick as Javanmard and Nazerzadeh [2019] and the regret bounds are the same.

Table 3.1: Summary of existing regret bounds and our results

Problem	**Linear Valuation (LV)**				**Linear Policy**	
Noise Assumption	No Noise	Known, Log-concave	Parametric	**Agnostic, Bounded**	**(LP)**	
Upper Bound	$O(d\log\log T)$	$O(d\log T)$	$\tilde{O}(d\sqrt{T})$	$\tilde{O}(T^{\frac{3}{4}}+d^{\frac{1}{2}}T^{\frac{5}{8}})$ **Our Work**	$\tilde{O}(d^{\frac{1}{3}}T^{\frac{2}{3}})$	**Our Work**
Lower Bound	$\Omega(d\log\log T)$	$\Omega(d\log T)$	$\Omega(d\sqrt{T})$	$\tilde{\Omega}(T^{\frac{2}{3}})$ **Our Work**	$\tilde{\Omega}(d^{\frac{1}{3}}T^{\frac{2}{3}})$	**Our Work**

with more details and examples. We emphasize that in both settings, the distributions of the valuation are unknown and non-parametric, and we are interested in designing no-regret algorithms and characterizing the complexity.

Summary of Results. Our contributions are threefold.

1. For the LP problem with adversarial x_t's, we present an algorithm "Linear-EXP4" that achieves $\tilde{O}(d^{\frac{1}{3}}T^{\frac{2}{3}})$ regret.

2. For the LV problem with adversarial x_t's, we present an algorithm "D2-EXP4" that achieves $\tilde{O}(T^{\frac{3}{4}}+d^{\frac{1}{2}}T^{\frac{5}{8}})$.

3. We present an $\tilde{\Omega}(d^{\frac{1}{3}}T^{\frac{2}{3}})$ regret lower bound for LP problem and an $\tilde{\Omega}(T^{\frac{2}{3}})$ for LV problem (even with stochastic x_t's, known θ^* and Lipschitz valuation distribution). The results indicate "Linear-EXP4" optimal up to logarithmic factors.

To the best of our knowledge, we are the first to study the LP problem and the version of the LV problem with no assumption on the noise. Comparing to the existing literature on this problem [Cohen et al., 2020, Javanmard and Nazerzadeh, 2019], our model makes fewer assumptions. Our results for LP is information-theoretically optimal, and our results in LV improve over the best known upper and lower bounds (from $\tilde{O}(T^{\frac{2}{3}\vee(1-\alpha)})$ on i.i.d. x_t's with an indeterministic α and $\Omega(T^{\frac{3}{5}})$ in Luo et al. [2021]).

Technical Novelty. We make use of the *half-Lipschitz* nature in pricing problems: the

probability of a price to be accepted will not decrease as the price decreases. This has been used in Kleinberg and Leighton [2003] and Cohen et al. [2020]. However, they directly applied this property in discretizing the action and policy spaces, which would lead to a linear regret in our LV problem setting. In our algorithm D2-EXP4, we settle this issue by also discretizing the noise distribution space and include these discretized CDF's as part of policy candidates. We also carefully adopt a conservative "markdown"[2] on the discretized output price to ensure a large-enough probability of acceptance. In this way, we get rid of all assumption on the noise distribution (even the basic Lipschitzness assumed by Luo et al. [2021]) while achieving a sub-linear regret. This discretization method, along with the price markdown, can be easily transferred to any pricing problem settings with unknown i.i.d. noise. For the lower bound proof, we adapt the nested intervals and bump functions introduced by Kleinberg [2004] for continuum bandits to our pricing problem models, and extend the $\Omega(T^{\frac{2}{3}})$ regret lower bound on non-continuous demand functions [Kleinberg and Leighton, 2003] to Lipschitz ones.

3.2 Related Works

In this section, we discuss how our work relates to the existing literature on (either contextual or non-contextual) pricing, bandits, and contextual search.

Non-Contextual Dynamic Pricing. Dynamic pricing was extensively studied under the single-product (non-contextual) setting [Kleinberg and Leighton, 2003, Besbes and Zeevi, 2009, 2012, Wang et al., 2014, Besbes and Zeevi, 2015, Chen et al., 2019b, Wang et al., 2021b]. The crux of pricing is to learn the demand curve (i.e., the noise distribution

[2]A price markdown is defined as a reduction on the selling price.

in our LP problem) from Boolean-censored feedback. Wang et al. [2021b] concludes existing results and characterizes the impact of different assumptions on the demand curve on the minimax regret. The problem of contextual dynamic pricing is more challenging mainly because we need to learn the valuation parameter θ^* and the noise distribution jointly. Knowing one would imply a learning algorithm for another [Javanmard and Nazerzadeh, 2019, Luo et al., 2021], but learning both together makes the problem highly nontrivial.

Contextual Dynamic Pricing. There is a growing body of recent works focusing on the LV model of the contextual dynamic pricing problem [Cohen et al., 2020, Javanmard and Nazerzadeh, 2019, Xu and Wang, 2021, Luo et al., 2021, Fan et al., 2021], but most of them make strong assumptions about the noise. Table 3.1 lists the best existing results under these assumptions. Besides these works, Cohen et al. [2020] also achieved an $O(d \log T)$ regret when the variance of the Sub-Gaussian noise is extremely small, i.e., $\tilde{O}(1/T)$. It is worth mentioning that our "Linear-EXP4" shares the same discretization factor with "ShallowPricing" algorithm in Cohen et al. [2020], but ours solves a different problem. The closest works to ours are the recent Luo et al. [2021] and Fan et al. [2021] that study the LV problem under only smoothness and log-concavity assumptions. In Luo et al. [2021], they develop a UCB-style algorithm that achieves $\tilde{O}(T^{\frac{2}{3}\vee(1-\alpha)})$ regret for noises with 2^{nd}-order smooth and log-concave CDF's, assuming the existence of a good-enough estimator that might approach θ^* with $O(T^{-\alpha})$ error only with the logged data. However, such an estimator was neither described nor trivial to construct with $\alpha > 0$. In Fan et al. [2021], they present a two-phase algorithm, with an exploration phase followed by an exploitation phase, and achieves $\tilde{O}((Td)^{\frac{2m+1}{4m-1}})$ regret for noises with m^{th}-order smooth ($m \geq 2$) and "well-behaved"[3] CDF's. In comparison, our "D2-EXP4" algorithm

[3]A property defined similarly as log-concavity.

achieves an $\tilde{O}(T^{\frac{3}{4}})$ regret with no distributional assumptions such as Lipschitzness or smoothness.

Bandits A multi-armed bandit (MAB) is an online learning model where one can only observe the feedback of the selected action at each time. Both LP and LV can be reduced to contextual bandits [Langford and Zhang, 2007, Agarwal et al., 2014] as long as the policies and prices are finite. In this work, we make use of an "EXP-4" algorithm [Auer et al., 2002b] in a new way: By carefully discretizing the parameter space and distribution functions, we enable EXP-4 agents to find out near-optimal policies among infinite continuum policy spaces. There exists another family of bandit problem: continuum-armed bandit (CAB) [Agrawal, 1995, Kleinberg, 2004, Auer et al., 2007], where the action space is continuum and the reward function is Lipschitz. We adapt the (bump functions, nested intervals) structures in Kleinberg [2004] to our lower bound proof. This adaptation is non-trivial since (1) their reward functions is not suitable for pricing problems, and (2) their feedback is not Boolean-censored.

Our results on the LP problem reveal that a reduction to contextual bandits is "tight" in regret bounds. A similar situation also occurs in Kleinberg and Leighton [2003] on non-contextual pricing. These results indicate a pricing feedback is not substantially richer than a bandit feedback in information theory, which is surprising as a pricing feedback indicates the potential feedback of a "halfspace" rather than a single point. However, does this imply we cannot get any extra information from a pricing feedback? Notice that we are matching a no-Lipschitz upper bound with a Lipschitz lower bound! In fact, a revenue curve is naturally "half Lipschitz", which helps us get rid of this assumption. We will discuss this property in Section 3.4.2.

Contextual search Contextual pricing is cohesively related to contextual search problems [Leme and Schneider, 2018, Lobel et al., 2018, Liu et al., 2021, Krishnamurthy et al., 2021] where they also learn from Boolean feedback and usually assume linear contexts. However, they are facing slightly different settings: Leme and Schneider [2018], Lobel et al. [2018] are noiseless and could achieve an optimal $O(\log \log T)$ regret; Liu et al. [2021] allows noises directly on customers' decisions instead of the valuations in our setting; Krishnamurthy et al. [2021] allows only small-variance valuation noises that is similar to Cohen et al. [2020].

3.3 Problem Setup

Symbols and Notations. Now we introduce the mathematical symbols and notations involved in the following pages. The game consists of T rounds. $x_t, \beta^*, \theta^* \in \mathbb{R}_+^d, y_t, N_t \in \mathbb{R}, v_t \in \mathbb{R}_+$[4], where $d \in \mathbb{Z}_+$. At each round, we receive a payoff (*reward*) $r_t = v_t \cdot \mathbb{1}_t$ where $\mathbb{1}_t := \mathbb{1}(v_t \le y_t)$ indicates the acceptance of v_t, i.e., $\mathbb{1}_t = 1$ if $v_t \le y_t$ and 0 otherwise. For LP problem, we denote $F_{LP}(v|x)$ as a *demand function*, i.e. the probability of price v being accepted given feature x. Therefore, $F_{LP}(v|x)$ is non-increasing with respect to v, for any $x \in \mathbb{R}^d$. For LV problem, we specifically denote $u_t = x_t^\top \theta^*$ as the *noiseless valuation* (or *expected valuation* for zero-mean noises), and denote F as its CDF. Finally, we define $h(v, x) = v \cdot F_{LP}(v|x)$ as an *expected revenue* function of price v given feature x in an LP problem, and $g(v, u, F) := v \cdot (1 - F(v - u))$ as an *expected revenue* function of price v given any noiseless valuation u and noise distribution F in an LV problem.

We may use discretization methods in the following sections. Here we adopt the notation

[4]We do not assume $y_t \ge 0$ since some customer would not buy anything despite the price.

in Cohen et al. [2020] by denoting

$$\lfloor x \rfloor_\gamma := \lfloor \frac{x}{\gamma} \rfloor \cdot \gamma, \lceil x \rceil_\gamma := \lceil \frac{x}{\gamma} \rceil \cdot \gamma. \tag{3.1}$$

as the γ-*lower/upper rounding* of x, which discretize x as its nearest smaller/larger integer multiples of γ. Similarly, for $\theta \in \mathbb{R}^d$, we may define $\lfloor \theta \rfloor_\gamma := [\lfloor \theta_1 \rfloor_\gamma, \lfloor \theta_2 \rfloor_\gamma, \ldots, \lfloor \theta_d \rfloor_\gamma]^\top$ and $\lceil \theta \rceil_\gamma := [\lceil \theta_1 \rceil_\gamma, \lceil \theta_2 \rceil_\gamma, \ldots, \lceil \theta_d \rceil_\gamma]^\top$. Based on this, we define a counting set $N_{\gamma,a} := \left\{0, 1, 2, \ldots, \lfloor \frac{a}{\gamma} \rfloor\right\}$.

Regret Definitions. Next we define the regrets in both problems.

Definition 3.3.1 (Regret in LP)**.** We define Reg_{LP} as the regret of the Linear Policy pricing problem.

$$Reg_{LP} := \max_\beta \sum_{t=1}^{T} h(x_t^\top \beta, x_t) - h(v_t, x_t). \tag{3.2}$$

Definition 3.3.2 (Regret in LV)**.** We define Reg_{LV} as the regret of the Linear Noisy Valuation problem.

$$Reg_{LV} := \sum_{t=1}^{T} \max_v g(v, u_t, F) - g(v_t, u_t, F). \tag{3.3}$$

Again, we aim at competing with the best fixed $\beta^* = \text{argmax}_\beta \sum_{t=1}^T h(x_t^\top \beta, x_t)$ in an LP problem, and with the global best pricing policy (maximizing expected revenue at every t) in an LV problem.

Summary of Assumptions We specify the problems by the following assumptions:

Assumption 3.3.3 (bounded features and parameters)**.** Without losing generality, we assume that $x_t, \beta^*, \theta^* \in \mathbb{R}_+^d$, $\|x_t\|_2 \leq B$, $\|\beta^*\|_2 \leq 1$, $\|\theta^*\|_2 \leq 1$, where $B \in \mathbb{Z}^+$ is a constant known to us in advance.

Assumption 3.3.4 (decreasing demand in LP)**.** In LP problem, assume that $F_{LP}(v|x)$ is non-increasing for any $v \geq 0, x \in \mathbb{R}_+^d$.

Assumption 3.3.5 (bounded noise)**.** In LV problem, assume that $N_t \in [-1, 1]$ that is i.i.d. sampled from a fixed unknown distribution $\mathbb{D}$.

These assumptions are mild and common for algorithm design. Based on these assumptions above, we only have to consider prices in $[0, B]$ for LP problems and $[0, B + 1]$ for LV problems. Besides, we assume that $T \geq d^4$ for a simplicity of comparing among different terms in regret bounds. In Section 3.5.2, we will introduce more assumptions to the distribution functions to demonstrate that our lower bounds hold *even if* those assumptions are made.

3.4 Algorithm

In this section, we propose two algorithms, Linear-EXP4 and D2-EXP4, for LP and LV problems respectively. Both of them are based on the EXP-4 algorithm [Auer et al., 2002b] along with discretized policy sets. First of all, we define these policy sets:

Definition 3.4.1 (parameter set)**.** For any small $0 < \Delta < 1$, we define a parameter set $\Omega_{\Delta,d} \subset \mathbb{R}^d$:

$$\Omega_{\Delta,d} := \left\{ \|\theta\|_2 \leq 1, \theta = [n_1\Delta, n_2\Delta, \ldots, n_d\Delta]^\top, n_1, n_2, \ldots, n_d \in N_{\Delta,1} \right\}$$

Definition 3.4.2 (CDF set)**.** For any small $0 < \gamma < 1$, we define a Cumulative Distribution Function (CDF) set $\mathcal{F}_\gamma$:

$$\mathcal{F}_\gamma := \left\{ \begin{array}{l} F : \mathbb{R} \to [0,1] \text{ non decreasing }, \\ \quad F(v) = 0 \text{ when } v < -1, \\ \quad F(v) = 1 \text{ when } v > 1, \\ \quad \dfrac{F(v)}{\gamma} \in N_{\gamma,1} \text{ when } \pm\dfrac{v}{\gamma} \in N_{\gamma,1}, \\ \quad F(v) = F(\lfloor v \rfloor_\gamma) + \dfrac{1}{\gamma}(F(\lfloor v \rfloor_\gamma + \gamma) - F(\lfloor v \rfloor_\gamma))(v - \lfloor v \rfloor_\gamma) \text{ otherwise} \end{array} \right\}.$$

Definition 3.4.1 is straightforward as we use Δ^d-grids to discretize the $[0,1]^d$ space. Definition 3.4.2 actually represents such a family of CDF: the random variable is defined on $[-1,1]$, and its CDF equals some integer multiple of γ when v (or $-v$) itself is an integer multiple of γ; for those v in between these grids, CDF connects the two endpoints as linear. In a word, each CDF in $\mathcal{F}_\gamma$ is a piecewise linear function with every integer-multiple-γ points valuating some integer-multiple-γ as well. From the definitions above, we know that $|\Omega_{\Delta,d}| = O\left((\frac{1}{\Delta})^d\right)$. Also, we have $|\mathcal{F}_\gamma| = \binom{\frac{3}{\gamma}}{\frac{1}{\gamma}} = O(2^{\frac{3}{\gamma}})$ according to a "balls into bins" model in combinatorial counting: At each point $\frac{\pm i}{\gamma}$ (for $i \in [\frac{2}{\gamma}]$) the CDF can increase by $j \cdot \gamma$, with j being a non-negative integer, and the summation of all increases is 1 (i.e., $\frac{1}{\gamma}$ of γ increments).

Finally we introduce the *EXP-4* algorithm [Auer et al., 2002b] for adversarial contextual bandits. With a finite action set A and policy set Π, the EXP-4 agent has a regret guarantee at $O(\sqrt{T|A|\log|\Pi|})$ in T rounds (comparing with the optimal policy in Π). The following is a simplified version of EXP-4 that illustrates its mechanism. For a more detailed introduction, please directly refer to Auer et al. [2002b].

EXP-4.

Input: Policy set Π, Action set A.

Initialize each policy i with weight w_i;

for $t = 1$ **to** T **do**

Set probability $p_j(t)$ for each action j according to weights of all policies;

Get a_t by Thompson sampling the action set A according to current probability $\{p_j(t)\}$;

Receive a reward r_t;

Construct an *Inverse Propensity Scoring (IPS)* estimator $\hat{r}_i(t)$ for the reward of each action i.

Update weights w_i's according to $\hat{r}_i(t)$.

end for

3.4.1 Linear-EXP4 for LP

Here we present our "Linear-EXP4" algorithm for the linear policy pricing problem. It takes $\Omega_{\Delta,d}$ as the policy set and plug it into EXP-4 algorithm, which is straightforward but significant in reducing the regret. The pseudo-code of Linear-EXP4 is summarized as Algorithm 3.

Algorithm 3 Linear-EXP4

Input: Parameter set $\Omega_{\Delta,d}$, Action set $A_\gamma = \{0, \gamma, 2\gamma, \ldots, \lfloor B \rfloor_\gamma\}$, parameters Δ, γ.

Set policy set $\Pi^{LP}_{\Delta,\gamma} = \{\pi_\beta(x) = \lfloor x^\top \beta \rfloor_\gamma, \beta \in \Omega_{\Delta,d}\}$

Initialize an EXP-4 agent $\mathcal{E}_{LP}$ with $\Pi^{LP}_{\Delta,\gamma}, A_\gamma$;

for $t = 1$ **to** T **do**

$\mathcal{E}_{LP}$ observe x_t;

$\mathcal{E}_{LP}$ choose an action (price) v_t;

Receive feedback $r_t = v_t \cdot \mathbb{1}_t$ and feed it into $\mathcal{E}_{LP}$;

end for

Here the EXP-4 agent $\mathcal{E}_{LP}$ would approach the best policy π^* in $\Pi^{LP}_{\Delta,\gamma}$ within a reasonable regret. Therefore, we have to carefully choose Δ and θ such that the regrets of both $\mathcal{E}_{LP}$ and π^* are well bounded.

3.4.2 Discrete-Distribution-EXP4 for LV

Here we present our "Discrete-Distribution-EXP-4" algorithm, or D2-EXP4 for the linear noisy valuation pricing problem. Though it originates EXP-4 as well as Linear-EXP4 above, the reduction is not as straightforward. In fact, the policy set is defined as follows:

$$\begin{aligned} \Pi^{LV}_{\Delta,\gamma} = \Big\{ & \pi | \pi(x; \hat{\theta}, \hat{F}) = \max\{\lfloor x^\top \hat{\theta} \rfloor_\gamma - (B+1)\gamma + \lfloor w^*(x) \rfloor_\gamma, 0\}, \\ & \text{where } \; w^*(x) = \operatorname*{argmax}_w g(u + w, x^\top \hat{\theta}, \hat{F}), \hat{\theta} \in \Omega_{\Delta,d}, \hat{F} \in \mathcal{F}_\gamma \Big\}. \end{aligned} \tag{3.4}$$

For each policy in $\Pi^{LV}_{\Delta,\gamma}$, it firstly takes a $\hat{\theta}$ from $\Omega_{\Delta,d}$ and a $\hat{F}$ from $\mathcal{F}_\gamma$, and then generate an "optimal incremental price" $w^*(x)$ greedily as if they are the true parameter θ^* and the true noise distribution F. Finally, the policy take an action (price) that is the summation of γ-lower roundings of $\hat{u} = x^\top \hat{\theta}$ and $w^*(x)$ to fit in the action set $A_\gamma := \{0, \gamma, 2\gamma, \ldots, \lfloor B + 1 \rfloor_\gamma\}$, and minus a $(B+1)\gamma$ amount. We know that $|\Pi^{LV}_{\Delta,\gamma}| =$

$|\Omega_{\Delta,d}| \cdot |\mathcal{F}_\gamma| = O((\frac{1}{\Delta})^d \cdot 2^{\frac{3}{\gamma}})$. We present the psuedo-code of D2-EXP4 as Algorithm 4.

Algorithm 4 Discrete-Distribution-EXP-4(D2-EXP4)

Input: Policy set $\Pi^{LV}_{\Delta,\gamma}$, Action set $A_\gamma = \{0, \gamma, 2\gamma, \ldots, \lfloor B+1 \rfloor_\gamma\}$, parameters Δ, γ.

Initialize an EXP-4 agent $\mathcal{E}_{LV}$ with $\Pi^{LV}_{\Delta,\gamma}, A_\gamma$;

for $t = 1$ **to** T **do**

$\mathcal{E}_{LV}$ observe x_t;

$\mathcal{E}_{LV}$ select an action(price) v_t;

Receive feedback $r_t = v_t \cdot \mathbb{1}_t$ and feed it into $\mathcal{E}_{LV}$;

end for

D2-EXP4 is straightforward that it takes the γ-rounding of a greedy price, except the $(B+1)\gamma$ price markdown. This is because we want a conservative price, and the $(B+1)\gamma$ markdown is to compensate the "exaggerate" $\lceil\theta\rceil_\gamma$ parameter we adopt in $\Pi^{LV}_{\Delta,\gamma}$. We will include more details in Section 3.4.2 below and in Section 3.5.1.

Adversarial Features and Agnostic Distributions Notice that both algorithms are suitable for adversarial x_t series, which is a property of EXP-4. It is worth mentioning that our Linear-EXP4 makes **no** assumptions on the distribution of y_t given x_t, and that D2-EXP4 assumes **no** pre-knowledge or technical assumptions on the noise distribution (despite that noises are bounded).

Conservative Pricing Strategy Both of our algorithms adopt a *conservative* strategy while pricing: In Linear-EXP4, a good-enough linear policy is the γ-lower rounding of parameter β^*; in D2-EXP4, we even define each policy by proposing a "greedy-and-safe" price which takes a $(B+1)\gamma$-markdown on the output of the optimal greedy pricing

policy. This is because of the "half-Lipschitz" nature of a demand curve: decreasing the price would at least maintain the chance of being accepted. Since we do not make any Lipschitz or smoothness assumptions on the distributions, these discretizations might marginally increase the price and cause drastic change of the expected revenue. In order to avoid this, it is always better to decrease the proposed price by an acceptable small amount as it guarantees the probability of acceptance.

Computational Efficiency Our algorithms require exponential computations w.r.t. dimension d since the EXP-4 agent requires exponential time to evaluate each policy in the policy set. An "optimization oracle"-efficient contextual bandit algorithm in Agarwal et al. [2014] can be used in place of EXP-4 to achieve a near-optimal regret (up to logarithmic factors), but it requires the input features x_t to be drawn from an unknown fixed distribution.

3.5 Regret Analysis

In this section, we analyze our Linear-EXP4 and D2-EXP4 algorithm and prove their $\tilde{O}(d^{\frac{1}{3}}T^{\frac{2}{3}})$ and $O(T^{\frac{3}{4}})$ regret bounds, respectively. Also, we present a scenario where a lower bound construction with $\tilde{\Omega}(T^{\frac{2}{3}})$ regret fits for both LP and LV problems, even under stronger assumptions including stochastic x_t's, Lipschitz distribution functions and unimodal demand curves.

3.5.1 Upper Bounds

Here we propose the following theorem as a regret bound of Linear-EXP4. This only requires the assumption that features x_t's and (potential) optimal parameter β^* is bounded by L_2-norm, without making any specifications on the feature-valuation mapping.

Theorem 3.5.1 (Regret of Linear-EXP4)**.** *In any LP problem, with Assumption 3.3.3, the expected regret of Linear-EXP4 does not exceed $O(d^{\frac{1}{3}}T^{\frac{2}{3}}\log dT)$ by setting $\Delta = T^{-\frac{1}{3}}d^{-\frac{1}{6}}$ and $\gamma = T^{-\frac{1}{3}}d^{\frac{1}{3}}$.*

Proof. We denote $\tilde{\beta}^* = \lfloor \beta^* \rfloor_\Delta$ and $\hat{\beta}^* := \text{argmax}_{\beta\in\Omega_{\Delta,d}} \sum_{t=1}^T \mathbb{E}[h(\pi_\beta(x_t), x_t)]$. Now we decompose the regret of LP problem as follows:

$$
\begin{aligned}
&\mathbb{E}[Reg_{LP}] \\
=&\sum_{t=1}^T \mathbb{E}[h(x_t^\top \beta^*, x_t) - h(v_t, x_t)] \\
=&\sum_{t=1}^T \mathbb{E}[h(x_t^\top \beta^*, x_t) - h(\pi_{\tilde{\beta}^*}(x_t), x_t)] + \mathbb{E}[h(\pi_{\tilde{\beta}^*}, x_t) - h(\pi_{\hat{\beta}^*}(x_t), x_t)] + \mathbb{E}[h(\pi_{\hat{\beta}^*}(x_t), x_t) - h(v_t, x_t)] \\
\leq&\sum_{t=1}^T (x_t^\top \beta^* - x_t^\top \tilde{\beta}^*) F_{LP}(x_t^\top \beta^* | x_t) + \mathbb{E}[h(\pi_{\tilde{\beta}^*}, x_t) - h(\pi_{\hat{\beta}^*}(x_t), x_t)] + \mathbb{E}[h(\pi_{\hat{\beta}^*}(x_t), x_t) - h(v_t, x_t)] \\
\leq&\sum_{t=1}^T B \cdot \Delta\sqrt{d} + 0 + \sqrt{T \cdot \frac{1}{\gamma} \cdot \log(\frac{1}{\Delta})^d} \\
=&O(d^{\frac{1}{3}}T^{\frac{2}{3}}\log dT).
\end{aligned}
\tag{3.5}
$$

Here the third row is because $\pi_{\tilde{\beta}^*}(x_t) = \lfloor x_t^\top \tilde{\beta}^* \rfloor_\gamma \leq x_t^\top \tilde{\beta}^* \leq x_t^\top \beta^*$ since $x_t, \beta \in \mathbb{R}_+^d$ (and thus $F_{LP}(x_t^\top \beta^*) \leq F_{LP}(x_t^\top \tilde{\beta}^*)$); The fourth row is because $(x_t^\top \beta^* - x_t^\top \tilde{\beta}^*) \leq \|x_t\|_2 \cdot \|\beta^* - \tilde{\beta}^*\| \leq B \cdot \Delta\sqrt{d}$, the optimality definition of $\hat{\beta}^*$ and the regret bound of EXP-4 from Auer et al. [2002b]; The last row is got by plugging in the value of Δ and γ. ∎

The proof of Theorem 3.5.1 is straightforward based on the existing $O(\sqrt{T|A|\log|\Pi|})$ bound of EXP-4. We only have to bound the error of the optimal policy in $\Pi_{\Delta,\gamma}$. Now

we present our result on D2-EXP4:

Theorem 3.5.2 (Regret of D2-EXP4)**.** *For any LV problem, with Assumption 3.3.3 Assumption 3.3.4, Assumption 3.3.5, our algorithm D2-EXP4 guarantees a regret no more than* $O(T^{\frac{3}{4}} + T^{\frac{2}{3}} d^{\frac{1}{2}} \log dT)$ *as we set* $\Delta = T^{-\frac{1}{4}} d^{-\frac{1}{2}}$ *and* $\gamma = T^{-\frac{1}{4}}$.

The proof of Theorem 3.5.2 is more sophisticated than that of Theorem 3.5.1, but they shares similar structures: we figure out one specific policy in $\Pi^{LV}_{\Delta,\gamma}$ that is close to the optimal policy of the LV problem. The main idea of this proof is to find out a tuple of $(\hat{\theta}, \hat{F})$ that approaches the true parameter and distribution, and to verify that the policy built on this approaching tuple is reliable only within small tractable error. The highlight is that we do not assume any Lipschitzness on the distribution, which is quite different from existing approximation methods. In fact, it is the natural property of pricing problems that enables this: for two prices $v_1 \geq v_2$, the probability of v_2 being accepted is greater (or equal) than that of v_1, and thus $(v_1 - v_2) \geq g(v_1, u, F) - g(v_2, u, F)$. We may call it a *Half-Lipschitz* property since it only upper bounds the increasing rates.

Here we show a proof sketch of Theorem 3.5.2, and leave the bulk to Section 3.9.1.

Proof Sketch. For any specific LV problem with linear parameter θ^* and noise CDF F, we define $\hat{\theta}^* := \lceil \theta^* \rceil_\Delta$ and $\hat{F}$:

$$\begin{aligned} \hat{F}(x) =& \lfloor F(x) \rfloor_\gamma \text{ when } x = i \cdot \gamma \text{ for } i \in \mathbb{Z}, \text{ and linearly connecting } \hat{F}(i\gamma) \text{ with } \hat{F}(i+1\gamma) \\ & \text{when } x \in (i\gamma, (i+1)\gamma). \end{aligned} \tag{3.6}$$

Our goal is to prove that $\pi(x; \hat{\theta}^*, \hat{F})$ performs well enough. We may furthermore define a few amounts:

(i) $\hat{u} = x^\top \hat{\theta}^*$;

(ii) $w^*(u) = \text{argmax}_w\, g(u + w, u, F)$;

(iii) $\hat{w}^*(u) = \text{argmax}_w\, g(u + w, u, F)$;

(iv) $\hat{w}(\hat{u}) = \text{argmax}_w\, g(\hat{u} + w, \hat{u}, \hat{F})$.

Therefore, the price our algorithm proposed for feature x is $\hat{v}(x) = \lfloor \hat{u} \rfloor_\gamma - (B+1)\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma$, and our goal is to prove that $g(\hat{v}, u, F) \geq g(u + w^*(u), u, F) - C \cdot \gamma$ for some constant C. Since $\gamma = T^{-\frac{1}{4}}$, this would upper bounds the optimality error up to $O(T \cdot \gamma) = O(T^{\frac{3}{4}})$. In fact, we have the following properties:

(i) $\hat{\theta}^* = \lceil \theta^* \rceil_\Delta$ (by definition);

(ii) $\|\theta^*\|_2 \leq \|\hat{\theta}^*\|_2 \leq \|\theta^*\|_2 + \Delta\sqrt{d} = \|\theta^*\|_2 + \gamma$;

(iii) $u - \gamma \leq \hat{u} - \gamma \leq \lfloor \hat{u} \rfloor_\gamma \leq \hat{u} \leq u + B\gamma$;

(iv) $\hat{F}(i\gamma) \leq F(i\gamma) \leq \hat{F}(i\gamma) + \gamma$.

According to these properties, we may derive:

$$
\begin{aligned}
& g(\lfloor \hat{u} \rfloor_\gamma - (B+1)\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma, u, F) \\
\geq & (u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - F(\lfloor \hat{w}(\hat{u}) \rfloor_\gamma)) - (B+2)\gamma \\
\geq & (u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - \hat{F}(\hat{w}(\hat{u}))) - (2B+3)\gamma \\
\geq & g(\hat{u} + \hat{w}(\hat{u}), \hat{u}, \hat{F}) - (3B+4)\gamma \\
\geq & g(u + \hat{w}^*(u), u, \hat{F}) - (3B+4)\gamma \\
\geq & g(u + w^*(u), u, F) - (3B+5)\gamma.
\end{aligned}
$$

The derivation of each step is shown in Section 3.9.1. With this policy-realizability error being bounded by $(3B+5)\gamma = O(T^{\frac{3}{4}})$ and the original regret of the EXP-4 agent

being $O(\sqrt{TK\log N}) = \tilde{O}(T^{\frac{3}{4}} + d^{\frac{1}{2}}T^{\frac{5}{8}})$, we may finally get a $\tilde{O}(T^{\frac{3}{4}} + d^{\frac{1}{2}}T^{\frac{5}{8}})$ upper regret bound. ■

3.5.2 Lower Bounds

In this part, we present an $\tilde{\Omega}(T^{\frac{2}{3}}d^{\frac{1}{3}})$ and an $\tilde{\Omega}(T^{\frac{2}{3}})$ regret lower bounds that hold for LP and LV problems respectively. We will firstly claim a lower bound for non-contextual pricing problem, and then generalize the result to LP and LV.

Theorem 3.5.3 (Lower bound for non-contextual pricing)**.** *For a non-contextual pricing problem where the valuation y_t's are generated independently and identically from a fixed unknown distribution satisfying (1) the CDF $F(y)$ is Lipschitz and (2) the revenue curve $g(v, F) = y \cdot (1 - F(v))$ is unimodal (i.e., non-decreasing on $(0, v_0)$ and non-increasing on $(v_0, +\infty)$ for some v_0), NO algorithm can achieve $O(T^{\frac{2}{3}-\delta})$ for any $\delta > 0$.*

The detailed proof of Theorem 3.5.3 is in Section 3.9.2, and in the main pages we briefly demonstrate the constructions of the subproblem family where we achieve this lower bound.

Here we take the idea of Kleinberg [2004] where they make use of bump functions and nested intervals to ensure Lipschitz continuity and unimodality, sequentially. Since that their model is not capturing a revenue curve and that their feedback is numerical instead of Boolean, we have to adjust their design to satisfy the pricing setting. On the one hand, the probability of a price to be accepted, i.e., the rate $\frac{\mathbb{E}[r(v)]}{v}$, is non-increasing as the prices increases, which is not guaranteed for that of a reward function of a continuum bandit (if we treat v as an action). In this proof, we adopt a series of transformations to convert the "bump function tower" into a revenue curve while keeping

all monotonically-increasing/decreasing intervals unchanged. On the other hand, we still use the KL-divergence to distinguish among distributions, but in a different way. As for Boolean feedback, we only need to calculate the KL-divergence of two Bernoulli random variables, which can be upper bounded by a quadratic term of their probabilistic difference.

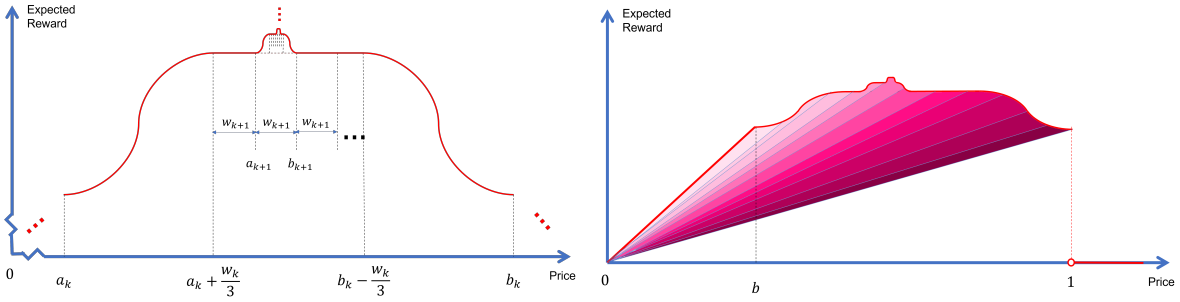


Figure 3.1: Structure of our lower bound function family. The left figure shows how we use bump functions to construct a reward function $f(v)$. Each bump function locates at $[a_k, b_k]$ with a length $w_k = 3^{-k!}$. Notice that the middle one-third of each bump is a plain divided into small intervals of length w_{k+1}, and we might randomly choose one to build up the $(k+1)^{\text{th}}$ bump. However, the rate $\frac{f(v)}{v}$ that indicates the probability of v to be accepted is not necessarily non-increasing, and therefore $f(v)$ cannot capture a revenue function for pricing. The right figure shows an ideal revenue curve $D(v)$ which equals v for $v \in [0, b]$ and equals $b + (1-b)(1 - \frac{1}{f(v)+1})$ for $v \in (b, 1]$. The slopes indicate that $\frac{D(v)}{v}$ is actually non-increasing. We draw the figures with exaggeration to show the hierarchical structures better.

The constructions of bump-based revenue curves are illustrated in Figure 3.1. Firstly, we define a nested-interval series $[0,1] = [a_0, b_0] \supset [a_1, b_1] \supset \ldots \supset [a_k, b_k] \supset \ldots$, where $b_k = a_k + w_k$, $w_k = 3^{-k!}$. We let a_k be chosen from the discrete set $\{a_{k-1} + \frac{w_{k-1}}{3} + i \cdot w_k, i = 0, 1, 2, \ldots, \frac{w_{k-1}}{3w_k}\}$. Secondly, we construct Lipschitz bump functions in each $[a_k, b_k]$ interval,

the middle one-third of which is a plain line Thirdly, we add all these bump function up, which forms a "tower" with its peak randomly generated by the series of tightening intervals $\{[a_k, b_k]\}$. Finally, it is transformed into a revenue curve after a series of operations.

If we treat this randomly-generated function a uniformly-distrbuted family of functions, then we can further prove our lower bound: On the one hand, we prove that the feedback cannot accurately locate where the "peak of the tower" is, from the perspective of information theory. In fact, any algorithm would have a constant chance of missing the peak. On the other hand, the cost of missing a peak can be lower bounded, and thus the expected regret is as well lower bounded by their product.

With this theorem holds, we can soon get the following two corollaries:

Corollary 3.5.4 (Lower bound of LP problem)**.** *The regret lower bound for LP problems is $\tilde{\Omega}(d^{\frac{1}{3}}T^{\frac{2}{3}})$, even with stochastic features and distributional properties same as those in Theorem 3.5.3.*

Proof. Here we construct the following LP problem: let $x_t = [0, \ldots, 0, 1, 0, \ldots, 0]^\top$ with only the i_t^{th} element being 1, where i_t is chosen from $\{1, 2, \ldots, d\}$ uniformly at random for each $t = 1, 2, ..., T$. As a result, the problem is split into d-subproblems with each of them a non-feature pricing problem in $\frac{T}{d}$ rounds in expectation (since the demand function $F_{LP}(y|x)$ can be totally different and independent for different x's). According to Theorem 3.5.3, the lower bound for this problem is $\tilde{\Omega}(d \cdot (\frac{T}{d})^{\frac{2}{3}}) = \tilde{\Omega}(d^{\frac{1}{3}}T^{\frac{2}{3}})$. ∎

Corollary 3.5.5 (Lower bound of LV problem)**.** *The regret lower bound for LV problems is $\tilde{\Omega}(T^{\frac{2}{3}})$, even with stochastic features and noise-distributional properties stated in Theorem 3.5.3.*

It is worth mentioning that the noise distribution is itself an (inversed) demand function on $(v-u)$, i.e., it is non-increasing as $(v-u)$ gets larger. Based on this insight, the derivation of Corollary 3.5.5 is straightforward: any non-feature pricing problem with bounded i.i.d. y_t's can be reduced to an LV problem up to constant coefficients. In fact, suppose $y_t \in [a,b], 0 \leq a < b$ in a non-feature pricing problem, and then we might define an LV problem by setting $d = 1, \theta^* = \frac{a+b}{b-a}$ and $x_t = 1, \forall t \in \mathbb{Z}_+$ since now $x_t^\top \theta^* + N_t \in [a,b]$. As long as the definition of LV problem does not specify the distributional properties (besides being bounded), the distribution family in the proof of Theorem 3.5.3 can be reduced to an LV problem as well. In this way, the $\tilde{\Omega}(T^{\frac{2}{3}})$ lower bounds are applicable to LV problems.

3.6 Numerical Experiments

In this section, we conduct numerical experiments to show the validity of Linear-EXP4. We assume $d = 2, B = 1$ as basic parameters, and assume a Gaussian noisy valuation model i.e., $y_t = u_t + N_t$ where $N_t \sim \mathcal{N}(0, \frac{1}{16})$ independently for all t. For the convenience of comparing with a fixed optimal linear policy β^*, we let $u_t = J^{-1}(x_t^\top \beta^*)$ for each t, where $J(u) = \operatorname{argmax}_v g(v, u, 1 - \Phi_{\mathcal{N}(0,\frac{1}{16})})$ is a greedy pricing function defined in Xu and Wang [2021][5]. In other words, the linear price $v_t^* = x_t^\top \beta^*$ always maximizes the expected reward for any t, and we may calculate the empirical *ex ante* regret (i.e., comparing the empirical performance with the maximizer of expected regret at each round) by comparing $v_t \cdot \mathbb{1}(v_t \leq y_t)$ with $x_t^\top \beta^* \cdot \mathbb{1}(x_t^\top \beta^* \leq y_t)$. According to Hoeffding's Inequality, the *ex post* regret that we adopt for the LP problem is only $\tilde{O}(\sqrt{T})$ different from the empirical *ex ante* regret. Given that the regret rate of Linear-EXP4 is $\tilde{\Theta}(T^{\frac{2}{3}})$, we may ignore this difference and only show the ex ante regret in our experiments. Since the

[5]They also show the existence of $J^{-1}(v)$ by showing that $J'(u) \in (0,1)$.

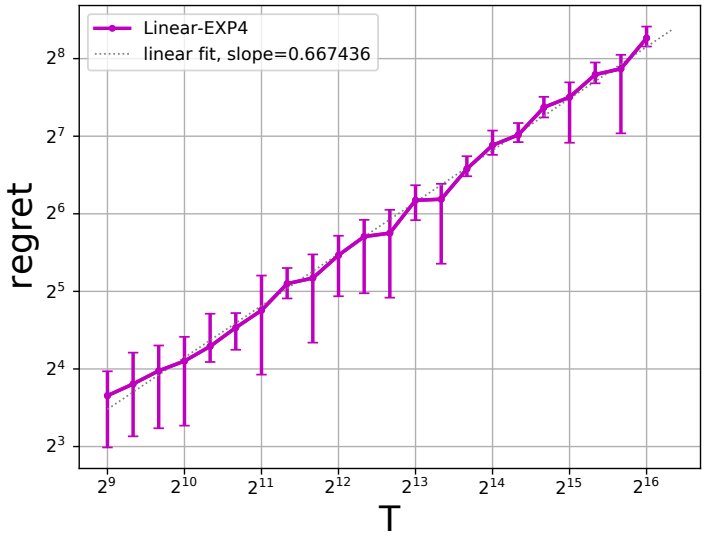


Figure 3.2: Regrets of Linear-EXP4 on simulated examples. The plot is on log-log scales to show the regret rate: a slope of α indicates an $O(T^\alpha)$ regret. Besides, we draw error bars with 0.95 coverage. Notice that the slope of its linear fit is 0.667, which matches the $\tilde{O}(T^{\frac{2}{3}})$ regret rate in theory.

EXP-4 learner requires pre-knowledge on T and is not an any-time algorithm (i.e., the cumulative regret is meaningful only at $t = T$), we execute Linear-EXP4 for a series of $T = \lfloor 2^{\frac{k}{3}} \rfloor$ for $k = 27, 28, \ldots, 48$. We repeat every experiment 20 times for each setting and then take an average. The results are shown in Figure 3.2

We were unable to conduct numerical experiments on D2-EXP4 due to the exponential time complexity of the EXP-4 learner along with the $2^{T^{\frac{1}{4}}}$-size policy set. We provide the code of D2-EXP4 in our supplementary materials.

3.7 Discussion

In this section, we discuss potential extensions of this work and our conjectures on the regret of LV problems.

From Linear to Non-Linear Both LP and LV problems are based on a linear principle of feature-price/valuation relationships, which is not reasonable in many real-world situations (for example, the price of a diamond). Based on our specifications on LP and LV problems, we may similarly define two corresponding problems: (1) We make no assumptions on the $x_t \to y_t$ mapping, but compare with the optimal policy in a parametric non-linear model space. (2) We directly assume that the $x_t \to y_t$ is a parametric non-linear function adding some unknown (and non-parametric) noise, and compare with the optimal price. We may slightly modify our Linear-EXP4 and D2-EXP4 to deal with these two problems by just replacing the linear discretized policy set with another non-linear one. However, we should be careful about any discretization involved: the γ-roundings of non-linear policy parameters do not necessarily lead to a slightly lower price (maybe either higher or much lower). Like what we designed in D2-EXP4, we still have to ensure the parametric optimal policy itself performs within a $[-O(\gamma), 0]$ range from the global optimal policy.

The Minimax Regret(s) of LV Existing works on solving LV have achieved various regret bounds with different assumptions. This is quite different from the linear regression problem where noise distributions do not significantly affect the result. To the best of our knowledge, we are the first to get rid of all assumptions (despite bounded-noise

assumption [6]). However, we did not close the regret gap in this setting. This problem is similar to a non-feature pricing problem as we adopt the same lower bound proof in this work, but the situations are entirely different: In non-feature pricing, we aim at a fixed optimal price, and we only have to know the valuation distribution around the optimal price. However, in an LV problem, we have to approach the exact linear valuation adding an optimal *increment* for each feature, and the optimal increments are **not** fixed for different valuations. As a result, we have to know the whole noise distribution. This drastically increases the hardness of LV, and we conjecture LV with a $\Theta(T^\alpha)$ regret where $\alpha > \frac{2}{3}$.

Dependence on Noise Scale R In this work we assume the noise $N_t \in [-1, 1]$. Based on this assumption, we construct a discrete noise CDF family $\mathcal{F}_\gamma$ whose size is $\binom{\frac{3}{\gamma}}{\frac{1}{\gamma}}$. When it changes to $N_t \in [-R, R]$ for larger R, the number of discrete CDF is $\binom{\frac{2R+1}{\gamma}}{\frac{1}{\gamma}} \leq (\frac{2R+1}{\gamma})^{\frac{1}{\gamma}}$. Also, this would increase the upper bound of prices from $(B+1)$ to $(B+R)$, which would increase the number of actions by $\frac{R}{\gamma}$. Recall that the regret of EXP-4 is $O(\sqrt{KT \log N})$ where K is the number of actions and N is the number of policies (i.e., # discrete θ times # discrete CDF). Therefore, the dependence on R is $O(\sqrt{R \log R})$.

Differences between LP and LV As we stated in Section 3.1, LP models our strategy while LV modes the nature. Also, a good (no-regret) LP algorithm approaches the best linear policy in total while a good LV algorithm approaches the global optimal price at each round. When we adopt a LP problem model, we indeed have very little information about the market valuation other than obvious features of the product to sell. In this situation, a linear pricing policy is tractable and transparent to the customers, but it

[6] If the noise is neither bounded nor parametrized, then any finite-time algorithm will suffer a linear regret when the noise is very large and prices are always being accepted.

is not guaranteed to present or approach the best price. When we adopt a LV problem model, it is assumed that we have already known all features of the selling session (not limited to the product itself), and the fluctuation caused by the market is independent to the product. Therefore, we may learn from the feature-pricing-feedback data over the time and estimate the noise distribution, which would help approaching the best price combining with a greedy policy. Here is a concrete example regarding vehicle owners, dealers and buyers that illustrates the difference between LP and LV:

1. In Session 1, suppose we are the owner and would like to sell our used car to a buyer/dealer. A 3rd-party evaluator will evaluate your car based on a few (but not all) factors, e.g., mileage, duration, condition and accident records, and then subtract a certain amount from the selling price of an identical new car. This amount is usually linearly or near-linearly dependent on these factors listed above. Remember that this selling price is proposed by we owners. In other words, we are the seller in this session, and the buyer/dealer would respond by accepting or declining the price we propose. Here we adopt a linear pricing policy because we do not have full information of the selling session, and therefore customers' valuation model is indeed unclear to us.

2. In Session 2, suppose we are the dealer and would like to sell a used car to a buyer. Car dealers usually have sufficient information on the vehicle and the market supply-demand relationship. At least, we know clearly about which features are related to customers' valuations. Therefore, it is reasonable for usr to assume a parametric noisy valuation model (possibly a LV model) on their customers, and we would optimize these parameters based on historical selling records. With the model being well-learned, we may approach the global optimal price every time. That we directly make assumptions on customers' valuation model is reasonable since we

dealers have sufficient information, but this could still be risky if the features we can observe are limited.

The Hardness of Pricing versus Bandits The generic feature-based dynamic pricing problem can be reduced to a contextual bandit problem with continuum action and infinite policy spaces, despite some literature that assumes a different acceptance/declination reward scope (see Bartók et al. [2014]). Therefore, the gap between a dynamic pricing problem and an ordinary (discrete-action and finite-policy) contextual bandit problem can be observed from three perspectives. Firstly, the pricing feedback contains more information than a bandit feedback: if v_t is accepted, then any $v \leq v_t$ would have been accepted if it were proposed. We call this a "half-space information". Secondly, a discrete action space might not contain the optimal or any near-optimal price that matches the minimax regret: the revenue curve can vary drastically with respect to the price (e.g., consider a noise whose pdf is a rescaled Weierstrass function). Thirdly, a finite policy space might not contain the global optimal or any near-optimal policy, either. This is possible even for a parametric policy space where the parameter space is infinite. Therefore, we cannot directly adopt the regret bounds of contextual bandits onto feature-based dynamic pricing problems unless there exists a rigorous reduction. However, we notice that the three perspectives above are pointing at different directions: the "half-space information" makes pricing easier than bandits, while the other two discretization issues makes it harder. In fact, we might partially offset the "continuum action" issue with the "half-space information" just like what we did in this paper: the revenue curve is actually "half Lipschitz" that $g(v_1, u, F) - g(v_2, u, F) \leq v_1 - v_2$ if $v_1 \geq v_2$. This helps our algorithms get rid of the Lipschitz assumption. However, this is not rich enough to substantially reduce the regret as we still use bandit algorithms to achieve a minimax

rate in an LP problem, where the lower bound holds even for Lipschitz revenue curve. Therefore, a very important question occurs to us: what else could a pricing feedback provide other than the "half Lipschitz"? Technically speaking, does a pricing feedback contain high-order information of the revenue curve? Besides, remember that we still do not have a unified approach toward a finite near-optimal policy set. In this work, we discretize the noise distribution by $\frac{1}{\gamma}$ grids, which indeed increases the regret bound. For more sophisticated feature-valuation mapping (e.g., a non-linear valuation model) that is hard to parameterize, maybe it is not suitable to just apply naive discretization methods. As a result, pricing problem is at least as hard as bandits, and it is still unclear whether the minimax regret of agnostic contextual pricing is obtained by contextual bandit methods.

Social Impacts In this work, we mainly focus on an online-fashion pricing problem where only one product is sold to one customer at each round (time spot). Therefore, it is not likely to commit a pricing discrimination according to its rigorous definition (since the price fluctuation over time should not be treated as discrimination). However, there exist chances that our algorithm could be misused. Notice that each item is characterized by a feature vector x_t, which might be used to capture more information, e.g., customers' behaviors. On the one hand, it is indeed a price discrimination if we propose differently-generated prices to customers with different personal features even at different time point as long as the market has not changed substantially. On the other hand, this would lead to a potential leakage of personal privacy. It is usually forbidden to collect and use personal information for commercial use, but the sellers would at least know what the customers have bought and how much they have paid. Even though the feature x_t can be encoded with cryptographic techniques such that it is still suitable for learning (e.g., a

"fully-homomorphic encryption", or FHE), at least the proposed prices are informative and might reveal the customer's behaviors. Indeed, auctions are a method to avoid any pricing discrimination, but it is not practical in most of the situations happening in our daily life.

3.8 Conclusion

In this chapter, we have studied two agnostic feature-based dynamic pricing problems: a linear pricing policy (LP) problem with no assumptions on feature-valuation mappings, and a linear noisy valuation (LV) problem with agnostic noise distributions. For the LP problem, we have presented a *Linear-EXP4* algorithm whose $\tilde{O}(T^{\frac{2}{3}}d^{\frac{1}{3}})$ regret matches the $\tilde{\Omega}(T^{\frac{2}{3}}d^{\frac{1}{3}})$ lower bound up to logarithmic factors. For the LV problem, we have proposed an $\tilde{O}(T^{\frac{3}{4}})$-regret algorithm *D2-EXP4* along with an $\tilde{\Omega}(T^{\frac{2}{3}})$ lower bound proof even with stochastic, Lipschitz and unimodal assumptions, and both of them substantially improve existing results from $O(T^{\frac{2}{3}\cup(1-\alpha)})$ (with smoothness assumptions and indeterministic α) and $\Omega(T^{\frac{3}{5}})$ respectively. Both Linear-EXP4 and D2-EXP4 allow adversarial features. Besides, we have discussed the prospective generalization of our results and the development of future research in feature-based dynamic pricing.

3.9 Proofs

Here we present the proof details of Theorem 3.5.2 and Theorem 3.5.3.

3.9.1 Proof of Regret of D2-EXP4: Theorem 3.5.2

Proof. For any specific LV problem that is defined with linear parameter θ^* and noise CDF F, we define another parameter $\hat{\theta}^* := \lceil \theta^* \rceil_\Delta$ and another CDF functions $\hat{F}$:

$$\hat{F}(x) = \lfloor F(x) \rfloor_\gamma \text{ when } x = i \cdot \gamma \text{ for } i \in \mathbb{Z}, \text{ and linearly connecting}$$
$$\hat{F}(i\gamma) \text{ with } \hat{F}(i+1\gamma) \text{ when } x \in (i\gamma, (i+1)\gamma).$$

Notice that $\hat{F} \in \mathcal{F}_\gamma, \hat{\theta}^* \in \Omega_{\Delta,d}$, and our goal is to prove that $\pi(x; \hat{\theta}^*, \hat{F})$ is good enough to mach the regret. With these two definitions, we might furthermore define a few amounts: $\hat{u} = x^\top \hat{\theta}^*, w^*(u) = \text{argmax}_w\, g(u+w, u, F), \hat{w}^*(u) = \text{argmax}_w\, g(u+w, u, F), \hat{w}(\hat{u}) = \text{argmax}_w\, g(\hat{u}+w, \hat{u}, \hat{F})$. Therefore, the price our algorithm proposed for feature x is $\hat{v}(x) = \lfloor \hat{u} \rfloor_\gamma - (B+1)\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma$, and our goal is to prove that $g(\hat{v}, u, F) \geq g(u+w^*(u), u, F) - C \cdot \gamma$ for some constant C. Since $\hat{\theta}^* := \lceil \theta^* \rceil_\Delta$, we have $\|\theta^*\|_2 \leq \|\hat{\theta}^*\|_2 \leq \|\theta^*\|_2 + \Delta\sqrt{d} = \|\theta^*\|_2 + \gamma$ and thus $u - \gamma \leq \hat{u} - \gamma \leq \lfloor \hat{u} \rfloor_\gamma \leq \hat{u} \leq u + B\gamma$. Based on this, we may get rid of $\lfloor \hat{u} \rfloor_\gamma$ as follows:

$$\begin{aligned}
& g(\lfloor \hat{u} \rfloor_\gamma - (B+1)\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma, u, F) \\
=& (\lfloor \hat{u} \rfloor_\gamma - (B+1)\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma) \cdot (1 - F(\lfloor \hat{u} \rfloor_\gamma - (B+1)\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma - u)) \\
\geq& (\lfloor \hat{u} \rfloor_\gamma + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma) \cdot (1 - F(\lfloor \hat{u} \rfloor_\gamma - (u + (B+1)\gamma) + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)) - (B+1)\gamma \\
\geq& (u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - F(\lfloor \hat{w}(\hat{u}) \rfloor_\gamma)) - (B+2)\gamma.
\end{aligned}$$

Now we target at $\lfloor \hat{w}(\hat{u}) \rfloor_\gamma$ that occurs in both of the price term and the probability term, and we will get rid of it by two steps. Since $\hat{F}(i\gamma) \leq F(i\gamma) \leq \hat{F}(i\gamma) + \gamma$, we have the first

step like:

$$\begin{aligned}
&(u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - F(\lfloor \hat{w}(\hat{u}) \rfloor_\gamma)) \\
\geq &(u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - \hat{F}(\lfloor \hat{w}(\hat{u}) \rfloor_\gamma) - \gamma) \\
\geq &(u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - \hat{F}(\lfloor \hat{w}(\hat{u}) \rfloor_\gamma)) - (B+1)\gamma \\
\geq &(u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - \hat{F}(\hat{w}(\hat{u}))) - (B+1)\gamma.
\end{aligned}$$

Here the second inequality comes from the $(B+1)$ natural bound of any price. Again, we apply $u \geq \hat{u} - B\gamma$ and get the second step:

$$\begin{aligned}
&(u + \lfloor \hat{w}(\hat{u}) \rfloor_\gamma)(1 - \hat{F}(\hat{w}(\hat{u}))) \\
\geq &(\hat{u} - B\gamma + \hat{w}(\hat{u}) - \gamma)(1 - \hat{F}(\hat{w}(\hat{u}))) \\
\geq &(\hat{u} + \hat{w}(\hat{u}))(1 - \hat{F}(\hat{w}(\hat{u}))) - (B+1)\gamma \\
= &g(\hat{u} + \hat{w}(\hat{u}), \hat{u}, \hat{F}) - (B+1)\gamma.
\end{aligned}$$

Now, there are only $\hat{\cdot}$'s instead of $\gamma-$roundings, and we will get rid of those $\hat{\cdot}$'s within some $C \cdot \gamma$ errors. According to the definition of $\hat{w}(\hat{u})$ that it optimizes $g(\hat{u} + w, \hat{u}, \hat{F})$, we further have:

$$\begin{aligned}
&g(\hat{u} + \hat{w}(\hat{u}), \hat{u}, \hat{F}) \\
\geq &g(\hat{u} + \hat{w}^*(u), \hat{u}, \hat{F}) \\
= &(\hat{u} + \hat{w}^*(u))(1 - \hat{F}(\hat{w}^*(u))) \\
\geq &(u + \hat{w}^*(u))(1 - \hat{F}(\hat{w}^*(u))) \\
= &g(u + \hat{w}^*(u), u, \hat{F})
\end{aligned}$$

Finally, according to the definition of $\hat{w}^*(u)$ that it optimizes $g(u+w, u, \hat{F})$, we have:

$$\begin{aligned}
& g(u+\hat{w}^*(u), u, \hat{F}) \\
\geq & g(u+\lfloor w^*(u) \rfloor_\gamma, u, \hat{F}) \\
= & (u+\lfloor w^*(u) \rfloor_\gamma)(1-\hat{F}(\lfloor w^*(u) \rfloor_\gamma)) \\
\geq & (u+\lfloor w^*(u) \rfloor_\gamma)(1-F(\lfloor w^*(u) \rfloor_\gamma)) \\
\geq & (u+w^*(u)-\gamma)(1-F(w^*(u))) \\
\geq & g(u+w^*(u), u, F)-\gamma.
\end{aligned}$$

Here the fourth line is again due to $\hat{F}(i\gamma) \leq F(i\gamma) \leq \hat{F}(i\gamma)+\gamma$ and the non-decreasing property of F. We make a tricky use of $\lfloor \cdot \rfloor_\gamma$ as a "ladder" helping us climb between F and $\hat{F}$, and the ladders only emerge on those $i\gamma$ places as $i \in \mathbb{Z}$. Therefore, we have $g(\hat{v}, u, F) \geq g(u+w^*(u), u, F) - (3B+5) \cdot \gamma$. Since $\gamma = T^{-\frac{1}{4}}$, this would upper bounds the optimality error up to $O(T \cdot \gamma) = O(T^{\frac{3}{4}})$. Also, the EXP-4 agent would cause a regret of $O(\sqrt{T|A| \log \Pi^{LV}_{\Delta,\gamma}}) = O(\sqrt{\frac{T}{\gamma^2} + \frac{Td \log dT}{\gamma}}) = O(T^{\frac{3}{4}} + T^{\frac{5}{8}} d^{\frac{1}{2}} \log dT)$. This completes the proof. ■

3.9.2 Proof of Lower Bound: Theorem 3.5.3

Before the proof begins, we make some necessary definitions. First of all, define a *bump function* as following:

Definition 3.9.1 (Bump function)**.** For $v \in \mathbb{R}+$, we define

$$B(v) = \begin{cases} 0 & v \in (-\infty, 0] \cup [1, +\infty) \\ \exp\{\frac{1}{(3v-1)^2-1}\} & v \in (0, 1/3) \\ 1 & v \in [1/3, 2/3] \\ \exp\{\frac{1}{(3v-2)^2-1}\} & v \in (2/3, 1) \\ 0 & v \in [1, +\infty) \end{cases}$$

as a basic bump function. Then we define a *rescaled bump function*:

$$B_{[a,b]}(v) = B(\frac{v-a}{b-a}).$$

Here we present a lemma on the Lipschitzness of $B(v)$:

Lemma 3.9.2 (Lipschitz continuity of $B(v)$)**.** *$B(v)$ is 6-Lipschitz, i.e., $|B'(v)| \leq 6$. Also, $|B'_{[a,b]}(v)| = |\frac{1}{b-a}B'(\frac{v-a}{b-a})| \leq \frac{6}{b-a}$.*

Proof. According to Definition 3.9.1, we have:

$$B'(v) = \begin{cases} 0 & v \in (-\infty, 0] \\ -\frac{1}{((3v-1)^2-1)^2} \cdot 6(3v-1)\exp\{\frac{1}{(3v-1)^2-1}\} & v \in (0, 1/3) \\ 0 & v \in [1/3, 2/3] \\ -\frac{1}{((3v-2)^2-1)^2} \cdot 6(3v-2)\exp\{\frac{1}{(3v-2)^2-1}\} & v \in (2/3, 1) \\ 0 & v \in [1, +\infty) \end{cases}$$

Now we propose a lemma:

Lemma 3.9.3. *For $t > 1$, we have $\frac{t^2}{e^t} \leq 1$.*

In fact, for both $1 < t \leq \sqrt{e}$ and $t \geq 2$, the inequality is trivial. For $t \in (\sqrt{e}, 2)$, we have $ln(e^t) > \ln(e^{\sqrt{e}}) = \sqrt{e} \cdot 1 > 1.6 > 2 \times 0.7 > 2\ln 2 > 2\ln t = \ln(t^2)$.

Now we denote $t_1 = -\frac{1}{(3v-1)^2-1}, t_2 = -\frac{1}{(3v-2)^2-1}$, and we know that $t_1 > 1$ for $v \in (0, \frac{1}{3})$ and $t_2 > 1$ for $v \in (\frac{2}{3}, 1)$

$$B'(v) = \begin{cases} 0 & v \in (-\infty, 0] \\ -t_1^2 \cdot 6(3v-1)\exp\{-t_1\} & v \in (0, 1/3) \\ 0 & v \in [1/3, 2/3] \\ -t_2^2 \cdot 6(3v-2)\exp\{-t_2\} & v \in (2/3, 1) \\ 0 & v \in [1, +\infty) \end{cases}$$

Given the lemma above, we can immediately see that $-6 \leq B'(v) \leq 6$. This ends the proof of Lemma 3.9.2. ∎

Secondly, we define a series of intervals $[0,1] = [a_0, b_0] \supset [a_1, b_1] \supset \ldots \supset [a_k, b_k] \supset \ldots$, where $b_k = a_k + w_k$, $w_k = 3^{-k!}$. Notice that w_k shrinks even faster than exponential series. Now we describe how to choose $[a_k, b_k]$ from $[a_{k-1}, b_{k-1}]$: We divide the range $[a_{k-1} + \frac{w_{k-1}}{3}, b_{k-1} + \frac{w_{k-1}}{3}]$ into $Q_k = \frac{w_{k-1}}{3w_k}$ sub-intervals of the same length w_k, and then we pick one of these sub-intervals uniformly at random and denote it as $[a_k, b_k]$. It is trivial to see that $[a_1, b_1] = [\frac{1}{3}, \frac{2}{3}]$, $[a_2, b_2] = [\frac{4}{9}, \frac{5}{9}]$.

Thirdly, we define a function:

$$f(v) := C_f \cdot \sum_{k=0}^{\infty} w_k \cdot B_{[a_k, b_k]}(v), \tag{3.7}$$

where $C_f > 0$ is a constant which we will determine later. There are a few properties of $f(v)$ shown in the following lemma:

Lemma 3.9.4. *Define $f(v)$ as Eq. (3.7), and we have:*

1. *There exists a unique $v^* \in [0,1]$ such that $f(v^*) = \max_{v \in [0,1]} f(v)$. In specific, $v^* = \bigcap_{k=1}^{\infty} [a_k, b_k]$.*

2. *$f(v)$ is unimodal.*

3. *For any $v \in [0,1]$, there exists at most one k, such that $B'_{[a_k,b_k]}(v) \neq 0$.*

4. *$f(v) \leq \frac{3}{2}C_f$.*

Proof. To prove 1, we first see that $v^* \in [a_k, b_k], k = 1, 2, \ldots$. Notice that $\lim_{k\to\infty} a_k$ exists (since $\{a_k\}_{k=0}^{\infty}$ is increasing and upper bounded) and that $\lim_{k\to\infty}(b_k - a_k) = \lim_{k\to\infty} 3^{-k!} = 0$. Therefore, $\Cap_{k=1}^{\infty}[a_k, b_k]$ is a unique real number within $[\frac{1}{3}, \frac{2}{3}]$.

To prove 2, notice that every $B_{[a_k,b_k]}(v)$ is non-decreasing in $[0, v^*]$ and non-increasing in $[v^*, 1]$.

To prove 3, consider the case when $B'_{[a_k,b_k]}(v) \neq 0$, and we know that: $(1) v \in [a_k, b_k] \subset [a_{k-1} + \frac{w_{k-1}}{3}, b_{k-1} - \frac{w_{k-1}}{3}] \subset [a_{k-2} + \frac{w_{k-2}}{3}, b_{k-2} - \frac{w_{k-2}}{3}] \subset \ldots \subset [a_0 + \frac{w_0}{3}, b_0 - \frac{w_0}{3}] = [\frac{1}{3}, \frac{2}{3}]$. Since $B'_{[a_j,b_j]}(v) = 0, v \in [a_j + \frac{w_j}{3}, b_j - \frac{w_j}{3}]$, we know that $B'_{[a_j,b_j]}(v) = 0, j = 0, 1, \ldots, k-1$. (2) $v \notin [a_{k+1}, b_{k+1}] \supset [a_{k+2}, b_{k+2}] \supset \ldots$, and we know that $B'_{[a_i,b_i]}(v) = 0, i = k+1, k+2, \ldots$.

To prove 4, just notice that $B(v) \leq 1$ and thus $f(v) \leq C_f \cdot \sum_{k=0}^{\infty} 3^{-k!} \leq C_f \cdot \sum_{k=0}^{\infty} 3^{-k!} = \frac{3}{2}C_f$.

■

According to Lemma 3.9.4 Property 3, we have:

$$\begin{aligned}
|f'(v)| &\leq C_f \max_{v\in[a_k,b_k],k=1,2,\ldots} |w_k \cdot B'_{[a_k,b_k]}(v)| \\
&= C_f \max_{v\in[a_k,b_k],k=1,2,\ldots} |w_k \cdot B'(\frac{v-a_k}{b_k-a_k}) \cdot \frac{1}{w_k}| \\
&\leq C_f \max_y |B'(y)| \\
&\leq 6C_f
\end{aligned}.$$

This holds for any $v \in [0, v^*) \cup (v^*, 1]$. Now we define another function $G(v)$[7]:

$$G(v) = 1 - \frac{1}{f(v) + 1}, v \in [0, 1]. \tag{3.8}$$

According to Lemma 3.9.4 that reveals the properties of $f(v)$, we have a similar lemma on $G(v)$:

Lemma 3.9.5. *Define $G(v)$ as Eq. (3.8), and we have the following properties:*

1. $G(0) = 0, G(1) = 0, 0 < G(v) < 1 for v \in (0, 1)$.
2. $G(v)$ *is unimodal in* $[0, 1]$.
3. $G'(v) = \frac{f'(v)}{(f(v)+1)^2} \Rightarrow |G'(v)| \leq 6C_f$.

The proof of Lemma 3.9.5 is trivial.

Notice that $G(v)$ is not necessarily a revenue curve, since $\frac{G(v)}{v}$ is not necessarily decreasing (and thus not a "survival function"). However, we can construct a revenue curve $D(v) : [0, 1] \to [0, 1]$ via an affine transformation:

$$D(v) = \begin{cases} v & v \in [0, b] \\ b + (1 - b)G(\frac{v-b}{1-b}) & v \in (b, 1]. \end{cases} \tag{3.9}$$

Here $b = \frac{6C_f+1}{2} \in (0, 1)$, and therefore $C_f < \frac{1}{6}$. An illustration of the transformation from $f(v)$ (the upper figure) to $D(v)$ (the lower figure) is shown in Figure 1. The monotonicity in each interval is not changed, while the rate of $\frac{\mathbb{E}[r(v)]}{v}$ is non-increasing after these transformations. As is mentioned above, the homothetic transformation with center $(1, 1)$ ensures a non-increasing property of $\frac{D(v)}{v}$. Here we denote $d(v) := \frac{D(v)}{v}$. To show that

[7]Here G stands for "gain", which is different from the revenue curve to be introduced later.

$D(v)$ is a revenue curve, we expand the definition of $d(v)$ to $\mathbb{R}$ as follows:

$$d(v) = \begin{cases} 1 & v \in (-\infty, 0] \\ \frac{D(v)}{v} & v \in (0, 1) \\ 0 & v \in [1, +\infty). \end{cases} \tag{3.10}$$

Now we claim that there exists a random variable $Y > 0$ such that $d(v) = \mathbb{P}[Y \geq v]$. To show this, it is sufficient to prove the following lemma:

Lemma 3.9.6. *For $d(v)$ defined in Eq. (3.10), we have the following properties:*

1. *$d(v)$ is non-increasing on $\mathbb{R}$.*
2. *$d(v)$ is continuous at $v = 0$, i.e. $d(0) = \lim_{v \to 0^+} d(v) = 1$.*
3. *$d(v) \geq 0, v \in [0, 1]$.*

Proof. According to Eq. (3.9) and Eq. (3.10), we have:

$$d(v) = \begin{cases} 1 & v \in (-\infty, b] \\ \frac{b}{v} + \frac{1-b}{v} \cdot G(\frac{v-b}{1-b}) & v \in (b, 1) \\ 0 & v \in [1, +\infty). \end{cases}$$

Therefore, we take the derivatives of $d(v)$ and get:

$$\frac{\partial d(v)}{v} = \begin{cases} o & v \in (-\infty, 0) \cup (0, b) \\ -\frac{b - v \cdot G'(\frac{v-b}{1-b}) + (1-b)G(\frac{v-b}{1-b})}{v^2} & v \in (b, 1) \\ 0 & v \in (1, +\infty). \end{cases}$$

Notice that

$$\begin{aligned} b - v \cdot G'(\frac{v-b}{1-b}) &\geq b - |v| \cdot |G'(\frac{v-b}{1-b})| \\ &\geq b - 1 \cdot 6C_f \\ &= \frac{6C_f + 1}{2} - 6C_f \\ &= \frac{1 - 6C_f}{2} \\ &> 0. \end{aligned}$$

The last inequality comes from the fact that $C_f < \frac{1}{6}$. Therefore, $d'(v)$ is non-positive in $(-\infty, 0) \cup (0, b) \cup (b, 1) \cup (1, +\infty)$. Also, $d(v)$ is continuous at $v = 0, v = b$ and $\lim_{v \to 1^-} = b > 0 = \lim_{v \to 1^+}$, we know that $d(v)$ is always non-increasing on $\mathbb{R}$. ■

Notice that there is a bijection between each $\{[a_k, b_k]\}_{k=0}^{\infty}$ series and each $f(v)$, and correspondingly each $G(v)$, $D(v)$ and $d(v)$. Still, a bijection lies between each $d(v)$ and the distribution $\mathbb{P}[Y \geq v]$ of the customers' valuation. Therefore, we will take $d(v)$ to represent this distribution.

With all preparations done above, we are now able to prove Theorem 3.5.3. Specifically, we will proof the theorem on an infinite series of $n_1, n_2, \ldots$, where $n_k = \lceil \frac{1}{k}(\frac{w_{k-1}}{w_k^3}) \rceil$. Consider the possible $Q_k = \frac{w_{k-1}}{3w_k}$ choices of $[a_k, b_k]$, and denote these intervals as $I_j, j = 1, 2, \ldots, Q_k$. If $[a_k, b_k] = I_j$, then we denote the corresponding $f(v), G(v), D(v), d(v)$ functions as $f_j(v), G_j(v), D_j(v)$ and $d_j(v)$ sequentially. Meanwhile, if we *do not* make any choice of $[a_k, b_k]$, and then we just have a finite series of intervals $[0, 1] = [a_0, b_0] \supset [a_1, b_1] \supset [a_2, b_2] \supset \ldots \supset [a_{k-1}, b_{k-1}]$, and then we can define a $f_0(v) = C_f \cdot \sum_{j=0}^{k-1} w_j \cdot B_{[a_j, b_j]}(v)$, and can also define corresponding $G_0(v), D_0(v), d_0(v)$ based on $f_0(v)$.

Now, consider the pricing feedbacks in total n rounds (where we denote n_k as n for simplicity). Define a *feedback vector* $\mathbf{r}_n \in \{0, 1\}^n$, denoting the outcome of a deterministic

policy interacting with the revenue curve. We claim that for $t = 1, 2, \ldots, n$, a vector $\mathbf{r}_t$ is sufficient for any deterministic policy to generate a price v_{t+1}, because $\mathbf{r}_i$ is a prefix of $\mathbf{r}_j$ when $i \leq j$. For any policy π, denote the probability of $\mathbf{r}_n$'s occurence as $\mathbb{P}_j(\mathbf{r}_n)$ under the distribution d_j, or $\mathbb{P}_0(\mathbf{r}_n)$ under the distribution d_0. Denote the series of prices that π has generated as $\{v_t\}, t = 1, 2, \ldots, n$, and we may assume $v_t \geq b$ without losing generality (as $0 \leq v < b$ is always suboptimal). Then, for any function $h : \{0,1\}^n \to [0, M]$, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_j}[h(\mathbf{r}_n)] - \mathbb{E}_{\mathbb{P}_0}[h(\mathbf{r}_n)] \\
=&\sum_{\mathbf{r}_n} h(\mathbf{r}_n) \cdot (\mathbb{P}_j[\mathbf{r}_n] - \mathbb{P}_0[\mathbf{r}_n]) \\
\leq&\sum_{\mathbf{r}_n : \mathbb{P}_j[\mathbf{r}_n] \geq \mathbb{P}_0[\mathbf{r}_n]} h(\mathbf{r}_n)(\mathbb{P}_j[\mathbf{r}_n] - \mathbb{P}_0[\mathbf{r}_n]) \\
\leq&M \cdot \sum_{\mathbf{r}_n : \mathbb{P}_j[\mathbf{r}_n] \geq \mathbb{P}_0[\mathbf{r}_n]} h(\mathbf{r}_n)(\mathbb{P}_j[\mathbf{r}_n] - \mathbb{P}_0[\mathbf{r}_n]) \\
=&\frac{M}{2}\|\mathbb{P}_j - \mathbb{P}_0\|_1 \\
\leq&\frac{M}{2}\sqrt{2\ln 2 \cdot KL(\mathbb{P}_0||\mathbb{P}_j)}.
\end{aligned}$$

The last line comes from Lemma 11.6.1 in Cover & Thomas, Elements of Information Theory, where KL stands for the KL-divergence. Since

$$\begin{aligned}
KL(\mathbb{P}_0(\mathbf{r}_n)||\mathbb{P}_j(\mathbf{r}_n)) =& \sum_{t=1}^{n} KL(\mathbb{P}_0[r_t|\mathbf{r}_{t-1}]||\mathbb{P}_j[r_t|\mathbf{r}_{t-1}]) \\
=& \sum_{t=1}^{t} \mathbb{P}_0(\frac{v_t - b}{1-b} \notin I_j) \cdot 0 + \mathbb{P}_0(\frac{v_t - b}{1-b} \in I_j) \cdot KL(\frac{D_0(v_t)}{v_t}||\frac{D_j(v_t)}{v_t}).
\end{aligned}$$

The first equality comes from the chain rule of decomposing a KL-divergence. The second equality is becauser_t is a Bernoulli random variable that satisfies $Ber(\frac{D_0(v_t)}{v_t})$ under $\mathbb{P}_0$, or $Ber(\frac{D_j(v_t)}{v_t})$ under $\mathbb{P}_j$. Denote $\mu_t := \frac{v_t - b}{1-b}$ for simplicity. Notice that if $v_t \in I_j$, then we

have:

$$\begin{aligned}
\frac{D_j(v_t)}{v_t} - \frac{D_0(v_t)}{v_t} &= \frac{(1-b)\cdot G_j(\mu_t)}{v_t} - \frac{(1-b)\cdot G_0(\mu_t)}{v_t} \\
&= \frac{1-b}{v_t}(G_j(\mu_t) - G_0(\mu_t)) \\
&= \frac{1-b}{v_t}\Big(\frac{1}{f_0(\mu_t)+1} - \frac{1}{f_j(\mu_t)+1}\Big) \\
&= \frac{1-b}{v_t}\frac{f_j(\mu_t) - f_0(\mu_t)}{(f_0(\mu_t)+1)(f_j(\mu_t)+1)} \\
&= \frac{1-b}{v_t}\frac{C_f\sum_{i=k}^{\infty} w_i \cdot B_{[a_i,b_i]}(\mu_t)}{(f_0(\mu_t)+1)(f_j(\mu_t)+1)} \\
&\le \frac{1-b}{b}\cdot\frac{C_f\cdot 2w_k}{1\times 1} \\
&\le 1\cdot 2C_f w_k \\
&\le \frac{w_k}{3}
\end{aligned}$$

Here the third last inequality comes from $v_t \ge b, f_0(v) \ge 0, f_j(v) \ge 0$, and the fact that

$$\sum_{i=k}^{\infty} w_i B_{[a_i,b_i]}(\mu_t) \le \sum_{i=k}^{\infty} 3^{-i!}\cdot 1 \le 3^{-k!}\sum_{i=0}^{\infty} 3^{-i} \le \frac{2}{3}\cdot 3^{-k!} < 2w_k.$$

The second last inequality comes from $b = \frac{6C_f+1}{2} \ge \frac{1}{2}$. The lastest inequality comes from the fact that $6C_f < 1$.

Now we propose a lemma:

Lemma 3.9.7. *For Bernoulli distributions $Ber(p)$ and $Ber(p+\epsilon)$ with $\frac{1}{2} \le p \le p+\epsilon \le \frac{1}{2} + C$, we have*

$$KL(p||p+\epsilon) \le \frac{1}{\ln 2}\frac{4}{1-4C^2}\epsilon^2.$$

Proof.

$$\begin{aligned}
KL(p||p+\epsilon) &= p\log(\frac{p}{p+\epsilon}) + (1-p)log(\frac{1-p}{1-p-\epsilon}) \\
&= \frac{1}{\ln 2}\cdot(p(-\ln(1+\frac{\epsilon}{p})) + (1-p)\ln(1+\frac{\epsilon}{1-p-\epsilon})) \\
&\leq \frac{1}{\ln 2}\cdot(p(-\frac{\epsilon}{p+\epsilon}) + (1-p)\frac{\epsilon}{1-p-\epsilon}) \\
&= \frac{1}{\ln 2}\cdot\frac{\epsilon^2}{(p+\epsilon)(1-p-\epsilon)} \\
&\leq \frac{1}{\ln 2}\cdot\frac{\epsilon^2}{(\frac{1}{2}+C)(\frac{1}{2}-C)} \\
&\leq \frac{1}{\ln 2}\cdot\frac{1}{\frac{1}{4}-C^2}\epsilon^2.
\end{aligned}$$

Here the third line comes from the fact that $\frac{v}{1+v} \leq \ln v \leq v$. ■

Let us come back to the proof of the theorem. Since $\frac{D_0(v_t)}{v_t} \geq b \geq \frac{1}{2}$, and the fact that

$$\begin{aligned}
\frac{D_j(v_t)}{v_t} &\leq \frac{D_j(b+\frac{1-b}{3})}{b+\frac{1-b}{3}} \\
&= 3\cdot\frac{b+(1-b)(1-G_j(\frac{1}{3}))}{1+2b} \\
&= 3\cdot\frac{b+(1-b)\frac{f_j(\frac{1}{3})}{f_j(\frac{1}{3})+1}}{1+2b} \\
&= 3\cdot\frac{b+(1-b)\frac{C_f}{C_f+1}}{1+2b}.
\end{aligned}$$

The first inequality is because $\frac{D(v)}{v}$ is non-increasing and the fact that $v_t \geq b+\frac{1-b}{3}$ if $v_t \in [a_k, b_k], k \geq 1$. The last equality comes from the fact that $f_j(\frac{1}{3}) = C_f$. Now we specify the constants: let $C_f = \frac{1}{60}, b = \frac{6C_f+1}{2} = \frac{11}{20}$. Plug in these constant values and we get:

$$\frac{D_j(v_t)}{v_t} \leq \frac{340}{427} < \frac{5}{6}.$$

According to Lemma 3.9.7, we have:

$$KL(\frac{D_0(v_t)}{v_t}||\frac{D_j(v_t)}{v_t}) \leq \frac{1}{\ln 2}\cdot\frac{4}{1-4\cdot(\frac{5}{6}-\frac{1}{2})^2}\cdot(\frac{w_k}{3})^2 = \frac{1}{\ln 2}\cdot\frac{36}{5}\cdot\frac{w_k^2}{9} = \frac{1}{\ln 2}\cdot\frac{4w_k^2}{5}.$$

. Recall that

$$\begin{aligned}
&KL(\mathbb{P}_0(\mathbf{r}_n)||\mathbb{P}_j(\mathbf{r}_n)) \\
=&\sum_{t=1}^{t}\mathbb{P}_0(\frac{v_t-b}{1-b}\in I_j)\cdot KL(\frac{D_0(v_t)}{v_t}||\frac{D_j(v_t)}{v_t}) \\
\leq&\frac{1}{\ln 2}\cdot\frac{4}{5}w_k^2\cdot\sum_{t=1}^{n}\mathbb{P}_0[\mu_t\in I_j].
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_j}[h(\mathbf{r}_n)]-\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{r}_n)] \\
\leq&\frac{M}{2}\sqrt{2\ln 2\cdot\frac{1}{\ln 2}\cdot\frac{4}{5}w_k^2\cdot\sum_{t=1}^{n}\mathbb{P}_0[\mu_t\in I_j]} \\
\leq&\frac{4M\cdot w_k}{5}\cdot\sqrt{\sum_{t=1}^{n}\mathbb{P}_0[\mu_t\in I_j]}.
\end{aligned}$$

Now , let $h(\mathbf{r}_n)$ be $N_j=|\{t|\mu_t\in I_j, t=1,2,\ldots,n_k\}|$, and we know that $M=n_k$. Since $n_k=\lceil\frac{1}{k}\frac{w_{k-1}}{w_k^3}\rceil$, we conduct the pricing for n_k times and have:

$$\mathbb{E}_{\mathbb{P}_j}[N_j]-\mathbb{E}_{\mathbb{P}_0}[N_j]\leq\frac{4M\cdot w_k}{5}\cdot\sqrt{\sum_{t=1}^{n}\mathbb{P}_0[\mu_t\in I_j]}=\frac{4M\cdot w_k}{5}\cdot\sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]}.$$

Sum over $j=1,2,\ldots,Q_k$ of the inequality above, and we take an average to get:

$$\begin{aligned}
\frac{1}{Q_k}\cdot\sum_{j=1}^{Q_k}\mathbb{E}_{\mathbb{P}_j}[N_j]&\leq\frac{1}{Q_k}\sum_{j=1}^{Q_k}\mathbb{E}_{\mathbb{P}_0}[N_j]+\frac{1}{Q_k}\frac{4}{5}n_k\cdot w_k\sum_{j=1}^{Q_k}\sqrt{\mathbb{E}_{\mathbb{P}_0}[N]} \\
&=\frac{1}{Q_k}\cdot n_k+\frac{1}{Q_k}\frac{4}{5}n_k\cdot w_k\sum_{j=1}^{Q_k}\sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]} \\
&\leq\frac{n_k}{Q_k}+\frac{4}{5}\frac{n_k}{Q_k}\cdot w_k\cdot\sqrt{Q_k\cdot\sum_{j=1}^{Q_k}\mathbb{E}_{\mathbb{P}_0}[N_j]} \\
&=\frac{n_k}{Q_k}+\frac{4}{5}\frac{n_k}{Q_k}\cdot w_k\cdot\sqrt{Q_kn_k} \\
&\leq\frac{3}{k}\cdot\frac{4}{5}\frac{3}{k}\frac{1}{w_k^2}\sqrt{\frac{3}{k^2}\cdot\frac{w_{k-1}^2}{w_k^4}} \\
&=\frac{3}{k}\frac{1}{w_k^2}+\frac{4\sqrt{3}}{5\sqrt{k}}\frac{1}{k}\frac{w_{k-1}}{w_k^3} \\
&\leq 0.9\cdot n_k,\ \text{for } k\geq 3.
\end{aligned}\tag{3.11}$$

In Eq. (3.11), the first line comes from the summation; the second (and the fourth) line is because $\sum_{j=1}^{Q_k} \mathbb{E}_{\mathbb{P}_0}[N_j] = \mathbb{E}_{\mathbb{P}_0}[\sum_{j=1}^{Q_k} N_j] = n_k$; the third line is an application of Cauchy-Schwartz's Inequality; the fifth line is derived by plugging in $Q_k = \frac{w_{k-1}}{3w_k}, n_k \leq \frac{1}{k} \cdot \frac{w_{k-1}}{w_k^3}, w_k = 3^{-k!}$; the last line is just calculations. Therefore, under distribution d_j, the policy π is expected to choose an $v_t \notin I_j$ for at least $0.1n_k$ times, which will bring a regret $0.1n_k \cdot C_j \cdot w_k = \frac{1}{600} n_k \cdot w_k = \frac{1}{k} \cdot w_k^{\frac{1}{k}-2}$. Since $n_k = \frac{1}{k} \cdot w_k^{\frac{1}{k}-3}$, we know that $Regret = \Omega((n_k)^{\frac{2}{3}-\frac{1}{3k}})$ up to logarithmic factors. Therefore, we claim that for any $\delta > 0$, no policy can achieve $o(n_k^{\frac{2}{3}-\delta})$ for sufficiently large k.

This ends the proof of Theorem 3.5.3.

Chapter 4

Pricing with Contextual Elasticity and Heteroscedastic Valuation

Price elasticity indicates how customers' demand responds to price changes for a specific product. However, our previous model that employ a linear noisy valuation with i.i.d. noise fails to capture the variability in elasticity across different products. This chapter introduces an advanced approach to modeling customer demand by integrating feature-based price elasticity, represented equivalently as a valuation with heteroscedastic noise. To solve the problem, we propose a computationally efficient algorithm called "Pricing with Perturbation (PwP)", which enjoys an $O(\sqrt{dT \log T})$ regret while allowing arbitrary adversarial input context sequences. We also prove a matching lower bound at $\Omega(\sqrt{dT})$ to show the optimality regarding d and T (up to $\log T$ factors). Our results shed light on the relationship between contextual elasticity and heteroscedastic valuation, offering valuable insights for developing effective and practical pricing strategies.

4.1 Introduction

Contextual pricing, a.k.a., Feature-based dynamic pricing, considers the problem of setting prices for a sequence of highly specialized or individualized products. With the growth of e-commerce and the increasing popularity of online retailers as well as customers, there has been a growing interest in this area [see, e.g., Amin et al., 2014, Qiang and Bayati, 2016, Javanmard and Nazerzadeh, 2019, Shah et al., 2019, Cohen et al., 2020, Xu and Wang, 2021, Bu et al., 2022].

Formulated as a learning problem, the seller has no prior knowledge of ideal prices but is expected to learn on the fly by exploring different prices and adjusting their pricing strategy after collecting every demand feedback from customers. Different from non-contextual dynamic pricing [Kleinberg and Leighton, 2003] where identical products are sold repeatedly, a contextual pricing agent is expected to generalize from one product to another in order to successfully price a previously-unseen product. A formal problem setup is described below:

Contextual pricing. For $t = 1, 2, ..., T$:

1. A product occurs, described by a context $x_t \in \mathbb{R}^d$.
2. The seller (we) proposes a price $p_t \geq 0$.
3. The customer reveals a demand $0 \leq D_t \leq 1$.
4. The seller gets a reward $r_t = p_t \cdot D_t$.

Here T is the time horizon, and the (random) demand D_t is drawn from a distribution determined by context (or feature) x_t and price p_t. The sequence of contexts $\{x_t\}$ can be either independently and identically distributed (iids) or chosen arbitrarily by an adversary. The seller's goal is to minimize the cumulative *regret* against the sequence of

optimal prices.

Existing works on contextual pricing usually assumes linearity on the demand, but they fall into two camps. On the one hand, the "linear demand" camp [Qiang and Bayati, 2016, Ban and Keskin, 2021, Bu et al., 2022] assumes the *demand* D_t as a (generalized) linear model. A typical model is $D_t = \lambda(\alpha p_t + x_t^T \beta) + \epsilon_t$. Here $\alpha < 0$ is a parameter closely related to the *price elasticity*. We will rigorously define a price elasticity in Section 4.8.1 according to Parkin et al. [2002], where we also show that α is the *coefficient of elasticity*. Besides of α, other parameters like $\beta \in \mathbb{R}^d$ captures the base demand of products with feature x_t, ϵ_t is a zero-mean demand noise, and λ is a known monotonically increasing link function. With this model, we have a noisy observation on the expected demand, which is reasonable as the same product is offered many times in period t. On the other hand, the "linear valuation" camp [Cohen et al., 2020, Javanmard and Nazerzadeh, 2019, Xu and Wang, 2021] models a buyer's *valuation* y_t as linear and assumes a binary demand $D_t = \mathbb{1}[p_t \leq y_t]$. All three works listed above assume a *linear-and-noisy* model with $y_t = x_t^\top \theta^* + N_t$, where $\theta^* \in \mathbb{R}^d$ is an unknown linear parameter that captures common valuations and N_t is an idiosyncratic noise assumed to be iid.

Interestingly, the seemingly different modeling principles are closely connected to each other. In the "linear valuation" camp, notice that a customer's probability of "buying" a product equals $\mathbb{E}[D_t]$, which is further given by

$$\mathbb{E}[D_t|p] = \mathbb{P}[y_t \geq p] := S(p - x_t^\top \theta^*),$$

where S is the survival function of N_t (i.e. $S(w) = 1 - \mathrm{CDF}(w)$ for $w \in \mathbb{R}$). This recovers a typical linear demand model by taking $\lambda(w) = S(-w)$ with $\alpha = -1$ and $\beta = \theta^*$. In

other words, the distribution of N_t completely characterizes the demand function $\lambda(\cdot)$ and vice versa.

However, the "linear demand" camp is not satisfied with a fixed $\alpha = -1$, while the "linear valuation" camp are skeptical about an observable demand D_t even with zero-mean iid noise. One common limitation to both models is that neither captures how feature x_t affects the price elasticity.

Our model. To address this issue, we propose a natural model that unifies the perspectives of both groups. Also, we resolve the common limitation by modeling *heteroscedasticity*, where we assume that the elasticity coefficient α is linearly dependent on feature x_t. This contextual modeling originates from the fact that different products have different price elasticities [Anderson et al., 1997].

In specific, we assume:

$$D_t \sim \mathrm{Ber}(S(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*)), \tag{4.1}$$

which adopts a generalized linear demand model (GLM) and a Boolean-censored feedback simultaneously. From the perspective of valuation model, it is *equivalent* to assume

$$D_t = \mathbb{1}[p_t \le y_t], \text{ where } y_t = \frac{1}{x_t^\top \eta^*} \cdot (x_t^\top \theta^* + N_t) \text{ and } \mathrm{CDF}_{N_t}(w) = 1 - S(w). \tag{4.2}$$

Although Eq. (4.1) seems more natural than Eq. (4.2), they are equivalent to each other (with reasonable assumptions on S). Notice that the random valuation y_t is *heteroscedastic*, which means its variance is not the same constant across a variety of x_t's. We provide a detailed interpretation of this linear fractional valuation model in appendix.

4.1.1 Contributions.

Our main results are twofold.

1. We propose a new demand model that assumes a feature-dependent price elasticity on every product. Equivalently, we model the heteroscedasticity on customers' valuations among different products. This model unifies the "linear demand" and "linear valuation" camps.

2. We propose a "Pricing with Perturbation (PwP)" algorithm that achieves $O(\sqrt{dT \log T})$ regret on this model, which is optimal up to $\log T$ factors. This regret upper bound holds for both i.i.d. and adversarial $\{x_t\}$ sequences.

4.1.2 Technical Novelty

To the best of our knowledge, we are the first to study a contextual pricing problem with heteroscedastic valuation and Boolean-censored feedback. Some existing works, including Javanmard and Nazerzadeh [2019], Miao et al. [2019], Ban and Keskin [2021], Wang et al. [2021a], focus on related topics and achieve theoretical guarantees. However, their methodologies are not applicable to our settings due to substantial obstacles, which we propose novel techniques to overcome.

Randomized surrogate regret. Xu and Wang [2021] solves the problem with $x_t^\top \eta^* = 1$, by taking the negative log-likelihood as a surrogate regret and running an optimization oracle that achieves a fast rate (i.e. an $O(\log T)$ regret). However, the log-likelihood is no longer a surrogate regret in our setting, since it is not "convex enough" and therefore cannot provide sufficient (Fisher) information. In this work, we overcome this challenge by constructing a *randomized* surrogate loss function, whose *expectation* is "strongly convex"

enough to upper bound the regret.

OCO for adversarial inputs. Javanmard and Nazerzadeh [2019] and Ban and Keskin [2021] study the problem with unknown or heterogeneous noise variances (i.e. elasticity coefficients), but their techniques highly rely on the distribution of the feature distributions. As a result, their algorithm could be easily attacked by an adversarial $\{x_t\}$ series. In our work, we settle this issue by conducting an online convex optimization (OCO) scheme while updating parameters. Instead of estimating from the history that requires sufficient randomness in the inputs, our algorithm can still work well for adversarial inputs.

In addition, our algorithm has more advanced properties such as computational efficiency and information-theoretical optimality. For more highlights of our algorithm, please refer to Section 4.4.1.

4.2 Related Works

Here we present a review of the pertinent literature on contextual pricing and heteroscedasticity in machine learning, aiming to position our work within the context of related studies. For more related works on non-contextual pricing, contextual pricing, contextual searching and contextual bandits, please refer to Wang et al. [2021b], Xu and Wang [2021], Krishnamurthy et al. [2021] and Zhou [2015] respectively.

Contextual Pricing. As we mentioned in Section 4.1.2, there are a large number of recent works on contextual dynamic pricing problems, and we refer to Ban and Keskin [2021] as a detailed introduction. On the one hand, Qiang and Bayati [2016], Nambiar et al. [2019], Miao et al. [2019], Wang et al. [2021a], Ban and Keskin [2021], Bu et al.

[2022] assume a (generalized) linear demand model with noise, i.e. $\mathbb{E}[D_t] = g(\alpha p_t - \beta^\top x_t)$. Among those papers, Miao et al. [2019] worksl with a fixed α while we assume α as context-dependent. Wang et al. [2021a] and Ban and Keskin [2021] are the closest to our problem setting, which consist of a generalized linear demand model and noisy observations. On the one hand, Ban and Keskin [2021] assumes independent add-on noises (while we allow binary martingale observations). With the development of a least-square estimator, they present an algorithm that achieves $\tilde{O}(s\sqrt{T})$ regret (with s being the sparsity factor). On the other hand, Wang et al. [2021a] further gets rid of the independence among noises and allow them to be idiosyncratic. They proposes a UCB-based algorithm with $\tilde{O}(d\sqrt{T})$ regret and another Thompson-Sampling-based algorithm with $\tilde{O}(d^{\frac{3}{2}}\sqrt{T})$ regret, both of which are sub-optimal in d. Moreover, all works mentioned above assume the context sequence $\{x_t\}$ to be i.i.d., whereas we consider it "too good to be true" and work towards an algorithm adaptive to adversarial input sequences. On the other hand, Golrezaei et al. [2019], Shah et al. [2019], Cohen et al. [2020], Javanmard and Nazerzadeh [2019], Xu and Wang [2021], Fan et al. [2021], Goyal and Perivier [2021], Luo et al. [2022] adopts the linear valuation model $y_t = x_t^\top \theta^* + N_t$, which is a special case of our model as $x_t^\top \eta^* = 1$. Specifically, both Javanmard and Nazerzadeh [2019] and Xu and Wang [2021] achieve an $O(d \log T)$ regret with N_t drawn from a known distribution with $x_t^\top \eta^* = -1$. Javanmard and Nazerzadeh [2019] also studies the setting when $x_t^\top \eta^*$ is fixed but unknown and achieves $O(d\sqrt{T})$ regret for stochastic $\{x_t\}$ sequences. In comparison, we achieve $O(\sqrt{dT \log T})$ on a more general problem and get rid of those assumptions. For a more detailed discussion, please refer to Section 4.2.

Heteroscedasticity. Since the valuation noise is scaled by a $\frac{1}{x_t^\top \eta^*}$ coefficient, the valuation is *heteroscedastic*, referring to a situation where the variance is not the same

	Known α	**Unknown fixed** α		**Heteroscedastic** $\alpha = x_t^\top \eta^*$	
Features	Stochastic & Adversarial	Stochastic	Adversarial	Stochastic	Adversarial
Upper Bound	$d \log T$ [XW21]	$d\sqrt{T}$ [JN19]	$? \Rightarrow \sqrt{dT}$ **Our Work**	$s\sqrt{T}$ (independent noises) [BK21] $d\sqrt{T}$ (idiosyncratic noises) [WTL21]	$? \Rightarrow \sqrt{dT}$ **Our Work**
Lower Bound	$d \log T$ [BR12]	$\sqrt{T}$ [JN19]		$\sqrt{T} \Rightarrow \sqrt{dT}$ **Our Work**	

Table 4.1: Existing related literature and results on regret bounds, with $\tilde{O}(\cdot)$ dropped. Note that each adversarial setting covers the stochastic setting under the same assumptions. Here [XW21] stands for Xu and Wang [2021], [JN19] for Javanmard and Nazerzadeh [2019], [BR12] for Broder and Rusmevichientong [2012], [BK21] for Ban and Keskin [2021], and [WTL21] for Wang et al. [2021a].

constant across all observations. Heteroscedasticity may lead to bias estimates or loss of sample information. There are several existing methods handling this problem, including weighted least squares method [Cunia, 1964], White's test [White, 1980] and Breusch-Pagan test [Breusch and Pagan, 1979]. Furthermore, Anava and Mannor [2016] and Chaudhuri et al. [2017] study online learning problems with heteroscedastic variances and provide regret bounds. For a formal and detailed introduction, we refer the audience to the textbook of Kaufman [2013].

4.3 Problem Setup

4.3.1 Notations

To formulate the problem, we firstly introduce necessary notations and symbols used in the following sections. The sales session contains T rounds with T known to the seller in advance[1]. At each time $t = 1, 2, \ldots, T$, a product with feature $x_t \in \mathbb{R}^d$ occurs and we propose a price $p_t \geq 0$. Then the nature draws a demand $D_t \sim \text{Ber}(S(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*))$, where $\theta^*, \eta^* \in \mathbb{R}^d$ are fixed unknown linear parameters and the link function $S : \mathbb{R} \to [0, 1]$ is non-increasing. By the end of time t, we receive a reward $r_t = p_t \cdot D_t$.

Equivalently, this customer has a valuation $y_t = \frac{x_t^\top \theta^* + N_t}{x_t^\top \eta^*}$ with noise $N_t \in \mathbb{R}$, and then make a decision $\mathbb{1}_t = \mathbb{1}[p_t \leq y_t] = D_t$ after seeing the price p_t. Similarly, we receive a reward $r_t = p_t \cdot \mathbb{1}_t$. Assume $N_t \sim \mathbb{D}_F$ is independently and identically distributed (i.i.d.), with cumulative distribution function (CDF) $F = 1 - S$. Denote $s := S'$ and $f := F'$.

4.3.2 Definitions

Here we define some key quantities. Firstly, we define an expected reward function.

Definition 4.3.1 (expected reward function)**.** Define

$$r(u, \beta, p) := \mathbb{E}[r_t | x_t^\top \theta^* = u, x_t^\top \eta^* = \beta, p_t = p] = p \cdot S(\beta \cdot p - u) \tag{4.3}$$

as the expected reward function.

Given this, we further define a greedy price function as the argmax of $r(u, \beta, p)$ over p.

[1]Here we assume T known for simplicity. For unknown T, we may apply a "doubling epoch" trick as Javanmard and Nazerzadeh [2019] without affecting the regret rate.

Definition 4.3.2 (greedy price function)**.** Define $J(u, \beta)$ as a greedy price function, i.e. the price that maximizes the expected reward given $u = x_t^\top \theta^*$ and $\beta = x_t^\top \eta^*$.

$$J(u, \beta) = \operatorname*{argmax}_{p \in \mathbb{R}} r(u, \beta, p) = \operatorname*{argmax}_{p \in \mathbb{R}} p \cdot S(\beta \cdot p - u) \tag{4.4}$$

Notice that

$$J(u, \beta) = \operatorname*{argmax}_{p} p \cdot S(\beta p - u) = \frac{1}{\beta} \cdot \operatorname*{argmax}_{\beta p} \beta p \cdot S(\beta p - u) = \frac{1}{\beta} J(u, 1). \tag{4.5}$$

According to Xu and Wang [2021, Section B.1], we have the following properties.

Lemma 4.3.3. *Denote* $\varphi(w) := -\frac{S(w)}{s(w)} - w = \frac{1-F(w)}{f(w)} - w$*, and we have* $J(u, \beta) = \frac{u + \varphi^{-1}(u)}{\beta}$*. Also, for* $u \geq 0$ *and* $\beta > 0$*, we have* $\frac{\partial J(u,\beta)}{\partial u} \in (0, 1)$*.*

Then we define a negative log-likelihood function of parameter hypothesis (θ, η) given the results at time t.

Definition 4.3.4 (log-likelihood functions)**.** Denote $\ell_t(\theta, \eta)$ as the negative log-likelihood at time t, and define $L_t(\theta, \eta)$ as their summations:

$$\begin{aligned} -\ell_t(\theta, \eta) =& \mathbb{1}_t \cdot \log S(x_t \top \eta \cdot p_t - x_t^\top \theta) + (1 - \mathbb{1}_t) \cdot \log(1 - S(x_t^\top \eta \cdot p_t - x_t^\top \theta)). \\ L_t(\theta, \eta) =& \sum_{\tau=1}^{t} \ell_t. \end{aligned} \tag{4.6}$$

Finally, we define a round-t expected regret and a cumulative expected regret.

Definition 4.3.5 (regrets)**.** Define

$$Reg_t(p_t) := r(x_t^\top \theta^*, x_t^\top \eta^*, J(x_t^\top \theta^*, x_t^\top \eta^*)) - r(x_t^\top \theta^*, x_t^\top \eta^*, p_t) \tag{4.7}$$

as the expected regret at round t, conditioning on price p_t. Also, define the cumulative regret as $Regret = \sum_{t=1}^{T} Reg_t(p_t)$.

4.3.3 Assumptions

We establish three technical assumptions to make our analysis and presentation clearer. Firstly, we assume that all feature and parameter vectors are bounded within a unit ball in Euclidean norm. This assumption is without loss of generality as it only rescales the problem.

Assumption 4.3.6 (bounded feature and parameter spaces)**.** Assume features $x_t \in \mathcal{H}_x$ and parameters $\theta \in \mathcal{H}_\theta, \eta \in \mathcal{H}_\eta$. Denote $U_p^d := \{x \in \mathbb{R}^d, \|x\|_p \leq 1\}$ as an L_p-norm unit ball in $\mathbb{R}^d$. Assume all $\mathcal{H}_x, \mathcal{H}_\theta, \mathcal{H}_\eta \in U_p^d$. Also, assume $x^\top \theta > 0, \forall x \in \mathcal{H}_x, \theta \in \mathcal{H}_\theta$ and $x^\top \eta > C_\beta > 0, \forall x \in \mathcal{H}_x, \eta \in \mathcal{H}_\eta$ for some constant $C_\beta \in (0, 1)$.

The positiveness of elasticity coefficient $x^\top \eta > 0$ comes from the *Law of Demand* [Gale, 1955, Hildenbrand, 1983], stating that the quantity purchased varies inversely with price. This is derived from the *Law of Diminishing Marginal Utilities* and has been widely accepted [Marshall, 2009]. We will show the necessity of assuming an elasticity lower bound C_β in Section 4.6. In specific, we claim that any algorithm will suffer a regret of $\Omega(\frac{1}{C_\beta})$. For the simplicity of notation, we denote $[\theta; \eta] := [\theta^\top, \eta^\top]^\top \in \mathbb{R}^{2d}$ as the combination of d-dimension column vectors θ and η. Since we know that $x_t^\top \theta \in [0, 1]$ and $x_t^\top \eta \in [C_\beta, 1]$, we have $J(x_t^\top \theta, x_t^\top \eta) \in [J(0, 1), J(1, C_\beta)]$. Later we will show that the price perturbation is no more than $\frac{J(0,1)}{10}$. Therefore, we may have the following assumption.

Assumption 4.3.7 (bounded prices)**.** For any price p_t at each time $t = 1, 2, \ldots, T$, we require $p_t \in [c_1, c_2]$, where $c_1 = \frac{J(0,1)}{2}$ and $c_2 = 2J(1, C_\beta)$.

Similar to Javanmard and Nazerzadeh [2019] and Xu and Wang [2021], we also assume a log-concavity on the noise CDF.

Assumption 4.3.8 (log-concavity)**.** Every D_t is independently sampled according to Eq. (4.1), with $S(\omega) \in [0,1]$ and $s(\omega) = S'(\omega) > 0, \forall \omega \in \mathbb{R}$. Equivalently, the valuation noise $N_t \sim \mathbb{D}_F$ is independently and identically distributed (i.i.d.), with CDF $F = 1 - S$. Assume that $S \in \mathbb{C}^2$, and S and $(1 - S)$ are strictly log-concave.

4.4 Main Results

To solve the contextual pricing problem with featurized elasticity, we propose our "Pricing with Perturbation (PwP)" algorithm. In the following, we firstly describe the algorithm and highlight its properties, then analyze (and bound) its cumulative regret, and finally prove a regret lower bound to show its optimality.

4.4.1 Algorithm

The pseudocode of PwP is displayed as Algorithm 5, which calls an ONS oracle (Algorithm 6).

At each time t, it inherits parameters θ_t and η_t from $(t-1)$ and takes in a context vector x_t. By trusting in θ_t and η_t, it calculates a greedy price $\hat{p}_t$ and outputs a perturbed version $p_t = \hat{p}_t + \Delta_t$. After seeing customer's decision $\mathbb{1}_t$, PwP calls an "Online Newton Step (ONS)" oracle (see Algorithm 6) to update the parameters as θ_{t+1} and η_{t+1} for future use.

Highlights

We highlight the achievements of the PwP algorithm in the following three aspects.

Algorithm 5 Pricing with Perturbation (PwP)

Input: parameter spaces $\mathcal{H}_\theta$, $\mathcal{H}_\eta$, link function S, time horizon T, dimension d

Initialization:parameters $\theta_1 \in \mathcal{H}_\theta$, $\eta_1 \in \mathcal{H}_\eta$, price perturbation Δ, cumulative likelihood $L_0 = 0$, matrix $A_0 = \epsilon \cdot I_{2d}$ and parameter ϵ, γ

for $t = 1, 2, \ldots, T$ **do**

Observe x_t;

Calculate greedy price $\hat{p}_t = J(x_t^\top \theta_t, x_t^\top \eta_t)$

Sample $\Delta_t = \Delta$ with $\Pr = 0.5$ and $\Delta_t = -\Delta$ with $\Pr = 0.5$;

Propose price $p_t = \hat{p}_t + \Delta_t$;

Receive the customer's decision $\mathbb{1}_t$;

Construct negative log-likelihood $\ell_t(\theta, \eta)$ and $L_t(\theta, \eta)$ as eq. (4.6);

Update parameters:

$$[\theta_{t+1}; \eta_{t+1}] \leftarrow ONS([\theta_t; \eta_t])$$

end for

In this pricing problem. As we mentioned in Section 4.1.2, the key to solving this contextual elasticity (or heteroscedastic valuation) pricing problem is to construct a surrogate loss function. Xu and Wang [2021] adopts negative log-likelihood in their setting, which does not work for ours since it is not "convex" enough. In our PwP algorithm, we overcome this challenge by introducing a perturbation Δ on the proposed greedy price. This idea originates from the observation that the *variance* of p_t contributes positively to the "convexity" of the expected log-likelihood, which helps "re-build" the upper-bound inequality.

In online optimization. PwP perturbs the greedy action (price) it should have taken. This idea is similar to a "Following the Perturbed Leader (FTPL)" algorithm [Hutter et al., 2005] that minimizes the summation of the empirical risk and a random loss function

Algorithm 6 Online Newton Step (ONS)

1: **Input: current parameter** $[\theta_t, \eta_t]$**, likelihood** $\ell_t(\theta, \eta)$**, matrix** A_t**, parameter** γ**, parameter spaces** $\mathcal{H}_\theta$ **and** $\mathcal{H}_\eta$**.**

2: Calculate $\nabla_t = \nabla \ell_t(\theta, \eta)$;

3: Rank-1 update: $A_t = A_{t-1} + \nabla_t \nabla_t^\top$;

4: Newton step: $[\hat{\theta}_{t+1}; \hat{\eta}_{t+1}] = [\hat{\theta}_t; \hat{\eta}_t] - \frac{1}{\gamma} A_t^{-1} \nabla_t$;

5: Projection: $[\theta_{t+1}; \eta_{t+1}] = \mathit{\Pi}_{\mathcal{H}_\theta \times \mathcal{H}_\eta}^{A_t}([\hat{\theta}_{t+1}; \hat{\eta}_{t+1}])$;

serving as a perturbation. However, this might lead to extra computational cost as the random perturbation is not necessarily smooth and therefore hard to optimize. In our work, PwP introduces a possible way to overcome this obstacle: Instead of perturbing the objective function, we may directly perturb the action to explore its neighborhood. Our regret analysis and results indicate the optimality of this method and imply a potentially wide application.

In information theory. We show the following fact in the regret analysis of PwP: By adding Δ perturbation on p_t, we may lose $O(\Delta^2)$ in reward but will gain $O(\Delta^2) \cdot I$ in Fisher information (i.e. the expected Hessian of negative log-likelihood function) in return. By Cramer-Rao Bound, this leads to $O(\frac{1}{\Delta^2})$ estimation error. In this way, we quantify the information (observing from exploration) on the scale of reward, which shares the same idea with the Upper Confidence Bound [Lai and Robbins, 1985] method that always maximizes the summation of empirical reward and information-traded reward.

Besides, PwP is computationally efficient as it only calls the ONS oracle for once. As for the ONS oracle, it updates an $A_t^{-1} = (A_{t-1} + \nabla_t \nabla_t^\top)^{-1}$ at each time t, which is with $O(d^2)$ time complexity according to the following *Woodbury matrix identity*

$$(A + xx^\top)^{-1} = A^{-1} - \frac{1}{1 + x^\top A^{-1} x} A^{-1} x (A^{-1} x)^\top. \tag{4.8}$$

4.4.2 Regret Upper Bound

Now we analyze the regret of PwP and propose an upper bound up to constant coefficients.

Theorem 4.4.1. *Under Assumption 4.3.6, Assumption 4.3.7 and Assumption 4.3.8, by taking* $\Delta = \min\left\{\left(\frac{d\log T}{T}\right)^{\frac{1}{4}}, \frac{J(0,1)}{10}, \frac{1}{10}\right\}$*, the algorithm PwP guarantees an expected regret at* $O(\sqrt{dT\log T})$.

In the following, we prove Theorem 4.4.1 by stating a thread of key lemmas. We leave the detailed proof of those lemmas to Section 4.8.

Proof. The proof overview can be displayed as the following roadmap of inequalities:

$$\begin{aligned}
\mathbb{E}[Regret] = \sum_{t=1}^{T} Reg_t(p_t) \leq & \mathbb{E}\left[\sum_{t=1}^{T} O\left((x_t^\top(\theta_t - \theta^*))^2 + (x_t^\top(\eta_t - \eta^*))^2 + \Delta^2\right)\right] \\
\leq & O\left(\frac{\sum_{t=1}^{T}\mathbb{E}\left[\ell_t(\theta_t, \eta_t) - \ell_t(\theta^*, \eta^*)\right]}{\Delta^2} + T\cdot\Delta^2\right) \\
\leq & O\left(\frac{d\log T}{\Delta^2} + T\cdot\Delta^2\right) = O(\sqrt{dT\log T}).
\end{aligned} \tag{4.9}$$

Here the first inequality is by the smoothness of regret function (see Lemma 4.4.2), the second inequality is by a special "strong convexity" of $\ell_t(\theta, \eta)$ that contributes to the surrogate loss (see Lemma 4.4.3), the third inequality is by Online Newton Step (see Lemma 4.4.4), and the last equality is by the value of Δ. A rigorous version of Eq. (4.9) can be found in Section 4.8.4.

We firstly show the smoothness of $Reg_t(p_t)$:

Lemma 4.4.2 (regret smoothness)**.** *Denote* $p_t^* := J(x_t^\top \theta^*, x_t^\top \eta^*)$*. There exists constants* $C_r > 0$ *and* $C_J > 0$ *such that*

$$Reg_t(p_t) \le C_r \cdot (p_t - p_t^*)^2 \le C_r \cdot 2\left(C_J \cdot \left[(x_t^\top(\theta_t - \theta^*))^2 + (x_t^\top(\eta_t - \eta^*))^2\right] + \Delta^2\right). \quad (4.10)$$

While the first inequality of Eq. (4.10) is from the smoothness, and the second inequality is by the Lipschitzness of function $J(u, \beta)$. Please refer to Section 4.8.2 for proof details. We then show the reason why the log-likelihood function can still be a surrogate loss with carefully randomized p_t.

Lemma 4.4.3 (surrogate expected regret)**.** *There exists a constant* $C_l > 0$ *such that* $\forall \theta \in \mathcal{H}_\theta, \eta \in \mathcal{H}_\eta$*, we have*

$$\begin{aligned}
&\mathbb{E}[\ell_t(\theta, \eta) - \ell_t(\theta^*, \eta^*) | \theta_t, \eta_t] \\
\ge &\frac{C_l \Delta^2}{10} [(\theta - \theta^*)^\top, (\eta - \eta^*)^\top] \begin{bmatrix} x_t x_t^\top & 0 \\ 0 & x_t x_t^\top \end{bmatrix} \begin{bmatrix} \theta - \theta^* \\ \eta - \eta^* \end{bmatrix} \\
= &\frac{C_l \cdot \Delta^2}{10} \left[\left(x_t^\top(\theta - \theta^*)\right)^2 + \left(x_t^\top(\eta - \eta^*)\right)^2\right].
\end{aligned} \quad (4.11)$$

This is the most important lemma in this chapter. We show a proof sketch here and defer the detailed proof to Section 4.8.3.

Proof sketch of Lemma 4.4.3. We show that there exist constants $C_l > 0, C_p > 0$ such that

1. $\nabla^2 \ell_t(\theta, \eta) \succeq C_l \cdot \begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & p_t^2 \cdot x_t x_t^\top \end{bmatrix}$, and
2. $\mathbb{E}\begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & p_t^2 \cdot x_t x_t^\top \end{bmatrix} | \theta_t, \eta_t \Big] \succeq C_p \Delta^2 \begin{bmatrix} x_t x_t^\top & 0 \\ 0 & x_t x_t^\top \end{bmatrix}$.

The first property above relies on the exp-concavity of ℓ_t. Notice that the second property does not hold without the $\mathbb{E}$ notation, as the left hand side is a $(a-b)^2$ form while the right hand side is in a (a^2+b^2) form. In general, there exist no constant $c>0$ such that $(a-b)^2 \geq c(a^2+b^2)$. However, due to the randomness of p_t, we have

$$\mathbb{E}[p_t^2|\hat{p}_t] = \mathbb{E}[p_t|\hat{p}_t]^2 + \Delta^2. \tag{4.12}$$

In this way, the *conditional expectation* of the left hand side turns to $(a-b)^2+\lambda\cdot b^2$ and we have

$$(a-b)^2+\lambda b^2 = (\frac{1}{\sqrt{1+\frac{\lambda}{2}}}\cdot a - \sqrt{1+\frac{\lambda}{2}}\cdot b)^2 + (1-\frac{1}{1+\frac{\lambda}{2}})a^2 + \frac{\lambda}{2}b^2 \geq \frac{\frac{\lambda}{2}}{1+\frac{\lambda}{2}}\cdot(a^2+b^2). \tag{4.13}$$

Similarly, we upper bound $\begin{bmatrix} x_t x_t^\top & 0 \\ 0 & x_t x_t^\top \end{bmatrix}$ with $\mathbb{E}[\nabla^2 \ell_t(\theta,\eta)|\theta_t,\eta_t]$ up to a $C_p\cdot\Delta^2$ coefficient.

With those two properties above, along with a property of likelihood function that $\mathbb{E}[\nabla \ell_t(\theta^*,\eta^*)] = 0$, we can prove Lemma 4.4.3 by taking a Taylor expansion of ℓ_t at $[\theta^*;\eta^*]$. ∎

Finally, we cite a theorem from Hazan [2016] as our Lemma 4.4.4 that reveals the surrogate regret rate on negative log-likelihood functions.

Lemma 4.4.4. *With parameters* $G = \sup_{\theta\in\mathcal{H}_\theta,\eta\in\mathcal{H}_\eta}\|\nabla l_t(\theta,\eta)\|_2$, $D = \sup\|[\theta_1;\eta_1] - [\theta_2;\theta_2]\| \leq 2$, $\alpha = C_e$, $\gamma = \frac{1}{2}\min\{\frac{1}{4GD},\alpha\}$ *and* $\epsilon = \frac{1}{\gamma^2 D^2}$ *and* $T > 4$, *Keep running Algorithm 6 for* $t = 1,2,\ldots,T$ *guarantees:*

$$\sup_{\{x_t\}}\left\{\sum_{t=1}^{T}\ell_t(\theta_t,\eta_t) - \min_{\theta\in\mathcal{H}_\theta,\eta\in\mathcal{H}_\eta}\sum_{t=1}^{T}\ell_t(\theta,\eta)\right\} \leq 5(\frac{1}{\alpha}+GD)d\log T. \tag{4.14}$$

With all these lemma above, we have proved every line of Eq. (4.9). ∎

4.4.3 Lower Bounds

We claim that PwP is near-optimal in information theory, by proposing a matching regret lower bound in Theorem 4.4.5. We present the proof with valuation model to match with existing results.

Theorem 4.4.5. *Consider the contextual pricing problem setting with Bernoulli demand model given in Eq.* (4.1)*. With all assumptions in Section 4.3 hold, any pricing algorithm has to suffer a* $\Omega(\sqrt{dT})$ *worst-case expected regret for* $T \geq 2d^2(1+\log d)$*, with* T *the time horizon and* d *the dimension of context.*

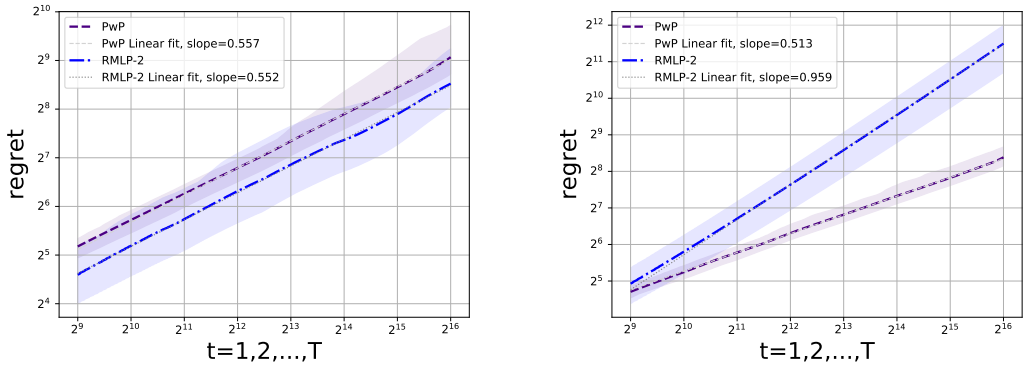
Proof Sketch. We defer the proof details to Section 4.8.5. The main idea is to reduce d numbers of 1-dimension problems to this problem setting. In fact, we may consider the following problem setting:

1. Construct set $X = \{e_i := [0, \ldots, 0, 1, 0, \ldots, 0]^\top \in \mathbb{R}^d$ with only i^{th} place being 1, $i = 1, 2, \ldots, d\}$.

2. Let $\theta^* = [\frac{u_1}{\sigma_1}, \frac{u_2}{\sigma_2}, \frac{u_3}{\sigma_3}, \ldots, \frac{u_d}{\sigma_d}]^\top, \eta^* = [\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \frac{1}{\sigma_3}, \ldots, \frac{1}{\sigma_d}]^\top$, and therefore we have $\frac{e_i^\top \theta^* + N_t}{e_i^\top \eta^*} = u_i + \sigma_i \cdot N_t$.

3. At each time $t = 1, 2, \ldots, T$, sample $x_t \sim X$ independently and uniformly at random.

In this way, we divide the whole time series T into d separated sub-problems, where the Sub-Problem i has a valuation model $y_t(i) = u_i + \sigma_i \cdot N_t$, for $i = 1, 2, \ldots, d$. Let $N_t \sim \mathcal{N}(0, 1), t = 1, 2, \ldots, T$, and $y_t(i) \sim \mathcal{N}(u_i, \sigma_i^2)$ are independent Gaussian random variables. For each Sub-Problem i, it has a time horizon $\frac{T}{2d}$ with high probability. According to Xu and Wang [2021, Theorem 12] (originated from Broder and Rusmevichientong [2012, Theorem 3.1]), the regret lower bound of each sub-problem is $\Omega(\sqrt{\frac{T}{d}})$. Therefore, the



(a) Stochastic $\{x_t\}$'s (b) Adversarial $\{x_t\}$'s

Figure 4.1: The regret of PwP algorithm and a modified RMLP-2 algorithm on simulation data (generated according to Eq. (4.1)), plotted in log-log scales to indicate the regret dependence on T. Figure 4.1a and Figure 4.1b are for stochastic and adversarial $\{x_t\}$ sequences respectively. We also plot linear fits for those regret curves, where a slope-α line indicates an $O(T^\alpha)$ regret. The error bands are drawn with 0.95 coverage using Wald's test. From the figures, we know that PwP performs closely to its $O(\sqrt{T \log T})$ regret regardless of the types of input context sequences, whereas RMLP-2 fails in the attack of adversarial input.

total regret lower bound is $\Omega(d \cdot \sqrt{\frac{T}{d}}) = \Omega(\sqrt{Td})$. ■

4.5 Numerical Experiments

Here we conduct numerical experiments to validate the low-regret performance of our algorithm PwP. We primarily validate the regret rate of our proposed PwP algorithm in well-modeled environments, and then test its adaptivity in homoscedastic and misspecified settings respectively.

4.5.1 Well-assumed Setting

Since we are the first to study this heteroscadestic valuation model, we do not have a baseline algorithm working for exactly the same problem. However, we can modify the RMLP-2 algorithm in Javanmard and Nazerzadeh [2019] by only replacing their max-likelihood estimator (MLE) for θ^* with a new MLE for both θ^* and η^*. This modified version of RMLP-2 does not have a regret guarantee in the current setting, but it still works as a baseline to compare with.

We test PwP and the modified RMLP-2 on the demand model assumed in Eq. (4.1) with both stochastic and adversarial $\{x_t\}$ sequences, respectively. Basically, we assume $T = 2^{16}$ $d = 2$, $N_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$, and we repeatedly run each algorithm for 20 times in each experiment setting. In order to show the regret dependence w.r.t. T, we plot all cumulative regret curves in log-log plots, where an α slope indicates an $O(T^\alpha)$ dependence.

Stochastic $\{x_t\}$. We implement and test PwP and RMLP-2 on stochastic $\{x_t\}$'s, where x_t are iid sampled from $\mathcal{N}(\mu_x, \Sigma_x)$ (for $\mu_x = [10, 10, \ldots, 10]^\top$ and some randomly sampled Σ_x) and then normalized s.t. $\|x_t\|_2 \leq 1$. The numerical results are shown in Figure 4.1a. Numerical results show that both algorithms achieve $\sim O(T^{0.56})$ regrets, which is close to the theoretic regret rate at $O(\sqrt{T \log T})$.

Adversarial $\{x_t\}$. Here we design an adversarial $\{x_t\}$ sequence to attack both algorithms. Since RMLP-2 divides the whole time horizon T into epochs with length $k = 1, 2, 3, \ldots$ sequentially and then does pure exploration at the beginning of each epoch, we may directly attack those pure-exploration rounds in the following way: (1) In each

pure-exploration round (i.e. when $t = 1, 3, 6, \ldots, \frac{k(k+1)}{2}, \ldots$), let the context be $x_t = [1, 0]^\top$; (2) In any other round, let the context be $x_t = [0, 1]^\top$. In this way, the RMLP-2 algorithm will never learn $\theta^*[2]$ and $\eta^*[2]$ since the inputs of pure-exploration rounds do not contain this information. Under this oblivious adversarial context sequence, we implement PwP and RMLP-2 and compare their performance. The results are shown in Figure 4.1b, indicating that PwP can still guarantee $O(T^{0.513})$ regret (close to $O(\sqrt{T \log T})$) while RMLP-2 runs into a linear regret.

As a high-level interpretation, the performance difference is because PwP adopts a "distributed" exploration at every time t while RMLP-2 makes it more "concentrated". Although both PwP and RMLP-2 take the same amount of exploration that optimally balance the reward loss and the information gain (and that is why they both perform well in stochastic inputs), randomly distributed exploration would save the algorithm from being "attacked" by oblivious adversary. In fact, this phenomenon is analog to ϵ-Greedy versus Exploration-first algorithms in multi-armed bandits. We will discuss more in Section 4.6.

In the following subsections, we also conduct experiments to show the robustness, where the true demand (or valuation) distribution is not necessarily assumed as Eq. (4.1) or Eq. (4.2). The numerical results are presented in Section 4.5.2 and Section 4.5.3.

4.5.2 Model Adaptivity

In this section, we show that it is necessary to model the heteroscedasticity. In specific, we compare PwP with the original RMLP-2 algorithm from Javanmard and Nazerzadeh [2019] that ignores heteroscedasticity in a heteroscedastic environment. We conduct both experiments for $T = 2^14$ rounds and repeat them for 10 epochs. The numerical results are

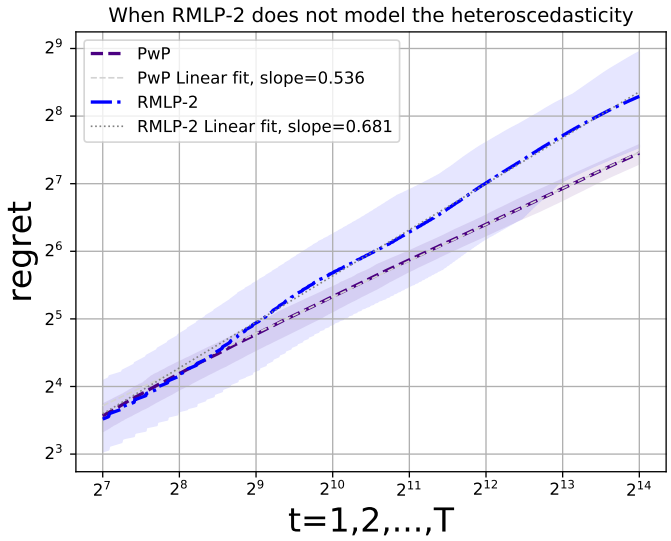


Figure 4.2: Regrets of PwP versus the original homoscedastic RMLP-2 algorithm. In this log-log diagram, a $O(T^\alpha)$ regret curve is shown as a straight line with slope α. From the figure, we notice that PwP is optimal while RMLP-2 is sub-optimal, indicating the necessity of modeling homoscedasticity to achieve optimal regrets.

displayed in the lower figure, plotted in log-log diagrams. From the figure, we notice that the regret of RMLP-2 is much larger than PwP. Also, the slope of regrets of RMLP-2 is $0.681 >> 0.5$, indicating that it does not guarantee a $O(\sqrt{T})$ regret. In comparison, PwP still performs well as it achieves a $\sim O(T^{0.536})$ regret. This indicates that the algorithmic adaptivity of PwP to both homoscedastic and heteroscedastic environments is highly non-trivial, and a failure of capturing it would result in a substantial sub-optimality.

4.5.3 Model Misspecification

In Section 4.5, we compare the cumulative regrets of our PwP algorithm with the (modified) RMLP-2 on the *linear demand* model (as Eq. (4.1) or equivalently, the linear fractional valuation model as Eq. (4.2)). In this section, we consider a model-misspecific setting,

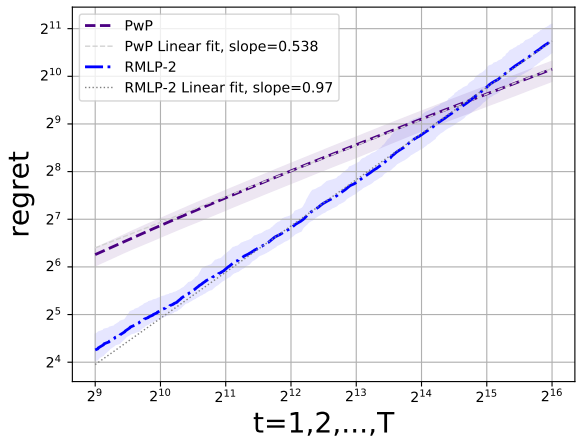


Figure 4.3: Regrets of misspecified PwP with expanded contexts, in comparison with a baseline RMLP-2 knowing the correct model. The results show that PwP still have a sub-linear regret in a certain period of time with context expansions, indicating that our linear demand model as Eq. (4.1) can be generalized to a linear valuation model as Eq. (4.15) in practice.

where customer's true valuation is given by the following equation

$$y_t = x_t^\top \theta^* + x_t^\top \eta^* \cdot N_t \tag{4.15}$$

and the demand $D_t = \mathbb{1}_t = \mathbb{1}[p_t \leq y_t]$. As a result, Eq. (4.15) captures a *linear valuation* model with heteroscedastic valuation.

Now, we design an experiment to show the generalizability of both our PwP algorithm and our demand model as Eq. (4.1). In specific, we run the PwP algorithm that still models a customer's valuation as $\tilde{y}_t = \frac{x_t^\top \tilde{\theta}^* + \tilde{N}_t}{\tilde{x}_t^\top \tilde{\eta}^*}$, where $\tilde{x}_t \in \mathbb{R}^q$ is an *expanded* version of the original context x_t (i.e. $\tilde{x}_t = \pi(x_t)$ for some fixed expanding policy π) and $\tilde{\theta}^*, \tilde{\eta}^* \in \mathbb{R}^q$ are some fixed parameters[2]. Therefore, PwP is trying to learn those misspecified $\tilde{\theta}^*$ and $\tilde{\eta}^*$ although there does not exist such an underground truth.

[2]We may assume $q \geq d$ without loss of generality.

We are curious whether the expansion of context (from x_t to $\tilde{x}_t$) would leverage the hardness of model misspecification. For $x = [x_1, x_2, \ldots, x_d]^\top$, denote $x^n := [x_1^n, x_2^n, \ldots, x_d^n]^\top$. Then for any context $x \in \mathbb{R}^d$, we specify each context-expanding policy as follows:

$$\pi(x; x_0, \mathbf{a}) := [x; (x - x_0)^{a_1}; (x - x_0)^{a_2}; \ldots; (x - x_0)^{a_m}]^\top \in \mathbb{R}^{(m+1)d}. \tag{4.16}$$

The policy π in Eq. (4.16) is a polynomial expansion of x with index list $\mathbf{a} = [a_1, a_2, \ldots, a_m] \in \mathbb{Z}^m$, where $x_0 \in \mathbb{R}^d$ is a fixed start point of this expansion.

Now we consider the baseline to compare with. We claim that it is very challenging to solve the contextual pricing problem with customers' valuations being Eq. (4.15) with theoretic regret guarantees (although the $\Omega(\sqrt{T})$ lower bound given by Javanmard and Nazerzadeh [2019] still holds), and there are no existing algorithms targeting at this problem setting. However, there are still some straightforward algorithms that might approach it: For example, a max-likelihood estimate (MLE) of θ^* and η^*. In fact, we may still reuse the framework of RMLP-2 by replacing its MLE oracle according to the distribution given by Eq. (4.15). In the following, we will compare the performances of

1. PwP algorithm with the misspecified linear demand model as Eq. (4.1), with expanded context $\{x_t\}$'s, and

2. RMLP-2 algorithm on the correct linear valuation model asEq. (4.15), with original context $\{x_t\}$'s.

We implement PwP and RMLP-2 on stochastic $\{x_t\}$ sequences (since RMLP-2 has already failed in the adversarial setting) and get numerical results shown as Figure 4.3. Here we choose $x_0 = [0.5, 0.5]^\top$ and $\mathbf{a} = [0, 1]$. For a model-misspecified online-learning algorithm, there generally exists an $O(\epsilon \cdot T)$ term in the regret rate, where ϵ is a parameter measuring

the distance between the global optimal policy and the best *proper* policy (i.e. the best policy in the hypothesis set). However, our numerical results imply that PwP may still achieve a sub-linear regret within a certain time horizon T, whereas the baseline RMLP-2 that takes the correct model has a much worse regret. It is worth mentioning that PwP may still run into $\Omega(T)$ regret as T gets sufficiently large, due to model misspecification. These results imply that

1. Our linear demand model Eq. (4.1) can be generalized to a linear valuation model as Eq. (4.15) in practice.

2. Our PwP algorithm can still perform well in model-misspecification settings, and even better than a baseline MLE algorithm with a correct model in a certain period of time.

For the first phenomenon that our demand model can be generalized with context expansion tricks, we may understand it as a Taylor expansion (and we take a linear approximation) at $x_0 = [0.5, 0.5]^\top$. For the second phenomenon that PwP outperforms RMLP-2, it might be caused by the non-convexity of the log-likelihood function of the valuation model specified in Eq. (4.15). As a result, while RMLP-2 is solving a non-convex MLE and getting estimates far from the true parameters, PwP instead works on an online convex optimization problem within a larger space (which probably contain the underground truth) due to context expansions. Unfortunately, we do not have a rigorous analysis of those two phenomenons.

4.6 Discussion

Here we discuss the motivations, justifications, limitations and extentions of our work.

Necessity of lower-bounding $x_t^\top \eta^*$ from 0. As we state in Assumption 4.3.6, the price elasticity coefficient $x_t^\top \eta^*$ is lower bounded by a constant $C_\beta > 0$. On the one hand, this is necessary since we cannot have an upper bound on the optimal price without this assumption. On the other hand, according to Eq. (4.3), we know that $r(u, \beta, p) = r(u, 1, \beta \cdot p) \cdot \frac{1}{\beta}$, which indicates that the reward is rescaled by $\frac{1}{\beta}$. As a result, the regret should be proportional to $\frac{1}{C_\beta}$. Although a larger (i.e. closer to 0) elasticity would lead to a more *smooth* demand curve, this actually reduce the information we could gather from customers' feedback and slow down the learning process. We look forward to future researches getting rid of this assumption and achieve more adaptive regret rates.

Assumption on lower-bounding elasticity as $C_\beta > 0$. Here we claim that the regret lower bound should have an $\Omega(\frac{1}{C_\beta})$ dependence on C_β. We prove this by contradiction. Without loss of generality, assume $C_\beta \in (0, 1)$. In specific, we construct a counter example to show it is impossible to have an $O(C_\beta^{-1+\alpha})$ regret for any $\alpha > 0$:

Firstly, let $\beta = C_\beta$. Suppose there exists an algorithm $\mathcal{A}$ that proposes a series of prices $\{p_t\}_{t=1}^T$ which achieve $O(C_\beta^{-1+\alpha})$ regret in any pricing problem instance under our assumptions.

Now, we consider another specific problem setting where $\beta = 1$ while all other quantities $\theta^*, \eta^*, \{x_t\}_{t=1}^T$ stay unchanged. Notice that the reward function has the following property:

$$r(u, \beta, p) = p \cdot S(\beta p - u) = \frac{1}{\beta} \cdot (\beta p) \cdot S(\beta p - u) = \frac{1}{\beta} \cdot r(u, 1, \beta p) \tag{4.17}$$

Therefore, we construct another algorithm $\mathcal{A}^*$ which proposes $C_\beta \cdot p_t$ at $t = 1, 2, \ldots, T$.

According to the $O(C_\beta^{-1+\alpha})$ regret bound of $\mathcal{A}$, we know that $\mathcal{A}^*$ will suffer $C_\beta \cdot O(C_\beta^{-1+\alpha}) = O(C_\beta^\alpha)$ regret. Let $C_\beta \to 0^+$ and observe the regret of $\mathcal{A}^*$ on the latter problem setting (where $\beta = 1$). On the one hand, this is a fixed problem setting with information-theoretic lower regret bound at $\Omega(\log T)$. On the other hand, the regret will be bounded by $\lim_{C_\beta \to 0^+} O(C_\beta^\alpha) = 0$. They are contradictory to each other. Given this, we know that there does not exist such an $\alpha > 0$ such that there exists an algorithm that can achieve $O(C_\beta^{-1+\alpha})$. As a result, it is necessary to lower bound the elasticities by C_β from 0.

Adversarial attacks. Our PwP algorithm achieves near-optimal regret even for adversarial context sequences, while the baseline (modified) RMLP-2 algorithm fails in an oblivious adversary and suffer a linear regret. This is mainly caused by the fact that RMLP-2 takes a pure-exploration step at a *fixed* time series, i.e. $t = 1, 1+2, 1+2+3, \ldots, \frac{k(k+1)}{2}$. This issue might be leveraged by randomizing the position of pure-exploration steps: In each Epoch $k = 1, 2, \ldots$, it may firstly sample one out of all k rounds in this epoch uniformly at random, and then propose a totally random price at this specific round. However, RMLP-2 still requires $\mathbb{E}[xx^\top] \succeq c \cdot I_d$ even with this trick.

Nonstationarity in Pricing Although our PwP algorithm is applicable on heteroscedastic valuations, we still benchmark with an optimal fixed pricing policy that knows η^* and θ^* in advance. In reality, customers' valuations and elasticities might fluctuate according to the market environment, causing θ_t^* and η^* different over $t \in [T]$. Existing works including Leme et al. [2021] and Baby et al. [2022] study similar settings but assume i.i.d. noises. It is worth to further investigate the setting when heteroscedasticity and nonstationarity occur simultaneously.

Regret lower bounds for fixed unknown noise distributions. We claim a $\Omega(\sqrt{dT})$ regret lower bound in Theorem 4.4.5 with customers' demand model being Eq. (4.1). However, this result does not imply a $\Omega(\sqrt{dT})$ regret lower bound for the contextual pricing problem with customers' valuation being $y_t = x_t^\top \theta^* + N_t$ adopted by Javanmard and Nazerzadeh [2019], Cohen et al. [2020], Xu and Wang [2021]. This is because our problem setting is more general than theirs, and our construction of $\Omega(\sqrt{dT})$ regret lower bounds are substantially beyond the scope of this specific subproblem. So far, the best existing regret lower bound for the linear noisy model ($y_t = x_t^\top \theta^* + N_t$) is still $\Omega(\sqrt{T})$. However, we conjecture that this should also be $\Omega(\sqrt{dT})$. The hardness of proving this lower bound comes from the fact that the noises are iid over time, and it is harder to be separated into several sub-sequences across d that are independent to each other.

Algorithm and analysis for unknown link function $S(\cdot)$. Unfortunately, our algorithm is unable to be generalized to the online contextual pricing problem with linear valuation and unknown noise distribution that has been studied by Fan et al. [2021]. Indeed, the problem becomes substantially harder when the noise distribution is unknown to the agent. Existing works usually adopt bandits or bandit-like algorithms to tackle that problem. For example, Fan et al. [2021] approaches it with a combination of exploration-first and kernel method (or equivalently, local polynomial), Luo et al. [2021] uses a UCB-styled algorithm, and Xu and Wang [2022] adopts a discrete EXP-4 algorithm. However, none of them close the regret gap even under the homoscedastic elasticity environment as they assumed, and the known lower bound is at least $\Omega(T^{\frac{2}{3}})$, or $\Omega(T^{\frac{m+1}{2m+1}})$ for smooth ones [Wang et al., 2021b]. On the other hand, we study a parametric model, and it is not quite suitable for a bandit algorithm to achieve optimality in regret. In a nutshell, these two problems (known vs unknown noise distributions), although seem

similar to each other, are indeed substantially different.

Linear demand model vs linear valuation model. In this chapter, we adopt a generalized linear demand model with Boolean feedback, as assumed in Eq. (4.1). As we have stated in Section 4.5.3, there exists a heteroscedastic linear valuation model as Eq. (4.15) that also captures a customer's behavior. However, this linear valuation model is actually harder to learn, as its log-likelihood function is non-convex. It is still an open problem to determine the minimax regret of an online contextual pricing problem with a valuation model like Eq. (4.15).

Ethic issues. Since we study a dynamic pricing problem, we have to consider the social impacts that our methodologies and results could have. The major concern in pricing is *fairness*, which attracts increasing research interests in recent years [Cohen et al., 2021, 2022, Xu et al., 2023, Chen et al., 2023b]. In general, we did not enforce or quantify the fairness of our algorithm. In fact, we might not guarantee an individual fairness since PwP proposes random prices, which means even the same input x_t's would lead to different output prices. Despite the perturbations Δ_t we add to the prices, the pricing model (i.e. the parameters θ^* and η^*) is updating adaptively over time. This indicates that customers arriving later would have relatively fairer prices, since the model is evolving drastically at the beginning rounds and is converging to (local) optimal after a sufficiently long time period. We claim that our PwP algorithm is still fairer than the baseline RMLP-2 algorithm we compare with, since RMLP-2 takes pure explorations at some specific time. As a result, those customers who are given a totally random price would have a either much higher or much lower expected price than those occurring in exploitation rounds. However, it is still worth mentioning that RMLP-2 satisfies individual fairness within each

pure-exploitation epoch, since it does not update parameters nor adding noises then.

4.7 Conclusion

In summary, our work focuses on the problem of contextual pricing with highly differentiated products. We propose a contextual elasticity model that unifies the "linear demand" and "linear valuation" camps and captures the price effect and heteroscedasticity. To solve this problem, we develop an algorithm PwP, which utilizes Online Newton Step (ONS) on a surrogate loss function and proposes perturbed prices for exploration. Our analysis show that it guarantees a $O(\sqrt{dT \log T})$ regret even for adversarial context sequences. We also provide a matching $\Omega(\sqrt{dT})$ regret lower bound to show its optimality (up to $\log T$ factors). Besides, our numerical experiments also validate the regret bounds of PwP and its advantage over existing method. We hope our results would shed lights on the research of contextual pricing as well as online decision-making problems.

4.8 Proofs

Here we show the proof details of the lemmas we have stated in Section 4.4.2. Before that, let us clarify some terminologies we mentioned in the main paper.

4.8.1 Supplementary Definitions

Firstly, we rigorously define the concept of *price elasticity* occurring in Section 4.1.

Definition 4.8.1 (Price Elasticity [Parkin et al., 2002])**.** Suppose $D(p)$ is a demand

function of price p. Then the price elasticity E_d of demand is defined as

$$E_D := \frac{\Delta D(p)/D(p)}{\Delta p/p} = \frac{\partial D(p)}{\partial p} \cdot \frac{p}{D(p)}. \tag{4.18}$$

With this definition, along with our generalized linear demand model given in Eq. (4.1), the price elasticity for the expected demand $S(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*)$ is

$$\begin{aligned} E_D =& \frac{\partial S(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*)}{\partial p_t} \cdot \frac{p_t}{S(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*)} \\ =& x_t^\top \eta^* \cdot \frac{s(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*)}{S(x_t^\top \eta^* \cdot p_t - x_t^\top \theta^*)} \cdot p_t. \end{aligned} \tag{4.19}$$

Here $s(\cdot) = S'(\cdot)$. Therefore, despite the effect of the link function and the price p_t, the price elasticity is proportional to the price coefficient $x_t^\top \eta^*$. This is why we call $x_t^\top \eta^*$ (or α in the general model $D(p) = \lambda(\alpha \cdot p + x_t^T \beta)$) the *elasticity coefficient* or *coefficient of elasticity* in Section 4.1.

4.8.2 Proof of Lemma 4.4.2

Proof. In order to prove Lemma 4.4.2, we show the following lemma that indicates the Lipschitzness of $J(u, \beta)$:

Lemma 4.8.2 (Lipschitz of optimal price)**.** *There exists a constant $C_J > 0$ such that*

$$|J(u_1, \beta_1) - J(u_2, \beta_2)| \leq C_J \cdot (|u_1 - u_2| + |\beta_1 - \beta_2|). \tag{4.20}$$

With this lemma, we get the second inequality of Eq. (4.10). We will prove this lemma

later in this subsection. Now, we focus on the first inequality. Notice that

$$\begin{aligned} Reg_t(p_t) =& r(x_t^\top \theta^*, x_t^\top \eta^*, p_t^*) - r(x_t^\top \theta^*, x_t^\top \eta^*, p_t) \\ \leq& -\frac{\partial r(u,\beta,p)}{\partial p}|_{u=x_t^\top\theta^*,\beta=x_t^\top\eta^*,p=p_t^*}(p_t^* - p_t) \\ & - \frac{1}{2}\cdot \inf_{p\in[c_1,c_2],\beta\in[C_\beta,1],u\in[0,1]} \frac{\partial^2 r(u,\beta,p)}{\partial p^2}|_{u=x_t^\top\theta^*,\beta=x_t^\top\eta^*,p=p_t^*}(p_t^* - p_t)^2 \\ =& 0 + \frac{1}{2}\cdot \sup_{p\in[c_1,c_2],\beta\in[C_\beta,1],u\in[0,1]} \{|2s(\beta\cdot p - u)\cdot\beta + p\cdot s'(\beta\cdot p-u)\cdot\beta^2|\}(p_t^* - p_t)^2. \end{aligned} \tag{4.21}$$

Here the first line is by the definition of $Reg_t(p_t)$, the second line is by smoothness, the third line is by the optimality of p_t^*, and the last line is by calculus. Since $|2s(\beta\cdot p-u)\cdot\beta + p\cdot s'(\beta\cdot p-u)\cdot\beta^2|$ is continuous on $p\in[c_1,c_2], \beta\in[C_\beta,1], u\in[0,1]$, we denote this maximum as $2C_r$, which proves the first inequality of Eq. (4.10). ∎

Now we show the proof of Lemma 4.8.2.

Proof of Lemma 4.8.2. Since $J(u,\beta) = \frac{u+\varphi^{-1}(u)}{\beta}$ where $\varphi(w) = -\frac{S(w)}{s(w)} - w$. Notice that

$$\varphi'(w) = -\frac{d\frac{S(w)}{s(w)}}{dw} - 1 = \frac{d^2\log(S(w))}{dw^2}\cdot\frac{S(w)^2}{s(w)^2} - 1 < -1 \tag{4.22}$$

since $S(w)$ is log-concave (as is assumed in Assumption 4.3.8). Given Eq. (4.22), we know that $\frac{d\varphi^{-1}(u)}{d(u)} = \frac{1}{\frac{d\varphi(w)}{dw}|_{w=\varphi^{-1}(u)}} \in (-1,0)$. Therefore, we have:

$$\begin{aligned} \frac{\partial J(u,\beta)}{\partial u} =& \frac{1+\frac{d\varphi^{-1}(u)}{du}}{\beta} \in (0, \frac{1}{C_\beta}) \\ \frac{\partial J(u,\beta)}{\partial \beta} =& \frac{\partial\frac{J(u,1)}{\beta}}{\partial\beta} = -\frac{J(u,1)}{\beta^2} \in [-\frac{c_2}{C_\beta}, -c_1]. \end{aligned} \tag{4.23}$$

Therefore, we know that $J(u,\beta)$ is Lipschitz with respect to u and β respectively. Take $C_J = \max\{\frac{1}{C_\beta}, \frac{c_2}{C_\beta}\}$ and we get Eq. (4.20). ∎

4.8.3 Proof of Lemma 4.4.3

Proof. We firstly show the convexity (and exp-concavity) of $\ell_t(\theta, \eta)$ by the following lemma.

Lemma 4.8.3 (exp-concavity)**.** *$\ell_t(\theta, \eta)$ is convex and C_e-exp-concave with respect to $[\theta; \eta]$, where $C_e > 0$ is a constant dependent on F and C_β. Equivalently, $\nabla^2 \ell_t(\theta, \eta) \succeq C_e \cdot \nabla \ell_t(\theta, \eta) \nabla \ell_t(\theta, \eta)^\top$. Also, we have $\nabla^2 \ell_t(\theta, \eta) \succeq C_l \cdot \begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & v_t^2 \cdot x_t x_t^\top \end{bmatrix}$ for some constant $C_l > 0$.*

The proof of Lemma 4.8.3 is mainly straightforward calculus, and we defer the proof to the end of this subsection. According to Lemma 4.8.3, we have $\nabla^2 \ell_t(\theta, \eta) \succeq C_l \cdot \begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & p_t^2 \cdot x_t x_t^\top \end{bmatrix}$. Therefore, we know that

$$\begin{aligned} \ell_t(\theta, \eta) \geq & \ell_t(\theta^*, \eta^*) + \nabla \ell_t(\theta^*, \eta^*)^\top \begin{bmatrix} \theta - \theta^* \\ \eta - \eta^* \end{bmatrix} \\ & + [(\theta - \theta^*)^\top, (\eta - \eta^*)^\top] C_l \begin{bmatrix} x_t x_t^\top & -p_t x_t x_t^\top \\ -p_t x_t x_t^\top & p_t^2 x_t x_t^\top \end{bmatrix} \begin{bmatrix} \theta - \theta^* \\ \eta - \eta^* \end{bmatrix} \end{aligned} \tag{4.24}$$

According to the property of likelihood, we have $\mathbb{E}[\nabla \ell_t(\theta^*, \eta^*) | \theta_t, \eta_t] = 0$ for any θ_t and η_t. Combining this with Eq. (4.24), we get

$$\begin{aligned} & \mathbb{E}[\ell_t(\theta, \eta) - \ell_t(\theta^*, \eta^*) | \theta_t, \eta_t] \\ \geq & C_l [(\theta - \theta^*)^\top, (\eta - \eta^*)^\top] \mathbb{E} \begin{bmatrix} x_t x_t^\top & -p_t x_t x_t^\top \\ -p_t x_t x_t^\top & p_t^2 x_t x_t^\top \end{bmatrix} | \theta_t, \eta_t \Bigg] \begin{bmatrix} \theta - \theta^* \\ \eta - \eta^* \end{bmatrix} \end{aligned} \tag{4.25}$$

Recall that $\hat{p}_t = J(x_t^\top \theta_t, x_t^\top \eta_t)$ and that $p_t = \hat{p}_t + \Delta_t$. Therefore, we know that the

conditional expectations $\mathbb{E}[p_t|\theta_t,\eta_t]=\hat{p}_t$ and $\mathbb{E}[p_t^2|\theta_t,\eta_t]=\hat{p}_t^2+\Delta^2$. Given this, we have

$$\begin{aligned}
&\mathbb{E}\begin{bmatrix} x_tx_t^\top & -p_tx_tx_t^\top \\ -p_tx_tx_t^\top & p_t^2x_tx_t^\top \end{bmatrix}|\theta_t,\eta_t\Bigg] \\
=&\begin{bmatrix} x_tx_t^\top & -\hat{p}_tx_tx_t^\top \\ -\hat{p}_tx_tx_t^\top & (\hat{p}_t^2+\Delta^2)x_tx_t^\top \end{bmatrix} \\
=&\begin{bmatrix} x_t \\ -\hat{p}_tx_t \end{bmatrix}\left[x_t^\top,-\hat{p}_tx_t^\top\right]+\begin{bmatrix} 0 & 0 \\ 0 & \Delta^2x_tx_t^\top \end{bmatrix} \\
=&\begin{bmatrix} \frac{1}{\sqrt{1+\frac{\Delta^2}{2}}}\cdot x_t \\ -\sqrt{1+\frac{\Delta^2}{2}}\hat{p}_t\cdot x_t \end{bmatrix}\left[\frac{1}{\sqrt{1+\frac{\Delta^2}{2}}}\cdot x_t^\top,-\sqrt{1+\frac{\Delta^2}{2}}\hat{p}_t\cdot x_t^\top\right] \\
&+\begin{bmatrix} (1-\frac{1}{1+\frac{\Delta^2}{2}})x_tx_t^\top & 0 \\ 0 & \frac{\Delta^2}{2}x_tx_t^\top \end{bmatrix}
\end{aligned} \tag{4.26}$$

Since $\Delta=\min\left\{\left(\frac{d\log T}{T}\right)^{\frac{1}{4}},\frac{J(0,1)}{10},\frac{1}{10}\right\}$, we have $1-\frac{1}{1+\frac{\Delta^2}{2}}=\frac{\frac{\Delta^2}{2}}{1+\frac{\Delta^2}{2}}\geq\frac{\Delta^2}{10}$. As a result, we have

$$\mathbb{E}\begin{bmatrix} x_tx_t^\top & -p_tx_tx_t^\top \\ -p_tx_tx_t^\top & p_t^2x_tx_t^\top \end{bmatrix}|\theta_t,\eta_t\Bigg]\geq\frac{\Delta^2}{10}\cdot\begin{bmatrix} x_tx_t^\top & 0 \\ 0 & x_tx_t^\top \end{bmatrix} \tag{4.27}$$

This proves the lemma. ■

Finally, we show the proof of Lemma 4.8.3.

Proof of Lemma 4.8.3. Recall that $\ell_t(\theta,\eta)=-\mathbb{1}_t\cdot\log(S(x_t^\top(p_t\eta-\theta)))-(1-\mathbb{1}_t)\cdot\log(1-S(x_t^\top(p_t\eta-\theta)))$. We first calculate the gradient and Hessian of $\ell_t(\theta,\eta)$ with respect to $[\theta;\eta]$. For notation simplicity, denote $w_t:=x_t^\top(p_t\eta-\theta)$ and

$$\nabla\ell_t:=-\left(\mathbb{1}_t\cdot\frac{s(w_t)}{S(w_t)}-(1-\mathbb{1}_t)\cdot\frac{s(w_t)}{1-S(w_t)}\right)\cdot\begin{bmatrix} -x_t \\ p_t\cdot x_t \end{bmatrix} \tag{4.28}$$

$$\nabla^2 \ell_t = -\left(\mathbb{1}_t \cdot \frac{s'(w_t)S(w_t) - s(w_t)^2}{S(w_t)^2} + (1-\mathbb{1}_t)\frac{-s'(w_t)(1-S(w_t)) - s(w_t)^2}{(1-S(w_t))^2}\right) \cdot \begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & p_t^2 \cdot x_t x_t^\top \end{bmatrix}. \tag{4.29}$$

According to Assumption 4.3.8, we know that $S(w)$ and $(1-S(w))$ are strictly log-concave, which indicates

$$\begin{aligned} \frac{d^2 \log(1-S(w))}{dw^2} =& \frac{-s'(w)(1-S(w)) - s(w)^2}{(1-S(w))^2} < 0 \\ \frac{d^2 \log(S(w))}{dw^2} =& \frac{s'(w)S(w) - s(w)^2}{S(w)^2} < 0, \forall w \in \mathbb{R}. \end{aligned} \tag{4.30}$$

Since $w_t = p_t \cdot x_t^\top \eta - x_t^\top \theta$ where $p_t \in [c_1, c_2]$, we know that $w_t \in [-1, c_2]$. Since $\frac{d^2 \log(S(w))}{dw^2}$ and $\frac{d^2 \log(1-S(w))}{dw^2}$ are continuous on $[-1, c_2]$, we know that

$$\begin{aligned} &\mathbb{1}_t \cdot \frac{s'(w_t)S(w_t) - s(w_t)^2}{S(w_t)^2} + (1-\mathbb{1}_t) \cdot \frac{-s'(w_t)(1-S(w_t)) - s(w_t)^2}{(1-S(w_t))^2} \\ \leq& \sup_{w \in [-1, c_2]} \max\left\{\frac{s'(w_t)S(w_t) - s(w_t)^2}{S(w_t)^2}, \frac{-s'(w_t)(1-S(w_t)) - s(w_t)^2}{(1-S(w_t))^2}\right\} < 0. \end{aligned} \tag{4.31}$$

Denote $C_l = -\sup_{w \in [-1, c_2]} \max\left\{\frac{s'(w_t)S(w_t) - s(w_t)^2}{S(w_t)^2}, \frac{-s'(w_t)(1-S(w_t)) - s(w_t)^2}{(1-S(w_t))^2}\right\} > 0$, and we know that

$$\nabla^2 \ell_t(\theta, \eta) \succeq C_l \cdot \begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & p_t^2 \cdot x_t x_t^\top \end{bmatrix}. \tag{4.32}$$

Similarly, we know that $\frac{s(w)}{S(w)}$ and $\frac{-s(w)}{1-S(w)}$ are continuous on $[-1, c_2]$. Therefore, we may denote $C_G = \sup_{w \in [-1, c_2]} \max\left\{\left|\frac{s(w)}{S(w)}\right|, \left|\frac{-s(w)}{1-S(w)}\right|\right\} > 0$ and get

$$\nabla \ell_t(\theta, \eta) \nabla \ell_t(\theta, \eta)^\top \preceq C_G^2 \cdot \begin{bmatrix} -x_t \\ p_t \cdot x_t \end{bmatrix} \left[-x_t^\top,\ p_t \cdot x_t^\top\right]. \tag{4.33}$$

Given all these above, we have

$$\begin{aligned}\nabla^2 \ell_t(\theta,\eta) &\succeq C_l \cdot \begin{bmatrix} x_t x_t^\top & -p_t \cdot x_t x_t^\top \\ -p_t \cdot x_t x_t^\top & p_t^2 \cdot x_t x_t^\top \end{bmatrix} \\ &= \frac{C_l}{C_G^2} \cdot C_G^2 \cdot \begin{bmatrix} -x_t \\ p_t \cdot x_t \end{bmatrix} \begin{bmatrix} -x_t^\top, & p_t \cdot x_t^\top \end{bmatrix} \\ &\succeq \frac{C_l}{C_G^2} \cdot \nabla \ell_t(\theta,\eta) \nabla \ell_t(\theta,\eta)^\top.\end{aligned} \tag{4.34}$$

Denote $C_e := \frac{C_l}{C_G^2}$ and we prove the lemma. ∎

4.8.4 Proof of Theorem 4.4.1

Proof. With all lemmas above, we have

$$\begin{aligned}\mathbb{E}[Regret] =& \mathbb{E}[\sum_{t=1}^{T} \mathbb{E}[Reg_t(p_t)|\theta_t,\eta_t]] \\ \leq& \mathbb{E}[\sum_{t=1}^{T} C_r \cdot 2 \cdot C_J \cdot \mathbb{E}[(x_t^\top(\theta_t-\theta^*))^2 + (x_T^\top(\eta_t-\eta^*))^2|\theta_t,\eta_t] + T \cdot C_r \cdot 2 \cdot \Delta^2] \\ \leq& \mathbb{E}[\sum_{t=1}^{T} 2C_r C_J \cdot \frac{10}{C_l \cdot \Delta^2} \cdot \mathbb{E}[\ell_t(\theta_t,\eta_t) - \ell_t(\theta^*,\eta^*)|\theta_t,\eta_t] + 2C_r T\Delta^2] \\ =& \frac{20C_r C_J}{C_l \Delta^2} \mathbb{E}[\sum_{t=1}^{T} \ell_t(\theta_t,\eta_t) - \ell_t(\theta^*,\eta^*)] + 2C_r T\Delta^2 \\ =& O(\frac{d\log T}{\Delta^2} + \Delta^2 T) \\ =& O(\sqrt{dT\log T}).\end{aligned} \tag{4.35}$$

Here the first line is by the law of total expectation, the second line is by Lemma 4.4.2, the third line is by Lemma 4.4.3, the fourth line is by equivalent transformation, the fifth line is by Lemma 4.4.4, and the sixth line is by the fact that $\Delta = \min\left\{\left(\frac{d\log T}{T}\right)^{\frac{1}{4}}, \frac{J(0,1)}{10}, \frac{1}{10}\right\}$. This holds the theorem. ∎

4.8.5 Proof of Theorem 4.4.5

Proof. Denote $\theta^* = [\theta_1, \theta_2, \ldots, \theta_d]^\top$ and $\eta^* = [\eta_1, \eta_2, \ldots, \eta_d]^\top$. We firstly construct a context set X as

$$X = \{e_i, i = 1, 2, \ldots, d | e_i = [0, \ldots, 0, 1, 0, \ldots, 0]^\top \in \mathbb{R}^d, e_i[i] = 1, e_i[j] \neq 1, \forall j \neq i\}. \tag{4.36}$$

Then we sample each $x_t \sim_{i.i.d.} \mathbb{U}_X$, where $\mathbb{U}_X$ is a uniform distribution defined on each element of X (i.e. $\Pr[x_t = e_i] = \frac{1}{d}, \forall i \in [d], t \in [T]$). Denote $i_t := i$ if $x_t = e_i$. Now we decompose the indexes set $[T]$ of series $\{x_t\}_{t=1}^T$ into d subsets:

$$S_i := \{t | x_t = e_i, t = 1, 2, \ldots, T\}, i = 1, 2, \ldots, d. \tag{4.37}$$

From the perspective of customers' valuations, we have $y_t = \frac{e_i^\top \theta^* + N_t}{e_i^\top \eta^*} = \frac{\theta_{i_t}}{\eta_{i_t}} + \frac{1}{\eta_{i_t}} \cdot N_t$ where N_t is an i.i.d. noise with known distribution. Let $N_t \sim_{i.i.d.} \mathcal{N}(0, 1)$ as standard Gaussian noises. Therefore, each $i \in [d]$ determines a sub-problem (denoted as $\mathcal{P}_i$) that only happens when $t \in S_i$ and has a fixed valuation distribution, i.e. $y_t = \frac{\theta_i}{\eta_i} + \frac{1}{\eta_i} \cdot N_t \sim_{i.i.d.} \mathcal{N}(\frac{\theta_i}{\eta_i}, \frac{1}{\eta_i^2})$. For any $t \notin S_i$, neither y_t nor $\mathbb{1}_t$ is dependent on θ_i or η_i, which enables us to separately consider each $\mathcal{P}_i$. Denote $T_i := |S_i|$. In the following, we bound the least possible regret of this sub-problem as $\Omega(\sqrt{T_i})$.

Let $\theta_i = \frac{u_i}{\sigma_i}$ and $\eta_i = \frac{1}{\sigma_i}$ where $u_i, \sigma_i > 0$ are unknown parameters to be determined. Given this, customers' valuation distribution is $y_t \sim \mathcal{N}(u_i, \sigma_i^2)$ for $t \in S_i$. According to Theorem 12 of Xu and Wang [2021], let $u_i = \sqrt{\frac{\pi}{2}}$ and $\sigma_i \in \{1, 1 - T_i^{-\frac{1}{4}}\}$, and any algorithm has to suffer at least $\frac{1}{24000} \cdot \sqrt{T_i}$ regret.

Then we show that $T_i \geq \frac{T}{2d}$ with high probability. Notice that

$$\begin{aligned}\mathbb{E}[T_i] =& \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}[x_t = e_i]] \\ =& \sum_{t=1}^{T} \Pr[x_t = e_i] \\ =& T \cdot \frac{1}{d} = \frac{T}{d}.\end{aligned} \tag{4.38}$$

According to Hoeffding's Inequality, we have:

$$\begin{aligned}\Pr[|\sum_{t=1}^{T} \mathbb{1}[x_t = e_i] - \frac{T}{d}| \geq \frac{T}{2d}] \leq 2\exp\{-2\frac{T^2}{4d^2} \cdot \frac{1}{T}\} \\ \Rightarrow \qquad \Pr[T_i \geq \frac{T}{2d}] \leq 2\exp\{-\frac{T}{2d^2}\}.\end{aligned} \tag{4.39}$$

According to a union bound on the failure probability, with $Pr \geq 1 - 2d\exp\{-\frac{T}{2d^2}\}$, we have $T_i \geq \frac{T}{2d}, \forall i \in [d]$. Therefore, the expected regret satisfies

$$\begin{aligned}\mathbb{E}[Regret] &= \mathbb{E}[\sum_{i=1}^{d} Regret(\mathcal{P}_i)] \\ &\geq \mathbb{E}[\sum_{i=1}^{d} \frac{1}{24000}\sqrt{T_i}] \\ &\geq \mathbb{E}[\sum_{i=1}^{d} \frac{1}{24000}\sqrt{\frac{T}{2d}} \cdot (1 - 2d\exp\{-\frac{T}{2d^2}\})] \\ &\geq \frac{1}{200000} \cdot \sqrt{Td}.\end{aligned} \tag{4.40}$$

Here the last inequality comes from the assumption that $T \geq 2d^2(1 + \log d)$ and therefore $1 - 2d\exp\{-\frac{T}{2d^2}\} \geq 1 - \frac{2}{e} > \frac{1}{4}$. ∎

Part II

Dynamic Pricing under Constraints

Overview

In this part, we study the problem of *pricing under constraints* which is prevalent in contemporary markets. Various constraints significantly influence the operations and decisions of market participants, altering the principles of pricing strategies compared to those in unconstrained environments.

Unlike Part I, in this part we focus on non-contextual pricing problems. As a result, the optimal target (i.e. the regret baseline) at each time period $t = 1, 2, \ldots, T$ is fixed. However, the introduction of constraints adds structural complexity to these problems, which we will address through our research in the upcoming chapters.

Pricing with Fairness Concerns. Customers often comparing with each other about their price offers on similar purchases. Despite the legality of personalized pricing in retail, customers are perceptive of unfair treatment in prices. In Chapter 5, we consider two types of perceived unfairness.

1. **Procedural unfairness.** This form of unfairness arises when *customers* receive significantly different price offers for the same product based on varied characteristics, backgrounds, or other information. It occurs solely within the pricing process and can manifest even before any transaction.

2. **Substantive unfairness.** This occurs post-transaction, where *buyers* end up paying different amounts for the same product based on distinctive demands or budgets. It is measured among customers who have already decided to purchase the product.

In Chapter 5, we quantitatively define these two concepts under the framework of *randomized* price. We aim at finding the best price under the constraints of eliminating both unfairness. Although a fixed-and-indifferent price is always fair enough, we demonstrate through an example that a randomized pricing approach can potentially yield higher revenue while satisfying fairness constraints. Based on this observation, we develop an algorithm that aims at learning the optimal fair (randomized) pricing policy, which achieves $\tilde{O}(\sqrt{T})$ optimal regret and $\tilde{O}(\sqrt{T})$ optimal unfairness. Moreover, we provide a trade-off lower bound on regret and unfairness, stating that any algorithm having $O(\sqrt{T})$ regret must incur at least $\Omega(\sqrt{T})$ in unfairness.

Pricing with Inventory-Censoring Effect. The crux of pricing strategy is to balance the price with the demand. However, in some scenarios where the seller has a limited *inventory*, the realized demand reflected through sales quantity might not represent the true market desire. Sometimes the seller is able to adjust their inventory level and co-optimize the price and inventory decision. However, in fixed inventory situations such as cinema ticket sales, any demand exceeding the number of seats will vanish without being observed.

In Chapter 6, we investigate a pricing model with a *fixed* inventory level and *unobservable* excess demand. We propose an algorithm that achieves $O(\sqrt{T})$ regret, and show its optimality by proving a matching lower bound.

Chapter 5

Pricing with Fairness Concerns

In this chapter, we study the problem of online dynamic pricing with two types of fairness constraints: a *procedural fairness* which requires the *proposed* prices to be equal in expectation among different groups, and a *substantive fairness* which requires the *accepted* prices to be equal in expectation among different groups. A policy that is simultaneously procedural and substantive fair is referred to as *doubly fair*. We show that a doubly fair policy must be random to have higher revenue than the best trivial policy that assigns the same price to different groups. In a two-group setting, we propose an online learning algorithm for the 2-group pricing problems that achieves $\tilde{O}(\sqrt{T})$ regret, zero procedural unfairness and $\tilde{O}(\sqrt{T})$ substantive unfairness over T rounds of learning. We also prove two lower bounds showing that these results on regret and unfairness are both information-theoretically optimal up to iterated logarithmic factors. To the best of our knowledge, this is the first dynamic pricing algorithm that learns to price while satisfying two fairness constraints at the same time.

5.1 Introduction

Pricing problems have been studied since Cournot [1897]. In a classical pricing problem setting such as Kleinberg and Leighton [2003], Broder and Rusmevichientong [2012], Besbes and Zeevi [2015], the seller (referred as "we") sells identical products in the following scheme.

Online pricing. For $t = 1, 2, \ldots, T$:

1. The customer valuates the product as y_t.
2. The seller proposes a price v_t concurrently without knowing y_t.
3. The customer makes a decision $\mathbf{1}_t = \mathbf{1}(v_t \leq y_t)$.
4. The seller receives a reward (revenue) $r_t = v_t \cdot \mathbf{1}_t$.

Here T is the time horizon known to the seller in advance[1], and y_t's are drawn from a fixed distribution independently. The goal is to approach an optimal price that maximizes the expected revenue-price function. In order to make this, we should learn gradually from the binary feedback and improve our knowledge on customers' valuation distribution (or so-called "demands" [Kleinberg and Leighton, 2003]).

In recent years, with the development of price discrimination and personalized pricing strategies, *fairness* issues on pricing arose social and academic concerns [Kaufmann et al., 1991, Chapuis, 2012, Richards et al., 2016, Eyster et al., 2021]. Customers are usually not satisfied with price discrimination, which would in turn undermine both their willing to purchase and the sellers' reputation. In the online pricing problem defined above, when we are selling identical items to customers from different *groups* (e.g., divided by gender, race, age, etc.), it can be unfair if we propose a specific optimal price for each group: These

[1]Here we assume T known for simplicity of notations. In fact, if T is unknown, then we may apply a "doubling epoch" trick as Javanmard and Nazerzadeh [2019] and the regret bounds are the same.

optimal prices in different groups are not necessarily the same, and unfairness occurs if different customers are provided or buying the same item with different prices. Inspired by the concept of *procedural and substantive unconscionability*[Elfin, 1988], we define a *procedural unfairness* measuring the difference of *proposed prices* between the two groups, and a *substantive unfairness* measuring that of *accepted prices* between the two groups. Given these notions, our goal is to approach the optimal pricing policy that maximizes the expected total revenue with no procedural and substantive unfairness.

The concept of procedural fairness has been well established in Cohen et al. [2022] as "price fairness", while the concept of the substantive fairness is new to this paper. In fact, both procedural and substantive fairness have significant impacts on customers' experience and social justice. For instance, these notions help prevent the following two scenarios:

- Perspective buyers who are women found that they are offered consistently higher average price than men for the same product.
- Women who have bought the product found that they paid a higher average price than men who have bought the product.

Therefore, a good pricing strategy has to satisfy both procedural and substantive fairness.

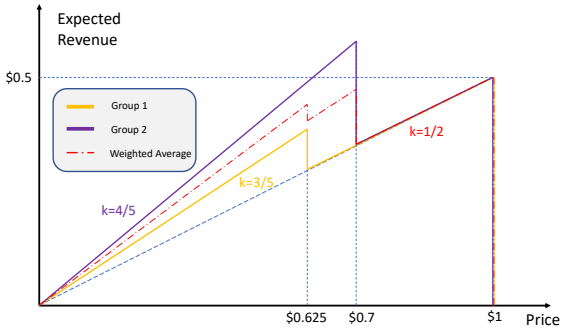
However, these constraints are very hard to satisfy even with full knowledge on customers' demands. If we want to fulfill those two sorts of fairness perfectly by proposing deterministic prices for different groups, the only thing we can do is to trivially set the same price in all groups and to maximize the weighted average revenue function by adjusting this uniformly fixed price with existing methods such as Kleinberg and Leighton [2003].

Consider the following example:

Example 5.1.1. Customers form two disjoint groups, where 30% customers are in Group 1 and the rest 70% are in Group 2.

For each price in $\{\$0.625, \$0.7, \$1\}$, customers in two groups have different acceptance rates:

Acceptance Rate	\$0.625	\$0.7	\$1
G_1 (30%)	3/5	1/2	1/2
G_2 (70%)	4/5	4/5	1/2



The right figure shows the expected revenue functions of prices in each group, where the red dashed line is their weighted average by population.

In Example 5.1.1, the only way to guarantee both fairness constraints is to propose the same price for both groups, and the optimal price is *\$1* whose expected revenue is *\$0.5* as is shown in the figure.

However, if we instead propose a *random price distribution* to each group and inspect those fairness notions *in expectation*, then there may exist price distributions that satisfy both fairness constraints and achieve higher expected revenue than any fixed-price strategy. Here a price distribution is the distribution over the prices for customers, and the exact price for each customer is *sampled* from this distribution *independently*. This random price sampling process can be implemented by marketing campaigns such as random discounts or randomly-distributed coupons. Again, we consider Example 5.1.1 and the

following random policy:

- For customers from G_1, propose *\$0.625* with probability $\frac{20}{29}$ and *\$1* with probability $\frac{9}{29}$.
- For customers from G_2, propose *\$0.7* with probability $\frac{25}{29}$ and *\$1* with probability $\frac{4}{29}$.

Under this policy, the expected proposed price and the expected accepted price in both groups are $\$\frac{43}{58}$ and $\$\frac{8}{11}$ respectively. Furthermore, the expected revenue is $\$\frac{74}{145} > \$0.5$, which means that this random policy performs better than the best fixed-price policy. It is worth mentioning that this is exactly the optimal random policy in this specific setting, but the proof of its optimality is highly non-trivial (and we put it in Section 5.8.3 as part of the proof of Theorem 5.5.4).

In this chapter, we consider a two-group setting and we denote a *policy* as the tuple of two price distributions over the two groups respectively. Therefore, we can formally define the optimal policy as follows:

$$\begin{aligned} \pi_* = & \operatorname*{argmax}_{\pi=(\pi^1,\pi^2)} q \cdot \mathop{\mathbb{E}}_{v_t^1\sim\pi^1, y_t^1\sim\mathbb{D}^1}[v_t^1 \cdot \mathbb{1}(v_t^1 \le y_t^1)] + (1-q) \cdot \mathop{\mathbb{E}}_{v_t^2\sim\pi^2, y_t^2\sim\mathbb{D}^2}[v_t^2 \cdot \mathbb{1}(v_t^2 \le y_t^2)] \\ s.t. \quad & \mathbb{E}_{\pi^1}[v_t^1] = \mathbb{E}_{\pi^2}[v_t^2] \\ & \mathbb{E}_{\pi^1,\mathbb{D}^1}[v_t^1 | \mathbb{1}(v_t^1 \le y_t^1) = 1] = \mathbb{E}_{\pi^2,\mathbb{D}^2}[v_t^2 | \mathbb{1}(v_t^2 \le y_t^2) = 1] \end{aligned} \tag{5.1}$$

Here $\pi^1, v_t^1, y_t^1, \mathbb{D}^1$ and $\pi^2, v_t^2, y_t^2, \mathbb{D}^2$ are the proposed price distributions, proposed prices, customer's valuations and valuation distributions of Group 1 and Group 2 respectively, and q is the proportion of Group 1 among all customers. From Eq. (5.1), the optimal policy

under the in-expectation fairness constraints should be random in general[2]. However, even we know the exact $\mathbb{D}^1$ and $\mathbb{D}^2$, it is still a very hard problem to get π_*: Both sides of the second constraint in Eq. (5.1) are conditional expectations (i.e., fractions of expected revenue over expected acceptance rate) and is thus not convex (and also not quasiconvex). To make it harder, the seller actually has no direct access to customers' demands $\mathbb{D}^1$ and $\mathbb{D}^2$ at the beginning. Therefore, in this chapter we consider a T-round *online* learning and pricing setting, where we could learn these demands from those *Boolean-censored* feedback (i.e., customers' decisions) and improve our pricing policy to approach π_* in Eq. (5.1).

In order to measure the performance of a specific policy, we define a *regret* metric that equals the expected revenue difference between this policy and the optimal policy. We also quantify the procedural and substantive unfairness that equals the absolute difference of expected proposed/accepted prices in two groups. We will establish a more detailed problem setting in Section 5.3.

Summary of Results Our contributions are threefold:

- We design an algorithm, FPA, that achieves an $O(\sqrt{T}d^{\frac{3}{2}}\log\frac{d\log T}{\epsilon})$ cumulative regret with 0 procedural unfairness and $O(\sqrt{T}d^{\frac{3}{2}}\log\frac{d\log T}{\epsilon})$ substantive unfairness, with probability at least $(1-\epsilon)$. Here d is the total number of prices allowed to be chosen from. These results indicate that our FPA is asymptotically no-regret and fair as T gets large.

- We show that the regret of FPA is optimal with respect to T, as it matches $\Omega(\sqrt{T})$ regret lower bound up to $\log\log T$ factors.

[2]Notice that a fixed-price policy can also be considered as "random".

- We show that the unfairness of FPA is also optimal with respect to T up to $\log \log T$ factors, as it has no procedural unfairness and its substantive unfairness matches the $\Omega(\sqrt{T})$ lower bound for any algorithm achieving an optimal $O(\sqrt{T})$ regret.

To the best of our knowledge, we are the first to study a pricing problem with multiple fairness constraints, where the optimal pricing policy is necessary to be random. We also develop an algorithm that is able to approach the best random pricing policy with high probability and at the least cost of revenue and fairness.

Technical Novelty. Our algorithm is a "conservative policy-elimination"-based strategy that runs in epochs with doubling batch sizes as in Auer et al. [2002a]. We cannot directly apply the action-elimination algorithm for multi-armed bandits as in Cesa-Bianchi et al. [2013], because the policy space is an infinite set and we cannot afford to try each one out. The fairness constraints further complicate things. Our solution is to work out just a few representative policies that are "good-and-exploratory", which can be used to evaluate the revenue and fairness of all other policies, then eliminate those that are unfair or have suboptimal revenue. Since we do not have direct access to the demand function, the estimated fairness constraints are changing over epochs due to estimation error, it is non-trivial to keep the target optimal policy inside our "good policy set" during iterations. We settle this issue by setting the criteria of a "good policy" conservatively.

Our lower bound is new too and it involves techniques that could be of independent interest to the machine learning theory community. Notice that it is possible to have a perfectly fair algorithm by trivially proposing the same fixed price for both groups. It is highly non-trivial to show the unfairness lower bound within the family of regret-optimal algorithms. We present our result in Section 5.5.3 by establishing two similar

problem settings that any algorithm cannot distinguish them efficiently and showing that a mismatch would cause a compatible amount of regret and substantive unfairness.

5.2 Related Works

Here we discuss existing literature on dynamic pricing, and fairness in machine learning, online learning and pricing. These aspects and works are closely related to this paper.

Dynamic Pricing Single product dynamic pricing problem has been well-studied through Kleinberg and Leighton [2003], Besbes and Zeevi [2009], Wang et al. [2014], Chen et al. [2019b], Wang et al. [2021b]. The crux is to learn and approach the optimal of a revenue curve from Boolean-censored feedback. In specific, Kleinberg and Leighton [2003] proves $\Theta(\log\log T), \Theta(\sqrt{T})$ and $\Theta(T^{\frac{2}{3}})$ minimax regret bounds under noise-free, infinitely smooth and stochastic/adversarial valuation assumptions, sequentially. Wang et al. [2021b] further shows a $\Theta(T^{\frac{K+1}{2K+1}})$ minimax regret bound for K^{th}-smooth revenue functions. In all these works, the decision space is continuous. In our problem setting, we fix the proposed prices to be chosen in a fixed set of d prices, and show a bandit-style $\Omega(\sqrt{dT})$ regret lower bound with a similar method to Auer et al. [2002b].

Fairness in Machine Learning Fairness is a long-existing topic that has been extensively studied. In machine learning society, fairness is defined from different perspectives. On the one hand, the concept of *group fairness* requires different groups to receive identical treatment in statistics. In a classification problem, for instance, there are mainly two different types of group fairness: (1) A "demographic parity" [Dwork et al., 2012] that requires the outcome of a classifier to be statistically independent to the group information,

and (2) an "equalized odds" (including "equal opportunity" as a relaxation) [Hardt et al., 2016] that requires the prediction of a classifier to be *conditionally* independent to the group information given the true label. In Agarwal et al. [2018], these probabilistic constraints are further modified as linear constraints, and therefore the fair classification problem is reduced to a cost-sensitive classification problem. It is worth mentioning that Agarwal et al. [2018] *allows* an ϵ_k-unfairness due to the learning error and *assumes* $\epsilon_k = O(n^{-\alpha})$ with some $\alpha \leq \frac{1}{2}$, while we *quantify* the learning-caused unfairness and *upper and lower bound* the cumulative unfairness without pre-assuming its scale.

On the other hand, [Dwork et al., 2012] also proposes the concept of "individual fairness" (or "Lipschitz property") where the difference of treatments toward two individuals should be upper bounded by a distance metric of their intrinsic features, i.e., $D(\mu_x, \mu_y) \leq d(x, y)$ where x, y are features and μ_x, μ_y are the distributions of actions onto x and y respectively. The notion "time fairness" is often considered as individual fairness as well. For a more inclusive review on different definitions of fairness in machine learning, please refer to Barocas et al. [2017].

Fairness in Online Learning Besides existing works on general machine learning fairness, there are some works that study online-learning or bandit problems. This is similar to our setting as we adopt an online pricing process. Among these works, Joseph et al. [2016] studies multi-armed and contextual bandits with fairness constraints. Their non-contextual setting is related to our works as our pricing problem can also be treated as a bandit. Their definition of δ-fairness is defined as comparisons among probabilities of taking actions, which is similar to our definition on procedural fairness. However, their fairness definitions are defined from the perspective of arms (i.e. actions): better actions worth larger probability to take. In comparison, our fairness definitions are more

on the results: different groups share the same expected prices. Bechavod et al. [2020] studies an online learning (in specific, an online classification) problem with unknown and non-parametric constraints on individual fairness at each round. They develop an adaptive algorithm that guarantees an $O(\sqrt{T})$ regret as well as an $O(\sqrt{T})$ cumulative fairness loss. However, their problem settings are quite different from ours. Primarily, they assume individual fairness as a constraint, while our fairness definitions are indeed group fairness. Also, their online classification problem is different from our online pricing problem as they have full access to the regret function while we even do not have full-information reward (i.e., which is Boolean-censored). Similarly, their fairness loss is accessible although the unfair pairs of (τ_1, τ_2) are not fully accessible, while in our settings we do not know the $S(\pi; F_1, F_2)$ function at all. Besides, we have to satisfy two constraints at one time and one of them (the substantive fairness) is highly non-convex. Gupta and Kamble [2021] studies an online learning problem with two different sorts of individual fairness constraints over time: a "fairness-across-time" (FT) and a "fairness-in-hindsight" (FH). They show that it is necessary to have a linear regret under FT constraints, and they also propose a CAFE algorithm that achieves an optimal regret under FH constraints.

Despite the specific properties of fairness constraints, we may also consider the framework of constraint online learning. Yu et al. [2017] studies an online convex optimization (OCO) problem with stochastic constraints, which might be applicable to online fair learning. However, their problem settings and methodologies are largely different from ours: Firstly, their constraints are assumed convex while our substantive fairness constraint (i.e., the $S(\pi; F_1, F_2)$ function) is highly non-convex. Also, they have a direct access to the realized objective function $f^t(x_t)$ at each time while our pricing problem only has a Boolean-censored feedback. More importantly, Yu et al. [2017] assumes the availability of unbiased samples on constraint-related variables. In specific, their constraints are roughly

$g_k(x) < 0$, and by the end of each period t they receive an unbiased sample of $g_k(x_t)$ for the x_t they have taken. On the contrary, we do not have any unbiased sample of $S(\pi, F_1, F_2)$ at each time, since there is only one customer from one of the two groups. Therefore, we cannot make use of their results in our problem setting.

Fairness in Pricing There are many recent works contributing to pricing fairness study [Kaufmann et al., 1991, Frey and Pommerehne, 1993, Chapuis, 2012, Richards et al., 2016, Priester et al., 2020, Eyster et al., 2021, Yang et al., 2022]. As is stated in Cohen et al. [2022], in a pricing problem with fairness concerns, the concept of fairness in existing works is modeled either as a utility or budget that trades-off the revenue or as a hard constraint that prevent us from taking the best action directly.Cohen et al. [2022] chooses the second model and defines four different types of fairness in pricing: price fairness, demand fairness, surplus fairness and no-purchase valuation fairness, each of which indicates the difference of prices, the acceptance rate, the surplus (i.e., (valuation − price) if bought and 0 otherwise) and the average valuation of not-purchasing customers in two groups is bounded, sequentially. They show that it is impossible to achieve any pair of different fairness notions simultaneously (with deterministic prices). In fact, this can be satisfied if they allow random pricing policies. Maestre et al. [2018] indeed builds their fairness definition upon random prices by introducing a "Jain's Index", which indicates the homogeneity of price distributions among different groups (i.e., our procedural fairness notion). They develop a reinforcement-learning-based algorithm to provide homogeneous prices, with no theoretic guarantees. Besides, Richards et al. [2016] discusses some fairness issues regarding personalized pricing from the perspective of econometrics. Eyster et al. [2021] studies a phenomenon where customers would mistakenly attribute the cost increases to a time unfairness, and they propose methods to release customer's feeling of such

unfairness by adjusting prices correspondingly. Chapuis [2012] looked into two fairness concerns called *price fairness* and *pricing fairness*, which indicates the *distributional* and *procedural* fairness of the pricing process respectively, from the **seller**'s perspective. In fact, their price fairness is more likely to be our *procedural fairness* definition although it is not in their paper. This is because that we are considering the fairness from **customers**' perspective, where their observations on prices serve as a procedure of their decision process and their decision on whether or not to buy is actually indicating the fairness of results. There are more interesting works as is listed in Section 5.2, and we refer the readers to Chen et al. [2021b] where there is a more comprehensive review on pricing and fairness.

Cohen et al. [2021] and Chen et al. [2021b] study the online-learning-fashion pricing problem as we do. Cohen et al. [2021] considers both group (price) fairness and individual (time) fairness, and their algorithm FaPU solves this problem with sublinear regret while guaranteeing fairness. They further study the pricing problem with demand fairness that are unknown and needs learning. In this setting, they propose another FaPD algorithm that achieves the optimal $\tilde{O}(\sqrt{T})$ regret and guarantees the demand fairness "almost surely", i.e., upper bounded by $\delta \cdot T$ as a budget. Chen et al. [2021b] considers two different sorts of fairness constraints: (1) Price fairness constraints (as in Cohen et al. [2022]) are enforced; (2)Price fairness constraints are generally defined (and maybe not accessible), where they adopt "soft fairness constraints" by adding the fairness violation to the regret with certain weights. In both cases, they achieve $\tilde{O}(T^{\frac{4}{5}})$ regrets. These learning-based fairness requirements are quite similar to our problem setting, but in our setting the fairness constraints are non-convex (while theirs are linear) and are also optimized to corresponding information-theoretic lower bounds without undermining the optimal regret.

5.3 Problem Setup

In this section, we describe the problem setting of online pricing, introduce new fairness definitions and set the goal of our algorithm design.

Problem Description. We start with the online pricing process. The whole selling session involves customers from two groups (G_1 and G_2) and lasts for T rounds. Prices are only allowed to be chosen from a *known* and *fixed* set of d prices: $\mathbf{V} = \{v_1, v_2, \ldots, v_d\}$, where $0 < v_1 < v_2 < \ldots < v_d \leq 1$. Denote $\Delta^d = \{x \in \mathbb{R}_+^d, \|x\|_1 = 1\}$ as the probabilistic simplex. At each time $t = 1, 2, \ldots, T$, we propose a pricing policy $\pi = (\pi^1, \pi^2)$ consisting of two probabilistic distributions $\pi^1, \pi^2 \in \Delta^d$ over all d prices. A customer then arrives with an observable group attribution G_e ($e \in \{1, 2\}$), and we propose a price by sampling a v_t^e from $\mathbf{V}$ according to distribution π^e. At the same time, the customer generates a valuation y_t^e *in secret*, where y_t^e is sampled independently and identically from some fixed unknown distribution $\mathbb{D}_e$. Afterward, we observe a feedback $\mathbf{1}_t^e = \mathbf{1}(v_t^e \leq y_t^e)$ and receive a reward(revenue) $r_t^e = \mathbf{1}_t^e \cdot v_t^e$.

Key Quantities. Here we define a few quantities and functions that is necessary to formulate the problem. Denote $\mathbf{v} := [v_1, v_2, \ldots, v_d]^\top$, $[d] := \{1, 2, \ldots, d\}$ and $\mathbb{1} := [1, 1, \ldots, 1]^\top \in \mathbb{R}^d$ for simplicity. Denote $F_e(i) := \Pr_{\mathbb{D}_e}[y_t^e \geq v_i], e = 1, 2, i \in [d]$ as the probability of price v_i being accepted in G_e, and we know that $F_e(1) \geq F_e(2) \geq \ldots \geq F_e(d)$. Notice that all $F_e(i)$'s are *unknown* to us. Define a matrix $F_e := \mathrm{diag}(F_e(1), F_e(2), \ldots, F_e(d))$.

As a result, for a customer from G_e ($e \in \{1, 2\}$), we know that

- The expected proposed price is $\mathbf{v}^\top \pi^e$.

- The expected reward(revenue) is $\mathbf{v}^\top F_e \pi^e$.
- The expected acceptance rate is $\mathbb{1}^\top F_e \pi^e$.
- The expected accepted price is $\frac{\mathbf{v}^\top F_e \pi^e}{\mathbb{1}^\top F_e \pi^e}$.

Denote the proportion of G_1 in all potential customers as q $(0 < q < 1)$ which is fixed and *known* to us, and we assume that every customer is chosen from all potential customers uniformly at random. As a consequence, we can define the expected revenue of a policy π.

Definition 5.3.1 (Expected Revenue)**.** For any pricing policy $\pi = (\pi^1, \pi^2) \in \mathit{\Pi}$, define its expected revenue (given F_1 and F_2) as the weighted average of the expected rewards of G_1 and G_2.

$$\begin{aligned} R(\pi; F_1, F_2) :=& \Pr[\text{Customer is from } G_1] \cdot \mathbb{E}[r_t^1] + \Pr[\text{Customer is from } G_2] \cdot \mathbb{E}[r_t^2] \\ =& q \cdot \mathbf{v}^\top F_1 \pi^1 + (1-q) \cdot \mathbf{v}^\top F_2 \pi^2 \end{aligned} \tag{5.2}$$

Also, we can define the two different unfairness notions based on these results above.

Definition 5.3.2 (Procedural Unfairness)**.** For any pricing policy $\pi \in \mathit{\Pi}$, define its procedural unfairness as the absolute difference between the expected *proposed* prices of two groups.

$$U(\pi) := |\mathbf{v}^\top \pi^1 - \mathbf{v}^\top \pi^2| = |\mathbf{v}^\top (\pi^1 - \pi^2)|. \tag{5.3}$$

Procedural unfairness is totally tractable as we have full access to $\mathbf{v}^\top$ and π. Therefore, we can define a policy family $\mathit{\Pi} := \{\pi = (\pi^1, \pi^2), U(\pi) = 0\}$ that contains all policies with no procedural unfairness. Now we define a substantive unfairness as another metric.

Definition 5.3.3 (Substantive Unfairness)**.** For any pricing policy $\pi \in \mathit{\Pi}$, define its substantive unfairness as the difference between the expected *accepted* prices of two groups.

$$\begin{aligned} S(\pi; F_1, F_2) :=& \Big| \mathbb{E}[v^1 | v^1 \sim \pi^1, v^1 \text{ being accepted }] - \mathbb{E}[v^2 | v^2 \sim \pi^2, v^2 \text{ being accepted }] \Big| \\ =& \left| \frac{\mathbf{v}^\top F_1 \pi^1}{\mathbb{1}^\top F_1 \pi^1} - \frac{\mathbf{v}^\top F_2 \pi^2}{\mathbb{1}^\top F_2 \pi^2} \right|. \end{aligned} \tag{5.4}$$

Substantive unfairness is not as tractable as procedural unfairness, as we have no direct access to the true F_1 and F_2. Ideally, the optimal policy that we would like to achieve is:

$$\begin{aligned} \pi_* = & \underset{\pi=(\pi^1,\pi^2)\in\mathit{\Pi}}{\operatorname{argmax}} \; R(\pi; F_1, F_2) \\ s.t. \quad & U(\pi) = 0, \quad S(\pi; F_1, F_2) = 0. \end{aligned} \tag{5.5}$$

The feasibility of this problem is trivial: policies such as $\pi^1 = \pi^2 = [0, \ldots, 0, 1, 0, \ldots, 0]^\top$ (i.e., proposing the same fixed price despite the customer's group attribution) are always feasible. However, this problem is in general highly non-convex and non-quasi-convex. Finally, we define a *(cumulative) regret* that measure the performance of any policy π:

Definition 5.3.4 (Regret)**.** For any algorithm $\mathcal{A}$, define its cumulative regret as follows:

$$Reg_T(\mathcal{A}) := \sum_{t=1}^{T} Reg(\pi_t; F_1, F_2) := \sum_{t=1}^{T} R(\pi_*; F_1, F_2) - R(\pi_t; F_1, F_2). \tag{5.6}$$

Here π_t is the policy proposed by $\mathcal{A}$ at time t.

Notice that we define the per-round regret by comparing the performance of π_t with the optimal policy π_* under constraints. Therefore, $Reg(\pi_t; F_1, F_2)$ is possible to be negative

if $\pi \in \Pi$ but $U(\pi_t) > 0$ or $S(\pi_t; F_1, F_2) > 0$. Similarly, we define a *cumulative substantive unfairness* as $S_T(\mathcal{A}) := \sum_{t=1}^{T} S(\pi; F_1, F_2)$.

Goal of Algorithm Design Our ultimate goal is to approach π_* in the performance. In the online pricing problem setting we adopt, however, we cannot guarantee $S(\pi_t; F_1, F_2) = 0$ for all π_t we propose at time $t = 1, 2, \ldots, T$ since we do not know F_1 and F_2 in advance. Instead, we may suffer a gradually vanishing unfairness as we learn F_1 and F_2 better. Therefore, our goal in this chapter is to design an algorithm that guarantees an optimal regret while suffering 0 cumulative procedural unfairness and the least cumulative substantive unfairness.

Technical Assumptions. Here we make some mild assumptions that help our analysis.

Assumption 5.3.5 (Least Probability of Acceptance)**.** There exists a fixed constant $F_{\min} > 0$ such that $F_e(d) \geq F_{\min}, e = 1, 2$.

Assumption 5.3.5 not only ensures the definition of expected accepted price to be sound (by ruling out these unacceptable prices), but also implies $S(\pi, F_1, F_2)$ to be Lipschitz. Besides, we can always achieve this by reducing v_d. We will provide a detailed discussion in Section 5.6.1.

Assumption 5.3.6 (Number of Possible Prices)**.** We treat d, the number of prices, as an amount independent from T. Also, we assume $d = O(T^{\frac{1}{3}})$.

Assumption 5.3.6 is a necessary condition of applying $\Omega(\sqrt{dT})$ regret lower bound, and here we make it to show the optimality of our algorithm w.r.t. T.

5.4 Algorithm

In this section, we propose our Fairly Pricing Algorithm (FPA) in Algorithm 1 and then discuss the techniques we develop and apply to achieve the "no-regret" and "no-unfairness" goal.

5.4.1 Algorithm Components

Algorithm 7 takes the following inputs: time horizon T, price set $\mathbf{V}$, error probability ϵ, a universal constant L as the coefficient of the performance-fairness tradeoff on constraint relaxations (see Lemma 5.8.5), and q as the proportion that G_1 takes. In the "before epochs" stage, we keep proposing the highest price v_d for $\tau_0 = O(\log T)$ rounds to estimate(lower-bound) the least acceptance rate $F_{\min}$. We also adopt the following techniques that serve as components of FPA and contribute to its no-regret and no-unfairness performance.

Doubling Epochs Despite the "before epochs" stage, we divide the whole time space into epochs $k = 1, 2, \ldots$, where each epoch k has a length $\tau_k = O(\sqrt{T} \cdot 2^k)$ that doubles that of epoch $(k-1)$. Within each epoch k, we run a set of "good-and-exploratory policies" (to be introduced in Section 5.4.1) with equal shares of τ_k. At the end of each epoch k, we update the estimates of F_1 and F_2, eliminate the sub-optimal policies and update the set of "good-and-exploratory policies" for the next epoch. Since the estimates of parameters get better as k increases, a doubling-epoch trick would ensure that we run better policies in longer epochs and therefore save the regret.

Algorithm 7 Fairly Pricing Algorithm (FPA)

Input: Time horizon T, prices set $\mathbf{V}$, error probability ϵ, proportion q, constant L.

Before Epochs: Keep proposing the highest price v_d for $\tau_0 = 2\log T \log \frac{16}{\epsilon}$ rounds. Estimate the average acceptance rates as $\bar{F}_d(1)$ and $\bar{F}_d(2)$. Take $\hat{F}_{\min} = 0.5\min\{\bar{F}_d(1), \bar{F}_d(2)\}$.

Initialization: Parameters C_q, c_t. Epoch length $\tau_k = O(d\sqrt{T}\cdot 2^k)$, reward uncertainty $\delta_{k,r}$ and unfairness uncertainty $\delta_{k,s}$ for $k = 1, 2, \ldots, O(\log T)$. Candidate policy set $\mathit{\Pi}_1 = \mathit{\Pi} := \{\pi = (\pi^1, \pi^2), U(\pi) = 0\}$ and price index set $I_0^1 = I_0^2 = [d]$.

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

  Set $A_k = \emptyset$, $I_k^1 = I_{k-1}^1$ and $I_k^2 = I_{k-1}^2$.

  for Group $e = 1, 2$ and for price index $i \in I_{k-1}^e$, **do**

    Get $\tilde{\pi}_{k,i,e} = \text{argmax}_{\pi\in\mathit{\Pi}_k} \pi^e(i)$. {Pick up policy maximizing each acceptance rate}

    If $\tilde{\pi}_{k,i,e}^e(i) \geq \frac{1}{\sqrt{T}}$, let $A_k = A_k \cup \{\tilde{\pi}_{k,i,e}\}$. Otherwise, remove i from I_k^e.

  end for

  Set $M_{k,e}(i) = N_{k,e}(i) = 0, \forall i \in [d], e = 1, 2$.

  for each policy $\pi \in A_k$, **do**

    Run π for $\frac{\tau_k}{|A_k|}$ rounds. {Sample random prices at $t = 1, 2, \ldots, \frac{\tau_k}{|A_k|}$ repeatedly.}

    For each time a price v_i is *proposed* in G_e, set $M_{k,e}(i) += 1$.

    For each time a price v_i is *accepted* in G_e, set $N_{k,e}(i) += 1$.

  end for

  For $e = 1, 2$, set $\bar{F}_{k,e}(i) = \max\{\frac{N_{k,e}(i)}{M_{k,e}(i)}, \hat{F}_{\min}\}$ for $i \in I_k^e$, and $\bar{F}_{k,e}(i) = \hat{F}_{\min}$ otherwise.

  Let $\hat{F}_{k,e} = \text{diag}(\bar{F}_{k,e}(1), \bar{F}_{k,e}(2), \ldots, \bar{F}_{k,e}(d)), e = 1, 2$.

  Solve the following optimization problem and get the empirical optimal policy $\hat{\pi}_{k,*}$.

$$\hat{\pi}_{k,*} = \underset{\pi\in\mathit{\Pi}_k}{\text{argmax}}\, R(\pi, \hat{F}_{k,1}, \hat{F}_{k,2}),\ \text{s.t. } S(\pi, \hat{F}_{k,1}, \hat{F}_{k,2}) \leq \delta_{k,s}. \tag{5.7}$$

  To eliminate largely suboptimal or unfair policies, construct

$$\mathit{\Pi}_{k+1} = \{\pi : \pi \in \mathit{\Pi}_k, S(\pi, \hat{F}_1, \hat{F}_2) \leq \delta_{k,s},\ R(\pi, \hat{F}_1, \hat{F}_2) \geq R(\hat{\pi}_{k,*}, \hat{F}_1, \hat{F}_2) - \delta_{k,r} - L\cdot\delta_{k,s}\}. \tag{5.8}$$

end for

Policy Eliminations At the end of each epoch k, we update the candidate policy set by eliminating those substantially sub-optimal policies: Firstly, we select an empirical optimal policy $\hat{\pi}_{k,*}$ that maximizes $R(\pi, \hat{F}_{k,1}, \hat{F}_{k,2})$ while guaranteeing $S(\pi, \hat{F}_{k,1}, \hat{F}_{k,2}) \leq \delta_{k,s}$. After that, we eliminate those policies that satisfy one of the following two criteria:

- Large unfairness: $S(\pi, \hat{F}_{k,1}, \hat{F}_{k,2}) > \delta_{k,s}$, or
- Large regret: $R(\pi, \hat{F}_{k,1}, \hat{F}_{K,2}) < R(\hat{\pi}_{k,*}, \hat{F}_{k,1}, \hat{F}_{k,2}) - \delta_{k,r} - L \cdot \delta_{k,s}$.

Here we adopt two subtractors on the regret criteria: $\delta_{k,r}$ for the estimation error in $R(\pi)$ caused by $\hat{F}_{k,e}$, and $L \cdot \delta_{k,s}$ for the possible increase of optimal reward by allowing $S(\pi) \leq \delta_{k,s}$ instead of $S(\pi) = 0$. In this way, we can always ensure the optimal policy π_* (i.e., the solution of Eq. (5.5)) to remain and also guarantee the other remaining policies perform similarly to π_*.

Good-and-Exploratory Policies Although all remaining policies perform good, not all of them are suitable of running in consideration of exploration. It is important to update the estimates of all $F_1(i)$ and $F_2(i)$ as they are required in the policy elimination. We solve this issue by keeping a set of *good-and-exploratory* policies: After eliminating sub-optimal policies at the end of previous epoch, for each price v_i in group G_e we find out a policy in the remaining policies that maximizes the probability of proposing v_i in G_e at the beginning of current epoch. The larger this probability is, the more times v_i can be chosen in G_e, which would lead to a better estimate of $F_e(i)$. Here we give up to estimate the acceptance probability of those v_i with $\leq \frac{1}{\sqrt{T}}$ to be chosen by the optimal policy π_*, as it would not affect the elimination process and the performance substantially.

5.4.2 Computational Cost

Our FPA algorithm is *oracle-efficient* due to the doubling-epoch design, as we only run each oracle and update each parameter for $O(\log T)$ times. However, the implementation of these oracles could be time-consuming: On the one hand, each set Π_k contains infinite policies, and a discretization would lead to exponential computational cost w.r.t. d. On the other hand, both Eq. (5.7) and Eq. (5.8) are highly non-convex on the constraints and are hard to solve with off-the-shelf methods.

5.5 Regret and Unfairness Analysis

In this section, we analyze the regret and unfairness of our FPA algorithm. We first present an $\tilde{O}(\sqrt{T}d^{\frac{3}{2}})$ regret upper bound along with an $\tilde{O}(\sqrt{T}d^{\frac{3}{2}})$ unfairness upper bound. Then we show both of them are optimal (w.r.t. T) up to $\log\log T$ factors by presenting matching lower bounds.

5.5.1 Regret Upper Bound

First of all, we propose the following theorem as the main results for our Algorithm 7 (FPA).

Theorem 5.5.1 (Regret and Unfairness)**.** *FPA guarantees an* $O(\sqrt{T}d^{\frac{3}{2}}\log\frac{d\log T}{\epsilon})$ *regret with no procedural unfairness and an* $O(\sqrt{T}d^{\frac{3}{2}}\log\frac{d\log T}{\epsilon})$ *substantive unfairness with probability* $1-\epsilon$.

Proof sketch. We prove this theorem by induction w.r.t. epoch index k. Firstly, we assume that $\pi_* \in \Pi_k$, which naturally holds as $k = 1$. Meanwhile, we show a high-probability

bound on the estimation error of each $F_e(i)$ for epoch k, according to concentration inequalities. Given this, we derive the estimation error bound of $R(\pi, F_1, F_2)$ and $S(\pi, F_1, F_2)$ for each policy $\pi \in \Pi_k$ in epoch k. After that, we bound the regret and unfairness of each policy remaining in Π_{k+1}, and therefore bound the regret and unfairness of epoch $(k+1)$ with high probability. Finally, we show that optimal fair policy π_* (defined in Eq. (5.5)) is also in Π_{k+1}, which matches the induction assumption for Epoch $(k+1)$. By adding up these performance over epochs, we get the cumulative regret and unfairness respectively. Please refer to Section 5.8.1 for a detailed proof. ■

Remark 5.5.2. Our algorithm guarantees $O(\sqrt{T}\log\log T)$ regret and unfairness simultaneously, whose average-over-time match the generic estimation error of $O(\frac{1}{\sqrt{T}})$. It implies that these fairness constraints do not bring informational obstacles to the learning process. In fact, these upper bounds are tight up to $O(\log\log T)$ factors, which are shown in Theorem 5.5.3 and Theorem 5.5.4.

5.5.2 Regret Lower Bound

Here we show the regret lower bound of this pricing problem.

Theorem 5.5.3 (Regret lower bound)**.** *Assume $d \le T^{\frac{1}{3}}$. Given the online two-group fair pricing problem and the regret definition as Eq.* (5.6)*, any algorithm would at least suffer an $\Omega(\sqrt{dT})$ regret.*

We may prove Theorem 5.5.3 by a reduction to online pricing problem with no fairness constraints: Given a problem setting where the two groups are identical, i.e. $F_1(i) = F_2(i), \forall i \in [d]$, and let $q = 0.5$. Notice that any policy satisfying $\pi^1 = \pi^2$ is fair, and the optimal policy is to always propose the best fixed price. Therefore, this can be reduced to

an online identical-product pricing problem, and we present a bandit-style lower bound proof in Section 5.8.2 inspired by Auer et al. [2002b].

5.5.3 Unfairness Lower Bound with Optimal Revenue

Here we show that any optimal algorithm has to suffer an $\Omega(\sqrt{T})$ substantive unfairness.

Theorem 5.5.4 (Substantive Unfairness Lower Bound)**.** *For any constant C_x, there exists constants $C_u > 0$ such that any algorithm with an $C_x \cdot T^{\frac{1}{2}}$ cumulative regret and zero procedural unfairness has to suffer an $C_u \cdot T^{\frac{1}{2}}$ substantive unfairness.*

It is worth mentioning that this result is different from ordinary lower bounds on the regret, as it also requires the algorithm to be optimal. In general, we propose 2 different problem settings, and we show the following four facts:

- No algorithm can perform well in both settings.
- Any algorithm cannot distinguish between the two settings very efficiently.
- Not trying to distinguish between them would suffer either a very large regret or a very large substantive unfairness, and therefore we cannot do this very often.
- Having tried but failed in distinguishing between them would definitely lead to a large substantive unfairness.

In order to prove this, we make use of Example 5.1.1 presented in Section 5.1. One of the settings is exactly Example 5.1.1, and the other one is identical to it except these 0.5 probabilities are now $(0.5 - \zeta)$ in both groups. We get close-form solutions to both

problem settings and show that they are indistinguishable in information theory. Please refer to Section 5.8.3 for more details.

5.6 Discussion

Here we primarily discuss some open issues and potential extensions of our results. We also remark on our problem settings, technique highlights and social impacts.

5.6.1 Extensions on Technical Assumptions.

In this chapter, we assumed the existence of a lower bound $F_{\min} > 0$ of the acceptance rate of all prices for both groups. This assumption is stronger than our expectation, as the seller would not know the highest price that customers would accept. We assume this for two reasons: (1) Without assuming $F_e(i) > 0$, the substantive unfairness function might be undefined. For instance, if a pricing policy is completely unacceptable in G_1 (with *no* accepted prices) but is acceptable in G_2, then is it a fair policy? (2) With a constantly large probability of acceptance, we can estimate every $F_e(i)$ and bound it away from 0 and therefore leads to the Lipschitzness of $S(\pi, \hat{F}_1, \hat{F}_2)$. However, there might exist an algorithm that works for $F_e(i) > 0$ generally and maintains these optimalities as well, which is an open problem to the future.

5.6.2 Feelings of Fairness in FPA.

In our FPA algorithm, notice that we run each $\tilde{\pi} \in A_k$ for a continuous batch of $\frac{\tau_k}{|A_k|} = \Omega(\sqrt{T} \cdot 2^k)$, which is long enough for customers to experience the fairness by comparing their proposed prices and accepted prices with customers from the other

group.

Relaxation on Substantive Fairness. Our algorithm approached the optimal policy as the solution of Eq. (5.5) through an online learning framework. This ensures an asymptotic fairness as $T \to +\infty$, but we still cannot guarantee an any-time fair algorithm precisely. Therefore, it is more practical to consider the following inequality-constraint optimization problem:

$$\pi_{\delta,*} = \underset{\pi=(\pi^1,\pi^2)\in\Pi}{\operatorname{argmax}} R(\pi; F_1, F_2) \qquad s.t. \quad U(\pi) = 0, \quad S(\pi; F_1, F_2) \le \delta. \tag{5.9}$$

Comparing Eq. (5.5) with Eq. (5.9), we know that $R(\pi_*) \le R(\pi_{\delta,*})$. According to Lemma 5.8.5, we further know that $R(\pi_*) \ge R(\pi_{\delta,*}) - L \cdot \delta$. Naturally, the substantive unfairness definition is now $\max\{0, S(\pi; F_1, F_2) - \delta\}$. If we still consider this problem under the framework of online learning, then two questions arose naturally: What are the optimal regret rate and (substantive) unfairness rate like? And how can we achieve them simultaneously? From our results in this chapter, we only know that (1) If $\delta = 0$, then both rates are $\Theta(\sqrt{T})$, and (2) if $\delta \ge 1$, then the optimal regret is $\Theta(\sqrt{T})$ and the optimal unfairness is 0 (as it is reduced to the unconstrained pricing problem). In fact, for $\delta = O(\sqrt{1/T})$, we may still achieve $O(\sqrt{T})$ regret and unfairness, but it is not clear if they are always optimal. For $\delta > \sqrt{1/T}$, we conjecture that the optimal regret is still $\Theta(\sqrt{T})$ and the optimal unfairness could be $\Theta(1/(\sqrt{T}\delta))$.

5.6.3 Optimal Policy on the Continuous Space.

We restrict our price choices in a fixed price set $\mathbf{V} = \{v_1, v_2, \ldots, v_d\}$ and aims at the optimal distributions on these v_i's. However, if we are allowed to propose any price within $[0, 1]$, then the optimal policy could be a tuple of two *continuous* distributions

that outperforms any policy restricted on $\mathbf{V}$. Even if we know that customers' valuations are all from $\mathbf{V}$, the optimal policy is not necessarily located inside $\mathbf{V}$ due to the fairness constraints. This optimization problem is even harder than Eq. (5.5), and the online-learning scheme further increases its hardness. Existing methods such as continuous distribution discretization [Xu and Wang, 2022] might work, but would definitely lead to an exponential time complexity.

5.6.4 Potential Generalizations of Current Problem Setting.

Currently we make a few technical assumptions that qualify the applications of our algorithm. In fact, these assumptions are mild and can be released by some tricks: On the one hand, we can always meet the requirement of Assumption 5.3.5 by reducing v_d. By running a binary-search algorithm for the highest acceptable price (with constant acceptance probability), we can find the feasible v_d within $O(\log T \log T)$ rounds (where $\log T$ for binary search and another $\log T$ for the concentration of a constant-expectation random variable, as we did in estimating $F_{\min}$). Since $O(\log T \log T)$ is much smaller than $O(\sqrt{T})$ as the optimal regret and unfairness, this would not harm the regret and unfairness substantially. On the other hand, we assume the prices to be chosen from a fixed and finite price set $\mathbf{V}$, which not only restricted our action but might lead to suboptimality from the perspective of a larger scope. In fact, if we allow the prices to be selected in the whole $[0, 1]$ range, a pricing policy can be a tuple of two continuous distributions over $[0, 1]$. To solve this problem, we may parametrize the distribution and learn the best parameters. We may also discretize the price space into small grids, i.e. prices are $\mathbf{V} = \{\gamma, 2\gamma, \ldots, (d-1)\gamma, d\gamma = 1\}$, where $\gamma = T^{-\alpha}$ with some constant α and d$= T^{\alpha}$ as a consequence. It is intrinsically a specific way of parametrization. According to the "half-Lipschitz" nature of pricing problem as well as our Lemma 5.8.5 along with

the Lipschiness of $S(\pi; F_1, F_2)$, we know that the per-round discretization error would be upper bounded by $O(T^{-\alpha})$. Let the cumulative discretization error $O(T^{\frac{1}{-}\alpha})$ balances the cumulative regret (or substantive unfairness), i.e., $O(d^{\frac{3}{2}}\sqrt{T}) = 0(T^{\frac{1}{2}+\frac{3\alpha}{2}})$, we can achieve an upper bound on both the regret at $O(T^{\frac{4}{5}})$ by letting $\alpha = T^{\frac{1}{5}}$. However, this is not optimal as we only match the upper and lower bounds w.r.t. T but not to d. Therefore, it would also be an interesting problem to see the minimax regret/unfairness dependent w.r.t. d.

Besides of the assumptions we have made, there are other notions regarding our problem setup that can be generalized. Firstly, we may generalize our problem setting from two groups to multiple $G \geq 3$ groups. Again, the feasible set is not empty as we can always propose the same fixed price to all groups. However, there is not a directly generalization of the fairness definition, which we defined as the difference of the expectation of certain amount between two groups. We might defined it as "pairwise unfairness" by comparing the same difference among each pair of groups and adding them up, but this is not rational: Consider the case when the expected proposed/accepted prices in $(G-1)$ groups are very high and the last one is very low, and compare this case with another case when the expected proposed/accepted prices in 1 groups are very high and the other $(G-1)$ groups has a very low expected prices. The unfairness in these two cases should be definitely different, as the first seems more acceptable (i.e., being kind to only the minority versus being kind to only the majority). However, their "pairwise unfairness" are exactly the same in these two cases. Therefore, a better notion of procedural/substantive unfairness should be established for multiple $G \geq 3$ groups.

Secondly, we may generalize the modeling on customers from i.i.d. to strategic. For example, what if a customer tries multiple times until getting the lowest price of the distribution for this group. This is an adaptive adversary and therefore very hard to deal

with even in the simplest decision-making process such as bandits.

Thirdly, we may also include more fairness concern. Currently we are considering the two types of fairness, but we define the cumulative fairness based as the summation of expected per-round unfairness. This definition does not take into consideration the changing of policies. For example, if we propose a fair policy at each round, but the policies over time changes drastically, then it is hard for the customers to feel or experience such a fairness. In our algorithm design of FPA, we always play the same policy for at least $\frac{\tau_k}{2d} = \Omega(\sqrt{T})$ rounds as a batch until the policy changes. This is a long enough time period for customers to experience fairness since at least a $Omega(\sqrt{T})$ number of customers from both groups would come and buy items under the same policy according to the Law of Large Numbers. However, this would still cause a feeling of unfairness for the two customers who are arriving almost simultaneously but the policy is just changed after the first customer buy or decide not to buy. Therefore, there exists necessity for us to consider the time/individual fairness under this online pricing problem scheme.

5.6.5 Potential Generalization of Techniques

Here we discuss a little bit more on the probable extension of the techniques we developed in our algorithm design and analysis.

From Two Groups to Multi Groups Our problem setting assumes that there are two groups of customers in total. We choose to study a two-group setting to simplify the presentation. In practice, however, it is very common that customers are coming from a number of groups with different valuations even on the same product. In fact, we believe it straightforward to extend our methodologies and results to multi-group settings, as

long as we determine a metric of multi-group unfairness. For instance, if we choose to define the multi-group unfairness as the summation of pairwise unfairness of all pairs of groups, we may adjust our algorithm by lengthening each epoch by $G/2$ times and keeping everything the same as in this paper. In this way, the upper regret bound would be $\tilde{O}(G^2\sqrt{T}d^{2/3})$, which is $O(G^2)$ times as we have shown in this paper. Therefore, it is still optimal w.r.t. T up to iterative-log factors.

A Good-and-exploratory Policy Set Our algorithm FPA maintains and updates a "good-and-exploratory" policy A_k in each epoch. Each policy in this set performs close to the optimal policy in both regret and unfairness reductions. A similar idea in reinforcement learning related research exists in Qiao et al. [2022] where they select policies that visit each (horizon, state, action) tuple most sufficiently while ensuring that the policy is low in regret. In fact, if we imagine an "exploratory" policy as the one that would elevate the most "information" (i.e., that would reduce the most uncertainty), then the "good-and-exploratory" policy-selection process is equivalent to an "Upper Confidence Bound" method Lai and Robbins [1985] where we always pull the arm with the highest upper confidence bound in a multi-armed bandit. The only difference is that: for traditional exploration-and-exploitation balancing algorithm, we only need to improve our estimation on the parameters of these optimal or near-optimal policies. However, in our problem setting, we have to guarantee a uniform error bound, i.e., we have to improve our estimation on all parameters instead of only those optimal-related ones. This is because that we have to improve the estimation on constraints as well as on the revenue function. In our algorithm design, we handle this problem by keeping eliminating a feasible policy set, which in turn releases the algorithm from estimating those unnecessary parameters. In a nutshell, our methods can be applied broadly in online-decision-making

problems.

Unfairness Lower Bound Proof on Optimal Algorithms The main idea of our proof of the $\Omega(\sqrt{T})$ unfairness lower bound on any algorithm with $O(\sqrt{T})$ optimal regret is to construct a trade-off on unfairness and regret between two adjacent problem settings. We first bound the "bad policies" away from each problem setting, to avoid those policies that are super fair in both setting but performs poor in both setting as well. Then we show that policies with small-enough regret and unfairness on one setting should suffer a large regret on the other. Finally we end the proof by showing that we will definitely make $\tilde{\Theta}(T)$ times of mistakes in expectation, according to information theory. We believe that this scheme can be used in proving a variety of trading-off lower bounds.

5.6.6 Social Impacts

In this work, we develop methods to prompt the procedural and substantive fairness of customers from all groups. We believe that our techniques and results would enhance the unity of people with different gender, race, age, cultural backgrounds, and so on. However, it is definitely correct that we have to *treat differently* to different group of people. In order to ensure the fairness from customers' perspective, the seller is required to behave unfairly. Of course we could partly get rid of this issue by leaving the generating process of a random price to the nature, i.e., we let each customer draw a coupon from a box randomly. However, this only means that the seller's pricing process is fair but not leads to a fair result, as customers' coupon varies a lot from person to person. This turn out to be the exact issue named as "pricing and price fairness" proposed in Chapuis [2012] regarding the fairness of a seller's behavior. Maybe in the future we could develop an algorithm that is not only profitable but also ensures the fairness from both the seller and

the customers' perspective, which could be a truly "doubly fair" dynamic pricing.

5.7 Conclusion

In this chapter, we studied the online pricing problem with fairness constraints. We introduced two fairness notions, a *procedural fairness* and a *substantive fairness* indicating the equality of proposed and accepted prices between two different groups respectively. In order to fulfill these two constraints simultaneously, we adopted *random* pricing policies and established the objective function and rewards in expectation. To solve this problem with unknown demands, we designed a policy-elimination-based algorithm, FPA, that achieves an $\tilde{O}(\sqrt{T})$ regret within an $\tilde{O}(\sqrt{T})$ unfairness. We showed that our algorithm is optimal up to $\log\log T$ factors by proving an $\Omega(\sqrt{T})$ regret lower bound and an $\Omega(\sqrt{T})$ unfairness lower bound for any optimal algorithm with an $O(\sqrt{T})$ regret.

5.8 Proofs

5.8.1 Proof of Theorem 5.5.1

Proof. First of all, we specify the parameters initialized in Algorithm 7: Let $C_q = 3\max\{\frac{1}{q}, \frac{1}{1-q}\}$ and $c_t = \max\{3, \sqrt{\frac{3}{\hat{F}_{\min}}}\}$. For $k = 1, 2, \ldots$, let $\tau_k = \frac{28C_q}{3} \cdot d\sqrt{T}\log(\frac{16d\log T}{\epsilon}) \cdot 2^k$, $\delta_{k,r} = 4c_t \log\frac{16d\log T}{\epsilon} d^{\frac{3}{2}}\sqrt{\frac{C_q}{\tau_k}}$, $\delta_{k,s} = \frac{32c_t}{(\hat{F}_{\min})^2}\log\frac{16d\log T}{\epsilon} d^{\frac{3}{2}}\sqrt{\frac{C_q}{\tau_k}}$. Now we prove that $\hat{F}_{\min} \leq F_{\min}$ with high probability. Recall than $C_q = 3\max\{\frac{1}{q}, \frac{1}{1-q}\}$. Recall that $\tau_0 = 2\log T \log\frac{16}{\epsilon}$.

According to Hoeffding's Inequality, we have:

$$\begin{aligned}\Pr[|\bar{F}_1(d) - F_1(d)| \geq \frac{F_1(d)}{2}] \leq & 2\exp\{-2(\frac{F_1(d)}{2})^2 \cdot \frac{1}{C_q}\tau_0\} \\ \Leftrightarrow \qquad \Pr[\frac{F_1(d)}{2} \leq \frac{3F_1(d)}{2}] \geq & 1 - 2\exp\{-(F_1(d))^2 \frac{1}{C_q}\log T \log\frac{16}{\epsilon}\} \\ \geq & 1 - \frac{\epsilon}{8}.\end{aligned} \tag{5.10}$$

Here the last inequality comes from Assumption 5.3.5 that $F_1(d) \geq F_{\min} > 0$ and therefore we have $(F_{\min})^2 \frac{1}{C_q}\log T \geq 1$ with large T. Therefore, we have $\frac{F_1(d)}{2}\bar{F}_1(d) \leq \frac{3F_1(d)}{2}$ with probability at least $1 - \frac{\epsilon}{8}$. Similarly, we have $\frac{F_2(d)}{2}\bar{F}_2(d) \leq \frac{3F_2(d)}{2}$ with probability at least $1 - \frac{\epsilon}{8}$. Therefore, with $\Pr \geq 1 - \frac{\epsilon}{4}$, we have $\hat{F}_{\min} = \frac{1}{2}\min\{\bar{F}_1(d), \bar{F}_2(d)\} \leq \min\{\frac{3F_1(d)}{4}, \frac{3F_2(d)}{4}\} = \frac{3}{4}F_{\min} < F_{\min}$.

We define some notations that are helpful to our proof. In epoch k, recall that we have:

$$\tilde{\pi}_{k,i,e} = \operatorname*{argmax}_{\pi \in \Pi_k} \pi^e(i). \tag{5.11}$$

For simplicity, for every $i \in I_k^e$, denote $\rho_{k,e}(i) := \tilde{\pi}^e_{k,i,e}(i)$ that is the largest probability of choosing price v_i in G_e among all policies in Π_k. For those $i \notin I_k^e$, we find out the largest k' such that $i \in I^e_{k'}$ and let $\rho_{k,e}(i) := \tilde{\pi}^e_{k',i,e}(i)$. According to the fact that $\Pi = \Pi_1 \supset \Pi_2 \supset \ldots \supset \Pi_k \supset \Pi_{k+1} \supset \ldots$, we have

$$\pi^e(i) \leq \rho_{k,e}(i), \forall \pi \in \Pi_k; e = 1, 2; i \in [d]. \tag{5.12}$$

Next, we prove the following lemmas together by induction over epoch index $k = 1, 2, \ldots$. We firstly state that

Lemma 5.8.1. *Recall the optimal policy π_* defined in Eq. (5.5). Before Epoch k, we have $\pi_* \in \Pi_k$ with high probability (the failure probability will be totally bounded at the end of this proof).*

which is natural at $k = 1$ as $\Pi_1 = \Pi$. Now, suppose Lemma 5.8.1 holds for $\leq k$, then we have:

Lemma 5.8.2 (Number of Choosing v_i in G_e)**.** *For $M_{k,e}(i)$ and $N_{k,e}(i)$ defined in Algorithm 7, for any $e = 1, 2; i \in I_k^e$, with $\Pr \geq 1 - \frac{\epsilon}{2 \log T}$ we have:*

$$\begin{aligned} \frac{\rho_{k,e}(i) \cdot \tau_k}{4d \cdot C_q} \leq & M_{k,e}(i), \\ |N_{k,e}(i) - M_{k,e}(i) \cdot F_e(i)| \leq & c_t \cdot \sqrt{F_e(i) \cdot M_{k,e}(i)} \cdot \log \frac{16d \log T}{\epsilon}. \end{aligned} \tag{5.13}$$

Here $c_t = \max\{3, \sqrt{\frac{3}{\hat{F}_{\min}}}\}$.

Proof of Lemma 5.8.2. For any $i \in I_k^e$, there exists a policy $\tilde{\pi}_{k,i,e}$ running in Epoch k for at least $\frac{\tau_k}{|A_k|}$ rounds, and. Therefore, we have $\mathbb{E}[M_{k,e}(i)] \geq \rho_{k,e}(i) \cdot \frac{\tau_k}{|A_k|} \cdot \min\{q, 1-q\} \geq \rho_{k,e}(i) \cdot \frac{\tau_k}{|A_k| \cdot C_q}$. According to Bernstein's Inequality, for any $e = 1, 2; i \in I_k^e$ we have:

$$
\begin{aligned}
&\Pr[|M_{k,e}(i)-\mathbb{E}[M_{k,e}(i)]| \le \frac{\mathbb{E}[M_{k,e}(i)]}{2}] \\
\ge& 1-2\exp\{-\frac{\frac{1}{2}(\frac{\mathbb{E}[M_{k,e}(i)]}{2})^2}{\sum_{t=1}^{\frac{\tau_k}{|A_k|}}\rho_{k,e}(i)\cdot\frac{1}{C_q}(1-\rho_{k,e}(i)\cdot\frac{1}{C_q})+\frac{1}{3}\cdot 1\cdot\frac{\mathbb{E}[M_{k,e}(i)]}{2}}\} \\
\ge& 1-2\exp\{-\frac{\frac{1}{8}(\mathbb{E}[M_{k,e}(i)])^2}{\rho_{k,e}(i)\cdot\frac{\tau_k}{|A_k|C_q}+\frac{1}{6}\cdot\mathbb{E}[M_{k,e}(i)]}\} \\
\ge& 1-2\exp\{-\frac{\frac{1}{8}(\mathbb{E}[M_{k,e}(i)])^2}{\mathbb{E}[M_{k,e}(i)]+\frac{1}{6}\cdot\mathbb{E}[M_{k,e}(i)]}\} \\
=& 1-2\exp\{-\frac{1}{8\cdot\frac{7}{6}}\mathbb{E}[M_{k,e}(i)]\} \\
\ge& 1-2\exp\{-\frac{3}{28}\rho_{k,e}(i)\cdot\frac{\tau_k}{|A_k|\cdot C_q}\} \\
=& 1-2\exp\{-\frac{3}{28}\rho_{k,e}(i)\cdot\frac{\frac{28}{3}\cdot d\sqrt{T}\log(\frac{16d\log T}{\epsilon})\cdot 2^k}{2d\cdot C_q}\} \\
=& 1-2\exp\{-\rho_{k,e}(i)\sqrt{T}\cdot\log(\frac{16d\log T}{\epsilon})\cdot\frac{d\cdot 2^k}{2d}\} \\
\ge& 1-2\exp\{-\log(\frac{16d\log T}{\epsilon})\} \\
=& 1-2\cdot\frac{\epsilon}{16d\log T} \\
=& 1-\frac{\epsilon}{8d\log T}.
\end{aligned}
\tag{5.14}
$$

Here the second line is because that

$$\sum_{t=1}^{\tau_k}\mathbb{E}[(\mathbf{1}(\text{choosing } v_i \text{ at time } t))^2] \le \sum_{t=1}^{\frac{\tau_k}{|A_k|}}\mathbb{E}[(\mathbf{1}(\text{ running } \tilde{\pi}_{k,i,e} \text{ and choosing } v_i \text{ at time } t))]$$

, the third line is for $1-\rho_{k,e}(i)\cdot\frac{1}{C_q}\le 1$, the fourth and sixth line are from $\mathbb{E}[M_{k,e}(i)]\ge\frac{\rho_{k,e}(i)\cdot\tau_k}{|A_k|\cdot C_q}$, the seventh line is by plugging in $\tau_k=\frac{28C_q}{3}\cdot d\sqrt{T}\log(\frac{16d\log T}{\epsilon})\cdot 2^k$, the eighth line is equivalent transformation and the ninth line is for $\rho_{k,e}(i)\ge\frac{1}{\sqrt{T}}$ according to Line 9 of Algorithm 7. As a result, with probability at least $1-\frac{\epsilon}{8d\log T}$, we have

$$M_{k,e}(i)\ge\frac{\mathbb{E}[M_{k,e}(i)]}{2}\ge\frac{\rho_{k,e}(i)\cdot\tau_k}{|A_k|\cdot C_q}\ge\frac{\rho_{k,e}(i)\cdot\tau_k}{4dC_q}. \tag{5.15}$$

Now, we analyze $N_{k,e}(i)$ for $i\in I_k^e$. Again, from Line 15 of Algorithm 7 we know that $N_{k,e}(i)=\sum_{t=1}^{M_{k,e}(i)}\mathbb{1}(v_t \text{ is accepted in } G_e)$. Therefore, we apply Bernstein's Inequality and

get:

$$\begin{aligned}
&\Pr[|N_{k,e}(i) - M_{k,e}(i)\cdot F_e(i)| \geq c_t \cdot \sqrt{M_{k,e}(i)\cdot F_e(i)}\log\frac{16d\log T}{\epsilon}]\\
\leq& 2\exp\{-\frac{\frac{1}{2}c_t^2\cdot M_{k,e}(i)F_e(i)(\log\frac{16d\log T}{\epsilon})^2}{M_{k,e}(i)F_e(i)\log\frac{16d\log T}{\epsilon}(1-F_e(i)) + \frac{1}{3}\cdot(c_t\cdot\sqrt{M_{k,e}(i)\cdot F_e(i)}\log\frac{16d\log T}{\epsilon})}\}\\
\leq& 2\exp\{-\frac{\frac{1}{2}c_t^2\log\frac{16d\log T}{\epsilon}}{1+\frac{c_t}{3}}\}\\
\leq& \frac{\epsilon}{8\log T}.
\end{aligned}$$

Here the last line is by $c_t = \max\{3, \sqrt{\frac{3}{\hat{F}_{\min}}}\} \geq 3$ and therefore $\frac{\frac{1}{2}c_t^2}{1+\frac{c_t}{3}} \geq 1$. As a result, for $e = 1, 2; i \in I_k^e$, with $\Pr \geq 1 - \frac{\epsilon}{8\log T}$ we have

$$|N_{k,e}(i) - M_{k,e}(i)\cdot F_e(i)| \leq c_t\cdot\sqrt{M_{k,e}(i)F_e(i)}\log\frac{16d\log T}{\epsilon}.$$

That is to say,

$$\begin{aligned}
|\bar{F}_{k,e}(i) - F_e(i)| &= |\max\{\frac{N_{k,e}(i)}{M_{k,e}(i)}, \hat{F}_{\min}\} - F_e(i)|\\
&\leq |\frac{N_{k,e}(i)}{M_{k,e}(i)}|\\
&\leq c_t\cdot\sqrt{\frac{F_e(i)}{M_{k,e}(i)}}\log\frac{16d\log T}{\epsilon}\\
&\leq c_t\log\frac{16d\log T}{\epsilon}\sqrt{F_e(i)}\cdot\sqrt{\frac{4dC_q}{\rho_{k,e}(i)\tau_k}}.
\end{aligned}$$

Here the first line is by definition of $\bar{F}_{k,e}$, the second line is because $\hat{F}_{\min} \leq F_{\min} \leq F_e(i)$, the third line is by the inequality above and the last line is by Eq. (5.15). Therefore, with $\Pr \geq 1 - \frac{\epsilon}{2\log T}$, Eq. (5.13) holds for $e = 1, 2$ and for $\forall i \in I_k^e$. ∎

Given Lemma 5.8.2, we have the following corollary directly:

Corollary 5.8.3 (Estimation Error of $\bar{F}_{k,e}(i)$)**.** *Assume that Lemma 5.8.2 holds. For*

$\bar{F}_{k,e}(i) = \max\{\frac{N_{k,e}(i)}{M_{k,e}(i)}, \hat{F}_{\min}\}$ *defined in Algorithm 7, for any* $e = 1, 2; i \in I_k^e$*, we have:*

$$|\bar{F}_{k,e}(i) - F_e(i)| \leq c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dF_e(i)C_q}{\rho_{k,e}(i)\tau_k}}. \tag{5.16}$$

For simplicity, denote$R(\pi) := R(\pi, F_1, F_2)$, $S(\pi) := S(\pi, F_1, F_2)$, $\hat{R}_k(\pi) := R(\pi, \bar{F}_{k,1}, \bar{F}_{k,2})$ and $\hat{S}_k(\pi) := S(\pi, \bar{F}_{k,1}, \bar{F}_{k,2})$. Based on Corollary 5.8.3, we can bound the estimation error of $\hat{R}_k(\pi)$ and $\hat{S}_k(\pi)$ by the following lemma:

Lemma 5.8.4 (Estimation Error of R and S Functions)**.** *Given Lemma 5.8.2, we have:*

$$\begin{aligned} |R(\pi) - \hat{R}_k(\pi)| \leq & \frac{\delta_{k,r}}{2}, \\ |S(\pi) - \hat{S}_k(\pi)| \leq & \frac{\delta_{k,s}}{2}. \end{aligned} \tag{5.17}$$

Here $\delta_{k,r} = 4c_t \log \frac{16d \log T}{\epsilon} d^{\frac{3}{2}} \sqrt{\frac{C_q}{\tau_k}}$ and $\delta_{k,s} = \frac{32c_t}{\hat{F}_{\min}^2} \log \frac{16d \log T}{\epsilon} d^{\frac{3}{2}} \sqrt{\frac{1}{\tau_k}}$ as is defined in Theorem 5.5.1.

Proof of Lemma 5.8.4. First of all, we show that for any $e = 1, 2; i = 1, 2, \ldots, d$ and for any $\pi \in \Pi_k$,

$$|\bar{F}_{k,e}(i) - F_e(i)| \cdot \pi^e(i) \leq c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q}{\tau_k}}. \tag{5.18}$$

In fact, when $i \in I_k^e$, according to Lemma 5.8.2 we have

$$\begin{aligned} |\bar{F}_{k,e}(i) - F_e(i)| \cdot \pi^e(i) \leq & |\bar{F}_{k,e}(i) - F_e(i)| \cdot \rho_{k,e}(i) \\ \leq & c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dF_e(i)C_q}{\rho_{k,e}(i)\tau_k}} \cdot \rho_{k,e}(i) \\ \leq & c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q \cdot (\rho_{k,e}(i))}{\tau_k}} \\ \leq & c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q}{\tau_k}}. \end{aligned} \tag{5.19}$$

When $i \notin I_k^e$, we know that $\rho_{k,e}(i) \leq \frac{1}{\sqrt{T}}$ and thus $\pi^e(i) \leq \frac{1}{\sqrt{T}}, \forall \pi \in \Pi_k$ according to Eq. (5.12). Also, since $\pi_* \in \Pi_k$ by induction, we know that $\pi_*^e \leq \rho_{k,e}(i) \leq \frac{1}{\sqrt{T}}$. Therefore, we have

$$\begin{aligned}
|\bar{F}_{k,e}(i) - F_e(i)| \cdot \pi^e(i) \leq & |\bar{F}_{k,e}(i) - F_e(i)| \cdot \rho_{k,e}(i) \\
\leq & |\bar{F}_{k,e}(i)\pi^e(i) - F_e(i)| \cdot \frac{1}{\sqrt{T}} \\
\leq & 1 \cdot \frac{1}{\sqrt{T}} \\
\leq & c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q}{\tau_k}}.
\end{aligned} \tag{5.20}$$

Here we assume that $\log \frac{16d \log T}{\epsilon} > 1$ without losing of generality (i.e., T is sufficiently large and ϵ can be arbitrarily close to zero), and the last inequality comes from $c_t \geq 3 > 1$ and $4d \geq 1$ and $\tau_k \leq T$. Combining Eq. (5.19) and Eq. (5.20), we know that Eq. (5.18) holds for all $e = 1, 2; i \in [d]$. Remember that $R(\pi) = q \cdot \sum_{i=1}^d F_1(i)\pi^1(i) + (1-q) \cdot \sum_{j=1}^d F_2(j)\pi^2(j)$ and that $\hat{R}_k(\pi) = q \cdot \sum_{i=1}^d \bar{F}_{k,1}(i)\pi^1(i) + (1-q) \cdot \sum_{j=1}^d \hat{F}_{k,2}(j)\pi^2(j)$. Therefore, we may bound the error between $\hat{R}_k(\pi)$ and $R(\pi)$. For $\forall \pi \in \Pi_k$, we have

$$\begin{aligned}
|\hat{R}_k(\pi) - R(\pi)| &\leq q \cdot \sum_{i=1}^d |\bar{F}_{k,1}(i) - F_1(i)| \cdot \pi^1(i) + (1-q) \cdot \sum_{j=1}^d |\bar{F}_{k,2}(j) - F_2(j)| \cdot \pi^2(i) \\
&\leq q \cdot \sum_{i=1}^d c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q}{\tau_k}} + (1-q) \cdot \sum_{j=1}^d c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q}{\tau_k}} \\
&= c_t \cdot \log \frac{16d \log T}{\epsilon} \sqrt{\frac{4dC_q}{\tau_k}} \cdot d \\
&= \frac{\delta_{k,r}}{2}.
\end{aligned} \tag{5.21}$$

Similarly, for the error between $\hat{S}_k(\pi)$ and $S(\pi)$ as $\pi \in \Pi_k$, we have:

$$
\begin{aligned}
&|\hat{S}_k(\pi) - S(\pi)| \\
=&||\frac{\mathbf{v}^\top \hat{F}_{k,1}\pi^1}{\mathbb{1}^\top \hat{F}_{k,1}\pi^1} - \frac{\mathbf{v}^\top \hat{F}_{k,2}\pi^2}{\mathbb{1}^\top \hat{F}_{k,2}\pi^2}| - |\frac{\mathbf{v}^\top F_1\pi^1}{\mathbb{1}^\top F_1\pi^1} - \frac{\mathbf{v}^\top F_2\pi^2}{\mathbb{1}^\top F_2\pi^2}|| \\
\leq&|\frac{\mathbf{v}^\top \hat{F}_{k,1}\pi^1}{\mathbb{1}^\top \hat{F}_{k,1}\pi^1} - \frac{\mathbf{v}^\top \hat{F}_{k,2}\pi^2}{\mathbb{1}^\top \hat{F}_{k,2}\pi^2} - (\frac{\mathbf{v}^\top F_1\pi^1}{\mathbb{1}^\top F_1\pi^1} - \frac{\mathbf{v}^\top F_2\pi^2}{\mathbb{1}^\top F_2\pi^2})| \\
\leq&|\frac{\mathbf{v}^\top \hat{F}_{k,1}\pi^1}{\mathbb{1}^\top \hat{F}_{k,1}\pi^1} - \frac{\mathbf{v}^\top F_1\pi^1}{\mathbb{1}^\top F_1\pi^1}| + |\frac{\mathbf{v}^\top \hat{F}_{k,2}\pi^2}{\mathbb{1}^\top \hat{F}_{k,2}\pi^2} - \frac{\mathbf{v}^\top F_2\pi^2}{\mathbb{1}^\top F_2\pi^2}| \\
=&\sum_{e=1}^{2}|\frac{\mathbf{v}^\top \hat{F}_{k,e}\pi^e}{\mathbb{1}^\top \hat{F}_{k,e}\pi^e} - \frac{\mathbf{v}^\top F_e\pi^e}{\mathbb{1}^\top F_e\pi^e}| \\
=&\sum_{e=1}^{2}|\frac{(\mathbf{v}^\top \hat{F}_{k,e}\pi^e)(\mathbb{1}^\top F_e\pi^e) - (\mathbf{v}^\top F_e\pi^e)(\mathbb{1}^\top \hat{F}_{k,e}\pi^e)}{(\mathbb{1}^\top \hat{F}_{k,e}\pi^e)(\mathbb{1}^\top F_e\pi^e)}| \\
=&\sum_{e=1}^{2}\frac{|(\pi^e)^\top(\hat{F}_{k,e} - F_e)\mathbf{v}\cdot(\mathbb{1}^\top F_e\pi^e) + (\mathbf{v}^\top F_e\pi^e)\mathbb{1}^\top(F_e - \hat{F}_{k,e})\pi^e|}{|(\mathbb{1}^\top \hat{F}_{k,e}\pi^e)|\cdot|(\mathbb{1}^\top F_e\pi^e)|} \\
\leq&\sum_{e=1}^{2}\frac{(\mathbb{1}^\top F_e\pi^e)\sum_{i=1}^{d}\pi^e(i)(\bar{F}_{k,e}(i) - F_e(i))v_i + (\mathbf{v}^\top F_e\pi^e)\sum_{j=1}^{d}1\cdot(F_e(j) - \hat{F}_{k,e}(j))\pi^e(j)}{|\hat{F}_{\min}|\cdot|\hat{F}_{\min}|} \\
\leq&\sum_{e=1}^{2}\frac{1\cdot\sum_{i=1}^{d}\pi^e(i)|\bar{F}_{k,e}(i) - F_e(i)|\cdot 1 + 1\cdot\sum_{j=1}^{d}1\cdot|F_e(j) - \hat{F}_{k,e}(j)|\pi^e(j)}{(\hat{F}_{\min})^2} \\
\leq&\frac{1}{(\hat{F}_{\min})^2}\sum_{e=1}^{2}(\sum_{i=1}^{d}c_t\cdot\log\frac{16d\log T}{\epsilon}\sqrt{\frac{4dC_q}{\tau_k}} + \sum_{j=1}^{d}c_t\cdot\log\frac{16d\log T}{\epsilon}\sqrt{\frac{4dC_q}{\tau_k}}) \\
=&\frac{1}{(\hat{F}_{\min})^2}\cdot 2d\cdot c_t\cdot\log\frac{16d\log T}{\epsilon}\sqrt{\frac{4dC_q}{\tau_k}} \\
\leq&\frac{\delta_{k,s}}{2}.
\end{aligned}
\tag{5.22}
$$

■

Since we have $\hat{S}_k(\pi) \leq \delta_{k,s}, \forall\pi \in \mathit{\Pi}_{k+1}$ by definition in Algorithm 7, we know that for any

policy $\pi \in \Pi_{k+1}$,

$$\begin{aligned} S(\pi) =& \hat{S}_k(\pi) + (S(\pi) - \hat{S}_k(\pi)) \\ \leq& \hat{S}_k(\pi) + |S(\pi) - \hat{S}_k(\pi)| \\ \leq& \delta_{k,s} + \frac{\delta_{k,s}}{2} \\ \leq& 2\delta_{k,s}. \end{aligned} \tag{5.23}$$

Therefore, any policy remaining in Π_{k+1} suffers at most $2\delta_{k,s}$ unfairness. Now let us bound the regret of any policy in Π_{k+1}. Here we firstly propose a lemma.

Lemma 5.8.5 (Small Relaxation Gain)**.** *Recall that π_* is the solution to Eq.* (5.5)*. Define a $\pi_{\delta,*}$ as follows.*

$$\begin{aligned} \pi_{\delta,*} =& \operatorname*{argmax}_{\pi \in \Pi} R(\pi) \\ s.t. \qquad & S(\pi) \leq \delta. \end{aligned} \tag{5.24}$$

Then there exists a constant $L \in \mathbb{R}_+$ such that $R(\pi_{\delta,*}) - R(\pi_*) \leq \frac{L}{2} \cdot \delta$.

We leave the proof of Lemma 5.8.5 to the end of this section. Given Lemma 5.8.5 and the previous Lemma 5.8.4, we have:

$$\begin{aligned} & \hat{R}(\hat{\pi}_{k,*}) - \hat{R}(\pi_*) \\ =& \hat{R}(\hat{\pi}_{k,*}) - R(\hat{\pi}_{k,*}) + R(\hat{\pi}_{k,*}) - R(\pi_{2\delta_{k,s},*}) + R(\pi_{2\delta_{k,s}},*) - R(\pi_*) + R(\pi_*) - \hat{R}(\pi_*) \\ \leq& |\hat{R}(\hat{\pi}_{k,*}) - R(\hat{\pi}_{k,*})| + (R(\hat{\pi}_{k,*}) - R(\pi_{2\delta_{k,s},*})) + (R(\pi_{2\delta_{k,s}},*) - R(\pi_*)) + |R(\pi_*) - \hat{R}(\pi_*)| \\ \leq& \frac{\delta_{k,r}}{2} + 0 + \frac{L}{2} \cdot 2\delta_{k,s} + \frac{\delta_{k,r}}{2} \\ =& \delta_{k,r} + L \cdot \delta_{k,s} \end{aligned} \tag{5.25}$$

By definition of Π_{k+1} at Eq. (5.8), we know that $\pi_* \in \Pi_{k+1}$, which holds Lemma 5.8.1 at $k+1$ and therefore completes the induction. As a result, all Lemma 5.8.1, Lemma 5.8.2,

Lemma 5.8.4 and Lemma 5.8.5 holds for all $k = 1, 2, \ldots$. As a result, we may calculate the total regret and substantive unfairness as follows.

For the regret, we may divide the whole time horizon T into three stages:

1. Stage 0: Before epochs where we propose v_d for $\tau_0 = 2 \log T \log \frac{16}{\epsilon}$ rounds in either G_1 or G_2. The regret for this stage is $O(\log T \log \frac{1}{\epsilon})$.

2. Stage 1: Epoch 1 where we try every price for $2 \cdot \frac{\tau_1}{2d}$ rounds in either G_1 or G_2. The regret for this stage is $O(\tau_1) = O(d\sqrt{T} \log \frac{\log T}{\epsilon})$.

3. Stage 2: Epoch $k = 2, 3, \ldots$. In each epoch k, every policy π we run satisfies $\pi \in \mathit{\Pi}_k$. Therefore, for any π running in Epoch $k = 2, 3, \ldots$, we have

$$\begin{aligned} R(\pi_*) - R(\pi) =& (R(\pi_*) - \hat{R}_{k-1}(\pi_*)) + (\hat{R}_{k-1}(\pi_*) - \hat{R}_{k-1}(\pi)) + (\hat{R}_{k-1}(\pi) - R(\pi)) \\ \leq& \frac{\delta_{k-1,r}}{2} + (\delta_{k-1,r} + L \cdot \delta_{k-1,s}) + \frac{\delta_{k-1,r}}{2} \\ =& 2\delta_{k-1,r} + L \cdot \delta_{k-1,s}. \end{aligned} \tag{5.26}$$

   The second line is by definition of $\mathit{\Pi}_k$ for $k \geq 2$ and by Lemma 5.8.4. Suppose there are K epochs in total, and then we know that:

$$T \geq \sum_{k=1}^{K} \tau_k = \frac{28 C_q}{3} \cdot d\sqrt{T} \cdot \log \frac{16 d \log T}{\epsilon} \cdot \sum_{k=1}^{K} 2^k.$$

   Solve the equaltion above and we get $K = O(\log \frac{\sqrt{T}}{d \log \frac{d \log T}{\epsilon}})$ and $K \leq \frac{1}{2} \log T$. Therefore, the total regret of Stage 2 is

$$Reg = O(\sum_{k=2}^{K} \tau_k \cdot (2\delta_{k-1,r} + L \cdot \delta_{k-1,s})) = O(\sqrt{T} \cdot d^{\frac{3}{2}} \log \frac{d \log T}{\epsilon}) \tag{5.27}$$

Add the regret of all three stages above, we get that the total regret is $O(\sqrt{T} \cdot d^{\frac{3}{2}} \log \frac{d \log T}{\epsilon})$.

For the unfairness, we derive it similarly in three stages:

1. Stage 0: Before epochs where we propose v_d for $\tau_0 = 2\log T \log \frac{16}{\epsilon}$ rounds in either G_1 or G_2. The unfairness for this stage is 0 as we always propose the same price to both groups.

2. Stage 1: Epoch 1 where we try every price for $2 \cdot \frac{\tau_1}{2d}$ rounds in either G_1 or G_2. The regret for this stage is 0 as well.

3. Stage 2: Epoch $k = 2, 3, \ldots$. In each epoch k, every policy π we run satisfies $\pi \in \Pi_k$. Therefore, for any π running in Epoch $k = 2, 3, \ldots$, we have

$$\begin{aligned} S(\pi) =& \hat{S}_{k-1}(\pi) + (S(\pi) - \hat{S}_{k-1}(\pi)) \\ \leq& \delta_{k-1,s} + \frac{\delta_{k-1,s}}{2} \\ \leq& \frac{3\delta_{k-1,s}}{2}. \end{aligned} \tag{5.28}$$

   Here the last line is by definition of Π_k for $k \geq 2$ and by Lemma 5.8.4. Therefore, the total unfairness of Stage 2 is

$$Unf \leq \sum_{k=2}^{K} \tau_k \cdot \frac{3\delta_{k-1,s}}{2} = O(\sqrt{T} \cdot d^{\frac{3}{2}} \log \frac{d \log T}{\epsilon}). \tag{5.29}$$

Therefore, the total substantive unfairness of all three stages is $O(\sqrt{T} \cdot d^{\frac{3}{2}} \log \frac{d \log T}{\epsilon})$ as well.

Finally, we count the probability of failure of all stages. For Stage 0, the failure probability is $\Pr_0 \leq \frac{\epsilon}{4}$. For each epoch, the failure probability is $\Pr_k \leq \frac{\epsilon}{2 \log T}$. Since there are $K \leq \frac{\log T}{2}$ epochs, the total failure probability is $\Pr_{failure} \leq \Pr_0 + K \cdot \Pr_k \leq \frac{\epsilon}{4} + K \cdot \frac{\epsilon}{2 \log T} \leq \frac{\epsilon}{2} < \epsilon$. That is to say, Theorem 5.5.1 holds with probability at least $\Pr \geq 1 - \epsilon$. ∎

At the end of this subsection, we prove Lemma 5.8.5 as we promised above.

Proof of Lemma 5.8.5. Denote any policy $\pi \in \mathit{\Pi}$ as $\pi = (\pi^1, \pi^2)$. For the simplicity of notation, we denote the following functions:

(a) Define $R_1(\pi^1) = \mathbf{v}^\top F_1 \pi^1$;

(b) Define $R_2(\pi^2) = \mathbf{v}^\top F_2 \pi^2$;

(c) Define $S_1(\pi^1) = \frac{\mathbf{v}^\top F_1 \pi^1}{\mathbb{1}^\top F_1 \pi^1}$;

(d) Define $S_2(\pi^2) - \frac{\mathbf{v}^\top F_2 \pi^2}{\mathbb{1}^\top F_2 \pi^2}$.

For $\pi_{\delta,*}$ defined in Eq. (5.9), denote $V_s := S_1(\pi^1_{\delta,*})$ and $z = S_2(\pi^2_{\delta,*}) - V_s$. Therefore, we know that $V_s \in [v_1, 1]$ (recalling that $v_1 > 0$) and $z \in [-\delta, \delta]$. According to the optimality of $\pi_{\delta,*}$, we have:

$$\begin{aligned} \pi_{\delta,*} = & \underset{\pi \in \mathit{\Pi}, V_s \in [v_1,1], z \in [-\delta,\delta]}{\operatorname{argmax}} qR_1(\pi^1) + (1-q)R_2(\pi^2) \\ s.t. \quad & S_1(\pi^1) = V_s \\ & S_2(\pi^2) = V_s + z \end{aligned} \tag{5.30}$$

. Consider the constraint $S_2(\pi^2) - V_s \in [-\delta, \delta]$, we can derive the following relaxation:

$$\begin{aligned} & S_2(\pi^2) - V_s \in [-\delta, \delta] \\ \Leftrightarrow \quad & -\delta \le \frac{\mathbf{v}^\top F_2 \pi^2}{\mathbb{1}^\top F_2 \pi^2} - V_s \le \delta \\ \Rightarrow \quad & -\delta(\mathbb{1}^\top F_2 \pi^2) \le \mathbf{v}^\top F_2 \pi^2 - V_s \cdot \mathbb{1}^\top F_2 \pi^2 \le \delta(\mathbb{1}^\top F_2 \pi^2) \\ \Rightarrow \quad & -\delta \le \mathbf{v}^\top F_2 \pi^2 - V_s \cdot \mathbb{1}^\top F_2 \pi^2 \le \delta. \end{aligned} \tag{5.31}$$

This is because $\mathbb{1}^\top F_2 \pi^2 \in [F_{\min}, 1] \subset (0, 1]$. Therefore, we may define $\theta_\delta = (\theta^1_\delta, \theta^2_\delta) \in \mathit{\Pi}$

such that

$$\begin{aligned}\theta_\delta := \underset{\theta\in\Pi, r, w\in[v_1\cdot F_{\min},1]}{\operatorname{argmax}} \quad & qR_1(\theta^1) + (1-q)R_2(\theta^2) \\ s.t. \qquad \mathbf{v}^\top F_1\theta^1 =& w \\ \cdot\mathbb{1}^\top F_1\theta^1 =& \frac{w}{V_s} \\ \mathbf{v}^\top F_2\theta^2 =& r \\ -\delta \le v_1\cdot\mathbb{1}^\top F_2\theta^2 - \frac{r\cdot v_1}{V_s} \le& \delta,\end{aligned} \tag{5.32}$$

for any $\theta \geq 0$. Here we make use of the fact that $V_s \in [v_1, 1]$. Notice that $[v_1 \cdot F_{\min}, 1]$ contains all possible r's due to the fact that $F_e(i) > F_{\min}$ and $v_i \geq v_1$ for any $i \in [d]$ and $e \in \{1, 2\}$, then we have $R_2(\theta^2_\delta) \geq R_2(\pi^2_{\delta,*})$ as a relaxation of conditions, which means that $R(\theta_\delta) \geq R(\pi_{\delta,*})$. Consider another policy π_{start}:

$$\begin{aligned}\pi_{start} := \underset{\pi\in\Pi}{\operatorname{argmax}}\, & qR_1(\pi^1) + (1-q)R_2(\pi^2) \\ s.t. \qquad S_1(\pi^1) =& V_s \\ S_2(\pi^2) =& V_s.\end{aligned} \tag{5.33}$$

Therefore, we know that when $\delta = 0$, we have $\theta_0 = \pi_{start}$ exactly. Also, since π_* can also be defined as follows:

$$\begin{aligned}\pi_* = \underset{\pi\in\Pi, v_s\in[v_1,1]}{\operatorname{argmax}} \quad & qR_1(\pi^1) + (1-q)R_2(\pi^2) \\ s.t. \qquad S_1(\pi^1) =& v_s \\ S_2(\pi^2) =& v_s.\end{aligned} \tag{5.34}$$

According to the optimality of π_* over all $v_s \in [v_1, 1]$ while π_{start} is restricted on a specific V_s, we have $R(\pi_*) \geq R(\pi_{start}) = R(\theta_0)$. Recall that we also have $R(\theta_\delta) \geq R(\pi_{\delta,*})$. Therefore, as long as we show that there exists a constant L such that $R(\theta_\delta) - R(\theta_0) \leq \frac{L}{2}\cdot\delta$, then it is sufficient to show that $R(\pi_{\delta,*}) - R(\pi_*) \leq \frac{L}{2}\cdot\delta$.

Denote

$$\tilde{\theta}_\delta = [(\theta_\delta^1)^\top, w, \frac{w \cdot V_1}{V_s}, (\theta_\delta^2)^\top, r, \frac{r \cdot v_1}{V_s}]^\top \in \mathbb{R}^{2d+4}. \tag{5.35}$$

Of course $\|\tilde{\theta}_\delta\|_1 \leq 1 + w + \frac{w}{V_s} + 1 + r + \frac{r}{V_s} \leq 4 + 2\frac{2}{v_1}$. Denote the domain of $\tilde{\theta}\delta$ as $\mathcal{D}(\delta)$. Therefore, we know that for any $\theta \in \mathcal{D}(\delta)$, we have

$$\begin{aligned}
\theta &\succeq 0 \\
[\mathbb{1}_d^\top, 0, 0, \ldots, 0]\theta &= 1 \\
[0, \ldots, 0, 0, 0, \mathbb{1}_d^\top, 0, 0]\theta &= 1 \\
[0, \ldots, 0, 1, 0, \ldots, 0]\theta &\leq 1 \text{ (for 1 in the } (d+1)^{th} \text{ place)} \\
[0, \ldots, 0, 1, 0, \ldots, 0]\theta &\leq 1 \text{ (for 1 in the } (d+2)^{th} \text{ place)} \\
[0, \ldots, 0, 1, 0]\theta &\leq 1 \\
[0, \ldots, 0, 0, 1]\theta &\leq 1
\end{aligned} \tag{5.36}$$

Denote $\tilde{\mathcal{D}}(\delta)$ as the space of all θ satisfying Eq. (5.36), and we know that $\tilde{\mathcal{D}}(\delta) \supseteq \mathcal{D}(\delta)$ and $\tilde{\mathcal{D}}(\delta)$ is a bounded, close and convex set with only linear boundaries. Also, denote the following fixed parameters:

$$\begin{aligned}
\mathbf{a} &:= [q \cdot (\mathbf{v}^\top F_1), 0, 0, (1-q) \cdot (\mathbf{v}^\top F_2), 0, 0] \in \mathbb{R}^{2d+4} \\
\mathbf{b}_1 &:= [\mathbf{v}^\top F_1, -1, 0, 0, \ldots, 0] \in \mathbb{R}^{2d+4} \\
\mathbf{b}_2 &:= [v_1 \mathbb{1}^\top F_1, 0, -1, 0, 0, \ldots, 0] \in \mathbb{R}^{2d+4} \\
\mathbf{g} &:= [0, \ldots, 0, 0, 0, \mathbf{v}^\top F_2, -1, 0] \in \mathbb{R}^{2d+4} \\
\mathbf{d} &:= [0, \ldots, 0, 0, 0, v_1 \cdot \mathbb{1}^\top F_2, 0, -1] \in \mathbb{R}^{2d+4}.
\end{aligned} \tag{5.37}$$

Again, these parameters are all constants under the same problem setting. Given these parameters, for the definition of θ_δ in Eq. (5.32), we may transform that definition into

the following one equivalently:

$$\begin{aligned} \tilde{\theta}_\delta := \underset{\theta \in \tilde{\mathcal{D}}(\delta)}{\operatorname{argmax}} \ & \mathbf{a}^\top \theta \\ s.t. \quad & \mathbf{b}_1^\top \theta = 0 \\ & \mathbf{b}_2^\top \theta = 0 \\ & \mathbf{g}^\top \theta = 0 \\ & \mathbf{d}^\top \theta \in [-\delta, \delta]. \end{aligned} \tag{5.38}$$

Since $\tilde{\mathcal{D}}(\delta) \supseteq \mathcal{D}(\delta)$, we know that $\mathbf{a}^\top \tilde{\theta}_\delta \geq R(\theta_\delta)$. Denote

$$\tilde{\mathcal{D}}_{abg}(\delta) := \{\theta | \theta \in \tilde{\mathcal{D}}(\delta), \mathbf{b}_1^\top \theta = 0, \mathbf{b}_2^\top \theta = 0, \mathbf{g}^\top \theta = 0\},$$

and we know that $\tilde{\mathcal{D}}_{abg}(\delta)$ is also a bounded close and convex set with only linear boundaries. Therefore, Eq. (5.38) is equivalent to the following definition:

$$\begin{aligned} \tilde{\theta}_\delta := \underset{\theta \in \tilde{\mathcal{D}}_{abg}(\delta)}{\operatorname{argmax}} \ & \mathbf{a}^\top \theta \\ s.t. \quad & \mathbf{d}^\top \theta \in [-\delta, \delta]. \end{aligned} \tag{5.39}$$

Now we present the following lemma.

Lemma 5.8.6 (Bounded Shifting)**.** *Given any space $Q \subset \mathbb{R}^n$ that is bounded, close and convex with only linear boundaries, consider the following subset $Q_0 := \{x \in Q, d^\top x = 0\} \neq \emptyset$. Then there exists a constant C_L such that for any $z \in \mathbb{R}$, $Q_z := \{x \in Q, d^\top x = z\}$ and any $\theta_z \in Q_z$, there always exists a $\theta_0 \in Q_0$ such that $\|\theta_z - \theta_0\|_2 \leq C_L \cdot |z|$.*

Proof of Lemma 5.8.6. Without loss of generality, we assume that $z > 0$. Denote $Q^+ = Q \cap \{x : d^\top x \geq 0\}$. Because Q is bounded, close and convex with only linear boundaries, the number of vertex of Q^+ must be finite. The vertex set of Q^+ can be decomposed as $V = V_0 + V_1$, where V_0 denotes the vertex such that $d^\top x = 0$ while V_1 denotes the vertex

such that $d^\top x > 0$. In addition, $Q_0 = Q^+ \cap \{x : d^\top x = 0\}$ is the cross section while we define $B = \{x : d^\top x = 0\}$.

For each point $x \in Q_0$, we define β_x to be min{The intersection angle between B and $\overrightarrow{xv}$, $v \in V_1$}. Due to the fact that β_x is continuous upon x, $\beta_x > 0$ and the domain Q_0 is bounded and close, there exists a $\beta_{\min} > 0$ such that $\beta_x \geq \beta_{\min}$, $\forall x \in Q_0$. Then we construct a corresponding cone $Cone_x$ for each $x \in Q_0$ such that $Cone_x = \{v : d^\top v \geq 0, \text{and the intersection angle between } B \text{ and } \overrightarrow{xv} \geq \beta_{\min}\}$.

Since Q^+ is bounded, close and convex with only linear boundaries, for any point $\theta_z \in Q_z$, there exits $v_1, v_2, \cdots, v_k$ and $a_1, a_2, \cdots, a_k$ such that $v_i \in V, a_i \geq 0, \forall i \in [k]$ and $\sum_{i=1}^k a_i = 1$ and it holds that $\theta_z = \sum_{i=1}^k a_i v_i$. Then according to our construction of the cones, for each selected vertex v_i, there exists a cone $Cone_{t_i}$ such that $t_i \in Q_0$ and $v_i \in Cone_{t_i}$. We claim that $Cone_{\sum_{i=1}^k a_i t_i} = \{\sum_{i=1}^k a_i f_i : f_i \in Cone_{t_i}\}$. Therefore, it holds that $\theta_z = \sum_{i=1}^k a_i v_i \in Cone_{\sum_{i=1}^k a_i t_i}$. Consider this $\theta_0 = \sum_{i=1}^k a_i t_i$, because Q_0 is convex, we have $\theta_0 \in Q_0$. In addition, $\|\theta_0 - \theta_z\|_2 \leq \frac{|z|}{\|d\|_2 \cdot \sin(\beta_{\min})}$, which means by choosing $C_L = \frac{1}{\|d\|_2 \cdot \sin(\beta_{\min})}$, the proof is complete. ∎

Denote $z_\delta := \tilde{\theta}_\delta$ and we know that $|z_\delta| \leq \delta$. In order to apply Lemma 5.8.6, we have to ensure that $\tilde{\mathcal{D}}_{abg}(0) \neq \emptyset$. In fact, notice that $\tilde{\theta}_0 \in \tilde{\mathcal{D}}_{abg} \cap \{\mathbf{d}^\top = 0\}$. With Lemma 5.8.6, there exists a $\hat{\theta}_0 \in \tilde{\mathcal{D}}_{abg}(0)$ such that $\|\tilde{\theta}_\delta - \hat{\theta}_0\|_2 \leq \frac{L}{2}|z| \leq \frac{L}{2}\delta$. As a result, we have:

$$\begin{aligned}
\mathbf{a}^\top \tilde{\theta}_\delta - \mathbf{a}^\top \hat{\theta}_0 &\leq \|a\|_2 \cdot \|\tilde{\theta}_\delta - \hat{\theta}_0\|_2 \\
&\leq \|a\|_1 \cdot C_L \cdot \delta \\
&\leq (q \cdot \mathbf{v}^\top F_1 \mathbb{1} + (1-q) \cdot \mathbf{v}^\top F_2 \mathbb{1}) \cdot C_L \cdot \delta \\
&:= C_a \cdot C_L \cdot \delta.
\end{aligned} \tag{5.40}$$

By definition of $\tilde{\theta}_\delta$, we know that $\tilde{\theta}_0$ maximizes $\mathbf{a}^\top \theta$ in $\tilde{\mathcal{D}}_{abg}(0)$, which means that

$\mathbf{a}^\top \tilde{\theta}_0 \geq \mathbf{a}^\top \hat{\theta}_0$. As a result, we have:

$$\begin{aligned} R(\theta_\delta) - R(\theta_0) =& \mathbf{a}^\top \tilde{\theta}_\delta - \mathbf{a}^\top \tilde{\theta}_0 \\ \leq& \mathbf{a}^\top \tilde{\theta}_\delta - \mathbf{a}^\top \hat{\theta}_0 \\ \leq& C_a \cdot C_L \cdot \delta. \end{aligned} \tag{5.41}$$

Therefore, we have $R(\pi_{\delta,*}) - R(\pi_*) \leq R(\theta_\delta) - R(\pi_{start}) = R(\theta_\delta) - R(\theta_0) \leq C_a \cdot C_L \cdot \delta$. Let $L := 2 \cdot C_a \cdot C_L$ and this holds the lemma. ∎

5.8.2 Proof of Theorem 5.5.3

As is stated in Section 5.5.2, we may reduce this fair pricing problem to an ordinary online pricing problem with no fairness constraints. Therefore, we only need to prove the following theorem.

Theorem 5.8.7 (Regret Lower Bound)**.** *Consider the online pricing problem with T rounds and d fixed prices in $[0, c]$ for $3 \leq d \leq T^{1/3}$ and some constant $c > 0$. Then any algorithm has to suffer at least $\Omega(\sqrt{dT})$ regret.*

Here we mainly adopt the proof roadmap of Kleinberg and Leighton [2003].

Proof. We let $c = 12$ without losing generality. Let $\epsilon = \sqrt{\frac{d}{T}}, l = 1, a_0 = 4l, a_i = (1 + \frac{\epsilon}{l})^i \cdot a_0, i = 1, 2, \ldots, d$, then we have: $4l = a_0 < a_1 < a_2 < \ldots < a_{d-1} < a_d < 12l$.

Define some distributions on the prices $\{a_i\}_{i=1}^d$:

- $\mathbb{P}_0$, with acceptance rates of each price: $P_0 = [\frac{l}{a_1}, \frac{l}{a_2}, \ldots, \frac{l}{a_{d-1}}, \frac{l}{a_d}]^T$, where $P_0(i) = \Pr[y \geq a_i] = \Pr[y > a_{i-1}] = \frac{l}{a_i} < \frac{1}{4}$.

- $\mathbb{P}_j$, with acceptance rates of each price: $P_j = [\frac{l}{a_1}, \frac{l}{a_2}, \ldots, \frac{l}{a_{j-1}}, \frac{l+\epsilon}{a_j}, \frac{l}{a_{j+1}}, \ldots, \frac{l}{a_{d-1}}, \frac{l}{a_d}]^T$, where $P_j(i) = \frac{l}{a_i} + \frac{\epsilon}{a_i} \cdot \mathbb{1}(i = j) \leq \frac{1}{4}$.

In the following part, we propose and prove the following lemma:

Lemma 5.8.8. *For any algorithm S, $\exists j \in \{1, 2, \ldots, d\}$, such that $Reg_{\mathbb{P}_j}(S) = \Omega(\sqrt{Td})$.*

Proof. Suppose f is a function: $\{0,1\}^T \to [0, M]$. Denote $\mathbf{r} = [\mathbb{1}_1, \mathbb{1}_2, \ldots, \mathbb{1}_T]^\top$ as a vector containing the customer's decisions in sequence. Then for any $j = 1, 2, \ldots, d$ we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_j}[f(\mathbf{r})] - \mathbb{E}_{\mathbb{P}_0}[f(\mathbf{r})] \\
=&\sum_{\mathbf{r}} f(\mathbf{r}) \cdot (\mathbb{P}_j[\mathbf{r}] - \mathbb{P}_0[\mathbf{r}]) \\
\leq&\sum_{\mathbf{r}:\mathbb{P}_j[\mathbf{r}] \geq \mathbb{P}_0[\mathbf{r}]} f(\mathbf{r})(\mathbb{P}_j[\mathbf{r}] - \mathbb{P}_0[\mathbf{r}]) \\
\leq&M \cdot \sum_{\mathbf{r}:\mathbb{P}_j[\mathbf{r}] \geq \mathbb{P}_0[\mathbf{r}]} f(\mathbf{r})(\mathbb{P}_j[\mathbf{r}] - \mathbb{P}_0[\mathbf{r}]) \\
=&\frac{M}{2}\|\mathbb{P}_j - \mathbb{P}_0\|_1.
\end{aligned} \tag{5.42}$$

Here we cite a lemma from *Cover and Thomas, Elements of Information theory*, Lemma 11.6.1.

Lemma 5.8.9.

$$KL(\mathbb{P}_1||\mathbb{P}_2) \geq \frac{1}{2\ln 2}\|\mathbb{P}_1 - \mathbb{P}_2\|_1^2.$$

Since

$$
\begin{aligned}
&KL(\mathbb{P}_0(\mathbf{r})||\mathbb{P}_j(\mathbf{r})) \\
=&\sum_{t=1}^{T} KL(\mathbb{P}_0[r_t|\mathbf{r}_{t-1}]||\mathbb{P}_j[r_t|\mathbf{r}_{t-1}]) \\
=&\sum_{t=1}^{t} \mathbb{P}_0(i_t \neq j)\cdot 0 + \mathbb{P}_0(i_t = j)\cdot KL(\frac{l}{a_j}||\frac{l}{a_j}+\frac{\epsilon}{a_j}).
\end{aligned} \tag{5.43}
$$

The first equality comes from the chain rule of decomposing a KL-divergence, and the second equality is because $\mathbb{1}_t$ satisfies a Bernoulli distribution $B(1, \frac{l}{a_{i_t}} + \frac{\epsilon}{a_{i_t}} \cdot \mathbb{1}(i_t = j))$. Now we propose another lemma:

Lemma 5.8.10. *If $\frac{1}{12} \leq p \leq \frac{1}{4}$, then we have: $KL(p, p+\epsilon) \leq 12\epsilon^2$ for sufficiently small ϵ.*

According to this Lemma Lemma 5.8.10, we have:

$$
\begin{aligned}
KL(\mathbb{P}_0(\mathbf{r})||\mathbb{P}_j(\mathbf{r})) &\leq \sum_{t=1}^{T} \mathbb{P}_0(i_t = j)\cdot 12\epsilon^2 \\
&\leq \sum_{t=1}^{T} \mathbb{P}_0(i_t = j)\cdot \frac{12}{16l^2}\epsilon^2.
\end{aligned} \tag{5.44}
$$

Therefore, we have:

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_j}[f(\mathbf{r})] - \mathbb{E}_{\mathbb{P}_0}[f(\mathbf{r})] \\
\leq&\frac{M}{2}\frac{2\ln 2}{\cdot}\sqrt{KL(\mathbb{P}_0(\mathbf{r})||\mathbb{P}_j(\mathbf{r}))} \\
\leq&\frac{\sqrt{6\ln 2}M}{4}\cdot(\sqrt{\sum_{t=1}^{T}\mathbb{P}_0(i_t = j)})\cdot\epsilon.
\end{aligned} \tag{5.45}
$$

Denote $N_j := \sum_{t=1}^{T} \mathbb{1}(i_t = j)$, and hence:

$$
\mathbb{E}_{\mathbb{P}_j}[f(\mathbf{r})] - \mathbb{E}_{\mathbb{P}_j}[f(\mathbf{r})] \leq \frac{\sqrt{6\ln 2}}{4} M(\sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]})\cdot\epsilon. \tag{5.46}
$$

Now let $f(\mathbf{r}) = N_j$, i.e., let function f simulate the algorithm which make choices of i_t's from historical results of $\{\mathbb{1}_1, \mathbb{1}_2, \ldots, \int_{t-1}\}$ (It is straightforward that $\{i_1, i_2, \ldots, i_{t-1}\}$ are also historical results crucial for deciding i_t. However, for a deterministic algorithm, it

can generate $i_1, i_2, \ldots, i_{t-1}$ directly from $\emptyset, \{\mathbb{1}_1\}, \{\mathbb{1}_1, \mathbb{1}_2\}, \ldots, \{\mathbb{1}_1, \mathbb{1}_2, \ldots, \mathbb{1}_{t-2}\}$.) Now, $0 \leq f(\mathbf{r}) \leq T$, which indicates that $M = T$.. Then it turns out that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_j}[N_j] - \mathbb{E}_{\mathbb{P}_0}[N_j] &\leq \frac{\sqrt{6\ln 2}}{4} \cdot T \cdot \epsilon \cdot \sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]} \\
\Rightarrow \frac{1}{d}\sum_{j=1}^{d} \mathbb{E}_{\mathbb{P}_j}[N_j] &\leq \frac{1}{d}\sum_{j=1}^{d}(\mathbb{E}_{\mathbb{P}_0}[N_j] + \frac{\sqrt{6\ln 2}}{4} \cdot T \cdot \epsilon \cdot \sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]}) \\
&= \frac{T}{d} + \frac{\sqrt{6\ln 2}}{4} \cdot \frac{T}{d} \cdot \epsilon \cdot \sum_{j=1}^{d} \sqrt{\mathbb{E}_{\mathbb{P}_)}[N_j]} \\
&\leq \frac{T}{d} + \frac{\sqrt{6\ln 2}}{4} \cdot \epsilon \cdot \frac{T}{d} \cdot \sqrt{Td} \\
&= \frac{T}{d} + \frac{\sqrt{6\ln 2}}{4} \cdot \sqrt{\frac{d}{T}} \cdot \frac{T}{d} \cdot \sqrt{Td} \\
&\leq \frac{T}{3} + 0.525 \cdot T \\
&\leq 0.9T
\end{aligned} \tag{5.47}$$

Here the second line is an average over all $j = 1, 2, \ldots, d$ of the first line, the third line uses the fact that $\sum_{j=1}^{d} \mathbb{E}[N_j] = T$, the fourth line applies a Cauchy-Schwarz Inequality that $Td = (\sum_{j=1}^{d} \mathbb{E}_{\mathbb{P}_0}[N_j])(d \cdot 1) \geq (\sum_{j=1}^{d} \sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]})^2$, the fifth line plugs in the values that $\epsilon = \sqrt{\frac{d}{T}}$, the sixth line uses the fact that $\ln 2 < 0.7$, and the last line holds for sufficient large T.

From Equation Eq. (5.47), we know that $\exists j \in \{1, 2, \ldots, d\}$ such that $\mathbb{E}_{\mathbb{P}_j}[N_j] \leq 0.9T$. As a result, we have:

$$\begin{aligned}
Reg_{\mathbb{P}_j}(S) \geq& (1 - 0.9)T(\frac{l + \epsilon}{a_j} \cdot a_j - \frac{l}{a_{i_t}} \cdot a_{i_t}),\ \forall i_t \neq j \\
=& 0.1T(l + \epsilon - l) \\
=& 0.1T\epsilon \\
=& 0.1\sqrt{Td}.
\end{aligned} \tag{5.48}$$

Therefore, the $\Omega(\sqrt{Td})$ regret bound holds. ∎

■

5.8.3 Proof of Theorem 5.5.4

Proof. Prior to our technical analysis, we briefly introduce the roadmap of proving the unfairness lower bound.

(i) We construct two different but very similar problem settings: one is exactly Example 5.1.1, the other is identical to Example 5.1.1 except all probabilities of 0.5 are now changed into $(0.5 - \zeta)$, where $\zeta = C \cdot T^{-\frac{1}{2}+\eta}$ for some super small constant $C \geq 0$ and some small $\eta \geq 0$. In the following, we may call them the "Problem 0" (or P_0) and the "Problem ζ" (or P_ζ) sequentially.

(ii) We derive the close-form solutions to both Problem 0 and Problem ζ, where we also parameterize the reward function with the expected proposed price V_r and the proposed accepted price V_s. Of course Problem ζ is more general and we may get the solutions to Problem 0 by simply let $\zeta = 0$.

(iii) We show that there does not exist any policy π that satisfies both of the following conditions simultaneously:

- π is within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-suboptimal (w.r.t. regret) and within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-unfair (w.r.t. fairness) in $P0$.
- π is within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-suboptimal (w.r.t. regret) and within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-unfair (w.r.t. fairness) in $P\zeta$.

(iv) We show that any algorithm have to distinguish P_0

According to the roadmap above, we firstly construct the following example as the problem setting for lower bound proof.

Example 5.8.11. Customers form 2 disjoint groups: Group 1 takes 30% proportion of customers, and Group 2 takes the rest 70%. In specific,

- In Group 1, 40% customers valuate the item as \$0, $(10\% + \zeta)$ valuate customers it as \$0.625, and $(50\% - \zeta)$ customers valuate it as \$1.
- In Group 2, 20% customers valuate it as \$0, $(30\% + \zeta)$ customers valuate it as \$0.7, and $(50\% - \zeta)$ customers valuate it as \$1.

Here $\zeta = C \cdot T^{-\frac{1}{2}+\eta}$ is a small amount, where $0 \leq C \leq 10^{-10}$. In other words, we have $\mathbf{v}^\top = [\frac{5}{8}, \frac{7}{10}, 1], F_1 = \text{diag}\{0.6, 0.5-\zeta, 0.5-\zeta\}, F_2 = \text{diag}\{0.8, 0.8, 0.5-\zeta\}$ and our policy $\pi = (\pi^1, \pi^2)$ where $\pi^1, \pi^2 \in \Delta^3$. Our goal is to approach the following optimal policy

$$\begin{aligned} \pi_{\zeta,*} = \underset{\pi=(\pi^1,\pi^2)\in\Pi}{\text{argmax}} \; & R(\pi; F_1, F_2) \\ s.t. \quad & U(\pi) = 0 \\ & S(\pi; F_1, F_2) = 0. \end{aligned} \tag{5.49}$$

For any policy π feasible to the constraints in Eq. (5.50), denote its expected accepted price as V_s (identical in both groups) and its expected proposed price as V_r (identical in both groups as well). Notice that $V_r \geq V_s \geq \frac{5}{8}$, we define $\alpha = V_r - V_s$ as their difference, and therefore we know that $\alpha \geq 0$. Again, we denote $R(\pi, F_1, F_2)$ as $R(\pi)$ without causing misunderstandings. Here we propose the following lemma regarding Example 5.8.11.

Lemma 5.8.12 (Close-form solution to Example 5.8.11)**.** *For the problem setting defined*

in Example 5.8.11, we have:

$$\begin{aligned}\pi^1_{\zeta,*} =&[\frac{20-40\zeta}{29-10\zeta}, 0, \frac{9+30\zeta}{29-10\zeta}]^\top,\\ \pi^2_{\zeta,*} =&[0, \frac{25-50\zeta}{29-10\zeta}, \frac{4+40\zeta}{29-10\zeta}]^\top, \forall \zeta \in [0, 10^{-10}].\end{aligned} \tag{5.50}$$

Besides, for any feasible policy π and its corresponding V_s and α, we have:

$$R(\pi) = \frac{71-30\zeta}{100} \cdot V_s + \frac{(100-60\zeta)-(142-60\zeta)V_s}{(8V_s-5)(1-V_s)25} \cdot V_s \cdot \alpha. \tag{5.51}$$

Proof of Lemma 5.8.12. For any feasible policy $\pi = (\pi^1, \pi^2)$, it has to satisfy the following equations for $e = 1, 2$:

$$\begin{cases} \mathbb{1}^\top \pi^e = 1 \\ \mathbf{v}^\top \pi^e = V_s + \alpha \\ \dfrac{\mathbf{v}^\top F_e \pi^e}{\mathbb{1}^\top F_e \pi^e} = V_s. \end{cases}$$

This is equivalent to the following linear equations system

$$\begin{cases} \mathbb{1}^\top \pi^e = 1 \\ \mathbf{v}^\top \pi^e = V_s + \alpha \\ (\mathbf{v} - V_s \cdot \mathbb{1})^\top F_e \pi^e = 0. \end{cases} \tag{5.52}$$

This is further equivalent to $A_1\pi^1 = [1, V_s+\alpha, 0]^\top$ and $A_2\pi^2 = [1, V_s+\alpha, 0]^\top$ where

$$A_1(V_s, \zeta) = \begin{bmatrix} 1 & 1 & 1 \\ \frac{5}{8} & \frac{7}{10} & 1 \\ (\frac{5}{8}-V_s)\cdot\frac{3}{5} & (\frac{7}{10}-V_s)\cdot(\frac{1}{2}-\zeta) & (1-V_s)\cdot(\frac{1}{2}-\zeta) \end{bmatrix}. \tag{5.53}$$

and

$$A_2(V_s, \zeta) = \begin{bmatrix} 1 & 1 & 1 \\ \frac{5}{8} & \frac{7}{10} & 1 \\ (\frac{5}{8}-V_s)\cdot\frac{4}{5} & (\frac{7}{10}-V_s)\cdot\frac{4}{5} & (1-V_s)\cdot(\frac{1}{2}-\zeta) \end{bmatrix}. \tag{5.54}$$

Here we may omit the parameters (V_s, ζ) without misunderstanding. For $V_s = \frac{5}{8}$, the only possible policy is to propose the lowest price $\frac{5}{8}$ for both groups, and the expected reward is $0.3 \times \frac{5}{8} \times \frac{3}{5} + 0.7 \times \frac{5}{8} \times \frac{4}{5} = 0.4625 < 0.5 - \zeta$. Therefore, it is suboptimal as its expected reward is less than that of a deterministic policy keep proposing 1 as a price (whose reward is $0.5 - \zeta$). In the following, we only consider the case when $V_s > \frac{5}{8}$. Solve these linear equation systems and get

$$\begin{aligned}\pi^1 =& A_1^{-1}[1, V_s + \alpha, 0]^\top \\ =& \frac{1}{3(8V_s - 5)(1 + 10\zeta)} \\ & \begin{bmatrix} 120\alpha(1 - 2\zeta) \\ -((1 + 10\zeta)8V_s + 10(1 - 8\zeta)) \cdot \alpha - (8(1 + 10\zeta)V_s^2 - 13(1 + 10\zeta)V_s + (1 + 10\zeta)5) \\ 10(8(1 + 10\zeta)V_s - 2(1 + 28\zeta)) \cdot \alpha + (1 + 10\zeta)(8V_s - 5)(10V_s - 7) \end{bmatrix}\end{aligned} \tag{5.55}$$

and

$$\begin{aligned}\pi^2 =& A_2^{-1}[1, V_s + \alpha, 0]^\top \\ =& \frac{1}{3(1 - V_s)(3 + 10\zeta)} \\ & \begin{bmatrix} 4(((3 + 10\zeta)10V_s - (6 + 100\zeta))\alpha - (3 + 10\zeta)(10V_s - 7)(V_s - 1)) \\ (-5) \cdot (((3 + 10\zeta)8V_s - 80\zeta)\alpha + (3 + 10\zeta)(8V_s - 5)(V_s - 1)) \\ 24\alpha \end{bmatrix} \cdot\end{aligned} \tag{5.56}$$

On the one hand, we can get the explicit form of $R(\pi)$ w.r.t. V_s and α:

$$\begin{aligned}R(\pi) =& q \cdot \mathbf{v}^\top F_1 \pi^1 + (1 - q) \cdot \mathbf{v}^\top F_2 \pi^2 \\ =& \frac{71 - 30\zeta}{100} V_s + \frac{(100 - 60\zeta) - (142 - 60\zeta)V_s}{(8V_s - 5)(1 - V_s)25} \cdot V_s \alpha.\end{aligned} \tag{5.57}$$

On the other hand, we have a few constraints to be applied. Since π is a probabilistic

distribution, we have $\pi^e(i) \geq 0, e = 1, 2; i = 1, 2, 3$, which lead to

$$\begin{cases} 120\alpha(1-2\zeta) \geq 0 \\ -((1+10\zeta)8V_s + 10(1-8\zeta)) \cdot \alpha - (8(1+10\zeta)V_s^2 - 13(1+10\zeta)V_s + (1+10\zeta)5) \geq 0 \\ 10(8(1+10\zeta)V_s - 2(1+28\zeta)) \cdot \alpha + (1+10\zeta)(8V_s-5)(10V_s-7) \geq 0 \\ 4(((3+10\zeta)10V_s - (6+100\zeta))\alpha - (3+10\zeta)(10V_s-7)(V_s-1)) \geq 0 \\ (-5) \cdot (((3+10\zeta)8V_s - 80\zeta)\alpha + (3+10\zeta)(8V_s-5)(V_s-1)) \geq 0 \\ 24\alpha \geq 0. \end{cases} \tag{5.58}$$

From Eq. (5.58), we may derive the following upper and lower bounds for α.

(a) The first line and the last line of Eq. (5.58) is naturally satisfied.

(b) From the second line, we have

$$\alpha \leq \frac{(1+10\zeta)(8V_s-5)(1-V_s)}{(1+10\zeta)8V_s + 10(1-8\zeta)} := B_1. \tag{5.59}$$

(c) From the third line, we have

$$\alpha \geq \frac{(1+10\zeta)(8 \cdot V_s - 5)(7-10V_s)}{(1+10\zeta)8 \cdot V_s - 2(1+28\zeta)} \cdot \frac{1}{10} := B_2. \tag{5.60}$$

(d) From the fourth line, we have

$$\alpha \geq \frac{(3+10\zeta)(10V_s-7)(1-V_s)}{(3+10\zeta)10V_s - (6+100\zeta)} := B_3. \tag{5.61}$$

(e) From the fifth line, we have

$$\alpha \leq \frac{(3+10\zeta)(8V_s-5)(1-V_s)}{(3+10\zeta)8 \cdot V_s - 80\zeta} := B_4. \tag{5.62}$$

We get four constraints on α as above, where Eq. (5.59) and Eq. (5.62) are upper bounds, and Eq. (5.60) and Eq. (5.61) are lower bounds. Compare B_1 with B_4, we notice that

$$\frac{B_1}{B_4} = \frac{\frac{80\zeta}{3+10\zeta}}{8V_s + 10 \cdot \frac{1-8\zeta}{1+10\zeta}} \leq 1. \tag{5.63}$$

Therefore, Eq. (5.59) is tighter than Eq. (5.62). For the comparison between B_2 and B_3, we notice that $B_2 < 0 < B_3$ when $V_s > \frac{7}{10}$ and $B_2 \geq 0 \geq B_3$ when $V_s \leq \frac{7}{10}$.

In the following part, we derive the optimal policy by cases.

(a) When $\frac{5}{8} < V_s \leq \frac{50-30\zeta}{71-30\zeta}$, we have

$$\begin{aligned}
R(\pi) =& \frac{71-30\zeta}{100}V_s + \frac{(100-60\zeta)-(142-60\zeta)V_s}{(8V_s-5)(1-V_s)25} \cdot V_s\alpha \\
\leq& \frac{71-30\zeta}{100}V_s + \frac{(100-60\zeta)-(142-60\zeta)V_s}{(8V_s-5)(1-V_s)25} \cdot V_s \cdot B_1 \\
=& \frac{71-30\zeta}{100}V_s + \frac{(100-60\zeta)-(142-60\zeta)V_s}{(8V_s-5)(1-V_s)25} \cdot \frac{(1+10\zeta)(8V_s-5)(1-V_s)}{(1+10\zeta)8V_s+10(1-8\zeta)}V_s \\
=& \frac{71-30\zeta}{100}V_S + \frac{100-142V_s-60\zeta(1-V_s)}{25(8V_s+10\cdot\frac{1-8\zeta}{1+10\zeta})} \cdot V_s \\
=& \frac{71-30\zeta}{100}V_S + \frac{100-142V_s-60\zeta(1-V_s)}{25(8V_s+10)-\frac{450\zeta}{1+10\zeta}} \cdot V_s \\
<& \frac{71-30\zeta}{100}V_S + \frac{100-142V_s-60\zeta(1-V_s)+\frac{450\zeta}{1+10\zeta}}{25(8V_s+10)} \cdot V_s \\
<& \frac{71}{100}V_s + \frac{100.1-142V_s}{25(8V_s+10)} \cdot V_s \\
=& \frac{71(8\cdot V_s+10)+(100.1-142\cdot V_s)\cdot 4}{100(8\cdot V_s+10)} \cdot V_s \\
=& \frac{11104\cdot V_s}{1000(8\cdot V_s+10)} \\
\leq& \frac{11104}{8000} - \frac{\frac{5}{4}\times 11104}{1000(8\times\frac{50}{71}+10)} \\
\leq& 0.50019
\end{aligned} \tag{5.64}$$

Here the first inequality (line 2) is by $(100-60\zeta)-(142-60\zeta)V_s \geq 0$ as $V_s \leq \frac{50-30\zeta}{71-30\zeta}$

and by $\alpha \le B_1$, the second inequality (line 6) is by the fact that V_s's coefficient is within $(0,1)$, the third inequality (line 7) is by $\zeta \le 10^{-10} < \frac{1}{4500}$, the fourth inequality (line 10) is by $V_s \le \frac{50}{71}$ and the last inequality (line 11) is by numerical computations. We will later show that 0.50019 is not optimal.

(b) When $V_s > \frac{50-30\zeta}{71-30\zeta}$, we know that $V_s > \frac{7}{10}$ as $\zeta < \frac{1}{30}$, and $B_2 < 0 < B_3$ and also

$$\alpha \ge B_3 = \frac{(3+10\zeta)(10V_s-7)(1-V_s)}{(3+10\zeta)10V_s-(6+100\zeta)}.$$

As a result, we have

$$\begin{aligned} R(\pi) &= \frac{71-30\zeta}{100}V_s - \frac{(142-60\zeta)V_s-(100-60\zeta)}{(8V_s-5)(1-V_s)25}\cdot V_s\alpha \\ &\le \frac{71-30\zeta}{100}V_s - \frac{(142-60\zeta)V_s-(100-60\zeta)}{(8V_s-5)(1-V_s)25}\cdot V_s\cdot B_3 \\ &\le \frac{71-30\zeta}{100}V_s - \frac{(142-60\zeta)V_s-(100-60\zeta)}{(8V_s-5)(1-V_s)25}\cdot V_s\cdot \frac{(3+10\zeta)(10V_s-7)(1-V_s)}{(3+10\zeta)10V_s-(6+100\zeta)} \end{aligned} \tag{5.65}$$

Also, we can derive an upper bound for V_s as $B_3 \le \alpha \le B_1$.

$$\begin{aligned} \frac{(3+10\zeta)(10V_s-7)(1-V_s)}{(3+10\zeta)10V_s-(6+100)} &\le \frac{(1+10\zeta)(8\cdot V_s-5)(1-V_s)}{(1+10\zeta)8\cdot V_s+10(1-8\zeta)} \\ \Leftrightarrow \qquad V_s &\le \frac{8+10\zeta}{11+10\zeta}. \end{aligned} \tag{5.66}$$

Then we solve the maximal of $R(\pi)$ on $\frac{50-30\zeta}{71-30\zeta} \le V_s \le \frac{8+10\zeta}{11+10\zeta}$ by combining Eq. (5.65).

$$\begin{aligned} \frac{\partial R(\pi)}{\partial V_s} =& \frac{(-3135+9870V_s-7491V_s^2)+\zeta(-46430+145660V_s-114638V_s^2)}{20(-5+8V_s)^2(-3-50\zeta+15V_s+50\zeta V_s)^2/3} \\ &+ \frac{\zeta^2(97900-315800V_s+251740V_s^2)+\zeta^3(15000-30000V_s+15000V_s^2)}{20(-5+8V_s)^2(-3-50\zeta+15V_s+50\zeta V_s)^2/3} \\ =& \frac{-22473(V_s-\frac{1645-6\sqrt{2685}}{2497})(V_s-\frac{1645+6\sqrt{2685}}{2497})}{20(-5+8V_s)^2(-3-50\zeta+15V_s+50\zeta V_s)^2} \\ &+ \frac{(-46430+145660V_s-114638V_s^2)\zeta+o(\zeta^2)}{20(-5+8V_s)^2(-3-50\zeta+15V_s+50\zeta V_s)^2/3} \\ \ge& \frac{-22473(V_s-0.5343)(v_s-0.7833)}{20(-5+8V_s)^2(-3-50\zeta+15V_s+50\zeta V_s)^2} > 0. \end{aligned} \tag{5.67}$$

Here the "$\gtrsim$" inequality is because the coefficient of ζ in any monomial above is within $\pm 10^6$, which indicates that any monomial containing ζ is within ± 0.0001 . The last line is because $\frac{7}{10} \leq \frac{50-30\zeta}{71-30\zeta} \leq V_s \leq \frac{8+10\zeta}{11+10\zeta} \leq \frac{3}{4}$ and therefore $(V_s - 0.5343)(v_s - 0.7833) < 0$. As a result, we know that $R(\pi)$ is monotonically increasing as V_s increases within the range above. Therefore, we have:

$$R(\pi) \leq R(\pi)|_{\alpha=B_3} \leq R(\pi)|_{\alpha=B_3 \text{ and } V_s=\frac{8+10\zeta}{11+10\zeta}} = \frac{37(1-2\zeta)(4+5\zeta)}{10(29-10\zeta)} \tag{5.68}$$

By plugging in $V_s = \frac{8+10\zeta}{11+10\zeta}$ into $\alpha = B_3$ and the close-form feasible solutions of π^1 and π^2 (i.e., Eq. (5.55) and Eq. (5.56)), we may get:

$$\begin{aligned} \alpha = B_3 &= \frac{3(1+10\zeta)(3+10\zeta)}{2(29-10\zeta)(11+10\zeta)} \\ \pi^1 &= [\frac{20-40\zeta}{29-10\zeta}, 0, \frac{9+30\zeta}{29-10\zeta}]^\top \\ \pi^2 &= [0, \frac{25-50\zeta}{29-10\zeta}, \frac{4+40\zeta}{29-10\zeta}]^\top. \end{aligned} \tag{5.69}$$

Pushing back Eq. (5.69) to Eq. (5.49), we verify that $R(\pi)_{\max} = \frac{37(1-2\zeta)(4+5\zeta)}{10(29-10\zeta)}$ and therefore all inequalities in Eq. (5.68) hold as equalities.

Notice that $\frac{37(1-2\zeta)(4+5\zeta)}{10(29-10\zeta)} > 0.50019$, and therefore the optimal policy $\pi_{\zeta,*}$ is what we derive in Eq. (5.69). This holds the lemma. ∎

With Lemma 5.8.12, we know that $\pi^1_{\zeta,*} = [\frac{20-40\zeta}{29-10\zeta}, 0, \frac{9+30\zeta}{29-10\zeta}]^\top$ and $\pi^2_{\zeta,*} = [0, \frac{25-50\zeta}{29-10\zeta}, \frac{4+40\zeta}{29-10\zeta}]^\top$. We denote $V^*_{s,\zeta} := \frac{8+10\zeta}{3+10\zeta}$ and $\alpha^*_\zeta = \frac{3(1+10\zeta)(3+10\zeta)}{2(29-10\zeta)(11+10\zeta)}$ for future use. We also know that the optimal policy for Example 5.1.1 (i.e., $\zeta = 0$) is exactly what we proposed, i.e., $\pi^1_* = [\frac{20}{29}, 0, \frac{9}{29}]^\top$ and $\pi^2_* = [0, \frac{25}{29}, \frac{4}{29}]^\top$.

Let us go back to the two problems: P_0 defined in Example 5.1.1 and P_ζ defined in Example 5.8.11, where we consider the following four conditions:

- π is within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-suboptimal (w.r.t. regret) in P_0 (denoted as Condition A).
- π is within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-suboptimal (w.r.t. regret) in P_ζ (denoted as Condition B).
- π is within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-unfair (w.r.t. fairness) in P_0 (denoted as Condition C).
- π is within $C_0 \cdot T^{-\frac{1}{2}+\eta}$-unfair (w.r.t. fairness) in P_ζ (denoted as Condition D).

According to our proof roadmap, we then prove the following lemma:

Lemma 5.8.13 (No policy fitting in P_0 and P_ζ)**.** *There exist constants $C_0 > 0$ such that there does not exist any policy $\pi \in \Pi$ that satisfies all of Condition $ABCD$ (denoting $A \wedge B \wedge C \wedge D$) simultaneously.*

Corollary 5.8.14. *The space of Π can be divided as the following 3 subspaces:*

1. *Policies satisfying Condition AC (denoted as Space AC).*
2. *Policies satisfying Condition BD (denoted as Space BD).*
3. *Policies satisfying Condition (denoted as Outer Spaces) $\bar{A}\bar{B} \vee \bar{C}\bar{D} \vee \bar{A}\bar{D} \vee \bar{B}\bar{C}$. and these three subspaces are pairwise disjoint.*

Proof of Lemma 5.8.13. Let $C_1 = \frac{C}{W}$ and $C_2 = \frac{C}{W \cdot L}$ where $L > 0$ is a constant from Lemma 5.8.5 and $W \geq 10$ to be specified later. Let $C_0 = \min\{C_1, C_2\}$, and we prove the lemma by contradiction. Suppose there exists a policy π satisfies the four conditions above, and then we denote the expected accepted prices in G_1 and G_2 in Problem ζ are $V_{s,\zeta}$ and $V_{s,\zeta} + \beta_\zeta$ sequentially, where $\beta_\zeta \in [0, C_2 T^{-\frac{1}{2}+\eta}]$. Here we assume $\beta \geq 0$ without losing generality as we will not use the specific property of G_1 versus G_2. Also, we denote

α_ζ as the difference between the expected proposed price in both groups (denoted as $V_{r,\zeta}$) and $V_{s,\zeta}$.

Now, consider a corresponding policy:

$$\check{\pi} := \begin{cases} G_1 : \mathbb{E}[\text{accepted price}] = V_{s,\zeta}, \mathbb{E}[\text{proposed price}] = V_{s,\zeta} + \alpha_\zeta \\ G_2 : \mathbb{E}[\text{accepted price}] = V_{s,\zeta}, \mathbb{E}[\text{proposed price}] = V_{s,\zeta} + \alpha_\zeta \end{cases} \tag{5.70}$$

According to Eq. (5.66), we know that $\frac{5}{8} \leq V_{s,\zeta} \leq V^*_{s,\zeta} = \frac{8+10\zeta}{11+10\zeta}$ and $R(\check{\pi}) \leq R(\pi_{\zeta,*})$. Therefore, we have:

$$\begin{aligned} R(\pi) &\leq R(\check{\pi}) + L \cdot \beta_\zeta \\ &\leq R(\pi_{\zeta,*}) - \min_{V_s \in [\frac{5}{8}, \frac{8+10\zeta}{11+10\zeta}]} \frac{\partial R(\pi)}{\partial V_s} \cdot (V^*_{s,\zeta} - V_{s,\zeta}) + L \cdot \beta_\zeta \\ &\leq R(\pi_{\zeta,*}) - \frac{1}{4} \cdot (V^*_{s,\zeta} - V_{s,\zeta}) + L \cdot \beta_\zeta. \end{aligned} \tag{5.71}$$

Here the first line comes from Lemma 5.8.5, the second line comes from the fact that $f(x_1) - f(x_2) \geq \min_x(f'(x))(x_1 - x_2)$ for $x_1 \geq x_2$ and $f'(x) > 0$, and the third line comes from the fact that $\frac{\partial R(\pi)}{\partial V_s} \geq \frac{1}{4}$ for $V_s \in [0.625, 0.728]$ as $0.728 > \frac{8+10\zeta}{11+10\zeta}$ for $\zeta \leq 10^{-10}$. Also, since π satisfies a low-regret condition, we have

$$R(\pi_{\zeta,*}) - R(\pi) \leq C_1 \cdot T^{-\frac{1}{2}+\eta}.$$

Combining with Eq. (5.71), we have

$$\begin{aligned} &\frac{1}{4}(V^*_{s,\zeta} - V_{s,\zeta}) - L \cdot \beta_\zeta \leq C_1 T^{-\frac{1}{2}+\eta} \\ \Rightarrow \quad &(V^*_{s,\zeta} - V_{s,\zeta}) \leq C_1 T^{-\frac{1}{2}+\eta} + L \cdot \beta_\zeta \\ &\leq (C_1 + LC_2) T^{-\frac{1}{2}+\eta}. \end{aligned} \tag{5.72}$$

Notice that this is suitable for any $\zeta \in [0, 10^-10]$, we may have the same result for both $\zeta = CT^{-\frac{1}{2}+\eta}$ and for $\zeta = 0$. We denote $\zeta_0 = 0$ and $\zeta_1 = CT^{-\frac{1}{2}+\eta}$ where $C = 10^{-10}$.

Therefore, we have:

$$\begin{aligned}(V^*_{s,\zeta_0} - V_{s,\zeta_0}) \leq & (C_1 + LC_2)T^{-\frac{1}{2}+\eta},\\ (V^*_{s,\zeta_1} - V_{s,\zeta_1}) \leq & (C_1 + LC_2)T^{-\frac{1}{2}+\eta}.\end{aligned} \tag{5.73}$$

Now let us bound $(\alpha^*_\zeta - \alpha_\zeta)$ for both ζ_0 and ζ_1. From Eq. (5.59) and Eq. (5.61), we have

$$\begin{aligned} & B_3|_{V_s=V_{s,\zeta}} \leq \alpha_\zeta \leq B_1|_{V_s=V_{s,\zeta}}\\ & B_3|_{V_s=V^*_{s,\zeta}} = \alpha^*_\zeta = B_1|_{V_s=V^*_{s,\zeta}}\\ \Rightarrow \min_{V_s\in[0.7,0.75]}&\{\frac{\partial B_3}{\partial V_s}, \frac{\partial B_1}{\partial V_s}\}(V^*_{s,\zeta} - V_{s,\zeta}) \leq (\alpha^*_\zeta - \alpha_\zeta) \leq \max_{V_S\in[0.7,0.75]}\{\frac{\partial B_3}{\partial V_s}, \frac{\partial B_1}{\partial V_s}\}(V^*_{s,\zeta} - V_{s,\zeta})\\ \Rightarrow \quad & 0 \leq 0.05(V^*_{s,\zeta} - V_{s,\zeta}) \leq (\alpha^*_\zeta - \alpha_\zeta) \leq 0.6(V^*_{s,\zeta} - V_{s,\zeta}).\end{aligned} \tag{5.74}$$

Therefore, we have:

$$0 \leq V^*_{r,\zeta} - V_{r,\zeta} = V^*_{s,\zeta} + \alpha^*_\zeta - (V_{s,\zeta} + \alpha_\zeta) \leq (1 + 0.6)(V^*_{s,\zeta} - V_{s,\zeta}) \leq \frac{8}{5} \cdot (C_1 + LC_2)T^{-\frac{1}{2}+\eta}.$$

Therefore, we know that

$$\begin{aligned} V^*_{r,\zeta_0} \geq & V_{r,\zeta_0} \geq V^*_{r,\zeta_0} - \frac{8}{5} \cdot (C_1 + LC_2)T^{-\frac{1}{2}+\eta},\\ V^*_{r,\zeta_1} \geq & V_{r,\zeta_1} \geq V^*_{r,\zeta_1} - \frac{8}{5} \cdot (C_1 + LC_2)T^{-\frac{1}{2}+\eta}.\\ \Rightarrow \quad & |V_{r,\zeta_1} - V_{r,\zeta_0}| \geq |V^*_{r,\zeta_0} - V^*_{r,\zeta_1}| - (C_1 + LC_2)T^{-\frac{1}{2}+\eta}.\end{aligned} \tag{5.75}$$

HOWEVER, we have $V_{r,\zeta_1} = V_{r,\zeta_0}$ since they are the expected proposed price of the same pricing policy π in P_0 and P_ζ where the prices sets are all the same! Therefore, we have $|V^*_{r,\zeta_0} - V^*_{r,\zeta_1}| - (C_1 + LC_2)T^{-\frac{1}{2}+\eta} \leq 0$. Since $V^*_{r,\zeta_0} = \frac{43}{58}$ and $V_{r,\zeta_1} = \frac{43+10\zeta_1}{58-20\zeta_1}$, we have $|V^*_{r,\zeta_0} - V^*_{r,\zeta_1}| = \frac{360\zeta}{29(29-10\zeta)} \geq \frac{C}{3} \cdot T^{-\frac{1}{2}+\eta}$. Since $C_1 = \frac{C}{W} \leq 10^{-11}$ and $C_2 = \frac{C}{W \cdot L} \leq \frac{1}{L} \times 10^{-}11$, we know that $|V^*_{r,\zeta_0} - V^*_{r,\zeta_1}| > (C_1 + LC_2)T^{-\frac{1}{2}+\eta}$, which contradicts to the inequality we derived. Therefore, the lemma is proved by contradiction. ■

In the following, we set $\zeta = \zeta_1 = CT^{-\frac{1}{2}+\eta}$ where $C = 10^-10$ as is defined in the proof of Lemma 5.8.13. Now let us go back to the main stream of proving Theorem 5.5.4. We also

make it by contradiction. For any given C_x, without loss of generality, we may assume that $C_u \leq C_x$ to be specified later. Define $x = \frac{C_x}{\log T}$, and therefore $C_x T^{\frac{1}{2}} = T^{\frac{1}{2}-x}$. We will make use of Example 5.1.1 and Example 5.8.11, and let $\eta > x$ to be specified later. Therefore, we have $C_u \cdot T^{\frac{1}{2}} \leq C_x \cdot T^{\frac{1}{2}} = T^{\frac{1}{2}-x}$, which means that the contradiction is a **sufficient condition** to the following result: Suppose there exists an $x > 0$ and an algorithm such that it can always achieve $O(T^{\frac{1}{2}+x})$ regret with zero procedural unfairness and $O(T^{\frac{1}{2}-x})$ substantive unfairness. According to Corollary 5.8.14, we know that any policy $\pi \in \mathit{\Pi}$ are in exact one of those three spaces. In our problem setting, denote the policy we take at time $t = 1, 2, \ldots, T$ as π_t. Now we show that: among all policies $\{\pi_t\}_{t=1}^T$ we have taken, there are at most $O(T^{1-\eta+x})$ policies in all T policies having been played belonging to the Outer Space defined in Corollary 5.8.14. In fact, for any policy π in the Outer Space, we have:

(i) When $\pi \in \bar{A}\bar{B}$, the policy π will definitely suffer a regret $C_0 \cdot T^{-\frac{1}{2}+\eta}$, no matter which the problem setting is (i.e., P_0 or P_ζ). In order to guarantee $O(T^{\frac{1}{2}+x})$ regret, there are at most $N_1 = O(T^{1-\eta+x}) = o(T)$ rounds to play a policy in $\bar{A}\bar{B}$.

(ii) When $\pi \in \bar{C}\bar{D}$, the policy π will definitely suffer a substantive unfairness $C_0 \cdot T^{-\frac{1}{2}+\eta}$, no matter which the problem setting is (i.e., P_0 or P_ζ). In order to guarantee $O(T^{\frac{1}{2}-x})$ regret, there are at most $N_2 = O(T^{1-\eta-x}) = o(T)$ rounds to play a policy in $\bar{C}\bar{D}$.

(iii) When $\pi \in \bar{A}\bar{D} \vee \bar{B}\bar{C}$, in either P_0 or P_ζ it suffers something (that could be either $C_0 \cdot T^{-\frac{1}{2}+\eta}$ regret or $C_0 \cdot T^{-\frac{1}{2}+\eta}$ unfairness). As we have to guarantee $O(T^{\frac{1}{2}+x})$ regret and $O(T^{\frac{1}{2}-x})$ substantive unfairness, there are still at most $N_3 = O(\max\{T^{1-\eta+x}, T^{1-\eta-x}\}) = O(T^{1-\eta+x}) = o(T)$.

Therefore, the number of rounds when we select and play a policy from Space AC or

Space BD is at least $T - o(T) \geq \frac{T}{2}$. Notice that if a policy in AC, then it performs well in P_0 but not necessarily in P_ζ. Similarly, if a policy in BD, then it performs well in P_ζ but not necessarily in P_0. Therefore, two questions emerges:

- How do policies in AC perform in P_ζ? and How do policies in BD perform in P_0? Specifically, we only care about the substantive fairness.
- How can we distinguish between P_ζ and P_0?

Denote $F_1(\zeta) = \text{diag}\{0.6, 0.5 - \zeta, 0.5 - \zeta\}$ and $F_2(\zeta) = \text{diag}\{0.8, 0.8, 0.5 - \zeta\}$. Also, denote $S_0(\pi) := S(\pi, F_1(0), F_2(0))|_{\zeta=0}$ and $S_\zeta(\pi) := S(\pi, F_1(\zeta), F_2(\zeta))$. In the following, we propose two lemmas that help us prove. The first lemma, Lemma 5.8.15, shows that failing to distinguish would lead to large substantive unfairness, which answers the first question above.

Lemma 5.8.15. *There exists a constant C_{ac} such that: for any policy $\pi \in AC$, we have $S_\zeta(\pi) > C_{ac} \cdot T^{-\frac{1}{2}+\eta}$. There also exists a constant C_{bd} such that: for any policy $\pi \in BD$, we have $S_0(\pi) > C_{bd} \cdot T^{-\frac{1}{2}+\eta}$.*

Proof of Lemma 5.8.15 . We firstly prove the first half of this lemma, and then demonstrate the second half (which can be proved in exact the same way.)

First of all, we have the close-form solution to both P_0 and P_ζ in Eq. (5.50). Therefore, we have

$$S_\zeta(\pi_{0,*}) = \frac{12\zeta(1-2\zeta)}{(11-6\zeta)(11-10\zeta)}. \tag{5.76}$$

Now, consider any policy $\pi \in AC$. Similar to the Proof of Lemma 5.8.13, we define its accepted prices in G_1 and G_2 are $V_{s,0}$ and $V_{s,0} + \beta$ where $\beta \in [0, C_2 T^{-frac12+\eta}]$. We also denote the expected proposed price in both group as $V_{r,0} = V_{s,0} + \alpha_0$. Also, define a

corresponding policy $\check{\pi}$:

$$\check{\pi} := \begin{cases} G_1 : \mathbb{E}[\text{accepted price}] = V_{s,0}, \mathbb{E}[\text{proposed price}] = V_{s,0} + \alpha_0 \\ G_2 : \mathbb{E}[\text{accepted price}] = V_{s,0}, \mathbb{E}[\text{proposed price}] = V_{s,0} + \alpha_0. \end{cases} \tag{5.77}$$

Notice that $\pi^1 = (A_1(V_{s,0}, 0))^{-1}[1, V_{r,0}, 0]^\top$ and $\pi^2 = (A_2(V_{s,0}, 0))^{-1}[1, V_{r,0}, \beta \cdot \mathbb{1}^\top F_2 \pi^2]^\top$. In comparison, we have $\check{\pi}^1 = (A_1(V_{s,0}, 0))^{-1}[1, V_{r,0}, 0]^\top$ and $\check{\pi}^2 = (A_2(V_{s,0}, 0))^{-1}[1, V_{r,0}, 0]^\top$. Therefore, we have:

$$\begin{aligned} \pi^1 &= \check{\pi}^1 \\ \|\pi^2 - \check{\pi}^2\|_1 &= \|(A_2(V_{s,0}, 0))^{-1}([1, V_{r,0}, \beta \cdot \mathbb{1}^\top F_2 \pi^2]^\top - [1, V_{r,0}, 0]^\top)\|_1 \\ &\leq \|(A_2(V_{s,0}, 0))^{-1}[0, 0, \beta]\|_1 \\ &= \|(A_2(V_{s,0}, 0))^{-1}_{[:,3]}\|_1 \cdot \beta \\ &\leq \frac{100\beta}{9(7V_s - 5)}. \end{aligned} \tag{5.78}$$

Also, since $F_{\min} \leq \mathbb{1}^\top F_2 \pi^2 \leq 1$ and $\|\mathbf{v}^\top F_2\|_1 \leq d$ always hold, we know that $\|\frac{\partial S_\zeta(\pi)}{\partial \pi^2}\| \leq \frac{d}{F_{\min}} \cdot \|\pi^2\|_1 = \frac{d}{F_{\min}}$. Since $V^*_{s,0} \approx \frac{8}{11}$ and all $V_{s,0}$ we consider are around it (According to Eq. (5.73)), we may assume that $(7V_s - 5) > \frac{1}{2} \cdot (\frac{8}{11} - \frac{5}{7}) > \frac{1}{200}$ Therefore, we have:

$$\begin{aligned} |S_\zeta(\check{\pi}) - S_\zeta(\pi)| &\leq \frac{d}{F_{\min}} \|\check{\pi}^2 - \pi^2\|_2 \\ &\leq \frac{100 d\beta}{9(7 \cdot V_s - 5) F_{\min}} \\ &\leq \frac{100 d C_2}{9(7 \cdot V_s - 5) F_{\min}} \cdot T^{-\frac{1}{2}+\eta} \\ &\leq \frac{3000 d C_2}{F_{\min}} \cdot T^{-\frac{1}{2}+\eta}. \end{aligned} \tag{5.79}$$

Also, according to the proof of Lemma 5.8.13, we know that $|V_{s,0} - V^*_{s,0}| \leq (C_1 + LC_2)T^{-\frac{1}{2}+\eta}$ and $|\alpha^*_0 - \alpha_0| \leq 0.6(C_1 + LC_2)T^{-\frac{1}{2}+\eta}$ (as $\zeta = 0$). Plugging in Eq. (5.55) and Eq. (5.56),

we have:

$$\begin{aligned}
\|\pi^1_{0,*} - \check{\pi}^1\|_1 \leq & 50 \cdot ((120 + 8 + 10 + 10 \times (8 + 2))|\alpha^*_0 - \alpha_0| + (13 + 106)|V_{s,0} - V^*_{s,0}|) \\
\leq & 1309(C_1 + LC_2)T^{-\frac{1}{2}+\eta} \\
\|\pi^2_{0,*} - \check{\pi}^2\|_1 \leq & \frac{1}{3 \cdot 0.2 \cdot 3}((144 + 120 + 24)|\alpha^*_0 - \alpha_0| + (120 + 51 + 120 + 39)|V_{s,0} - V^*_{s,0}|) \\
\leq & 350(C_1 + LC_2)T^{-\frac{1}{2}+\eta}
\end{aligned} \tag{5.80}$$

Therefore, we have:

$$\begin{aligned}
|S_\zeta(\pi_{0,*}) - S_\zeta(\check{\pi}) \leq & \frac{d}{F_{\min}}\|\pi_{0,*} - \check{\pi}\|_2 \\
= & \frac{d}{F_{\min}}(\|\pi^1_{0,*} - \check{\pi}^1\|_2 + \|\pi^2_{0,*} - \check{\pi}^2\|_2) \\
\leq & \frac{d}{F_{\min}}(\|\pi^1_{0,*} - \check{\pi}^1\|_1 + \|\pi^2_{0,*} - \check{\pi}^2\|_1) \\
\leq & \frac{d}{F_{\min}} \cdot (1309(C_1 + LC_2)T^{-\frac{1}{2}+\eta} + 350(C_1 + LC_2)T^{-\frac{1}{2}+\eta}) \\
\leq & \frac{d}{F_{\min}} 2000(C_1 + LC_2)T^{-\frac{1}{2}+\eta}.
\end{aligned} \tag{5.81}$$

Recall that $C_1 = \frac{C}{W}$ and $C_2 = \frac{C}{W \ldots L}$. Now, we let $W = 10^6 \frac{d}{F_{\min}}$. Therefore, we have:

$$\begin{aligned}
S_\zeta(\pi) = & S_\zeta(\pi) - S_\zeta(\check{\pi}) + S_\zeta(\check{\pi}) - S_\zeta(\pi_{0,*}) + S_\zeta(\pi_{0,*}) \\
\geq & S_\zeta(\pi_{0,*}) - |S_\zeta(\pi) - S_\zeta(\check{\pi})| - |S_\zeta(\check{\pi}) - S_\zeta(\pi_{0,*})| \\
\geq & \frac{12\zeta(1 - 2\zeta)}{(11 - 2\zeta)(11 - 6\zeta)} - \frac{3000dC_2}{F_{\min}} \cdot T^{-\frac{1}{2}+\eta} - \frac{d}{F_{\min}} 2000(C_1 + LC_2)T^{-\frac{1}{2}+\eta} \\
\geq & \frac{1}{20}\zeta - \frac{5000d(C_1 + LC_2)}{F_{\min}} \cdot T^{-\frac{1}{2}+\eta} \\
= & \frac{1}{20}C \cdot T^{-\frac{1}{2}+\eta} - \frac{5000dC}{F_{\min} \cdot W} \cdot T^{-\frac{1}{2}+\eta} \\
\geq & \frac{1}{20}C \cdot T^{-\frac{1}{2}+\eta} - \frac{1}{200}C \cdot T^{-\frac{1}{2}+\eta} \\
\geq & \frac{1}{30}C \cdot T^{-\frac{1}{2}+\eta}.
\end{aligned} \tag{5.82}$$

Let $C_{ac} = \frac{1}{30} \cdot C$ and this lemma holds. ■

Define $\mathbb{P}_{P_0}$ and $\mathbb{P}_{P_\zeta}$ as the probabilistic distribution of customer's feedback at each round. In order to increase the information for distinguishing between two problem settings, we assume that a customer would always tell us whether or not she accept the price \$1, at each time $t = 1, 2, \ldots, T$. Therefore, both $\mathbb{P}_{P_0}$ and $\mathbb{P}_{P_\zeta}$ are binomial distributions $B(T, 0.5)$ and $B(T, 0.5 - \zeta)$. Here we present another lemma, the Lemma 5.8.16, that indicates the hardness of distinguishing the two settings.

Lemma 5.8.16. *Consider the $N \geq \frac{T}{2}$ rounds when we play a policy in $AC \vee BD$. For any algorithm ϕ, denote $\phi_t = 1$ if $\pi_t \in AC$ and $\phi_t = 0$ if $\pi_t \in BD$. Then we have:*

$$\max\{\mathbb{E}_{P_0}[\sum_{t=1}^{N} \phi_t], \mathbb{E}_{P_\zeta}[\sum_{t=1}^{N}(1 - \phi_t)]\} \geq \frac{1}{8} T \cdot \exp(-T^{2\eta}). \tag{5.83}$$

Proof of Lemma 5.8.16 . In fact, we have:

$$\begin{aligned}
\max\{\mathbb{E}_{P_0}[\sum_{t=1}^{N} \phi_t], \mathbb{E}_{P_\zeta}[\sum_{t=1}^{N}(1 - \phi_t)]\} &\geq \frac{\mathbb{E}_{P_0}[\sum_{t=1}^{N} \phi_t] + \mathbb{E}_{P_\zeta}[\sum_{t=1}^{N}(1 - \phi_t)]}{2} \\
&= N \cdot \frac{\mathbb{P}_{P_0}[\phi_t == 1] + \mathbb{P}_{P_\zeta}[\phi_t == 0]}{2} \\
&\geq \frac{T}{4} \cdot (\mathbb{P}_{P_0}[\phi_t == 1] + \mathbb{P}_{P_\zeta}[\phi_t == 0]) \\
&\geq \frac{T}{8} \cdot \exp(-N \cdot KL(\mathbb{P}_{P_0} || \mathbb{P}_{P_\zeta})) \\
&\geq \frac{T}{8} \cdot \exp(-N \cdot KL(Ber(0.5) || Ber(0.5 - \zeta))) \\
&\geq \frac{T}{8} \cdot \exp(-N \cdot 12\zeta^2) \\
&= \frac{T}{8} \cdot \exp(-N \cdot 12(C \cdot T^{-\frac{1}{2}+\eta})^2) \\
&\geq \frac{T}{8} \cdot \exp(-12C^2 T^{2\eta}) \\
&\geq \frac{T}{8} \cdot \exp(-T^{2\eta}).
\end{aligned} \tag{5.84}$$

Here the first line is for max $\geq$ average, the second is by definition of ϕ_t, the third line is for $N \geq \frac{T}{2}$, the fourth line is from Fano's Inequality that $\mathbb{P}_0[\phi == 1] + \mathbb{P}_1[\phi == 0] \geq$

$\frac{1}{2} \cdot \exp\{-N \cdot KL(\mathbb{P}_0||\mathbb{P}_1)\}$ for any distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, the fifth line is by definition of P_0 and P_ζ that they are only different in the customers' feedback satisfying $Ber(0.5)$ and $\Pr = 0.5 - \zeta$ for some actions, respectively, the sixth line is from Lemma 5.8.10, the seventh line is for $\zeta = C \cdot T^{-\frac{1}{2}+\eta}$, the eighth line is for $N \leq T$, and the last line is for $12C^2 \leq 1$. ■

With the two lemma above, we know that

- For any algorithm ϕ, we either run at least $\frac{T}{8} \cdot \exp(-T^{2\eta})$ rounds with some $\pi_t \in AC$ when the problem setting is P_ζ, or run at least $\frac{T}{8} \cdot \exp(-T^{2\eta})$ rounds with some $\pi_t \in BD$ when the problem setting is P_0, according to Lemma 5.8.16.
- For each round we mismatching the problem setting, we will suffer a $\min\{C_{ac}, C_{bd}\} \cdot T^{-\frac{1}{2}+\eta}$ unfairness, according to Lemma 5.8.15.

Given these two facts, denote $C_{\min} := \frac{1}{8}\min\{C_{ac}, C_{bd}\}$ and we at least have $C_{\min}T \cdot \exp(-T^{2\eta}) \cdot T^{-\frac{1}{2}+\eta}$ unfairness. For $x = \frac{C_x}{\log T}$ with any constant C_x, we let $\eta = \frac{3x}{2} = \frac{3C_x}{2\log T}$ and therefore $\eta > x$. As a result, we have

$$\begin{aligned} C_{\min}T \cdot \exp(-T^{2\eta}) \cdot T^{-\frac{1}{2}+\eta} &= C_{\min}\exp(-T^{2\eta} + \eta \log T)T^{\frac{1}{2}} \\ &= C_{\min}\exp(-T^{2\cdot\frac{C_x}{\log T}} + \frac{3C_x}{2\log T} \cdot \log T)T^{\frac{1}{2}} \\ &= C_{\min}\exp(-\exp(2 \cdot \frac{C_x}{\log T} \cdot \log T) + \frac{3C_x}{2})T^{\frac{1}{2}} \\ &= C_{\min}\exp(-\exp(2C_x) + \frac{3C_x}{2})T^{\frac{1}{2}} \end{aligned} \tag{5.85}$$

Let $C_u = \frac{C_{\min}\exp(-\exp(2C_x)+\frac{3C_x}{2})}{2}$, and then the result of the equation above contradicts with the suppose that the unfairness does not exceed $C_u \cdot T^{\frac{1}{2}}$. Therefore, we have proved the theorem.

■

Chapter 6

Pricing with Inventory-Censoring Effect

Consider an online pricing problem where the *potential* demand is linearly dependent on the price proposed by the seller at each time period $t = 1, 2, \ldots, T$. However, a fixed perishable inventory is imposed at every time t, *censoring* the potential demand if it exceeds the inventory level. To address this challenge, we introduce a novel and efficient pricing algorithm that achieves $\tilde{O}(\sqrt{T})$ regret for a linear noisy demand model. Furthermore, we show the optimality of our algorithm by deriving a matching $\Omega(\sqrt{T})$ lower bound. Our findings advance the state-of-the-art in online decision-making problems with censored feedback, offering a theoretically optimal solution that can be broadly applied.

6.1 Introduction

The problem of dynamic pricing, where the seller proposes and adjusts their prices over time, has been studied since the seminal work of Cournot [1897]. The crux to pricing is to balance the profit of sales per unit with the quantity of sales. Therefore, it is imperative for the seller to learn customers' demand as a function of price (commonly known as the *demand curve*) on the fly. However, the demand can often be obfuscated by the observed quantity of sales, especially when *censored* by *inventory* stockouts. Such instances severely impede the seller from learning the underlying demand distributions, thereby hindering our pursuit of the optimal price.

Existing literature has devoted considerable effort to the intersection of pricing and inventory decisions. Such works often consider scenarios with indirectly observable lost demands [Keskin et al., 2022], recoverable leftover demands Chen et al. [2019a], or controllable inventory level [Chen et al., 2023a]. However,these assumptions do not always align with the realities faced in various common business environments. To illustrate, we present two pertinent examples:

Example 6.1.1 (Theater Tickets). Suppose we are operators of a theater that has recently undertaken a series of diverse performances, and thus we are commencing ticket sales and determining ticket prices. If the ticket price is set too high, it may lead to insufficient attendance, adversely affecting revenue. Conversely, if the ticket price is too low, we stand to lose potential audience members who wish to attend the performance but are unable to purchase tickets due to high demand. On the one hand, we are unaware of the exact number of audience who attempt to purchase tickets but are unsuccessful. On the other hand, since these are varied performances, there is no guarantee that individuals who miss out on purchasing tickets for one show will opt to buy tickets for a subsequent one.

Additionally, the number of seats in the theater is fixed, precluding our ability to freely adjust the supply of seating.

Example 6.1.2 (Fruit Retails). Sweetsop (*Annona squamosa*, or so-called "sugar apple") is a tropical fruit that is particularly perishable: Ripe sweetsops may only be stored for 2 to 4 days [Crane et al., 2005]. Suppose we are proprietors of a fruit shop. We have entered into a long-term contract with a nearby farm to supply us with a fixed quantity of sweetsops every three days during the harvest season. Consequently, we are compelled to sell the entire stock from the previous delivery before the arrival of the next supply. Otherwise, the remaining sweetsops will blacken and rot. However, if we exhaust our inventory ahead of time, customers will turn to other fruit shops to make their purchases rather than waiting for our next restock.

Products in the two instances above have the following properties,

1. Inventory level is determined by pre-established objective factors, and is fixed for every individual time period.

2. Products are perishable and salable only within a single time period.

In this chapter, we study a dynamic pricing problem where the products possess these properties. The problem model is defined as follows. At each time $t = 1, 2, \ldots, T$, we firstly propose a price p_t, and then a price-dependent *potential demand* occurs as d_t. However, we might have no access to d_t as it is censored by a *fixed* inventory level γ_0. Instead, we observe a censored demand $D_t = \min\{\gamma_0, d_t\}$ and receive the revenue r_t as a reward at t. Our goal is to learn and approach the optimal price, thereby maximizing the cumulative revenue.

Dynamic pricing with inventory constraint. For $t = 1, 2, ..., T$:

1. The seller (we) receives γ_0 identical products.
2. The seller proposes a price $p_t \geq 0$.
3. The customers generate an invisible potential demand $d_t \geq 0$, dependent on p_t.
4. The market reveals an inventory-censored demand $D_t = \min\{\gamma_0, d_t\}$.
5. The seller gets a reward $r_t = p_t \cdot D_t$.
6. All unsold products perish before $t + 1$.

6.1.1 Summary of Contributions

We consider the problem setting displayed above and assume the potential demand $d_t = a - bp_t + N_t$ is *linear* and *noisy*. Here $a, b \in \mathbb{R}^+$ are fixed unknown parameters and N_t is an *unknown* i.i.d.[1] noise with zero mean. Under this premise, the key to deriving the optimal price is to accurately learn the expected reward function $r(p)$, which is equivalent to learning the linear parameters $[a, b]$ and the noise distribution. We are confronted by two principal challenges:

1. The absence of unbiased observations of the potential demand or its derivatives with respect to p, which prevents us from estimating $[a, b]$ directly.
2. The dependence of the optimal price on the noise distribution, which is assumed to be unknown and partially censored.

In this paper, we introduce an algorithm that employs innovative techniques to resolve the aforementioned challenges. On the one hand, we devise a pure-exploration phase that bypasses the censoring effect and obtains an unbiased estimator of $\frac{1}{b}$. This is founded

[1]Independently and identically distributed

on two insights: (1) While the inventory level γ_0 is fixed, we can still set a arbitrary *observation thresholds* at any $\gamma \leq \gamma_0$ for statistic purposes. (2) When Y is uniform distributed on a closed interval $[L, R]$ and an independent X lies in the same range, we have $\Pr[Y \geq X] = \frac{\mathbb{E}[X]-L}{R-L}$ according to the Law of Total Expectation, leading to an unbiased estimator of $\mathbb{E}[X]$ without observing any realized X. On the other hand, we design a searching method that relies on a biased estimator of the derivatives $r'(p)$, circumventing the need of learning the noise distribution and constructing an estimator of $r(p)$. Provided that we find a $\hat{p}$ such that $\hat{r}'(\hat{p})$ approximates 0, we may assert the near-optimality of $\hat{p}$ with high probability. These methods are not only pivotal to our study but also hold the potential for broad application in a variety of online decision-making scenarios with censored feedback.

Our algorithm attains a regret guarantee of $\tilde{O}(\sqrt{T}\log T)$, which is near-optimal as it matches the $\Omega(\sqrt{T})$ information-theoretic lower bound up to $O(\log T)$ factors. It is worth noting that this lower bound is applicable even in scenarios without inventory censoring. Hence, our findings suggest that the presence of inventory censoring does not substantially increase the hardness of pricing measured by regret.

6.1.2 Paper Structure

The rest of this paper is organized as follows. We discuss and compare with related works in Section 6.2, and then describe the problem setting in Section 6.3. Our primary technical contributions will be displayed in Section 6.4, consisting of algorithmic design, regret analysis and a lower bound proof. We further discuss the limitations and potential extensions of our methodologies in Section 6.5, followed by a brief conclusion in Section 6.6.

6.2 Related Works

There exists a large volume of literature related to the problem we study in this chapter. Here we discuss them in the following categories.

Data-driven dynamic pricing Dynamic pricing for identical products is a well-established research area, starting with Kleinberg and Leighton [2003] and continuing through seminal works by Besbes and Zeevi [2009], Broder and Rusmevichientong [2012], Wang et al. [2014, 2021b]. The standard approach involves learning a demand curve from price-sensitive demand arriving in real-time, aiming to approximate the optimal price. Kleinberg and Leighton [2003] provided algorithms with regret bounds of $O(T^{\frac{2}{3}})$ and $O(\sqrt{T})$ for arbitrary and infinitely smooth demand curves, respectively. Wang et al. [2021b] refined this further, offering an $O(T^{\frac{k+1}{2k+1}})$ regret for k-times continuously differentiable demand curves. This line of inquiry is also intricately linked to the multi-armed bandit problems [Lai and Robbins, 1985, Auer et al., 2002b] and continuum-armed bandits [Kleinberg, 2004], where each action taken reveals a reward without insight into the foregone rewards of other actions.

Contextual Pricing A surge of research has delved into *feature-based* dynamic pricing [Cohen et al., 2020] or *pricing with contexts/covariates* [Amin et al., 2014, Miao et al., 2019, Liu et al., 2021]. These works considered situations where each pricing period was preceded by a context, influencing both the demand curve and noise distribution. Specifically, Cohen et al. [2020], Javanmard and Nazerzadeh [2019], Xu and Wang [2021] explored a linear valuation framework with known distribution noise, leading to binary customer demand outcomes based on price comparisons to their valuations. Expanding on this, Golrezaei et al. [2019], Fan et al. [2021], Luo et al. [2021] examined similar

models but with unknown noise distributions. In another vein, Ban and Keskin [2021] and Wang et al. [2021a] investigated personalized pricing where demand was modeled as a generalized linear function sensitive to contextual price elasticity. Many of these works on valuation-based contextual pricing also assume a censored demand: The seller only observes a binary feedback determined by a comparison of price with valuation, instead of observing the valuation directly. However, it was important to differentiate between the linear (potential) demand model we assumed and their linear valuation models, and there exists no inclusive relationship.

Pricing with inventory concerns Dynamic pricing problems began to incorporate inventory constraints with the work of Besbes and Zeevi [2009], which assumed a fixed initial stock available at the start of the selling period. They introduced near-optimal algorithms for both parametric and non-parametric demand distributions, operating under the assumption that the inventory was non-replenishable and non-perishable. Wang et al. [2014] adopted a comparable framework but allowed for customer arrivals to follow a Poisson process. In these earlier works, the actual demand is fully disclosed until the inventory is depleted. Subsequent research allowed inventory replenishment, with the seller's decisions encompassing both pricing and restocking at each time interval. Chen et al. [2019a] proposed a demand model subject to additive/multiplicative noise and developed a policy that achieved $O(\sqrt{T})$ regret. More recent studies, such as those by Chen et al. [2020], Keskin et al. [2022] explored the dynamic pricing of perishable goods where unsold inventory would expire. However, the uncensored demand is observable as assumed in both works. Specifically, Chen et al. [2020] allowed recouping backlogged demand, albeit at a cost, and introduced an algorithm with optimal regret. Keskin et al. [2022] focused on cases where both fulfilled demand and lost sales were observable.

Chen et al. [2021a] and their subsequent work, Chen et al. [2023a], are the related works most closely aligned with our problem settings, where the demand is *censored* by the inventory level and any leftover inventory or lost sales disappear at the end of each period. With the assumption of concave reward functions and the restriction of at most m price changes, Chen et al. [2021a] proposed MLE-based algorithms that attain a regret of $\tilde{O}(T^{\frac{1}{m+1}})$ in the well-separated case and $\tilde{O}(T^{\frac{1}{2}+\epsilon})$ for some $\epsilon = o(1)$ as $T \to \infty$ in the general case. Under similar assumptions (except infinite-order smoothness), Chen et al. [2023a] designed a trinary-search algorithm based on a reward-difference estimator. With this algorithm, they not only enhanced the prior result for concave reward functions to $\tilde{O}(\sqrt{T})$, but also obtained a general $\tilde{O}(T^{2/3})$ regret for non-concave reward functions. Our problem model mirrors theirs in the sense that we lack access to both the uncensored demand and its gradient. However, in their models, sellers have the flexibility to determine inventory levels, unlike in our case where the inventory is predetermined by nature. An adversarial selection of the inventory level could impede us from learning the optimal price in the worst-case scenarios.It is important to note that neither our results nor theirs are mutually inclusive.

6.3 Problem Setup

We study the following online non-contextual dynamic pricing problem. At each time step $t = 1, 2, \ldots, T$, the seller (we) proposes a price p_t and receives an inventory-censored demand $D_t = \min\{d_t, \gamma_0\}$. Here $d_t = a - b \cdot p_t + N_t$ is a potential linear demand, where a, b are unknown parameters, N_t is a demand noise, and γ_0 is a *fixed* and *known* inventory level at every time period t.

6.3.1 Definitions

Here we define some key quantities that are involved in the algorithm design and analysis. Firstly, there are different types of demand functions.

Definition 6.3.1 (Demand functions)**.** Denote $d_t(p) := a - bp + N_t$ as the potential demand function, and $d(p) := a - bp$ as the expected potential demand function. Denote $D_t(p) := \min\{\gamma_0, d_t(p)\}$ as the *censored demand* function.

Moreover, we define some distributional functions of the demand noise N_t.

Definition 6.3.2 (Distributional functions)**.** For N_t as the demand noise, denote $F(x)$ as its *cumulative distribution function* (CDF) and $f(x)$ as its *probabilistic density function* (PDF), $x \in \mathbb{R}$. Also, denote

$$G(x) := \int_{-\infty}^{x} F(\omega)d\omega, x \in \mathbb{R} \tag{6.1}$$

as the *integrated CDF*.

We will make more assumptions on the noise distribution later. Finally, we may define the revenue function and the regret.

Definition 6.3.3 (Revenue function)**.** Denote $r(p)$ as the expected revenue function of price p, satisfying

$$r(p) := p \cdot \mathbb{E}[D_t | p_t = p], p \geq 0. \tag{6.2}$$

Definition 6.3.4 (Regret)**.** Denote

$$Regret := \sum_{t=1}^{T} r(p^*) - r(p_t) \tag{6.3}$$

as the *cumulative regret* (or *regret*) of the price sequence $\{p_t\}_{t=1}^{T}$.

6.3.2 Assumptions

We make reasonable and mild assumptions as follows.

Assumption 6.3.5 (Boundedness)**.** Assume $0 < a \leq a_{\max}$, $0 < b_{\min} \leq b \leq b_{\max}$, $N_t \in [-c, c]$ for some *known finite* constants $a_{\max}, b_{\min}, b_{\max}, c > 0$. Also, we restrict the proposed price p_t at any $t = 1, 2, \ldots, T$ satisfies $0 \leq p_t \leq p_{\max}$ with a *known finite* constant $p_{\max} > 0$.

Assumption 6.3.6 (Noise Distribution)**.** Each N_t is drawn from an *unknown* independent and identical distribution (i.i.d.) with CDF $F(x)$ and PDF $f(x), x \in \mathbb{R}$, satisfying $\mathbb{E}[N_t] = 0$ and $f(x) \in [0, f_{\max}]$ for $\forall x \in [-c, c]$ with some *known finite* constant $f_{\max} > 0$.

Assumption 6.3.7 (Inequalities of Parameters)**.** All parameters and constants satisfy the following inequalities:

1. $a - c > \gamma_0$. Demands at $p_t = 0$ must be censored.
2. $\gamma_0 > 2c$. Inventory level exceeds noise support.
3. $a - bp_{\max} - c > 0$. Demands must be positive.
4. $a_{\max} - b_{\min} p_{\max} + c < \gamma_0$. Demands at $p_t = p_{\max}$ must be uncensored.

6.4 Main Results

In this section, we introduce our pricing algorithm and analyze its regret guarantee. Furthermore, we prove a matching lower bound on a simplified problem setting, demonstrating the information-theoretic hardness of this problem and the optimality of our algorithm.

6.4.1 Algorithm

We propose an Algorithm 8 that addresses the inventory-censored pricing problem. It is inspired by a *binary search* strategy with respect to the *derivative* of revenue. We show that the expected revenue function $r(p)$ is unimodal and smooth on price p. Consequently, there exists a unique optimal price p^* satisfying $r'(p^*) = 0$. We deduce that for any price p_1 with $r'(p_1) > 0$, p_1 is less than p^*, while any p_2 with $r'(p_2) < 0$ exceeds p^*.

However, constructing an unbiased estimator of the derivative $r'(p)$ is notably challenging due to the inventory-censored feedback. Our algorithm circumvents this by utilizing a slightly biased estimator: We initially estimate the parameters a and b through a pure-exploration phase with adequate time horizon, and then incorporate these plug-in estimators as basis of the derivative estimations in the following epochs.

Algorithm Design

Our algorithm has three phases:

1. **Exploring**: In Epoch 0, the seller proposes uniformly random prices in the range of $[0, p_{\max}]$ and obtains $\hat{a}$ and $\hat{b}$ as estimators of a and b, respectively. By the end of Epoch 0, we assign $[u_1, v_1] = [0, p_{\max}]$ as the initial range of searching.

2. **Searching**: In every Epoch $k = 1, 2, \ldots$, we repeatedly propose the price $p_k = \frac{u_k = v_k}{2}$ for $O(\sqrt{T})$ times. With the demand feedback, we calculate $\hat{r}'_k$ as an estimator of $r'(p_k)$. If $\hat{r}'_k$ is significantly positive, then we narrow our search to $[u_{k+1}, v_{k+1}] = [u_k, p_k]$; If $\hat{r}'_k$ is significantly negative, then we adapt to $[u_{k+1}, v_{k+1}] = [p_k, v_k]$. Repeat those routines until $|\hat{r}'_k|$ becomes negligible (or until $t = T$ if arrived earlier).

Algorithm 8 Pricing Algorithm under a Fixed Inventory Constraint

Input: Inventory constraint γ_0, time horizon T, constants $p_{\max}, a_{\max}, b_{\min}, b_{\max}, c, f_{\max}$, epoch length τ, parameters C_K.
Epoch 0 Choose $\gamma_1, \gamma_2, \gamma_3$ such that $a_{\max} - b_{\min} p_{\max} + c < \gamma_1 < \gamma_2 < \gamma_3 < \gamma_0$.
for $t = 1, 2, \ldots, \tau$ **do**
  Sample price $p_{0,t}$ from uniform distribution $U[0, p_{\max}]$ and propose it.
  Receive demand $D_{0,t}$ and signals of $e_{i,t} = \mathbb{1}[D_{0,t} \geq \gamma_i], i = 1, 2, 3$.
end for
Let

$$\begin{aligned} \hat{b} =& \frac{\gamma_2 - \gamma_1}{p_{\max}} \cdot \frac{1}{\frac{1}{\tau}\sum_{t=1}^{\tau} e_{1,t} - e_{2,t}} \\ \hat{a} =& p_{\max} \cdot \hat{b} \cdot (\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t}) + \gamma_3 \end{aligned} \tag{6.4}$$

Assign $u_1 = 0, v_1 = p_{\max}$.
for Epoch $k = 1, 2, \ldots$ **do**
  Let $p_k = \frac{u_1 + v_1}{2}$
  for $t = 1, 2, \ldots, \tau$ **do**
    Propose price $p_{k,t} = p_k$.
    Receive demand $D_{k,t}$ and signal $\mathbb{1}_{k,t} = \mathbb{1}[D_{k,t} < \gamma_0]$.
  end for
  Denote

$$\hat{r}'_k := \frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t} - p_k \cdot \hat{b} \cdot \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{1}_{k,t}. \tag{6.5}$$

  if $\hat{r}'_k > 2C_K \cdot \frac{1}{\sqrt{\tau}}$ **then**
    Let $[u_{k+1}, v_{k+1}] \leftarrow [u_k, p_k]$.
  else if $\hat{r}'_k < -2C_K \cdot \frac{1}{\sqrt{\tau}}$ **then**
    Let $[u_{k+1}, v_{k+1}] \leftarrow [p_k, v_{k+1}]$.
  else
    Let $\hat{p}^* := p_k$ and $K := k$.
    Break.
  end if
end for
Keep proposing $\hat{p}^*$ at each time step t until $t = T$.

3. **Exploiting**: Should time permit (i.e., $t < T$), we continue to offer the most recent price p_k until the period concludes at $t = T$.

Algorithm 8 exhibits several advantageous properties. It consumes *linear* time complexity and requires only *constant* extra space. Additionally, it is suitable for processing streaming data as the constructions of $\hat{a}, \hat{b}, \hat{r}'_k$ are updated *incrementally* with each new observation (including $e_{i,t}, D_k, t, \mathbb{1}_{k,t}$) without the need of revisiting any historical data. A potential risk of computation might arise on the calculation of $\hat{b}$, where $\sum_{t=1}^{\tau} e_{1,t} - e_{2,t}$ can be 0 with a small but nonzero probability. Although this event does not undermine the high-probability regret guarantee, it might still be harmful to the computational system for numerical experiments. To mitigate this incident in practice, we may either extend Epoch 0 until one non-zero $e_{1,t} - e_{2,t} = 1$ is observed, or restart Epoch 0 at $t = \tau$.

Highlighted Techniques: Uniform Exploration

We incorporate a uniform-exploration phase for estimating a and b in our algorithm, bypassing the obstacle brought by demand censoring. This approach is supported by the following insight: When Y is a uniformly distributed random variable within a closed interval $[L, R]$, and X is another random variable, independent to Y and also distributed within $[L, R]$, we have:

$$\mathbb{E}[\mathbb{1}[Y \geq X]] = \Pr[Y \geq X] = \mathbb{E}[\Pr[Y \geq X|X]] = \mathbb{E}[\frac{X - L}{R - L}] = \frac{\mathbb{E}[X] - L}{R - L}. \tag{6.6}$$

Here the second step uses the Law of Total Expectation. Eq. (6.6) indicates that we can derive an unbiased estimator of $\mathbb{E}[X]$ through $\mathbb{1}[Y \geq X]$ even in the absense of any direct observation of X. Looking back to our algorithm, when $p_t \sim U[0, p_{\max}]$, we have

$$\mathbb{E}[e_{i,t}] = \mathbb{E}[\mathbb{1}[a - bp_t + N_t \geq \gamma_i]] = \mathbb{E}[\mathbb{E}[N_t \geq \gamma_i - a + bp_t|N_t]] = \mathbb{E}[\frac{N_t - \gamma_i + a}{bp_{\max}}] = \frac{a - \gamma_i}{bp_{\max}}. \tag{6.7}$$

The last equality comes from $\mathbb{E}[N_t] = 0$. By deploying different γ_i at $i = 1, 2, 3$, we can estimate a and b through the observations of $e_{i,t}$, effectively circumventing the censoring effect. A similar technique has been utilized by Fan et al. [2021] to construct an unbiased estimator of the *valuations* instead of the demands as we concern. However, their application of uniform exploration might be sub-optimal as they adopt an *exploration-then-exploitation* design in each epoch. On the contrary, our algorithm uses this uniform exploration merely as a *trigger* of further learning. Our tight regret bound indicates that uniform exploration can still contribute to an optimal algorithm for a broad range of online learning instances.

6.4.2 Regret Analysis

In this section, we analyze the cumulative regret of our algorithm and show a $\tilde{O}(\sqrt{T})$ regret guarantee with high probability. We leave all of the proof details to Section 6.7, and here we only display proof sketches. We firstly present our main theorem.

Theorem 6.4.1 (Regret)**.** *In Algorithm 8, let* $\gamma_1 = \frac{a_{\max} - b_{\min} p_{\max} + c + \gamma_0}{2}, \gamma_2 = \frac{\gamma_1 + \gamma_0}{2}, \gamma_3 = \frac{\gamma_2 + \gamma_0}{2}$, $\tau = \sqrt{T}$, $C_K = (\gamma_0 + 2b_{\max} p_{\max} + \frac{2b_{\max}^2 p_{\max}}{\gamma_2 - \gamma_1}) \cdot \sqrt{\frac{1}{2} \cdot \log \frac{2}{\eta\delta}}$ *where* $\eta = \frac{1}{4\sqrt{T}}$*, it holds with* $\Pr > 1 - \delta$ *that*

$$Regret := \sum_{t=1}^{T} r(p^*) - r(p_t) = O(\sqrt{T} \log \frac{T}{\delta}) \tag{6.8}$$

for sufficiently large T*.*

Before getting into the proof of Theorem 6.4.1, we propose a lemma regarding the first- and second-order derivatives of the revenue function $r(p)$.

Lemma 6.4.2 (revenue function)**.** *For the expected revenue function* $r(p)$ *in Eq.* (6.2)*,*

we have:

$$\begin{aligned} r(p) &= p(\gamma_0 - c + G(c) - G(\gamma_0 - a + bp)) \\ r'(p) &= \gamma_0 - c + G(c) - G(\gamma_0 - a + bp) - bp \cdot F(\gamma_0 - a + bp) \\ r''(p) &= -2b \cdot F(\gamma_0 - a + bp) - b^2 p \cdot f(\gamma_0 - a + bp). \end{aligned} \tag{6.9}$$

Please refer to Section 6.7.1 for a detailed proof of Lemma 6.4.2. With this lemma, we further notice the following properties of $r(p)$.

Lemma 6.4.3. *$r(p)$ has the following properties:*

1. *There exists $p^* \in [0, \frac{a}{b}]$ such that $r'(p) = 0$, and $r(p)$ monotonically increase in $[0, p^*]$ and decrease in $[p^*, \frac{a}{b}]$. Notice that $\frac{a}{b} > p_{\max}$ according to Assumption 6.3.7.*

2. *For any $p \in [0, p_{\max}]$, denote $C_s := 2b_{\max} + b_{\max}^2 p_{\max} f_{\max}$, and it holds that*

$$-C_s \leq r''(p) \leq 0. \tag{6.10}$$

3. *There exists finite constants $\Delta > 0, C_\Delta > 0$ such that $r''(p) \leq -C_\Delta, \forall p \in [p^* - \Delta, p^* + \Delta]$.*

Please refer to Section 6.7.2 for a detailed proof of Lemma 6.4.3, and here we provide a proof sketch.

Proof sketch of Lemma 6.4.3. 1. From Lemma 6.4.2, we have $r''(p) \leq 0$, which indicates $r'(p)$ is non-increasing on $[0, \frac{a}{b}]$. Also, notice that $r'(0) > 0$ and $r'(\frac{a}{b}) < 0$. Therefore, $\exists p^* \in (0, \frac{a}{b})$ such that $r'(p^*) = 0$. Then we show p^* is unique with a proof by contradiction. Given this, we know that $r'(p) > 0$ for $p \in (0, p^*)$ and $r'(p) < 0$ for $p \in (p^*, \frac{a}{b})$.

2. Given $0 \leq F(x) \leq 1$ and $0 \leq f(x) \leq f_{\max}, \forall x \in \mathbb{R}$, we have $r''(p) \geq -2b_{\max} - b_{\max}^2 p_{\max} f_{\max} = -C_s$.

3. Let $C_\Delta = bF(\gamma_0 - a + bp^*)$ and $\Delta = \frac{F(\gamma_0 - a + bp^*)}{2f_{\max} b_{\max}} > 0$, and the inequality holds.

■

Now we may continue the proof of Theorem 6.4.1. Similarly, we provide a proof sketch here and defer the details to Section 6.7.4.

Proof sketch of Theorem 6.4.1. We firstly bound the estimation error of $\hat{a}$ and $\hat{b}$ respectively. Notice that for any $\gamma \in [a_{\max} - b_{\min} p_{\max} + c, \gamma_0]$, we have $\mathbb{E}[\mathbb{1}[D_{0,t} \geq \gamma]] = \frac{a-\gamma}{bp_{\max}}$. Then we have $\mathbb{E}[e_{1,t} - e_{2,t}] = \frac{\gamma_2 - \gamma_1}{bp_{\max}}$ and $\mathbb{E}[e_{3,t}] = \frac{a-\gamma_3}{bp_{\max}}$. According to Hoeffing's Inequality, it holds with $\Pr > 1 - 2\eta\delta$ that

$$\begin{aligned} |\frac{1}{\tau}\sum_{t=1}^{\tau} e_{1,t} - e_{2,t} - \frac{\gamma_2 - \gamma_1}{bp_{\max}}| \leq & \sqrt{\frac{1}{2} \cdot \log\frac{2}{\eta\delta}} \cdot \frac{1}{\sqrt{\tau}} \\ |\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t} - \frac{a - \gamma_3}{bp_{\max}}| \leq & \sqrt{\frac{1}{2} \cdot \log\frac{2}{\eta\delta}} \cdot \frac{1}{\sqrt{\tau}}. \end{aligned} \tag{6.11}$$

With Eq. (6.11) we may bound the estimation errors of $\hat{b}$ and $\hat{a}$ by

$$\begin{aligned} |\hat{b} - b| \leq & C_b \cdot \frac{1}{\sqrt{\tau}}, \\ |\hat{a} - a| \leq & C_a \cdot \frac{1}{\sqrt{\tau}}, \end{aligned} \tag{6.12}$$

where $C_a = p_{\max}(b_{\max} + \frac{(a_{\max} + \gamma_3) \cdot b_{\max}^2}{b_{\min}(\gamma_2 - \gamma_1)} \cdot \sqrt{2\log\frac{1}{\eta\delta}}$ and $C_b = \frac{b_{\max}^2 p_{\max}}{\gamma_2 - \gamma_1} \cdot \sqrt{2\log\frac{2}{\eta\delta}}$. Now, we calculate the cumulative regret in Epoch $k = 1, 2, \ldots$. According to Lemma 6.4.3, we know that

$$r(p^*) - r(p) \leq C_s \cdot (p^* - p)^2. \tag{6.13}$$

Then we propose a lemma implying the estimation error of the derivatives.

Lemma 6.4.4 (Estimation error of $\hat{r}'_k$)**.** *Denote* $C_K := (\gamma_0 + 2b_{\max}p_{\max} + \frac{2b_{\max}^2 p_{\max}}{\gamma_2 - \gamma_1}) \cdot \sqrt{\frac{1}{2} \cdot \log \frac{2}{\eta\delta}}$. *With* $\Pr > 1 - 2K\eta\delta$*, it holds that*

$$|\hat{r}'_k - r'(p_k)| \leq C_K \cdot \frac{1}{\sqrt{\tau}}, k = 1, 2, \ldots, K \leq \frac{T}{\tau}. \tag{6.14}$$

We defer the proof of Lemma 6.4.4 to Section 6.7.3, and here we provide a proof sketch.

Proof sketch of Lemma 6.4.4. From Eq. (6.9), we know that $r'(p_t)$ has two components:

1. $\gamma_0 - c + G(c) - G(\gamma_0 - a + bp_t)$, which equals $\mathbb{E}[D_t]$. According to Hoeffding's Inequality, we have $|\gamma_0 - c + G(c) - G(\gamma_0 - a + bp_k) - \frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t}| = \tilde{O}(\frac{1}{\sqrt{\tau}})$.

2. $bp_t \cdot F(\gamma_0 - a + bp_t)$, which equals $bp_t \cdot \mathbb{E}[\mathbb{1}_{k_t}]$. According to Hoeffding's Inequality, we have $|F(\gamma_0 - a + bp_k) - \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{1}_{k,t}| = \tilde{O}(\frac{1}{\sqrt{\tau}})$.

Since we already have $|\hat{b} - b| = O(\frac{1}{\sqrt{\tau}})$, it holds that

$$\begin{aligned} |\hat{r}'_k - r'(p_k)| = |(&\frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t} - p_k \cdot \hat{b} \cdot \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{1}_{k,t}) \\ &- (\gamma_0 - c + G(c) - G(\gamma_0 - a + bp_k) - bp_k \cdot F(\gamma_0 - a + bp_k))| \\ &= \tilde{O}(\frac{1}{\sqrt{\tau}}). \end{aligned} \tag{6.15}$$

Here $\tilde{O}(\cdot)$ omits the dependence on $\log(T)$ and $\log\delta$. ∎

From Lemma 6.4.4, we immediately get to the following corollary.

Corollary 6.4.5 (derivative indicator)**.** *With probability* $\Pr > 1 - 2K\eta\delta$*,* $\hat{r}'_k > 2C_K \cdot \frac{1}{\sqrt{\tau}}$ *is sufficient for* $p_k \leq p^*$*, and* $\hat{r}'_k < -2C_K \cdot \frac{1}{\sqrt{\tau}}$ *is sufficient for* $p_k \geq p^*$*.*

Corollary 6.4.5 indicates that our binary searching is correct with high probability. Therefore, with $\Pr > 1 - 2K \cdot \eta\delta$ it holds that $p^* \in [u_k, v_k], \forall k = 1, 2, \ldots$. Since

$u_k - v_k = \frac{p_{\max}}{2^{k-1}}$ according to the binary search principle, we have:

$$Regret_{search} = \sum_{k=1}^{K}\sum_{t=1}^{\tau} r(p^*) - r(p_k) \leq \sum_{k=1}^{K} \tau C_s (p^* - p_k)^2 \leq \sum_{k=1}^{K} \tau C_s (\frac{p_{\max}}{2^{k-1}})^2 \leq \frac{4}{3} C_s p_{\max}^2 \tau. \tag{6.16}$$

Now we upper bound the distance between p_K and p^* if $\hat{r}'_k \in [-2C_K \cdot \frac{1}{\sqrt{\tau}}, +2C_K \cdot \frac{1}{\sqrt{\tau}}]$. According to Lemma 6.4.4, in this case we have

$$|r'(p_k)| \leq |\hat{r}'_k - r'(p_k)| + |\hat{r}'_k| \leq C_K \frac{1}{\sqrt{\tau}} + 2C_K \frac{1}{\sqrt{\tau}} = 3C_K \frac{1}{\sqrt{\tau}}. \tag{6.17}$$

According to Lemma 6.4.3 Property 3, when $p \in [p^* - \Delta, p^* + \Delta]$, we have $|r'(p)| = |r'(p) - r'(p^*)| \geq C_\Delta |p - p^*|$. Also, since $r''(p) \leq 0$, indicating a monotonic decrease of $r'(p)$, we know that

$$\begin{aligned} r'(p) \geq r'(p^* - \Delta) \geq & C_\Delta |(p^* - \Delta) - p^*| = C_\Delta \cdot \Delta, \forall p \in [0, p^* - \Delta] \\ r'(p) \leq r'(p^* + \Delta) \leq & - C_\Delta |p^* - (p^* + \Delta)| = -C_\Delta \cdot \Delta, \forall p \in [p^* + \Delta, p_{\max}]. \end{aligned} \tag{6.18}$$

From Eq. (6.18), we know that $r'(p) \geq C_\Delta \cdot \Delta$ for $|p - p^*| \geq \Delta$. Therefore, if $p \in [0, p_{\max}]$ such that $|r'(p)| < C_\Delta \cdot \Delta$, then $p \in [p^* - \Delta, p^* + \Delta]$. According to Eq. (6.17), we have $|r'(p_k)| \leq \frac{C_\Delta \cdot \Delta}{2} < C_\Delta \cdot \Delta$ if $\hat{r}'_k \in [-2C_K \cdot \frac{1}{\sqrt{\tau}}, +2C_K \cdot \frac{1}{\sqrt{\tau}}]$. Since we let $\hat{p}^* = p_k$ when $\hat{r}'_k \in [-2C_K \cdot \frac{1}{\sqrt{\tau}}, +2C_K \cdot \frac{1}{\sqrt{\tau}}]$, we know that $\hat{p}^* \in [p^* - \Delta, p^* + \Delta]$ and therefore

$$|\hat{p}^* - p^*| \leq \frac{1}{C_\Delta} \cdot |r'(\hat{p}^*)| \leq \frac{3C_K \cdot \frac{1}{\sqrt{\tau}}}{C_\Delta} = \frac{3C_K}{C_\Delta \sqrt{\tau}}. \tag{6.19}$$

By the time we determine $\hat{p}^*$, we have already proposed $(K+1)\tau$ prices. As a result, there are still $(T - (K+1)\tau)$ time steps left, where we keep proposing $\hat{p}^*$. According to Eq. (6.13) and Eq. (6.19), the cumulative regret of this period should be:

$$Regret_{exploiting} = \sum_{t=1}^{T-(K+1)\tau} r(p^*) - r(\hat{p}^*) \leq \sum_{t=1}^{T} C_s (p^* - \hat{p}^*)^2 \leq T C_s \frac{9C_K^2}{C_\Delta^2} \frac{1}{\tau} = \frac{9C_s C_K^2}{C_\Delta^2} \cdot \frac{T}{\tau}. \tag{6.20}$$

Finally, the regret of Epoch 0 cannot exceed $\gamma_0 p_{\max} \cdot \tau$. Combining with Eq. (6.16) and Eq. (6.20), we may bound the total regret with $\Pr > 1 - (2K+2)\eta\delta$ as

$$\begin{aligned} Regret =& Regret_{exploring} + Regret_{searching} + Regret_{exploiting} \\ \leq& \gamma_0 p_{\max}\tau + \frac{4}{3} C_s p_{\max}^2 \cdot \tau + \frac{9 C_s C_K^2}{C_\Delta^2} \cdot \frac{T}{\tau} \end{aligned} \tag{6.21}$$

.

Plug in $\tau = \sqrt{T}$ and $\eta = \frac{1}{4\sqrt{T}} = \frac{\tau}{4T} \leq \frac{1}{2K+2}$, and we get $Regret = O(\sqrt{T}\log\frac{T}{\delta})$ with Probability $\Pr \geq 1 - \frac{2K+2}{4K} \cdot \delta \geq 1 - \delta$. ■

6.4.3 Lower Bound

"The inventory-censoring effect complicates the estimation of demand curves and the determination of the optimal price. Also, the demand-censored pricing problem includes a subproblem where the inventory is sufficient and no censoring actually happens, suggesting a comparison of their difficulties. However, it is not necessary that demand censoring leads to a substantial higher regret rate. To demonstrate this, we prove an $\Omega(\sqrt{T})$ regret lower bound for pricing with uncensored demand. Based on this, we claim the pricing problem with censored demands is as hard as that without it, measured by the worst-case minimax regret. Furthermore, the matching of upper-and-lower bounds also affirms the optimality of our algorithm (up to $O(\log T)$ factors).

Theorem 6.4.6 (Lower Bound)**.** *Assume the realized demand $D_t = a - bp_t + N_t$, where (a, b) are fixed unknown parameters, p_t is the price and $N_t \sim_{i.i.d.} \mathbb{D}_N$ is a zero-mean noise at each time period $t = 1, 2, \ldots, T$. Denote $\mathcal{H}_\tau := \{(p_t, D_t)\}_{t=1}^{\tau}$ as the historical prices and demands before time periods τ, and denote $\mathcal{A} := (\pi_1, \pi_2, \ldots, \pi_T)$ (where $\pi_t(\mathcal{H}_{t-1}) = p_t$ is a pricing policy) as an algorithm. There exists a constant C_{LB} such that for any algorithm*

$\mathcal{A}$, there exists a problem setting $((a,b),\mathbb{D}_N)$ such that

$$\inf_{\mathcal{A}} \max_{(a,b),\mathbb{D}_N} \mathbb{E}[\sum_{t=1}^{T} \max_{p^*} r(p^*) - r(\mathcal{A}(\mathcal{H}_{t-1}))] \geq C_{LB} \cdot \sqrt{T}. \tag{6.22}$$

The key to proving this lower bound lies in a construction of two similar problem instances, whose demand curves intersect at the optimal price of one instance. In this way, a price far away from the intersection could be very sub-optimal and cause a large regret, but a price very close to the intersection might provide limited information for distinguishing the two instances and lead to a large regret as well. We defer the proof details to Section 6.7.5. A similar high-level idea of proving this lower bound can be found in the work of Broder and Rusmevichientong [2012].

6.5 Discussions

Here we discuss the limitations, potential extensions and impacts of our work.

6.5.1 Generalization to Unbounded Noises

We assume the noise is bounded in a constant-width range. This assumption streamlines the pure-exploration phase and facilitates the estimation of the parameters b and a. While our methods and results can be extended to unbounded $O(\frac{1}{\log T})$-subGaussian noises by simple truncation, challenges remain for handling generic unbounded noises. Moreover, the problem can be more sophisticated with *dual-censoring*, both from above by inventory–as we have discussed– and from below by 0, especially when considering unbounded noises.

6.5.2 Extensions to Adversarial Series of Inventory

Imagine a scenario where the inventory level at each time t is set to a variable γ_t by adversary, instead of being a fixed γ_0. Our uniform-exploration is still applicable to estimate a and b. However, the searching and exploiting phases might struggle in this setting, since the derivatives estimate at one time period is not valid accross all periods (unless further imposing i.i.d. assumptions on γ_t). In the face of adversarial $\{\gamma_t\}_{t=1}^{\top}$ sequences, we conjecture an Online Gradient Descent (OGD) method [Biehl and Schwarze, 1995] with epoch-based updates might serve as a possible alternative.

6.5.3 Extensions to Contextual Pricing

In this chapter, we assume a and b are static, which may not hold in many real scenarios. Example 6.1.1 serves as a good instance, showcasing significant fluctuations in popularity across different performances. A reasonable extension of our work would be modeling a and b as *contextual* parameters. Similar modelings have been adopt by Wang et al. [2021a] and Ban and Keskin [2021] in the realm of personalized pricing research.

6.5.4 Societal Impacts

Our research primarily addresses a non-contextual pricing model that does not incorporate personal or group-specific data, thereby adhering to conventional fairness standards relating to temporal, group, demand, and utility discrepancies as outlined by Cohen et al. [2022] and Chen et al. [2023b]. However, it is crucial to remain vigilant about the possible extension of our methodologies to diverse customer groups, which may exhibit different market noise distributions. Such application could result in varying *fulfillment rate* across groups, i.e. the proportion of satisfied demand might be different at the optimal price in

each group. This raises concern regarding unfairness in fulfillment rate [Spiliotopoulou and Conte, 2022] particularly on product of significant social and individual importance.

6.6 Conclusions

In this paper, we studied the online dynamic pricing problem with a fixed inventory constraint imposed on each time period. We introduced an exploring-then-searching algorithm that is capable of deducing the optimal price from censored demands. Our algorithm enjoys a regret guarantee of $\tilde{O}(\sqrt{T}\log T)$, which is (near) optimal as it matches the $\Omega(\sqrt{T})$ lower bound we proved. To the best of our knowledge, we are the first to address this fixed-inventory pricing problem, and our results indicate that the associated type of demand censoring does not substantially increase the hardness of pricing in terms of minimax regret.

6.7 Proofs

6.7.1 Proof of Lemma 6.4.2

Proof. For $r(p)$, we have

$$
\begin{aligned}
r(p) &= \mathbb{E}[p_t \cdot D_t | p_t = p] \\
&= p \cdot \mathbb{E}[\min\{\gamma_0, a - bp_t + N_t\} | p_t = p] \\
&= p \cdot \mathbb{E}[\mathbb{1}[a - bp + N_t \le \gamma_0] \cdot (a - bp + N_t) + \mathbb{1}[a - bp + N_t > \gamma_0] \cdot \gamma_0] \\
&= p \cdot \mathbb{E}[\mathbb{1}[N_t \le \gamma_0 - a + bp] \cdot (a - bp + N_t) + \mathbb{1}[N_t > \gamma_0 - a + bp] \cdot \gamma_0] \\
&= p \left(\int_{-c}^{\gamma_0 - a + bp} (a - bp + x) f(x) dx + \int_{\gamma_0 - a + bp}^{c} \gamma_0 f(x) dx \right) \\
&= p \left(\int_{-c}^{c} (a - bp + x) f(x) dx + \int_{\gamma_0 - a + bp}^{c} (\gamma_0 - (a - bp + x)) f(x) dx \right) \\
&= p \Bigg((a - bp) \cdot \int_{-c}^{c} f(x) dx + \int_{-c}^{c} x f(x) dx \\
&\qquad + (\gamma_0 - a + bp) \int_{\gamma_0 - a + bp}^{c} f(x) dx - \int_{\gamma_0 - a + bp}^{c} x f(x) dx \Bigg) \\
&= p(a - bp) + 0 + p(\gamma_0 - a + bp)(1 - F(\gamma_0 - a + bp)) - p \cdot (xF(x) - G(x))|_{\gamma_0 - a + bp}^{c} \\
&= p\gamma_0 - p(\gamma_0 - a + bp) F(\gamma_0 - a + bp) \\
&\quad - p(c - G(c) - F(\gamma_0 - a + bp) \cdot (\gamma_0 - a + bp) + G(\gamma_0 - a + bp)) \\
&= p(\gamma_0 - c + G(c) - G(\gamma_0 - a + bp)).
\end{aligned}
\tag{6.23}
$$

Here the eighth line comes from $\int_{-c}^{c} f(x)dx = F(c) - F(-c) = 1$ and $\int_{-c}^{c} xf(x)dx = \mathbb{E}[x] = 0$. Given the close form of $r(p)$, we may derive those of $r'(p)$ and $r''(p)$. ∎

6.7.2 Proof of Lemma 6.4.3

Proof. First of all, notice that $r''(p) \le 0$ holds for any $p \ge 0$. Therefore, $r'(p)$ is monotonically non-increasing on $p \in [0, \frac{a}{b}]$. Also, $F(x) = 0$ for $x < -c$ and $F(x) = 1$

for $x > c$. Since we define $G(x) = \int_{-c}^{x} F(\omega)d\omega$, we know that $G(x)$ is monotonically non-increasing on $x \in [-c, c]$, and that $G(c + x) = G(c) + x$ for $x \geq 0$.

1. Notice that

$$\begin{aligned} r'(0) =& \gamma_0 - c + G(c) - G(\gamma_0 - a) - 0 \\ >& \gamma_0 - c + G(c) - G(0) \\ >& 0. \end{aligned}$$

The second line is due to $\gamma_0 > 2c > c$ and $\gamma_0 < a - c < a$ according to Assumption 6.3.7, and the third line is due to the monotonicity of $G(x)$. Also, we notice that

$$\begin{aligned} r'(\frac{a}{b}) =& \gamma_0 - c + G(c) - G(\gamma_0) - b \cdot \frac{a}{b} \cdot F(\gamma_0) \\ =& \gamma_0 - c + G(c) - G(c + (\gamma_0 - c)) - a \cdot 1 \\ =& \gamma_0 - c + G(c) - G(c) - (\gamma_0 - c) - a \\ =& - a < 0. \end{aligned}$$

Also, we have $r'(p)$ monotonically non-increasing on $[0, \frac{a}{b}]$. According to the intermediate value theorem, there should exist a $p^* \in (0, \frac{a}{b})$ such that $r'(p^*) = 0$. Now we prove that p^* is unique by contradiction. If there exists $0 < p_1^* < p_2^* < \frac{a}{b}$ such that $r'(p_1^*) = r'(p_2^*) = 0$, we know that $r''(p) = 0, \forall p \in (p_1^*, p_2^*)$, which indicates $F(\gamma_0 - a + bp) = 0$ and $f(\gamma_0 - a + bp) = 0, \forall p \in (p_1^*, p_2^*)$. Since $F(\gamma_0 - a + bp)$ is non-decreasing, we know that $F(\gamma_0 - a + bp) = 0, \forall p < p_2^*$ and therefore $G(\gamma_0 - a + bp_1^*) = 0 = F(\gamma_0 - a + bp_1^*)$. Given this, we have $r'(p_1^*) = \gamma_0 - c + G(c) - 0 > 0$ (since $\gamma_0 > 2c > c$ according to Assumption 6.3.7), which is contradictory to our assumption that $r'(p_1^*) = 0$. Therefore, p^* such that $r'(p^*) = 0$ is unique. Due to the non-increasing property of $r'(p)$, we know that $r'(p) > 0$ for $p \in (0, p^*)$ and $r'(p) < 0$ for $p \in (p^*, \frac{a}{b}$, and therefore $r(p)$ monotonically increases on $p \in (0, p^*)$ and decreases on $p \in (p^*, \frac{a}{b})$.

2. Given that $0 \leq F(x) \leq 1$ and $0 \leq f(x) \leq f_{\max}, \forall x \in \mathbb{R}$, we have $r''(p) \geq -2b_{\max} - b_{\max}^2 p_{\max} f_{\max} = -C_s$.

3. According to the proof of Property 1, we know that $F(\gamma_0 - a + bp^*) > 0$, or otherwise $r'(p^*) > 0$ leading to contradiction. Denote $h(p) = F(\gamma_0 - a + bp)$, and for $\Delta = \frac{F(\gamma_0 - a + bp^*)}{2 f_{\max} b_{\max}} > 0$ we have

$$\begin{aligned} F(\gamma_0 - a + b(p^* - \Delta)) &\geq F(\gamma_0 - a + bp^*) - f_{\max} \cdot b\Delta \\ &\geq F(\gamma_0 - a + bp^*) - f_{\max} b_{\max} \cdot \frac{F(\gamma_0 - a + bp^*)}{2 f_{\max} b_{\max}} \\ &= \frac{F(\gamma_0 - a + bp^*)}{2}. \end{aligned} \tag{6.24}$$

Let $C_\Delta = bF(\gamma_0 - a + bp^*) > 0$, and for any $p \in [p^* - \Delta, p^* + \Delta]$ we have:

$$\begin{aligned} r''(p) &= -2b \cdot F(\gamma_0 - a + bp) - b^2 p \cdot f(\gamma_0 - a + bp) \\ &\leq -2b \cdot F(\gamma_0 - a + bp) \\ &\leq -2b \cdot F(\gamma_0 - a + b(p - \Delta)) \\ &\leq -2b \cdot \frac{F(\gamma_0 - a + bp^*)}{2} \\ &= -C_\Delta. \end{aligned} \tag{6.25}$$

■

6.7.3 Proof of Lemma 6.4.4

Proof. Recall that $\hat{r}'_k = \frac{1}{\tau} \sum_{t=1}^{\tau} D_{k,t} - p_k \cdot \hat{b} \cdot \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{1}_{k,t}$. Notice that

$$\begin{aligned} \mathbb{E}[D_{k,t}] &= \frac{r(p)}{p} = \gamma_0 - c + G(c) - G(\gamma_0 - a + bp) \\ \mathbb{E}[\mathbb{1}_{k,t}] &= \mathbb{E}[\mathbb{1}[a - bp + N_t < \gamma_0]] = F(\gamma_0 - a + bp). \end{aligned} \tag{6.26}$$

Also, $0 \leq D_{k,t} \leq \gamma_0$ and $0 \leq \mathbb{1}_{k,t} \leq 1$. According to Hoeffding's Inequality, with $\Pr \geq 1 - \eta\delta$ it holds that

$$|\frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t} - (\gamma_0 - c + G(c) - G(\gamma_0 - a + bp))| \leq \sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\gamma_0\cdot\frac{1}{\sqrt{\tau}}. \tag{6.27}$$

Also, with $\Pr \geq 1 - \eta\delta$ it holds that

$$|\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{1}_{k,t} - F(\gamma_0 - a + bp_k)| \leq \sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}}. \tag{6.28}$$

Therefore, with $\Pr \geq 1 - 2K\eta\delta$ it holds that

$$\begin{aligned}
|\hat{r}'_k - r'(p_k)| &= |(\frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t} - p_k\cdot\hat{b}\cdot\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{1}_{k,t}) \\
&\quad - (\gamma_0 - c + G(c) - G(\gamma_0 - a + bp_k) - bp_k\cdot F(\gamma_0 - a + bp_k))| \\
&= |(\frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t} - p_k\cdot\hat{b}\cdot\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{1}_{k,t}) + p_k\hat{b}F(\gamma_0 - a + bp_k) - p_k\hat{b}F(\gamma_0 - a + bp_k) \\
&\quad - (\gamma_0 - c + G(c) - G(\gamma_0 - a + bp) - bp\cdot F(\gamma_0 - a + bp))| \\
&\leq |\frac{1}{\tau}\sum_{t=1}^{\tau} D_{k,t} - (\gamma_0 - c + G(c) - G(\gamma_0 - a + bp))| + p_k|\hat{b} - b|F(\gamma_0 - a + bp_k) \\
&\quad + p_k\hat{b}\cdot|\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{1}_{k,t} - F(\gamma_0 - a + bp_k| \\
&\leq \sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\gamma_0\cdot\frac{1}{\sqrt{\tau}} + p_{\max}\cdot C_b\cdot\frac{1}{\sqrt{\tau}}\cdot 1 + p_{\max}\hat{b}\cdot\sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}} \\
&\leq \sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\gamma_0\cdot\frac{1}{\sqrt{\tau}} + p_{\max}\cdot\frac{b_{\max}^2 p_{\max}}{\gamma_2 - \gamma_1}\cdot\sqrt{2\log\frac{2}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}} \\
&\quad + 2p_{\max}b_{\max}\cdot\sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}} \\
&= C_K\cdot\frac{1}{\sqrt{\tau}}
\end{aligned} \tag{6.29}$$

Here the fifth (in)equality relies on $\hat{b} \leq 2b_{\max}$, which is a consequence of $|\hat{b} - b| \leq C_b\cdot\frac{1}{\sqrt{\tau}}$ and $\tau = \sqrt{T} \geq \left(\frac{C_b}{b_{\max}}\right)^2 = \left(\frac{b_{\max}p_{\max}}{\gamma_2 - \gamma_1}\right)^2$. ∎

6.7.4 Proof details of Theorem 6.4.1

Proof. We firstly bound the estimation error of $\hat{a}$ and $\hat{b}$ respectively. Notice that for any $\gamma \in [a_{\max} - b_{\min} p_{\max} + c, \gamma_0]$,

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[D_{0,t} \geq \gamma]] &= \mathbb{E}_{N_t}[\Pr_{p_t \sim U[0,p_{\max}]}[D_{0,t} \geq \gamma | N_t]] \\
&= \mathbb{E}_{N_t}[\Pr_{p_t \sim U[0,p_{\max}]}[a - b p_t + N_t \geq \gamma | N_t]] \\
&= \mathbb{E}_{N_t}[\frac{a + N_t - \gamma}{b p_{\max}}] \\
&= \frac{a - \gamma}{b p_{\max}}.
\end{aligned} \tag{6.30}$$

Given this, it holds

$$\begin{aligned}
\mathbb{E}[e_{1,t} - e_{2,t}] &= \frac{\gamma_2 - \gamma_1}{b p_{\max}} \\
\mathbb{E}[e_{3,t}] &= \frac{a - \gamma_3}{b p_{\max}}
\end{aligned} \tag{6.31}$$

Notice that $e_{1,t} \geq e_{2,t}$. According to Hoeffing's Inequality, it holds with $\Pr > 1 - \eta\delta$ that

$$|\frac{1}{\tau}\sum_{t=1}^{\tau} e_{1,t} - e_{2,t} - \frac{\gamma_2 - \gamma_1}{b p_{\max}}| \leq \sqrt{\frac{1}{2} \cdot \log \frac{2}{\eta\delta}} \cdot \frac{1}{\sqrt{\tau}}. \tag{6.32}$$

According to the definition of $\hat{b}$ in Eq. (6.4), we have:

$$\begin{aligned}
|\hat{b} - b| &= \frac{\gamma_2 - \gamma_1}{p_{\max}} \cdot |\frac{1}{\frac{1}{\tau}\sum_{t=1}^{\tau} e_{1,t} - e_{2,t}} - \frac{b p_{\max}}{\gamma_2 - \gamma_1}| \\
&= \frac{\gamma_2 - \gamma_1}{p_{\max}} \cdot \frac{|\frac{1}{\tau}\sum_{t=1}^{\tau} e_{1,t} - e_{2,t} - \frac{\gamma_2 - \gamma_1}{b p_{\max}}|}{(\frac{1}{\tau}\sum_{t=1}^{\tau} e_{1,t} - e_{2,t}) \cdot \frac{\gamma_2 - \gamma_1}{b p_{\max}}} \\
&\leq \frac{(\gamma_2 - \gamma_1)\sqrt{\frac{1}{2} \cdot \log \frac{2}{\eta\delta}} \cdot \frac{1}{\sqrt{\tau}}}{p_{\max} \cdot \frac{1}{2} \cdot \frac{\gamma_2 - \gamma_1}{b p_{\max}} \cdot \frac{\gamma_2 - \gamma_1}{b p_{\max}}} \\
&= \frac{b^2 p_{\max}}{\gamma_2 - \gamma_1} \cdot \sqrt{2 \log \frac{2}{\eta\delta}} \cdot \frac{1}{\sqrt{\tau}} \\
&\leq \frac{b_{\max}^2 p_{\max}}{\gamma_2 - \gamma_1} \cdot \sqrt{2 \log \frac{2}{\eta\delta}} \cdot \frac{1}{\sqrt{\tau}}.
\end{aligned} \tag{6.33}$$

Here the third row holds if $\tau \geq \frac{1}{2} \cdot \log \frac{2}{\eta\delta} \cdot (\frac{\gamma_2 - \gamma_1}{2 b p_{\max}})^2$. Again, according to Hoeffding's Inequality, it holds with $\Pr > 1 - \eta\delta$ that

$$|\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t} - \frac{a-\gamma_3}{bp_{\max}}| \le \sqrt{\frac{1}{2}\cdot\log\frac{2}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}}, \tag{6.34}$$

and we correspondingly bound the estimation error of $\hat{a}$ according to its definition in Eq. (6.4):

$$\begin{aligned}
|\hat{a}-a| =& |p_{\max}\cdot\hat{b}\cdot(\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t}) + \gamma_3 - a| \\
=& |p_{\max}\cdot\hat{b}\cdot(\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t} - \frac{a-\gamma_3}{bp_{\max}} + \frac{a-\gamma_3}{bp_{\max}}) + \gamma_3 - a| \\
=& |p_{\max}\cdot\hat{b}\cdot(\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t} - \frac{a-\gamma_3}{bp_{\max}}) + \frac{\hat{b}-b}{b}\cdot(a+\gamma_3)| \\
\le& p_{\max}\cdot\hat{b}\cdot|\frac{1}{\tau}\sum_{t=1}^{\tau} e_{3,t} - \frac{a-\gamma_3}{bp_{\max}}| + |\frac{\hat{b}-b}{b}|(a+\gamma_3) \\
\le& p_{\max}\cdot 2b_{\max}\cdot\sqrt{\frac{1}{2}\cdot\log\frac{1}{\eta\delta}}\frac{1}{\sqrt{\tau}} + \frac{a_{\max}+\gamma_3}{b_{\min}}\cdot\frac{b_{\max}^2 p_{\max}}{\gamma_2-\gamma_1}\cdot\sqrt{2\log\frac{1}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}} \\
=& p_{\max}(b_{\max} + \frac{(a_{\max}+\gamma_3)\cdot b_{\max}^2}{b_{\min}(\gamma_2-\gamma_1)})\cdot\sqrt{2\log\frac{1}{\eta\delta}}\cdot\frac{1}{\sqrt{\tau}}.
\end{aligned} \tag{6.35}$$

Here the fourth row holds if $\tau \ge 2\log\frac{2}{\eta\delta}\cdot(\frac{\gamma_2-\gamma_1}{bp_{\max}})^2$. Denote $C_a = p_{\max}(b_{\max} + \frac{(a_{\max}+\gamma_3)\cdot b_{\max}^2}{b_{\min}(\gamma_2-\gamma_1)})\cdot\sqrt{2\log\frac{1}{\eta\delta}}$ and $C_b = \frac{b_{\max}^2 p_{\max}}{\gamma_2-\gamma_1}\cdot\sqrt{2\log\frac{2}{\eta\delta}}$, and with $\Pr > 1-2\eta\delta$ we have

$$|\hat{a}-a| \le C_a\cdot\frac{1}{\sqrt{\tau}}, \quad |\hat{b}-b| \le C_b\cdot\frac{1}{\sqrt{\tau}}.$$

Now, we consider the cumulative regret in Epoch $k = 1, 2, \ldots$. According to Lemma 6.4.3, we know that

$$r(p^*) - r(p) \le C_s\cdot(p^*-p)^2. \tag{6.36}$$

According to Lemma 6.4.4 and Corollary 6.4.5 our binary searching is correct with high probability. Therefore, with $\Pr > 1-\log T\cdot\eta\delta$ it holds that $p^* \in [u_k, v_k], \forall k = 1, 2, \ldots$.

Notice that $u_k - v_k = \frac{p_{\max}}{2^{k-1}}$. Denote the index of the last epoch is K, and we have:

$$\begin{aligned} Regret_{search} =& \sum_{k=1}^{K}\sum_{t=1}^{\tau} r(p^*) - r(p_k) \\ \leq& \sum_{k=1}^{K} \tau \cdot C_s \cdot (p^* - p_k)^2 \\ \leq& \sum_{k=1}^{K} \tau \cdot C_s \cdot (v_k - u_k)^2 \\ =& \sum_{k=1}^{K} \tau \cdot C_s \cdot (\frac{p_{\max}}{2^{k-1}})^2 \\ \leq& \frac{4}{3} C_s \cdot p_{\max}^2 \cdot \tau. \end{aligned} \tag{6.37}$$

Now we upper bound the distance between p_K and p^* if $\hat{r}'_k \in [-2C_K \cdot \frac{1}{\sqrt{\tau}}, +2C_K \cdot \frac{1}{\sqrt{\tau}}]$. According to Lemma 6.4.4, in this case we have

$$|r'(p_k)| \leq |\hat{r}'_k - r'(p_k)| + |\hat{r}'_k| \leq C_K \cdot \frac{1}{\sqrt{\tau}} + 2C_K \cdot \frac{1}{\sqrt{\tau}} = 3C_K \cdot \frac{1}{\sqrt{\tau}}. \tag{6.38}$$

According to Lemma 6.4.3 Property 3, when $p \in [p^* - \Delta, p^* + \Delta]$, we have $|r'(p)| = |r'(p) - r'(p^*)| \geq C_\Delta |p - p^*|$. Also, since $r''(p) \leq 0$, indicating a monotonic decrease of $r'(p)$, we know that

$$\begin{aligned} r'(p) \geq r'(p^* - \Delta) \geq& C_\Delta |(p^* - \Delta) - p^*| = C_\Delta \cdot \Delta, \forall p \in [0, p^* - \Delta] \\ r'(p) \leq r'(p^* + \Delta) \leq& - C_\Delta |p^* - (p^* + \Delta)| = -C_\Delta \cdot \Delta, \forall p \in [p^* + \Delta, p_{\max}]. \end{aligned} \tag{6.39}$$

From Eq. (6.39), we know that $r'(p) \geq C_\Delta \cdot \Delta$ for $|p - p^*| \geq \Delta$. Therefore, if $p \in [0, p_{\max}]$ such that $|r'(p)| < C_\Delta \cdot \Delta$, then $p \in [p^* - \Delta, p^* + \Delta]$. Notice that $\tau = \sqrt{T} \geq (\frac{6C_K}{C_\Delta \cdot \Delta})^2$, and according to Eq. (6.38) we have $|r'(p_k)| \leq \frac{C_\Delta \cdot \Delta}{2} < C_\Delta \cdot \Delta$ if $\hat{r}'_k \in [-2C_K \cdot \frac{1}{\sqrt{\tau}}, +2C_K \cdot \frac{1}{\sqrt{\tau}}]$. Since we let $\hat{p}^* = p_k$ when $\hat{r}'_k \in [-2C_K \cdot \frac{1}{\sqrt{\tau}}, +2C_K \cdot \frac{1}{\sqrt{\tau}}]$, we know that $\hat{p}^* \in [p^* - \Delta, p^* + \Delta]$

and therefore

$$\begin{aligned}|\hat{p}^* - p^*| \leq & \frac{1}{C_\Delta} \cdot |r'(\hat{p}^*)| \\ \leq & \frac{3C_K \cdot \frac{1}{\sqrt{\tau}}}{C_\Delta} \\ = & \frac{3C_K}{C_\Delta \sqrt{\tau}}.\end{aligned} \tag{6.40}$$

By the time we determine $\hat{p}^*$, we have already proposed $(K+1)\tau$ prices. As a result, there are still $(T-(K+1)\tau)$ time steps left, where we keep proposing $\hat{p}^*$. According to Eq. (6.36) and Eq. (6.40), the cumulative regret of this period should be:

$$\begin{aligned}Regret_{exploiting} = & \sum_{t=1}^{T-(K+1)\tau} r(p^*) - r(\hat{p}^*) \\ \leq & \sum_{t=1}^{T} C_s \cdot (p^* - \hat{p}^*)^2 \\ \leq & T \cdot C_s \cdot \frac{9C_K^2}{C_\Delta^2} \cdot \frac{1}{\tau} \\ = & \frac{9C_s C_K^2}{C_\Delta^2} \cdot \frac{T}{\tau}.\end{aligned} \tag{6.41}$$

Finally, the regret of Epoch 0 cannot exceed $\gamma_0 p_{\max} \cdot \tau$. Combining with Eq. (6.37) and Eq. (6.41), we may bound the total regret with $\Pr > 1 - (2K+2)\eta\delta$ as

$$\begin{aligned}Regret = & Regret_{exploring} + Regret_{searching} + Regret_{exploiting} \\ \leq & \gamma_0 p_{\max} \cdot \tau + \frac{4}{3} C_s \cdot p_{\max}^2 \cdot \tau + \frac{9C_s C_K^2}{C_\Delta^2} \cdot \frac{T}{\tau}\end{aligned} \tag{6.42}$$

.

Since $\tau = \sqrt{T}$ and $\eta = \frac{1}{4\sqrt{T}} = \frac{\tau}{4T} \leq \frac{1}{2K+2}$, we have

$$\begin{aligned}Regret \leq & \gamma_0 p_{\max} \cdot \tau + \frac{4}{3} C_s \cdot p_{\max}^2 \cdot \tau + \frac{9C_s C_K^2}{C_\Delta^2} \cdot \frac{T}{\tau} \\ \leq & \gamma_0 p_{\max} \sqrt{T} + \frac{4C_s p_{\max}^2}{3} \cdot \sqrt{T} \\ & + \frac{9C_s}{(bF(\gamma_0 - a + bp))^2} \cdot \left(\gamma_0 + 2b_{\max} p_{\max} + \frac{2b_{\max}^2 p_{\max}}{\gamma_2 - \gamma_1} \right) \cdot \frac{1}{2} \cdot \log \frac{8\sqrt{T}}{\delta} \cdot \sqrt{T} \\ = & O(\sqrt{T} \log \frac{T}{\delta})\end{aligned} \tag{6.43}$$

with $\Pr \geq 1 - \frac{2K+2}{4K} \cdot \delta \geq 1 - \delta$. ■

6.7.5 Proof of Theorem 6.4.6

Proof. We firstly define a few functions:

1. Denote $Reg(p; a, b) := \max_{p^*} r(p^*) - r(p) = b(p - \frac{a}{2b})^2$ as the regret of price p under the true parameter (a, b).

2. Denote $\mathbb{D}(p; a, b) := a - bp + \mathbb{D}_N$ as the demand distribution when price p is proposed.

3. Denote $Q_\tau^{\mathcal{A}}(a, b)$ as the joint distribution of demands over time periods $t = 1, 2, \ldots, \tau$ by running Algorithm $\mathcal{A}$.

4. Denote $Reg((a, b), \tau, \mathcal{A})$ as the cumulative regret over time periods $t = 1, 2, \ldots, \tau$ by running Algorithm $\mathcal{A}$.

Now we specify the key quantities of problem setting as follows:

1. Let the noise distribution $\mathbb{D}_N$ be standard Gaussian distribution $\mathcal{N}(0, 1)$.

2. Let $(a_0, b_0) = (2, 1)$ be the "basic" problem setting. In this setting, the optimal price $p_0^* = 1$.

3. Let $(a_1, b_1) = (2 - \Delta, 1 - \Delta)$ be the "deviated" problem setting, where $\Delta \in (0, \frac{1}{4})$ is a small quantity to be specified later. In this setting, the optimal price $p_1^* = 1 + \frac{\Delta}{2(1-\Delta)}$.

Then we propose a lemma:

Lemma 6.7.1. *For any algorithm $\mathcal{A}$ and any $t \in [T]$, it holds*

$$KL(Q_t^{\mathcal{A}}(a_0, b_0)||Q_t^{\mathcal{A}}(a_1, b_1)) \leq \frac{1}{2} \cdot \Delta^2 \cdot Reg((a_0, b_0), t, \mathcal{A}). \tag{6.44}$$

Proof of Lemma 6.7.1. According to Duchi [2007], the KL divergence of two univariate Gaussian distributions holds

$$KL(\mathcal{N}(\mu_1, \sigma_1^2)||\mathcal{N}(\mu_2, \sigma_2^2)) = \log(\frac{\sigma_2}{\sigma_1}) + \frac{\sigma_1^2 - \sigma_2^2}{2\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}. \tag{6.45}$$

We will frequently apply Eq. (6.45) in the following proof. According to the Chain rule of KL Divergence, we have:

$$KL(Q_t^{\mathcal{A}}(a_0, b_0)||Q_t^{\mathcal{A}}(a_1, b_1)) = \sum_{s=1}^{t} KL(Q_s^{\mathcal{A}}(a_0, b_0)||Q_s^{\mathcal{A}}(a_1, b_1)|\mathcal{H}_{s-1}). \tag{6.46}$$

Notice that

$$\begin{aligned}
&KL(Q_s^{\mathcal{A}}(a_0, b_0)||Q_s^{\mathcal{A}}(a_1, b_1)|\mathcal{H}_{s-1}) \\
=&KL(\mathbb{D}(\pi_s(\mathcal{H}_{s-1}); a_0, b_0)||\mathbb{D}(\pi_s(\mathcal{H}_{s-1}); a_1, b_1)) \\
=&KL(\mathcal{N}(a_0 - b_0 \cdot \pi_s(\mathcal{H}_{s-1}), 1)||\mathcal{N}(a_1 - b_1 \cdot \pi_s(\mathcal{H}_{s-1}), 1)) \\
=&\frac{1}{2} \cdot ((a_0 - b_0 \cdot \pi_s(\mathcal{H}_{s-1})) - (a_1 - b_1 \cdot \pi_s(\mathcal{H}_{s-1})))^2 \\
=&\frac{1}{2}((a_0 - a_1) - (b_0 - b_1)\pi_s(\mathcal{H}_{s-1}))^2 \\
=&\frac{1}{2}(\Delta - \Delta\pi_s(\mathcal{H}_{s-1}))^2 \\
=&\frac{1}{2} \cdot \Delta^2(1 - \pi_s(\mathcal{H}_{s-1}))^2 \\
=&\frac{\Delta^2}{2}(p_0^* - \pi_s(\mathcal{H}_{s-1}))^2
\end{aligned} \tag{6.47}$$

Therefore, we have

$$\begin{aligned}
KL(Q_t^{\mathcal{A}}(a_0, b_0)||Q_t^{\mathcal{A}}(a_1, b_1)) =&\frac{\Delta^2}{2} \cdot \sum_{s=1}^{t}(p_0^* - \pi_s(\mathcal{H}_{s-1}))^2 \\
=&\frac{\Delta^2}{2} \cdot \sum_{s=1}^{t} b_0(p_0^* - \pi_s(\mathcal{H}_{s-1}))^2 \\
=&\frac{\Delta^2}{2} \cdot Reg((a_0, b_0), t, \mathcal{A}).
\end{aligned} \tag{6.48}$$

■

Now we propose another lemma:

Lemma 6.7.2. *We have*

$$Reg((a_0,b_0),T,\mathcal{A})+Reg((a_1,b_1),T,\mathcal{A})\geq\frac{\Delta^2}{64}\cdot T\cdot\exp\{-KL(Q_T^{\mathcal{A}}(a_0,b_0)||Q_T^{\mathcal{A}}(a_1,b_1))\}. \tag{6.49}$$

Proof of Lemma 6.7.2. For an algorithm $\mathcal{A}$, denote $\mathbf{v}:=[\pi_1(\mathcal{H}_0),\pi_2(\mathcal{H}_1),\ldots,\pi_T(\mathcal{H}_{T-1})]^\top\in\mathbb{R}^T$. Also, denote $\mathbf{v}_0:=[\frac{a_0}{2b_0},\frac{a_0}{2b_0},\ldots,\frac{a_0}{2b_0}]^\top\in\mathbb{R}^T$ and $\mathbf{v}_1:=[\frac{a_1}{2b_1},\frac{a_1}{2b_1},\ldots,\frac{a_1}{2b_1}]^\top\in\mathbb{R}^T$. Define a metric $d(\mathbf{x},\mathbf{y}):=\|x-y\|_2^2$. On the one hand, we have

$$\begin{aligned}Reg((a_0,b_0),T,\mathcal{A})+Reg((a_1,b_1),T,\mathcal{A})\geq&\sum_{t=1}^{T}b_0(\pi_t(\mathcal{H}_{t-1})-\frac{a_0}{2b_0})^2+b_1(\pi_t(\mathcal{H}_{t-1})-\frac{a_1}{2b_1})^2\\=&b_0\cdot d(\mathbf{v},\mathbf{v}_0)+b_1\cdot d(\mathbf{v},\mathbf{v}_1)\\\geq&\max_{i\in\{0,1\}}b_i\cdot d(\mathbf{v},\mathbf{v}_i)\\\geq&\max_{i\in\{0,1\}}\frac{1}{2}\cdot d(\mathbf{v},\mathbf{v}_i).\end{aligned} \tag{6.50}$$

The last inequality comes from $\Delta\in[0,\frac{1}{4}]$. Le Cam's Theorem [See Le Cam and Yang, 2000] states that

$$\inf_{\hat{\theta}}\sup_{\mathcal{P}\in\{\mathcal{P}_0,\mathcal{P}_1\}}\mathbb{E}[d(\hat{\theta},\theta(\mathcal{P})]\geq\frac{S}{8}\exp\{-KL(\mathcal{P}_0||,\mathcal{P}_1)\}, \tag{6.51}$$

where $S=d(\theta(\mathcal{P}_0),\theta(\mathcal{P}_1))$. Therefore, we have

$$\begin{aligned}&Reg((a_0,b_0),T,\mathcal{A})+Reg((a_1,b_1),T,\mathcal{A})\\\geq&\frac{1}{2}\cdot\frac{1}{8}\cdot d(\mathbf{v}_0,\mathbf{v}_1)\cdot\exp\{-KL(Q_T^{\mathcal{A}}(a_0,b_0)||Q_T^{\mathcal{A}}(a_1,b_1))\}\\=&\frac{1}{16}\cdot T\cdot(\frac{a_0}{2b_0}-\frac{a_1}{2b_1})^2\exp\{-KL(Q_T^{\mathcal{A}}(a_0,b_0)||Q_T^{\mathcal{A}}(a_1,b_1))\}\\=&\frac{1}{16}\cdot T(\frac{\Delta}{2(1-\Delta)})^2\exp\{-KL(Q_T^{\mathcal{A}}(a_0,b_0)||Q_T^{\mathcal{A}}(a_1,b_1))\}\\\geq&\frac{1}{64}T\Delta^2\exp\{-KL(Q_T^{\mathcal{A}}(a_0,b_0)||Q_T^{\mathcal{A}}(a_1,b_1))\}.\end{aligned} \tag{6.52}$$

■

Combine Lemma 6.7.1 and Lemma 6.7.2 and let $\Delta = \frac{1}{4} \cdot T^{-\frac{1}{4}}$, we have:

$$\begin{aligned}
&2(Reg((a_0, b_0), T, \mathcal{A}) + Reg((a_1, b_1), T, \mathcal{A})) \\
\geq &(Reg((a_0, b_0), T, \mathcal{A}) + Reg((a_1, b_1), T, \mathcal{A})) + Reg((a_0, b_0), T, \mathcal{A}) \\
\geq &\frac{1}{64} T\Delta^2 \exp\{-KL(Q_T^{\mathcal{A}}(a_0, b_0)||Q_T^{\mathcal{A}}(a_1, b_1))\} + \frac{2}{\Delta^2} \cdot KL(Q_T^{\mathcal{A}}(a_0, b_0)||Q_T^{\mathcal{A}}(a_1, b_1)) \\
= &\frac{1}{1024} \cdot \sqrt{T} \exp -KL(Q_T^{\mathcal{A}}(a_0, b_0)||Q_T^{\mathcal{A}}(a_1, b_1)) + 32\sqrt{T} \cdot KL(Q_T^{\mathcal{A}}(a_0, b_0)||Q_T^{\mathcal{A}}(a_1, b_1)) \\
\geq &\frac{1}{1024} \cdot \sqrt{T}(e^{-KL(Q_T^{\mathcal{A}}(a_0, b_0)||Q_T^{\mathcal{A}}(a_1, b_1))} + KL(Q_T^{\mathcal{A}}(a_0, b_0)||Q_T^{\mathcal{A}}(a_1, b_1))) \\
\geq &\frac{1}{1024} \sqrt{T}.
\end{aligned} \tag{6.53}$$

The last equation comes from the fact that $e^x \geq x + 1$ for any $x \in \mathbb{R}$. ■

Chapter 7

Conclusion and Discussion

In this thesis, we have explored a diverse array of topics within the realm of dynamic pricing, under the framework of online decision-making problems. Our investigation encompasses two primary areas: Feature-based Dynamic Pricing (Part I) and Pricing under Constraints (Part II). This final chapter aims to summarize our high-level observations and to propose several potential directions extended from this thesis toward further research.

7.1 Summary of Observations and Insights

Here we summarize our insights derived from our investigation into dynamic pricing, which also hold relevance for a broader spectrum of online learning problems.

Log-likelihood function as surrogate loss. The ONSP algorithm introduced in Chapter 2 employs a (negative) log-likelihood function as a surrogate loss. By running an online convex optimization algorithm (ONS) on this surrogate loss, we indirectly reduce

the regret and achieve an optimal rate of $O(d \log T)$. This approach is predicated on two conditions:

i The regret is *smooth* with respect to the parameters.

ii The negative log-likelihood function is *exp-concave* (i.e. strongly convex on empirical norm $\mathbb{E}[xx^\top]$). This is reflected in its Hessian matrix, or the so-called *Fisher Information* matrix, which must dominate $\mathbb{E}[xx^\top]$.

The first condition bounds the regret by the squared estimation error, while the second condition ensures a fast rate on optimizing the log-likelihood. Together, these conditions validate the use of this surrogate loss. The underlying intuition is that the Fisher Information encapsulated by the log-likelihood represents the maximum quantity of information accessible through observation. As long as our observation provides sufficient information, it become feasible to transit the "signal" (our observation) through our algorithmic "channel", thereby enabling an effective estimator for minimizing regret.

Half-Lipschitz nature of revenue function. In Chapter 3, we introduce the D2-EXP4 algorithm that adopts a discretization framework without assuming any continuity. This design choice is informed by the inherent characteristics of the revenue function: While an *increase* in price may lead to abrupt reduction on demand, a *decrease* in price typically results in a non-decreasing demand. Consequently, the potential loss from reducing the price by δ is at most $1 \cdot \delta$ for any single item sold. This observation leads us to identify the proximal gradient at each point on the revenue curve as residing within the range of $(-\infty, 1]$, a property we denote as *Half-Lipschitzness*. Leveraging this property allows for discretizations without the necessity of assuming continuity/Lipschitzness in dynamic pricing. Instead, we just need to set *conservative* prices every time a discretized

approximation is applied.

The hardness of pricing versus bandits. We establish a $\tilde{\Omega}(T^{\frac{2}{3}})$ lower bound under the assumption of Lipschitz demand in Chapter 3. On the one hand, this matches with the upper bound in Kleinberg [2004], indicating that the hardness of pricing is comparable to continuous bandits under Lipschitz assumptions. Combining with the results in Wang et al. [2021b], we know that under m^{th}-order smooth assumption for each $m = 1, 2, \ldots$, the minimax regret for (non-contextual) dynamic pricing is the same as that for continuum-armed bandits, which is $O(T^{\frac{m+1}{2m+1}})$. On the other hand, when the demand curve lacks Lipschitz continuity, the regret stays as $O(T^{\frac{2}{3}})$. However, for continuum-armed bandits with non-continuous or non-Lipschitz reward functions, achieving sublinear regret is impossible. As a conclusion, the inherent hardness of continuum-armed bandit problem is at least equivalent to, if not harder than, the dynamic pricing problem.

Perturbation as a regularizer. As previously discussed, two prerequisites enable the log-likelihood to function effectively as a surrogate loss: (i) Smoothness of the true loss (regret), and (ii) Exp-concavity of the log-likelihood function. However, Condition (ii) is not satisfied in the problem setting established in Chapter 4. To address this, we add a zero-mean perturbation on each proposed price. Given that the Fisher Information is the covariance matrix of the score function, this perturbation increases the variance and therefore enhance the Fisher Information. This strategy is analogous to incorporating a quadratic regularizer into the objective function to increase its convexity. However, in order to satisfy Condition (i), the ideal regularizer in our problem would be $\|\theta^* - \hat{\theta}\|_2^2$, to which we have no access, instead of $\|\hat{\theta}\|_2^2$. In conclusion, the perturbation performs as an indirect regularizer we cannot explicitly construct. We believe that the technique of

integrating perturbations into decisions holds significant promise for broader applications in online convex optimization.

Trade-off between revenue and fairness. In addition to analyzing algorithmic performance in Chapter 5, we establish a trade-off lower bound between regret and unfairness. Our findings indicate any optimal algorithm that achieves $C \cdot \sqrt{T}$ regret has to suffer at least $f(C) \cdot \sqrt{T}$ (substantive) unfairness. This relationship introduces a *"unit cost"* of pricing fairness quantified as $\frac{f(C)}{C}$. This phenomenon primarily stems from the high non-convexity of the revenue function with respect to pricing policies. On the one hand, it consumes sufficient sample complexity to learn and approach the optimal fair policy. On the other hand, the fixed-same price policy that ensures 0-unfairness is much sub-optimal. The inherent non-convexity of the problem precludes a natural equilibrium between fairness and profitability. This insight offers a significant contribution to the fields of machine learning and operations management, highlighting the complex interplay between economic efficiency and ethical considerations.

Pure-exploration to "see through" censoring. The challenge of inventory censoring, as modeled in Chapter 6, significantly complicates the estimation of demand curves. Despite these difficulties, we successfully obtain unbiased estimates for linear parameters through the use of *uniform exploration*. This technique was initially introduced by Fan et al. [2021] and subsequently adopted by Luo et al. [2022], both of whom explored settings where linear valuations are obscured by binary feedback, as detailed in Chapter 3. By adapting their uniform exploration approach to a new context, we not only address the specific challenges posed by inventory censoring but also demonstrate the potential of this strategy to advance research across a wider range of applications.

7.2 Future Directions

In the near future, my research will focus on online learning problems with *structural observations*. I anticipate that our findings and insights in dynamic pricing will contribute toward a deeper understanding of the interplay among decisions, information, and rewards within the broader framework of decision-making.

Here I outline some prospective projects that have been considered and discussed but not yet thoroughly investigated.

Online Decision-Making with Adversarial Censoring In our work presented in Chapter 6, we develop an algorithm for dynamic pricing under censored observations. This scenario arises because the seller cannot control inventory levels, presenting a unique challenge of non-controllable censoring effects. Therefore, a natural extension of this project could be *dynamic pricing under adversarial inventory constraints*. Consider, for example, a tropical fruit shop where supply is contingent on nearby farm production, which in turn is influenced by unpredictable natural factors such as temperature, rainfall, and the growth conditions of trees or crops. In such an adversarial inventory scenario, where the censoring threshold fluctuated and the optimal price varies over time, our previous see-through method (i.e. uniform exploration) for parameter estimation is still viable. However, we can no longer apply the search-based framework as we do in Chapter 6. Given the dependency of optimal pricing on inventory levels, it is necessary to construct *distributional* estimates of potential demands at each proposed price point. We conjecture that a UCB-styled algorithm would be suitable for this challenge. Notably, as the upper bound achieved in Chapter 6 matches the lower bound of pricing without censoring, implying that this censoring does not substantially increase the sample complexity, we

also conjecture the minimax regret is $O(\sqrt{T})$ in this adversarial setting.

Market Making with Mutually-Censoring Effect Previously we have discussed about methodologies to incorporate censored feedback in dynamic pricing, primarily focus on the demand side of market scenarios while often overlooking supply dynamics and associated costs. As we look into a two-sided online market such as Uber/Lyft, we notice there are two prices involved: (i) The amount which a customer pays for a unit of service, and (ii) The amount which a supplier earns by providing a unit of service. Based on this mechanism, the market maker, who sets these two prices, profits from the margin and volume of successful transactions. However, different from traditional market-making problems where bid (demand) and ask (supply) quantities are publicly known, in a contemporary online market with the most efficiency, the exceeded demand or supply might be fulfilled on competitors' platforms without being observed by the principle market maker. We can describe this dynamic as follows:

$$r(p_a, p_b) := \min\{D(p_a), S(p_b)\} \cdot (p_a - p_b).$$

Here p_a, p_b are the ask price and bid price, respectively; $r(\cdot), D(\cdot), S(\cdot)$ are the functions for revenue, demand, supply functions, sequentially. For each unit deal, the market maker pays p_b to the seller and earns p_a from the buyer, with $\min\{D(p_a), S(p_b)\}$ representing the volume of successful deals, which is the only observation that the market maker has access to. The market maker's goal is to learn from this censored observation and approach the optimal (p_a^*, p_b^*) with least cumulative regret. This problem could be very challenging as there exist no straightforward assumptions leading to its convexity. We conjecture that the solution to this problem depends on the strategies developed for "dynamic pricing under adversarial inventory constraints": By treating each side of the market individually while treating the other side as an adversarial environment, we are

hopeful to approach the optimal (p_a^*, p_b^*) by alternatively updating either side under a primal-dual framework.

Bibliography

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML-14)*, pages 1638–1646, 2014.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

R. Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

J. H. Aldrich, F. D. Nelson, and E. S. Adler. *Linear Probability, Logit, and Probit Models*. Number 45. Sage, 1984.

K. Amin, A. Rostamizadeh, and U. Syed. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems (NIPS-14)*, pages 622–630, 2014.

O. Anava and S. Mannor. Heteroscedastic sequences: beyond gaussianity. In *International Conference on Machine Learning*, pages 755–763. PMLR, 2016.

P. L. Anderson, R. D. McLellan, J. P. Overton, and G. L. Wolfram. Price elasticity of demand. *McKinac Center for Public Policy. Accessed October*, 13(2), 1997.

V. F. Araman and R. Caldentey. Dynamic pricing for nonperishable products with demand learning. *Operations Research*, 57(5):1169–1188, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

G. Aydin and S. Ziya. Personalized dynamic pricing of limited inventories. *Operations Research*, 57(6):1523–1531, 2009.

D. Baby, J. Xu, and Y.-X. Wang. Non-stationary contextual pricing with safety constraints. *Transactions on Machine Learning Research*, 2022.

G.-Y. Ban and N. B. Keskin. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 67(9): 5549–5568, 2021.

S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *Nips tutorial*, 1: 2, 2017.

G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Y. Bechavod, C. Jung, and S. Z. Wu. Metric-free individual fairness in online learning. *Advances in neural information processing systems*, 33:11214–11225, 2020.

D. Besanko, S. Gupta, and D. Jain. Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Science*, 44(11-part-1):1533–1547, 1998.

D. Besanko, J.-P. Dubé, and S. Gupta. Competitive price discrimination strategies in a vertical channel using aggregate retail data. *Management Science*, 49(9):1121–1138, 2003.

O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

O. Besbes and A. Zeevi. Blind network revenue management. *Operations research*, 60(6): 1537–1550, 2012.

O. Besbes and A. Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

M. Biehl and H. Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and general*, 28(3):643, 1995.

T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, pages 1287–1294, 1979.

J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

J. Bu, D. Simchi-Levi, and C. Wang. Context-based dynamic pricing with partially linear demand model. In *Advances in Neural Information Processing Systems*, 2022.

N. Cesa-Bianchi, O. Dekel, and O. Shamir. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26, 2013.

T. Chan, V. Kadiyali, and P. Xiao. Structural models of pricing. *Handbook of pricing research in marketing*, pages 108–131, 2009.

J. M. Chapuis. Price fairness versus pricing fairness. *Revenue & Yield Management eJournal*, page 12, 2012.

K. Chaudhuri, P. Jain, and N. Natarajan. Active heteroscedastic regression. In *International Conference on Machine Learning*, pages 694–702. PMLR, 2017.

B. Chen, X. Chao, and H.-S. Ahn. Coordinating pricing and inventory replenishment with nonparametric demand learning. *Operations Research*, 67(4):1035–1052, 2019a.

B. Chen, X. Chao, and Y. Wang. Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Operations Research*, 68(5):1445–1456, 2020.

B. Chen, X. Chao, and C. Shi. Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. *Mathematics of Operations Research*, 46(2):726–756, 2021a.

B. Chen, Y. Wang, and Y. Zhou. Optimal policies for dynamic pricing and inventory control with nonparametric censored demands. *Management Science*, 2023a.

N. Chen and G. Gallego. A primal-dual learning algorithm for personalized dynamic pricing with an inventory constraint. *Mathematics of Operations Research*, 2021.

Q. Chen, S. Jasin, and I. Duenyas. Nonparametric self-adjusting control for joint learning and optimization of multiproduct pricing with finite resource capacity. *Mathematics of Operations Research*, 44(2):601–631, 2019b.

X. Chen, X. Zhang, and Y. Zhou. Fairness-aware online price discrimination with nonparametric demand models. *arXiv preprint arXiv:2111.08221*, 2021b.

X. Chen, D. Simchi-Levi, and Y. Wang. Utility fairness in contextual dynamic pricing with demand learning. *arXiv preprint arXiv:2311.16528*, 2023b.

Y. Chen and V. F. Farias. Simple policies for dynamic pricing with imperfect forecasts. *Operations Research*, 61(3):612–624, 2013.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, pages 208–214, 2011.

M. C. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

M. C. Cohen, S. Miao, and Y. Wang. Dynamic pricing with fairness constraints. *Available at SSRN 3930622*, 2021.

M. C. Cohen, A. N. Elmachtoub, and X. Lei. Price discrimination with fairness constraints. *Management Science*, 2022.

A. A. Cournot. *Researches into the Mathematical Principles of the Theory of Wealth*. Macmillan, 1897.

J. H. Crane, C. F. Balerdi, and I. Maguire. Sugar apple growing in the florida home landscape. *Gainesville: University of Florida*, 2005.

T. Cunia. Weighted least squares method and construction of volume tables. *Forest Science*, 10(2):180–191, 1964.

A. V. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18, 2015.

M. Draganska and D. C. Jain. Consumer preferences and product-line pricing strategies: An empirical analysis. *Marketing science*, 25(2):164–174, 2006.

J. Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3(1): 2325–5870, 2007.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

R. Elfin. The future use of unconscionability and impracticability as contract doctrines. *Mercer L. Rev.*, 40:937, 1988.

G. C. Evans. The dynamics of monopoly. *The American Mathematical Monthly*, 31(2): 77–83, 1924.

E. Eyster, K. Madarász, and P. Michaillat. Pricing under fairness concerns. *Journal of the European Economic Association*, 19(3):1853–1898, 2021.

J. Fan, Y. Guo, and M. Yu. Policy optimization using semiparametric models for dynamic pricing. *arXiv preprint arXiv:2109.06368*, 2021.

B. S. Frey and W. W. Pommerehne. On the fairness of pricing—an empirical survey among the general population. *Journal of Economic Behavior & Organization*, 20(3): 295–307, 1993.

D. Gale. The law of supply and demand. *Mathematica scandinavica*, pages 155–169, 1955.

N. Golrezaei, P. Jaillet, and J. C. N. Liang. Incentive-aware contextual pricing with non-parametric market noise. *arXiv preprint arXiv:1911.03508*, 2019.

V. Goyal and N. Perivier. Dynamic pricing and assortment under a contextual mnl demand. *arXiv preprint arXiv:2110.10018*, 2021.

S. Gupta and V. Kamble. Individual fairness in hindsight. *J. Mach. Learn. Res.*, 22(144): 1–35, 2021.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

W. Hildenbrand. On the" law of demand". *Econometrica: Journal of the Econometric Society*, pages 997–1019, 1983.

M. Hutter, J. Poland, and M. Warmuth. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6(4), 2005.

R. Iyengar, A. Ansari, and S. Gupta. A model of consumer learning for service quality and usage. *Journal of Marketing Research*, 44(4):529–544, 2007.

A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

P. L. Joskow and C. D. Wolfram. Dynamic pricing of electricity. *American Economic Review*, 102(3):381–85, 2012.

V. Kadiyali, N. J. Vilcassim, and P. K. Chintagunta. Empirical analysis of competitive product line pricing decisions: Lead, follow, or move together? *Journal of Business*, pages 459–487, 1996.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

R. L. Kaufman. *Heteroskedasticity in regression: Detection and correction*. Sage Publications, 2013.

P. J. Kaufmann, G. Ortmeyer, and N. C. Smith. Fairness in consumer pricing. *Journal of Consumer Policy*, 14(2):117–140, 1991.

N. B. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.

N. B. Keskin, Y. Li, and J.-S. Song. Data-driven dynamic pricing and ordering with perishable inventory in a changing environment. *Management Science*, 68(3):1938–1958, 2022.

W. Kincaid and D. Darling. An inventory pricing problem. *Journal of Mathematical Analysis and Applications*, 7:183–208, 1963.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17:697–704, 2004.

R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symposium on Foundations of Computer Science (FOCS-03)*, pages 594–605. IEEE, 2003.

T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems (NIPS-15)*, pages 1477–1485, 2015.

A. Krämer, M. Friesen, and T. Shelton. Are airline passengers ready for personalized dynamic pricing? a study of german consumers. *Journal of Revenue and Pricing Management*, 17(2):115–120, 2018.

A. Krishnamurthy, T. Lykouris, C. Podimata, and R. Schapire. Contextual search in the presence of irrational agents. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC-21)*, pages 910–918, 2021.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

A. Lambrecht, K. Seim, and B. Skiera. Does uncertainty matter? consumer behavior under three-part tariffs. *Marketing Science*, 26(5):698–710, 2007.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS-07)*, pages 817–824, 2007.

L. M. Le Cam and G. L. Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

R. P. Leme and J. Schneider. Contextual search via intrinsic volumes. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS-18)*, pages 268–282. IEEE, 2018.

R. P. Leme, B. Sivan, Y. Teng, and P. Worah. Learning to price against a moving target. In *International Conference on Machine Learning*, pages 6223–6232. PMLR, 2021.

A. Liu, R. P. Leme, and J. Schneider. Optimal contextual pricing and extensions. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA-21)*, pages 1059–1078. SIAM, 2021.

I. Lobel, R. P. Leme, and A. Vladu. Multidimensional binary search for contextual decision-making. *Operations Research*, 66(5):1346–1361, 2018.

Y. Luo, W. W. Sun, et al. Distribution-free contextual dynamic pricing. *arXiv preprint arXiv:2109.07340*, 2021.

Y. Luo, W. W. Sun, and Y. Liu. Contextual dynamic pricing with unknown noise: Explore-then-ucb strategy and improved regrets. In *Advances in Neural Information Processing Systems*, 2022.

R. Maestre, J. Duque, A. Rubio, and J. Arévalo. Reinforcement learning for fair dynamic pricing. In *Proceedings of SAI Intelligent Systems Conference*, pages 120–135. Springer, 2018.

J. Mao, R. P. Leme, and J. Schneider. Contextual pricing for lipschitz buyers. In *NeurIPS*, pages 5648–5656, 2018.

A. Marshall. *Principles of economics: unabridged eighth edition*. Cosimo, Inc., 2009.

T. Mazumdar, S. P. Raj, and I. Sinha. Reference price research: Review and propositions. *Journal of Marketing*, 69(4):84–102, 2005.

S. Miao, X. Chen, X. Chao, J. Liu, and Y. Zhang. Context-based dynamic pricing with online clustering. *arXiv preprint arXiv:1902.06199*, 2019.

J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019.

K. P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

M. Nambiar, D. Simchi-Levi, and H. Wang. Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000, 2019.

M. Parkin, M. Powell, and K. Matthews. *Economics*. Addison-Wesley, Harlow, 2002.

A. Priester, T. Robbert, and S. Roth. A special price just for you: Effects of personalized dynamic pricing on consumer fairness perceptions. *Journal of Revenue and Pricing Management*, 19(2):99–112, 2020.

M. L. Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

S. Qiang and M. Bayati. Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*, 2016.

D. Qiao, M. Yin, M. Min, and Y.-X. Wang. Sample-efficient reinforcement learning with loglog (T) switching cost. *arXiv preprint arXiv:2202.06385*, 2022.

T. J. Richards, J. Liaukonyte, and N. A. Streletskaya. Personalized pricing and price fairness. *International Journal of Industrial Organization*, 44:138–153, 2016.

H. Schultz et al. *Theory and Measurement of Demand*. The University of Chicago Press, 1938.

V. Shah, R. Johari, and J. Blanchet. Semi-parametric dynamic contextual pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

E. Spiliotopoulou and A. Conte. Fairness ideals in inventory allocation. *Decision Sciences*, 53(6):985–1002, 2022.

C.-H. Wang, Z. Wang, W. W. Sun, and G. Cheng. Online regularization for high-dimensional dynamic pricing algorithms. *arXiv preprint arXiv:2007.02470*, 2020.

H. Wang, K. Talluri, and X. Li. On dynamic pricing with covariates. *arXiv preprint arXiv:2112.13254*, 2021a.

Y. Wang, B. Chen, and D. Simchi-Levi. Multimodal dynamic pricing. *Management Science*, 2021b.

Z. Wang, S. Deng, and Y. Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

P. Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):143–149, 1980.

R. E. Wright. Logistic regression. 1995.

J. Xu and Y.-X. Wang. Logarithmic regret in feature-based dynamic pricing. *Advances in Neural Information Processing Systems*, 34, 2021.

J. Xu and Y.-X. Wang. Towards agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

J. Xu, D. Qiao, and Y.-X. Wang. Doubly fair dynamic pricing. In *International Conference on Artificial Intelligence and Statistics*, pages 9941–9975. PMLR, 2023.

Z. Yang, X. Fu, P. Gao, and Y.-J. Chen. Fairness regulation of prices in competitive markets. *Available at SSRN*, 2022.

H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30, 2017.

L. Zhou. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*, 2015.