# Data-Driven Dynamic Decision Making: Algorithms, Structures, and Complexity Analysis

Yunzong Xu

Massachusetts Institute of Technology

May 2023

**Data-Driven Dynamic Decision Making: Algorithms, Structures, and Complexity Analysis**

by

Yunzong Xu

Submitted to the Institute for Data, Systems, and Society
on May 5, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

This thesis aims to advance the theory and practice of data-driven dynamic decision making, by synergizing ideas from machine learning and operations research. Throughout this thesis, we focus on three aspects: (i) developing new, practical *algorithms* that systematically empower data-driven dynamic decision making, (ii) identifying and utilizing key problem *structures* that lead to statistical and computational efficiency, and (iii) contributing to a general understanding of the statistical and computational *complexity* of data-driven dynamic decision making, which parallels our understanding of supervised machine learning and also accounts for the crucial roles of model structures and constraints for decision making.

Specifically, the thesis consists of three parts.

Part I of this thesis develops methodologies that *automatically* translate advances in supervised learning into effective dynamic decision making. Focusing on contextual bandits, a core class of online decision-making problems, we present the first optimal and efficient reduction from contextual bandits to offline regression. A remarkable consequence of our results is that advances in offline regression immediately translate to contextual bandits, statistically and computationally. We illustrate the advantages of our results through new guarantees in complex operational environments and experiments on real-world datasets. We also extend our results to more challenging setups, including reinforcement learning in large state spaces. Beyond the positive results, we establish new fundamental limits for general, unstructured reinforcement learning, emphasizing the importance of problem structures in reinforcement learning.

Part II of this thesis develops a framework that incorporates offline data into online decision making, motivated by practical challenges in business and operations. In the context of dynamic pricing, the framework allows us to rigorously characterize the value of data and the synergy between online and offline learning in data-driven decision making. The theory provides important insights for practice.

Part III of this thesis studies classical online decision-making problems in new settings where the decision maker may face a variety of long-term constraints. Such constraints are motivated by societal and operational considerations, and may limit the decision maker's ability to switch between actions, consume resources, or query accumulated data. We characterize the statistical and computational consequences brought by such long-term constraints, i.e., how the complexity of the problem changes with respect to different levels of constraints. The results provide precise characterizations on various intriguing trade-offs in data-driven dynamic decision making.

Thesis Supervisor: David Simchi-Levi
Title: Professor of Engineering Systems

Thesis Committee Member: Alexander Rakhlin
Title: Professor of Brain and Cognitive Sciences

Thesis Committee Member: John N. Tsitsiklis
Title: C. J. LeBel Professor of Electrical Engineering and Computer Science

## Acknowledgments

You are about to begin reading the Ph.D. thesis of Yunzong Xu, a young man who spent five years working on his Ph.D. research, exploring the connections between learning, inference, decision making, algorithms, structures, complexity, and related topics. You are curious to know more about him, his motivations, his experiences, his passions. You turn to the acknowledgments page, expecting to find a list of names and institutions that supported his work. But as you start to read, you realize that this is not a conventional acknowledgments section.

So what is this section about, for the author? First, it is about his gratitude. He wants to introduce some important people to you, who have made this work possible: his advisor, his committee members, his colleagues, his friends, his family. He wants to tell you how much he learned from them, how much he enjoyed interacting with them, how much he owes them. He wants to make you feel that this thesis is not only his achievement, but also theirs.

But he also believes that he is writing something like a love poem addressed to the five-year Ph.D. journey, which has taught him a lot about himself — that he is braver than he believes, stronger than he seems, and smarter than he thinks. The journey has been like exploring a labyrinth of infinite possibilities, a city of endless connections, and a forest of branching paths.

And he does not believe that every journey must have an end. That's why he also wants to thank you, for choosing to read his thesis. He knows that every reader makes his journey continue in a new way. He hopes that you will enjoy this experience and encounter something memorable. He knows that acknowledgments are usually boring and formal, and he doesn't want to bore you or be formal. He wants to surprise you with something special. He wants to make you smile, chuckle, giggle.

So here you are now, ready to be part of this journey. Relax. Concentrate. Dispel every other thought. The author—"he"—is starting to tell his stories.

He remembers his first meeting with Professor David Simchi-Levi, who would later become his advisor and thesis supervisor. He had made David wait for a long time due to a horrible traffic jam, which added to his nervousness when he finally showed up. But David greeted him warmly and kindly, and soon he felt like he was talking to a mentor. He describes how David's eyes shone with enthusiasm and excitement when listening to him, and

how that encouraged him to confidently share his own vision. He boldly proposed several trending topics as potential research directions for his Ph.D., but none of them turned out to be his actual focus. He says that what he ended up doing and achieving during his Ph.D. journey was far beyond his initial expectations. And he thanks David for that, for always supporting him to pursue his passions and explore his own paths, even when they seemed to lead nowhere or everywhere. He feels so fortunate to have David as his advisor, who gave him unwavering trust in his potential (even though he often doubted himself), who gave him boundless freedom to choose his research topics (even during times when securing sufficient funding for the group was challenging). It makes such a difference, he says, to have someone who believes in you, understands you, values you, supports you, and is proud of you. He does not know if he can live up to David's expectations in the future, but it always cheers him up whenever David tells him, "You make me proud."

He then expresses his gratitude to his other committee members; each of them has shaped his growth and goals in a unique way. He wants to thank Professor Sasha Rakhlin, who has always inspired him to pursue the depth and the truth. He says Sasha helped him to appreciate the beauty and elegance of learning theory and statistics, and influenced him with a very high standard on good research, good papers, good talks, and good mathematics. He also wants to thank Professor John Tsitsiklis, who has shown him new horizons and landscapes. He says John taught him to search for meaning and principles in a world of chaos and uncertainty, and provided him invaluable advice on various aspects of his research and career. He is grateful for the encouragement and support he received from these role models; they inspire him to be serious about the type of research he wants to do; they motivate him to become the type of person he wants to be.

He then thanks to his collaborators who have contributed to parts of this thesis: Jinzhi Bu, Dylan Foster, Akshay Krishnamurthy, and Jinglong Zhao. They have been his companions, who have shared with him their optimism and concentration, who have inspired him with their curiosity and creativity. He remembers the late nights they spent working on their manuscripts before submission deadlines, the Zoom and Microsoft Teams meetings they had to chat about research and life, and the excitement they shared when discovering something new. He also wants to thank his mentors who guided him before he started his Ph.D. journey: Yong Liang, Zizhuo Wang, and Fuqiang Zhang. They taught him how to do research when he was an undergraduate student, for which he feels grateful forever.

He owes special thanks to his friends at MIT: Jinzhi, Dylan, Jinglong, Michael Beeler, Marie Charpignon, Hongyu Chen, Louis Chen, Elaheh Fata, Evelyn Gong, Cate Heine, Yiqun Hu, Lei Huang, Yan Jin, Janet Kerrigan, Dimitris Konomis, Kirby Ledvina, Hanwei Li, Menglong Li, Jian Qian, Hanzhang Qin, Beth Milnes, Mila Nambiar, Junyi Sha, Ali Shameli, Rui Sun, Prem Talwai, Renfei Tan, Chonghuan Wang, Li Wang, Xinshang Wang, Xirui Wang, Yuhao Wang, Yuting Wang, Michelle Wu, Qi Yang, Leon Yao, Jiaheng Yu, Tiancheng Yu, Sabrina Zhai, Jiawei Zhang, Peter Zhang, Feng Zhu, Ruihao Zhu... (the list is too long to fit here). They shared with him their joys and sorrows, their ups and downs. They helped him cope with the stress and challenges of graduate school life. They also enriched his life with their diverse backgrounds and interests. They introduced him to new hobbies and habits (even when he did not seem interested in them at the beginning). They invited him to gather and play board games (even though he was not good at remembering complicated rules). They drove him to try different restaurants and cuisines (even though he could not drive). They also joined him in many adventures inside and outside the school: exploring the MIT tunnels (and getting lost); sitting in difficult classes outside their fields (and getting lost again); competing on the fields (and getting beaten); "lying flat" on the grass (and getting bitten again); sailing in the Boston Harbor (and getting cold); kayaking in the Charles River (and getting arm pain); going to BSO concerts (and getting peace); traveling to large conferences (and getting lost yet again). They made these years unforgettable.

Now, he wants to thank his parents, who have given him life and education; who have nurtured him and established his values and principles; who have supported him in every major decision he made (even when they had a different opinion); who have believed in him and encouraged him (even when they did not understand what he was doing); who have sacrificed so much for him (even when he was not aware); who have always been there for him (even when he was far away).

Then he wants to thank his twin brother Yunbei. Yunbei has been his best friend since childhood. They have shared their dreams and passions (some of which still inspire them). They have also shared their difficulties and worries (some of which still challenge them). They have always stood by each other (even when they fell). They have always guided each other (even when they strayed). They have always admired each other (even though they seldom express it). They have always influenced each other (even when they did not realize it). They have always been each other's first reader, first referee; first fan, first rock star.

And finally, he wants to thank his partner Xin, who has been by his side throughout this journey (and before); who has understood him completely (even when he was incomprehensible); who has supported him wholeheartedly (even when he was frustrated); who has made him happy beyond words (even though he is good with words).

He dedicates this thesis to his family, without whom none of this thesis would have been possible (or meaningful).

The author has told his stories, and now he pauses, looking at you, the reader. He believes that his journey does not end with his thesis; it goes on with every reader who joins him; it lives with every reader who feels him. He wants to thank you for your attention and curiosity; for your patience and generosity; for your time and presence.

He concludes with a simple but heartfelt sentence: "Thank you for enriching the meaning of my journey."

Yunzong Xu

Cambridge, MA

May 2023

# Contents

**1 Introduction** **17**

1.1 Overview of Part I (Chapters 2 to 4) . . . . . 19

1.2 Overview of Part II (Chapter 5) . . . . . 22

1.3 Overview of Part III (Chapters 6 to 7) . . . . . 23

**2 Optimal and Efficient Reduction from Contextual Bandits to Offline Regression** **25**

2.1 Introduction . . . . . 25

2.2 Algorithm and Guarantees . . . . . 37

2.3 General Offline Regression Oracles . . . . . 42

2.4 Regret Analysis . . . . . 49

2.5 Concluding Remarks . . . . . 59

**3 Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning** **61**

3.1 Introduction . . . . . 61

3.2 Instance-Dependent Complexity of Contextual Bandits . . . . . 65

3.3 Instance-Dependent Complexity of Reinforcement Learning . . . . . 74

3.4 Concluding Remarks . . . . . 77

**4 Fundamental Barriers for Offline Reinforcement Learning with Value Function Approximation** **79**

4.1 Introduction . . . . . 79

4.2 Fundamental Barriers for Offline Reinforcement Learning . . . . . 88

4.3 Proof Overview for Theorem 4.2 . . . . . 101

4.4 Concluding Remarks . . . 103

**5 Online Pricing with Offline Data 105**

5.1 Introduction . . . 105

5.2 Related Literature . . . 112

5.3 Single Historical Price . . . 115

5.4 Multiple Historical Prices . . . 125

5.5 Numerical Experiments . . . 131

5.6 Further Discussion: Offline Data and Self-Exploration . . . 136

5.7 Concluding Remarks . . . 138

**6 Bandits with Switching Constraints 141**

6.1 Introduction . . . 141

6.2 Related Literature . . . 152

6.3 Unit Switching Costs . . . 155

6.4 General Switching Costs . . . 171

6.5 Numerical Experiments . . . 175

6.6 Concluding Remarks . . . 178

**7 Blind Network Revenue Management and Bandits with Knapsacks Under Limited Switches 179**

7.1 Introduction . . . 179

7.2 Problem Formulation . . . 187

7.3 Warm-Up: Network Revenue Management Under Limited Switches . . . 190

7.4 Blind Network Revenue Management Under Limited Switches . . . 194

7.5 Concluding Remarks . . . 205

**8 Conclusion and Future Plan 207**

**A Supplementary Material for Chapter 2 209**

A.1 Proof of Theorems 2.1 and 2.2 . . . 209

**B Supplemental Material for Chapter 3 223**

B.1 Details for Results of Contextual Bandits . . . 223

B.2 Details for Results of Reinforcement Learning . . . 239

**C Suppmentary Material for Chapter 4** **247**

C.1 General Scheme to Construct Hard Families of Instances . . . . . 247

C.2 Computation of Value Functions (Proposition 4.1) . . . . . 248

C.3 Proof of Lemma 4.1 . . . . . 249

C.4 Proof of Lemma 4.2 . . . . . 250

C.5 Theorem 4.2: Lower Bound Construction and Proof . . . . . 259

C.6 Proof of Lemma C.8 . . . . . 264

C.7 Proof of Lemma C.9 . . . . . 265

C.8 Proofs of Propositions C.1 and C.2 . . . . . 273

**D Supplementary Material for Chapter 5** **277**

D.1 Proofs of Statements in Section 5.3 . . . . . 277

D.2 Proofs of Statements in Section 5.4 . . . . . 288

D.3 Proof of Proposition 5.1 in Section 5.6 . . . . . 298

D.4 On the Definition of the Optimal Regret . . . . . 300

D.5 Extension to Generalized Linear Model . . . . . 304

D.6 Extension to Adaptive Offline Data . . . . . 307

D.7 Multi-Armed Bandits with Offline Data . . . . . 310

D.8 Tables in Sections 5.3 and 5.4 . . . . . 312

**E Supplementary Material for Chapter 6** **315**

E.1 Results on Distribution-Dependent Regret Bounds . . . . . 315

E.2 Rounding Issues of Algorithms . . . . . 317

E.3 Illustration of AdaLS . . . . . 318

E.4 Reverse Fano-Type Inequalities and Lower Bound Analysis . . . . . 318

E.5 Explanations for Section 6.4.1 . . . . . 322

E.6 Additional Numerical Experiments . . . . . 323

E.7 Proof of Proposition 6.1 . . . . . 326

E.8 Proof of Theorem 6.1 . . . . . 329

E.9 Proof of Theorem 6.3 . . . . . 342

E.10 Proof of the Upper Bound in Theorem 6.5 . . . . . 343

E.11 Information-Theoretic Tools . . . . . 344

E.12 Proof of Theorem 6.2 . . . . . 348

E.13 Proof of Theorem 6.4 . . . . . 367
E.14 Proof of the Lower Bound in Theorem 6.5 . . . . . 375

**F Supplementary Material for Chapter 7** **383**
F.1 Extended Models in the Bandits with Knapsacks Setup . . . . . 383
F.2 Bandits with Knapsacks Under Limited Switches . . . . . 385

# List of Figures


3-1 Relationship between complexity measures. . . . . 64

4-1 The MDPs in $\mathcal{M}$ are parametrized by three scalars $\alpha, \beta, w$ and a subset of states $I$. The state space consists of an *initial state* $\mathfrak{s}$, a large number of *intermediate states* $\mathcal{S}^1$, and four self-looping *terminal states* $\{W, X, Y, Z\}$. From the initial state $\mathfrak{s}$, action 1 (in red) transitions to state $W$, while action 2 (in blue) transitions to a subset of intermediate states $I \subset \mathcal{S}^1$ with equal probability. In all intermediate states and terminal states, actions 1 and 2 have the same effect, with transitions denoted in black. Among the intermediate states, $I \subset \mathcal{S}^1$ (the gray ones) are the *planted states* which transition with probability $\alpha$ to state $X$ and $1-\alpha$ to state $Y$, and the remaining $\mathcal{S}^1 \setminus I$ (the striped ones) are the *unplanted states* which transition with probability $\beta$ to $Z$ and $(1-\beta)$ to $Y$. There are combinatorially many choices for $I$. Only terminal states can generate non-zero rewards: the rewards of the states $W$, $X$, $Y$ and $Z$ are $w$, 1, 0 and $\alpha/\beta$, respectively. . . . . 89

4-2 Illustration of the MDP family used to prove Theorem 4.2, with $L = 3$ layers of the planted subset structure (note that in general, we take $L > 3$). The rewards of the states $W$, $X$, $Y$ and $Z$ are $w$, 1, 0 and $\alpha/(1-L\alpha)$, respectively, where $w$ and $\alpha$ are parameters of the MDP family. . . . . 102

5-1 Revenue curves under three different parameters (blue), and the optimistic revenue (red). . . . . 117

5-2 Phase transitions for the single-historical-price setting with constant $\delta$. . . . . 124

5-3 Phase transitions for the single-historical-price setting with general $\delta$. Left figure: $\delta \gtrsim T^{-\frac{1}{4}}$; right figure: $\delta \lesssim T^{-\frac{1}{4}}$. . . . . 124

5-4 Multiple-historical-price setting with $\delta \gtrsim T^{-\frac{1}{4}}$ and different $\sigma$. . . . . 130

5-5 Phase transitions for the multiple-historical-price setting with $\delta \gtrsim T^{-\frac{1}{4}}$. Left figure: $\sigma = o(\delta)$; right figure: $\sigma = \Omega(\delta)$. . . . . 130
5-6 Phase transitions for the multiple-historical-price setting with $\delta \lesssim T^{-\frac{1}{4}}$. . . . 132
5-7 Comparison between O3FU and CILS when there are no offline data. . . . . 133
5-8 Comparison between O3FU and CILS when there are $n = 1000$ offline demand data. . . . . 133
5-9 95% confidence-region comparison between O3FU and CILS with $\kappa = 0.5$. . . 134
5-10 $T = 10^4$-period relative regret for the single-historical-price setting with different $n$. . . . . 135
5-11 $T = 10^4$-period relative regret for the single-historical-price setting with different $\delta$. . . . . 135
5-12 $T = 10^4$-period relative regret for the multiple-historical-price setting with different $\sigma$. . . . . 135

6-1 Empirical average-case regret v.s. the switching budget $S$, for the four considered algorithms. The regret of UCB has to be plotted separately because it is too large. . . . . 177

7-1 Optimal regret rate exponent $\lim_{T\to\infty} \log R^\star(T)/\log T$ as a function of switching budget $s$ in the BNRM-LS problem. Here $R^\star(T)$ stands for the optimal (i.e., minimax) regret. . . . . 182

C-1 Illustration of the average MDP with $L = 3$. . . . . 267

D-1 Phase transition in $K$-armed bandits with offline data when $\Delta = \Omega(T^{-\frac{1}{2}})$. . . 311

E-1 Empirical average-case regret v.s. the switching budget $S$, under $K = 4$. The regret of UCB has to be plotted separately because it is too large. . . . . 324
E-2 Empirical average-case regret v.s. the switching budget $S$, under $K = 16$. The regret of UCB has to be plotted separately because it is too large. . . . . 325

# List of Tables


2.1 Algorithms' performance with general finite $\mathcal{F}$ and i.i.d. contexts. Advantages are marked in bold. . . . . 28

5.1 Optimal regret for the single-historical-price setting. . . . . 111

5.2 Optimal regret for the multiple-historical-price setting. . . . . 111

6.1 Regret of LS-SE and AdaLS under different switching budgets. Here $\epsilon \in (0, 1)$ is an arbitrary constant independent of $K$ and $T$ (it can be arbitrarily close to 0, as long as it is fixed). . . . . 163

6.2 Optimal regret rate under different switching budgets for a fixed $K$. . . . . 167

6.3 Optimal regret rate under different BAR $\theta$ when $K$ grows as $T^{\alpha}$. . . . . 169

6.4 Optimal regret rate (of D-BwSC) under different switching budgets for fixed $K$ and $\boldsymbol{c}$. . . . . 175

C.1 Value of $\frac{P_I(s'|s,2)P_{I'}(s'|s,2)}{P_0(s'|s,2)}$ for all possible pairs $(s, s')$. . . . . 254

C.2 Value of $\frac{P_{\boldsymbol{I}}(s'|s,\mathfrak{a})P_{\boldsymbol{I}'}(s'|s,\mathfrak{a})}{P_0(s'|s,\mathfrak{a})}$ for all possible pairs $(s, s')$. . . . . 270

D.1 Regret upper bound in Theorem 5.1 for the single-historical-price setting. . . 312

D.2 Regret lower bound in Theorem 5.2 for the single-historical-price setting. . . . 313

D.3 Regret upper bound in Theorem 5.3 for the multiple-historical-price setting. . 313

D.4 Regret lower bound in Theorem 5.4 for the multiple-historical-price setting. . 314

# Chapter 1

# Introduction

The increasing availability of data and advances in machine learning have the potential to revolutionize the way organizations make decisions. However, such potential is heavily constrained now. While supervised machine learning traditionally excels at making *predictions* based on *passively observed, well-distributed* data, modern organizations often face tasks that require *dynamic decision making* over time (e.g., online recommendation and resource allocation tasks in operations research), which typically generate *sequential* data that exhibit *distribution shift* over time, making the effective integration of machine learning and decision making difficult. Moreover, many real-world organizations have to operate under various regulations and practical constraints, which cannot be captured by classical machine learning frameworks. Such discrepancies present significant challenges in the development of modern data-driven dynamic decision-making systems, both in theory and practice.

The high-level goal of this thesis is to address these challenges to advance the development of modern data-driven dynamic decision-making systems. This requires extending the boundaries of machine learning by making the advances therein (e.g., function approximation, scalable computing, state-of-the-art estimators) more applicable to decision making, as well as expanding the scope of operations research by developing new decision-making models that incorporate modern machine learning and its salient features as key components, such that the use and value of data and learning can be systematically analyzed and optimized.

Motivated by the high-level goal, this thesis seeks to advance the theory and practice of data-driven dynamic decision making, by synergizing ideas from machine learning and operations research. Throughout this thesis, we focus on three aspects: (i) developing general, flexible *algorithms* that systematically empower data-driven dynamic decision making, (ii)

identifying and utilizing key problem *structures* that lead to statistical and computational efficiency, and (iii) contributing to a general understanding of the statistical and computational *complexity* of data-driven dynamic decision making, which parallels our understanding of supervised machine learning and also accounts for the crucial roles of model structures and constraints for decision making.

Specifically, the thesis consists of three parts.

Part I of this thesis (Chapters 2 to 4) develops methodologies that *automatically* translate advances in supervised learning into effective dynamic decision making. Focusing on contextual bandits, a core class of online decision-making problems, we present the first optimal and efficient reduction from contextual bandits to offline regression. A remarkable consequence of our results is that advances in offline regression immediately translate to contextual bandits, statistically and computationally. We illustrate the advantages of our results through new guarantees in complex operational environments and experiments on real-world datasets. We also extend our results to more challenging setups, including reinforcement learning in large state spaces. Beyond the positive results, we establish new fundamental limits for general, unstructured reinforcement learning, emphasizing the importance of problem structures in reinforcement learning.

Part II of this thesis (Chapter 5) develops a framework that incorporates offline data into online decision making, motivated by practical challenges in business and operations. In the context of dynamic pricing, the framework allows us to rigorously characterize the *value of data* and the *synergy between online and offline learning* in data-driven decision making. The theory provides important insights for practice.

Part III of this thesis (Chapters 6 and 7) studies classical online decision-making problems in new settings where the decision maker may face a variety of long-term constraints. Such constraints are motivated by societal and operational considerations, and may limit the decision maker's ability to switch between actions, consume resources, or query accumulated data. We characterize the statistical and computational consequences brought by such long-term constraints, i.e., how the complexity of the problem changes with respect to different levels of constraints. The results provide precise characterizations on various intriguing trade-offs in data-driven dynamic decision making.

## 1.1 Overview of Part I (Chapters 2 to 4)

Machine learning, from its foundation, has largely focused on *(offline) supervised learning*, i.e., making *predictions* based on *passively observed, well-distributed* data. Many modern decision-making tasks (e.g., online product recommendation, personalized medicine), however, require making dynamic, effective *decisions* based on *sequential* data which *critically depend on the (limited) feedback of the decisions*. Such discrepancy presents significant challenges in the development of more powerful and impactful data-driven decision-making systems, both in theory and practice.

The stream of works in Part I of this thesis seeks to address the complications of data-driven decision making beyond offline supervised learning, and provide practical algorithms to systematically empower data-driven decision making by utilizing the advances in offline supervised learning. Fundamental limits are established when positive results are not possible. The results in this stream of works resolve several important open problems in the (broad) field of reinforcement learning (RL).

**Chapter 2: The first optimal reduction from general contextual bandits to offline regression** The contextual bandit problem is a fundamental framework for online data-driven decision making (as well as an important special case of online RL), with diverse applications ranging from electronic commerce to healthcare; see Li et al. (2010), Tewari and Murphy (2017) for illustrations on its practical importance. Focusing on contextual bandits, Chapter 2 develops general, principled approaches that enable the simple *plug-in* of *any* efficient offline supervised learning estimators to improve the statistical and computational efficiency of contextual bandit algorithms, in a provably optimal manner.

In more detail, Chapter 2 studies the general (stochastic) contextual bandit problem under the realizability assumption, i.e., the expected reward, as a function of contexts and actions, belongs to a general function class $\mathcal{F}$. Building on an intriguing algorithmic strategy called *inverse gap weighting* (Abe and Long 1999, Foster and Rakhlin 2020), we present a fast and simple algorithm for the general contextual bandit problem, which utilizes access to an *offline regression oracle* that is capable of solving the offline prediction/estimation problem associated with the function class $\mathcal{F}$. We show that the algorithm achieves the statistically optimal regret with only $O(\log T)$ calls to the offline regression oracle across all $T$ rounds (whenever the offline regression oracle attains the optimal offline estimation error). The

number of oracle calls can be further reduced to $O(\log \log T)$ if $T$ is known in advance. Our results provide the first universal and optimal reduction from contextual bandits to offline regression, solving an important open problem in the contextual bandit literature. A direct consequence of our results is that any advances in offline regression immediately translate to contextual bandits, statistically and computationally. This leads to faster algorithms and improved regret guarantees for broader classes of contextual bandit problems.

Chapter 2 is based on the following paper:

- David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3), 1904-1931, 2022.

**Chapter 3: Refined (more adaptive) algorithms and guarantees, with extensions to RL** In the classical multi-armed bandit problem, *instance-dependent* algorithms attain improved performance on "easy" problems with a gap between the best and second-best arm. Such adaptive guarantees are favored in various practical applications, where the potentially "nice" structures of the underlying models can be utilized to accelerate learning. Are similar guarantees possible for contextual bandits? In Chapter 3 (a follow-up work of both (Foster and Rakhlin 2020) and (Simchi-Levi and Xu 2022)), we extend the inverse gap weighting techniques and the general analysis based on offline regression oracles to obtain refined instance-dependent algorithms and guarantees for contextual bandits. We conduct extensive experiments on over 500 real-world datasets, and find that the refined algorithms typically enjoy superior performance (compared with existing benchmarks), especially on challenging datasets with many actions.

Beyond the algorithmic contributions, we introduce a family of complexity measures that are both sufficient and necessary for contextual bandit models to allow for sharp instance-dependent guarantees. Turning our focus to reinforcement learning with function approximation, we develop new oracle-efficient algorithms for (structured) reinforcement learning with rich observations that obtain optimal gap-dependent sample complexity.

Chapter 3 is based on the following paper:

- Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

(Extended abstract appeared in Conference on Learning Theory 2021)

**Chapter 4: Fundamental limits for offline RL with value function approximation** The previous examples illustrate that for certain classes of data-driven decision-making problems (which are special cases of structured RL problems), the simple/nice structures of the underlying models allow us to reduce the decision-making problems to offline supervised learning problems (and to achieve refined guarantees when possible). The structures of models are in fact crucial — in a general RL problem where the underlying Markov decision process (MDP) does not admit simple or nice structures, RL can be fundamentally harder than supervised learning.

Recently, there is a growing interest in establishing *information-theoretic* hardness results for RL, which can help us understand what elements of RL could make it fundamentally difficult. Chapter 4 contributes to this research direction by establishing a strong hardness result for *offline RL*; the offline RL setting is different from the online RL setting discussed previously, but is equally important as it finds broad applications in safety-critical domains like healthcare and autonomous driving. Our result shows that in the value function approximation setting, offline (unstructured) RL is fundamentally harder than supervised learning, resolving a well-known open problem in the field (Chen and Jiang 2019). Technically, this negative result is not directly comparable to the positive results discussed before (due to the different RL settings). But conceptually, it highlights the insights that the structures of underlying models are crucial to make data-driven decision-making problems tractable (and to make the powerful reduction to supervised learning possible).

Chapter 4 is based on the following paper:

- Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.
  (Extended abstract appeared in Conference on Learning Theory 2022)

The importance of model structures also motivates me to utilize the expertise of operation research to advance the development of reinforcement learning, which I will describe in Chapter 8 (future research).

## 1.2 Overview of Part II (Chapter 5)

A central goal of the previous research theme (Section 1.1) is to make online decision making as easy as offline supervised learning. In the second research theme, we aim to *synergize* online decision making and offline supervised learning to improve upon both.

As we mentioned, offline supervised learning deals with the problem of finding a predictive function based on the entire training data set. In contrast to the offline learning setting where the entire training data set is directly available before the algorithm is applied, online decision making deals with a setting where data become available in a sequential manner that may depend on the actions taken by the algorithm. While offline learning assumes access to offline data (but not online data) and online decision making assumes access to online data (but not offline data), in reality, a broad class of real-world problems incorporate both aspects: there is an offline historical data set (based on historical actions) at the time that the decision maker starts an online decision-making process.

**Chapter 5: Online pricing with offline data — new framework and insights** Currently, there is no standard framework for the above type of "hybrid" learning problems, as classical online and offline learning theory have different settings and goals. While establishing a framework that bridges all aspects of offline learning and online decision making is generally a very complicated task, in Chapter 5, we propose a new framework that bridges the gap between offline learning and online decision making in the context of dynamic pricing — a representative revenue management setting. As we will discuss in Chapter 5, our framework captures the essence of many dynamic pricing problems that sellers face in reality, and is highly relevant to revenue management practice.

Through our framework, we characterize the joint effect of the *size*, *location* and *dispersion* of the offline data on the optimal *instance-dependent* regret of the online pricing process. Specifically, the *size*, *location* and *dispersion* of the offline data are measured by the number of historical samples, the distance between the average historical price and the optimal price, and the standard deviation of the historical prices, respectively. We also design an adaptive algorithm to achieve the optimal instance-dependent regret. Our results reveal surprising transformations of the optimal regret rate with respect to the size of the offline data, which we refer to as *phase transitions* — the phenomena provide insights on the *value of (more) offline data*. In addition, our results demonstrate that the location and dispersion of the

offline data have an intrinsic effect on the optimal regret, and we quantify this effect via the *inverse-square law* — the law provides insights on the *synergy of online and offline learning*. Numerical experiments demonstrate the promising empirical performance of our algorithms, as well as the value of pre-existing offline data for dynamic pricing.

Chapter 5 is based on the following paper:

- Jinzhi Bu, David Simchi-Levi, and Yunzong Xu. Online pricing with offline data: Phase transition and inverse square law. *Management Science*, 68(12), 8568-8588, 2022. (Preliminary version appeared in International Conference on Machine Learning 2020)

## 1.3 Overview of Part III (Chapters 6 to 7)

Another research theme of this thesis is to study classical online decision-making problems in new settings where the decision-making process is subject to a variety of long-term budget constraints. Such budget constraints are motivated by operational practice, and may limit the policy's ability to switch between actions, consume resources, or query accumulated data. We are particularly interested in the *statistical consequences* brought by such budget constraints, i.e., how the *statistical complexity* (measured by the optimal regret rate) of the problem changes with respect to different budget levels.

**Chapter 6: Phase transitions in bandits with switching constraints** In Chapter 6, we consider the classical stochastic multi-armed bandit problem with a constraint that limits the total cost incurred by switching between actions to be no larger than a given switching budget. This model, referred to as *bandits with switching constraints*, is highly relevant to real-word applications where there are strict limits on the learner's switching behavior. A concrete application in practice is dynamic pricing with demand learning, where sellers often limit the number of price changes — sellers limit the number of price changes either because of implementation constraints, or for fear of confusing customers and receiving negative customer feedback.

We prove matching upper and lower bounds on the optimal (i.e., minimax) regret, and provide efficient rate-optimal algorithms. Surprisingly, the optimal regret of this problem exhibits a non-conventional growth rate in terms of the time horizon and the number of arms. Consequently, we discover surprising *phase transitions* regarding how the optimal

regret rate changes with respect to the switching budget: when the number of arms is fixed, there are equal-length phases, where the optimal regret rate remains (almost) the same within each phase and exhibits abrupt changes between phases; when the number of arms grows with the time horizon, such abrupt changes become subtler and may disappear, but a generalized notion of phase transitions involving certain new measurements still exists. The results enable us to fully characterize the trade-off between the regret rate and the incurred switching cost in the stochastic multi-armed bandit problem, contributing new insights to this fundamental problem. Under the general switching cost structure, the results reveal interesting connections between bandit problems and graph traversal problems, such as the shortest Hamiltonian path problem.

Numerical experiments demonstrate the practicality and effectiveness of our algorithms.

Chapter 6 is based on the following paper:

- David Simchi-Levi and Yunzong Xu. Phase transitions in bandits with switching constraints. *Management Science*, forthcoming, 2023.
  (Preliminary version appeared in Neural Information Processing Systems 2019)

**Chapter 7: Extensions to bandits with knapsacks and network revenue management** In Chapter 7, we consider an additional type of budget constraints — the (multi-dimensional) knapsack constraints introduced in the *blind network revenue management* and the *bandits with knapsacks* frameworks. We generalize previous results by considering the "integrated" impact of switching constraints and knapsack constraints on the stochastic bandit problem, and obtain interesting findings: a piecewise-constant function of the switching budget proves to completely characterize the optimal regret rate, which surprisingly depends on the dimension of knapsack constraints.

Chapter 7 is based on the following paper:

- David Simchi-Levi, Yunzong Xu, and Jinglong Zhao. Blind network revenue management and bandits with knapsacks under limited switches. *Available at SSRN 3479477*, 2019.
  (Latest version revised in 2023)

# Chapter 2

# Optimal and Efficient Reduction from Contextual Bandits to Offline Regression

## 2.1 Introduction

The contextual bandit problem is a fundamental framework for online decision making and interactive machine learning, with diverse applications ranging from healthcare (Tewari and Murphy 2017, Bastani and Bayati 2020) to electronic commerce (Li et al. 2010, Agarwal et al. 2016). It has been extensively studied in computer science, operations research, and statistics literature.

Broadly speaking, approaches to contextual bandits can be classified into two categories (see Foster et al. 2018): *realizability-based* approaches which rely on weak or strong assumptions on the model representation, and *agnostic* approaches which are completely model-free. While many different contextual bandit algorithms (realizability-based or agnostic) have been proposed over the past twenty years, most of them suffer from either theoretical or practical issues (see Bietti et al. 2018). Existing realizability-based algorithms building on upper confidence bounds (e.g., Filippi et al. 2010, Abbasi-Yadkori et al. 2011, Chu et al. 2011, Li et al. 2017) and Thompson sampling (e.g., Agrawal and Goyal 2013, Russo et al. 2018) rely on strong assumptions on the model representation and are only tractable for specific parametrized families of models like generalized linear models. Meanwhile, agnostic algorithms that make no assumption on the model representation (e.g., Dudík et al. 2011, Agarwal et al. 2014) may lead to overly conservative exploration in practice (Bietti et al. 2018), and their reliance on an *offline cost-sensitive classification oracle* as a subroutine

typically causes implementation difficulties as the oracle itself is computationally intractable in general. At this moment, designing a provably optimal contextual bandit algorithm that is applicable for large-scale real-world deployments is still widely deemed a very challenging task (see Agarwal et al. 2016, Foster and Rakhlin 2020).

Recently, Foster et al. (2018) propose an approach to solve contextual bandits with general model representations (i.e., general function classes) using an *offline regression oracle* — an oracle that can typically be implemented efficiently and has wide availability for numerous function classes due to its core role in modern machine learning. Specifically, motivated by the work of Krishnamurthy et al. (2019) which initiates such a key idea, Foster et al. (2018) assume access to a *weighted least squares regression oracle*, which is deemed highly practical as it has a strongly convex loss function and is amenable to gradient-based methods. As Foster et al. (2018) point out, designing offline-regression-oracle-based algorithms is a promising direction for making contextual bandits practical, as they seem to combine the advantages of both realizability-based and agnostic algorithms: they are general and flexible enough to work with any given function class, while using a more realistic and reasonable oracle than the computationally-expensive classification oracle. Indeed, according to multiple experiments and extensive empirical evaluations conducted by Foster et al. (2018) and Bietti et al. (2018), the algorithm of Foster et al. (2018) "works the best overall" among existing contextual bandit approaches.

Despite its empirical success, the algorithm of Foster et al. (2018) is, however, theoretically sub-optimal — it could incur $\widetilde{\Omega}(T)$ regret in the worst case. Whether the optimal regret of contextual bandits can be attained via an offline-regression-oracle-based algorithm is listed as an open problem in Foster et al. (2018). In fact, this problem has been open to the bandit community since 2012 — it dates back to Agarwal et al. (2012), where the authors propose a computationally *inefficient* contextual bandit algorithm that achieves the optimal $\widetilde{O}(\sqrt{KT \log |\mathcal{F}|})$ regret for a general *finite* function class $\mathcal{F}$, but leave designing computationally tractable algorithms as an open problem.

More recently, Foster and Rakhlin (2020) propose an algorithm that achieves the optimal regret for contextual bandits by assuming access to an *online regression oracle* (which is not an offline oracle and has to work with an adaptive adversary). Their finding that contextual bandits can be reduced to online regression is novel and important, and their result is also very general: it requires only the minimal realizability assumption, and holds true even when

the contexts are chosen adversarially. However, compared with access to an offline regression oracle, access to an online regression oracle is a much stronger (and relatively restrictive) assumption. In particular, practical algorithms for online regression are only known for specific function classes. Whether the optimal regret of contextual bandits can be attained via a reduction to an offline regression oracle is listed as an open problem again in Foster and Rakhlin (2020).

### 2.1.1 Our Contributions

In this paper, we study the following question repeatedly mentioned in the contextual bandit literature (Agarwal et al. 2012, Foster et al. 2018, Foster and Rakhlin 2020): *Is there an offline-regression-oracle-based algorithm that achieves the optimal regret for general (stochastic) contextual bandits?*

We answer this question in the affirmative by providing the first optimal black-box reduction from contextual bandits to offline regression, with only the minimal realizability assumption. The significance of this result is that it reduces contextual bandits, a prominent online decision-making problem, to offline regression, a very basic and common supervised learning task that serves as the building block of modern machine learning. A consequence of this result is that any advances in solving offline regression problems translate to contextual bandits, statistically and computationally. Note that such online-to-offline reductions are highly nontrivial for online learning problems in general; in fact, a generic reduction from fully adversarial online learning to offline learning is not possible (Hazan and Koren 2016).

Our reduction is accomplished by providing a surprisingly fast and simple algorithm (which builds on and connects the approaches of Abe and Long 1999, Agarwal et al. 2014, Foster and Rakhlin 2020) and proving strong theoretical guarantees for this algorithm. For a general finite function class $\mathcal{F}$, our algorithm achieves the optimal $\widetilde{O}(\sqrt{KT\log|\mathcal{F}|})$ regret with only $O(\log T)$ calls to an offline least squares regression oracle over $T$ rounds. The number of oracle calls can be further reduced to $O(\log\log T)$ if $T$ is known. Notably, this can be understood as a "triply exponential" improvement over previous work: (i) compared with the previously known regret-optimal algorithm of Agarwal et al. (2012) for this setting, which requires enumerating over $\mathcal{F}$ at each round, our algorithm accesses the function class only through a least squares regression oracle, thus typically avoids an exponential computational cost at each round; (ii) compared with the classification-oracle-based algorithm

of Agarwal et al. (2014) which requires $\widetilde{O}(\sqrt{KT/\log|\mathcal{F}|})$ calls to a computationally expensive classification oracle, our algorithm requires only $O(\log T)$ calls to a simple regression oracle, which implies an exponential improvement (in the number of oracle calls) over existing provably optimal oracle-efficient algorithms, even when we ignore the difference between regression and classification oracles; (iii) when the number of rounds $T$ is known in advance, our algorithm can further reduce the number of oracle calls to $O(\log\log T)$, which is an exponential improvement by itself. Our algorithm is thus highly practical; see Table 2.1 for a detailed comparison with existing work.

Table 2.1: Algorithms' performance with general finite $\mathcal{F}$ and i.i.d. contexts. Advantages are marked in bold.

| Algorithm | Regret rate | Computational complexity |
|---|---|---|
| `Regressor Elimination` (Agarwal et al. 2012) | **optimal** | $\Omega(\lvert\mathcal{F}\rvert)$ intractable |
| `ILOVETOCONBANDITS` (Agarwal et al. 2014) | **optimal** | $\widetilde{O}(\sqrt{KT/\log\lvert\mathcal{F}\rvert})$ calls to an offline classification oracle |
| `RegCB` (Foster et al. 2018) | suboptimal | $O(T^{3/2})$ calls to an offline least squares oracle |
| `SquareCB` (Foster and Rakhlin 2020) | **optimal** | $O(T)$ calls to an online regression oracle |
| `FALCON` / `FALCON+` (this paper) | **optimal** | $\boldsymbol{O(\log T)}$ or $\boldsymbol{O(\log\log T)}$ calls to an **offline regression** oracle* |

* Not restricted to least squares; strictly easier to solve than online regression. See §2.3 for details.

We then extend all of the above results to the general setting where (i) the function class $\mathcal{F}$ can be infinite, and (ii) the offline regression oracle is not necessarily a least squares oracle. For this general setting, our reduction can be stated as follows: for any function class $\mathcal{F}$, given an arbitrary offline regression oracle with an arbitrary offline *estimation error* (or *excess risk*) guarantee, we provide a fast and simple contextual bandit algorithm whose regret can be bounded by a function of the offline estimation error, through only $O(\log T)$ calls (or $O(\log\log T)$ calls if $T$ is known) to the offline regression oracle. We show that our algorithm is statistically optimal as long as the offline regression oracle is statistically optimal. Notably, the above results provide a universal and optimal "converter" from results of offline regression with general function classes to results of contextual bandits with general function classes. This leads to improved algorithms with tighter regret bounds for many existing contextual

bandit problems, as well as practical algorithms for many new contextual bandit problems, e.g., contextual bandits with certain types of neural networks, and contextual bandits with heavy-tailed rewards.

The analysis of our algorithm is particularly interesting. Unlike existing analysis of other realizability-based algorithms in the literature, we do not directly analyze the decision outcomes of our algorithm — instead, we find a dual interpretation of our algorithm as sequentially maintaining a *dense* distribution over *all* (possibly *improper*) policies, where a policy is defined as a deterministic decision function mapping contexts to actions. We analyze how the realizability assumption enables us to establish uniform-convergence-type results for some *implicit* quantities in the universal policy space, regardless of the huge capacity of the universal policy space. Note that while the dual interpretation itself is not easy to compute in the universal policy space, it is only applied for the purpose of analysis and has nothing to do with our original algorithm's implementation. Through this lens, we find that our algorithm's dual interpretation satisfies a series of sufficient conditions for optimal contextual bandit learning. Our identified sufficient conditions for optimal contextual bandit learning in the universal policy space build on the previous work of Dudík et al. (2011), Agarwal et al. (2012) and Agarwal et al. (2014) — the first one is colloquially referred to as the "monster paper" by its authors due to its complexity, and the third one is titled as "taming the monster" by its authors due to its improved computational efficiency. Since our algorithm achieves all the conditions required for regret optimality in the universal policy space in a completely *implicit* way (which means that all the requirements are automatically satisfied without explicit computation), our algorithm comes with significantly reduced computational cost compared with previous work (thanks to the realizability assumption), and we thus title our paper as "bypassing the monster."

Overall, our algorithm is fast and simple, and our analysis is quite general. We believe that the algorithm has the potential to be implemented on a large scale, and our analysis may contribute to deeper understanding of contextual bandits. We will go over the details in the rest of this article.

### 2.1.2 Learning Model

The general stochastic contextual bandit problem can be stated as follows. Let $\mathcal{A}$ be a finite set of $K$ actions and $\mathcal{X}$ be an arbitrary space of contexts (e.g., a feature space). The

interaction between the learner and nature happens over $T$ rounds, where $T$ is possibly unknown. At each round $t$, nature samples a context $x_t \in \mathcal{X}$ and a context-dependent reward vector $r_t \in [0,1]^{\mathcal{A}}$ according to a fixed but unknown (joint) distribution $\mathcal{D}$, with component $r_t(a)$ denoting the reward for action $a \in \mathcal{A}$; the learner observes $x_t$, picks an action $a_t \in \mathcal{A}$, and observes the reward for her action $r_t(a_t)$. Notably, the learner's reward $r_t(a_t)$ depends on both the context $x_t$ and her action $a_t$, and is a partial observation of the full reward vector $r_t$. Depending on whether there is an assumption about nature's reward model, prior literature studies the contextual bandit problem in two different but closely related settings.

**Agnostic setting.** Let $\Pi \subset \mathcal{A}^{\mathcal{X}}$ be a class of *policies* (i.e., decision functions) that map contexts $x \in \mathcal{X}$ to actions $a \in \mathcal{A}$, and $\pi_\star = \arg\max_{\pi \in \Pi} \mathbb{E}_{(x,r)\sim\mathcal{D}}[r(\pi(x))]$ be the optimal policy in $\Pi$ that maximizes the expected reward. The learner's goal is to compete with the (in-class) optimal policy $\pi_\star$ and minimize her (empirical cumulative) *regret* after $T$ rounds, which is defined as

$$\sum_{t=1}^{T} (r_t(\pi_\star(x_t)) - r_t(a_t)).$$

The above setting is called *agnostic* in the sense that it imposes no assumption on nature.

**Realizable setting.** Let $\mathcal{F}$ be a class of *predictors* (i.e., reward functions), where each predictor is a function $f : \mathcal{X} \times \mathcal{A} \to [0,1]$ trying to approximate the *true reward function* $f^\star$ defined by $f^\star(x,a) = \mathbb{E}[r_t(a) \mid x_t = x]$, $\forall x \in \mathcal{X}, a \in \mathcal{A}$. The standard realizability assumption (Chu et al. 2011, Agarwal et al. 2012, Foster et al. 2018) is as follows:

**Assumption 2.1** (Realizability)**.** *The true reward function is contained in $\mathcal{F}$, i.e., $f^\star \in \mathcal{F}$.*

Given a predictor $f \in \mathcal{F}$, the associated reward-maximizing policy $\pi_f$ always picks the action with the highest predicted reward, i.e., $\pi_f(x) = \arg\max_{a \in \mathcal{A}} f(x,a)$. The learner's goal is to compete with the *globally optimal policy* $\pi_{f^\star}$ and minimizes her (empirical cumulative) *regret* after $T$ rounds, which is defined as

$$\sum_{t=1}^{T} (r_t(\pi_{f^\star}(x_t)) - r_t(a_t)).$$

The above setting is called *realizable* in the sense that it assumes that nature can be well-specified by a predictor in $\mathcal{F}$. In this paper, we consider a general $\mathcal{F}$, which can be a class of parametric functions, nonparametric functions, regression trees, neural networks, etc.

We make some remarks on the above two settings from a pure modeling perspective. First,

the agnostic setting does not require realizability and is more general than the realizable setting. Indeed, given any function class $\mathcal{F}$, one can construct an *induced* policy class $\Pi_{\mathcal{F}} = \{\pi_f \mid f \in \mathcal{F}\}$, thus any realizable contextual bandit problem can be reduced to an agnostic contextual bandit problem. Second, the realizable setting has its own merit, as the additional realizability assumption enables stronger performance guarantees: once the realizability assumption holds, the learner's competing policy $\pi_{f^\star}$ is guaranteed to be *globally optimal* (i.e., no policy can be better than $\pi_{f^\star}$), thus small regret necessarily means large total reward. By contrast, in the no-realizability agnostic setting, the "optimal policy in $\Pi$" is not necessarily an effective policy if there are significantly more effective polices outside of $\Pi$. More comparisons between the two settings regarding theoretical tractability, computational efficiency and practical implementability will be provided in §2.1.3.

### 2.1.3 Related Work

Contextual bandits have been extensively studied for two decades; see Chapter 18 of Lattimore and Szepesvári (2020) and Chapter 8 of Slivkins (2019) for detailed surveys. Here we mention some important and closely related work.

#### Agnostic Approaches

Papers studying contextual bandits in the agnostic setting aim to design general-purpose and computationally-tractable algorithms that are provably efficient for any given policy class $\Pi$ while avoiding the computational complexity of enumerating over $\Pi$ (as the size of $\Pi$ is usually extremely large). The primary focus of prior literature is on the case of general finite $\Pi$, as this is the starting point for further studies of infinite (parametric or nonparametric) $\Pi$. For this case, the EXP4-family algorithms (Auer et al. 2002b, McMahan and Streeter 2009, Beygelzimer et al. 2011) achieve the optimal $O(\sqrt{KT \log |\Pi|})$ regret but requires $\Omega(|\Pi|)$ running time at each round, which makes the algorithms intractable for large $\Pi$. In order to circumvent the $\Omega(|\Pi|)$ running time barrier, researchers (e.g., Langford and Zhang 2008, Dudík et al. 2011, Agarwal et al. 2014) restrict their attention to *oracle-based* algorithms that access the policy space only through an *offline optimization oracle* — specifically, an

offline cost-sensitive classification oracle that solves

$$\arg\max_{\pi\in\Pi}\sum_{s=1}^{t}\tilde{r}_s(\pi(x_s)) \tag{2.1}$$

for any given sequence of context and reward vectors $(x_1,\tilde{r}_1),\cdots,(x_t,\tilde{r}_t)\in\mathcal{X}\times\mathbb{R}_+^{\mathcal{A}}$. An *oracle-efficient* algorithm refers to an algorithm whose number of oracle calls is polynomial in $T$ over $T$ rounds.

The first provably optimal oracle-efficient algorithm is the `Randomized UCB` algorithm of Dudík et al. (2011), which achieves the optimal regret with $\widetilde{O}(T^6)$ calls to the cost-sensitive classification oracle. A breakthrough is achieved by the `ILOVETOCONBANDITS` algorithm in the celebrated work of Agarwal et al. (2014), where the number of oracle calls is significantly reduced to $\widetilde{O}(\sqrt{KT/\log|\Pi|})$. The above results are fascinating in theory because they enable a "online-to-offline reduction" from contextual bandits to cost-sensitive classification, which is highly non-trivial for online learning problems in general. However, the practicability of the above algorithms is heavily restricted due to their reliance on the cost-sensitive classification oracle (2.1), as this task is computationally intractable even for simple policy classes (Klivans and Sherstov 2009, Agrawal and Devanur 2016) and typically involves solving NP-hard problems. As a result, the practical implementations of the above classification-oracle-based algorithms typically resort to heuristics (Agarwal et al. 2014, Bietti et al. 2018). Moreover, the above algorithms are memory hungry: since they must feed *augmented* versions of the dataset (rather than the original version of the dataset) into the oracle, they have to repeatedly create auxiliary data and store them in memory. Therefore, these approaches may not perform well in practice (Bietti et al. 2018), and are generally impractical for large-scale real-world deployments (Foster et al. 2018, Foster and Rakhlin 2020).

### Realizibility-Based Approaches

In contrast to the agnostic setting where research primarily focuses on designing general-purpose algorithms that work for any given $\Pi$, a majority of research in the realizable setting tends to design specialized algorithms that work well for a particular parametrized family of $\mathcal{F}$. Two of the dominant strategies for the realizable setting are upper confidence bounds (e.g., Filippi et al. 2010, Abbasi-Yadkori et al. 2011, Chu et al. 2011, Li et al. 2017, 2019) and Thompson sampling (e.g., Agrawal and Goyal 2013, Russo et al. 2018).

While these approaches have been practically successful in several scenarios (Li et al. 2010), their theoretical guarantees and computational tractability critically rely on their strong assumptions on $\mathcal{F}$, which restrict their usage in other scenarios (Bietti et al. 2018).

To our knowledge, Agarwal et al. (2012) is the first paper studying contextual bandits with a general finite $\mathcal{F}$, under the minimal realizability assumption. They propose a elimination-based algorithm, called `Regressor Elimination`, that achieves the optimal $\widetilde{O}(\sqrt{KT \log |\mathcal{F}|})$ regret. However, their algorithm is computational inefficient, as it enumerates over the whole function class and requires $\Omega(|\mathcal{F}|)$ computational cost at each round (note that the size of $\mathcal{F}$ is typically extremely large). The computational issues of Agarwal et al. (2012) are addressed by Foster et al. (2018), who propose an oracle-efficient contextual bandit algorithm `RegCB`, which always accesses the function class through a weighted least squares regression oracle that solves

$$\arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t} w_s (f(x_s, a_s) - y_s)^2 \tag{2.2}$$

for any given input sequence $(w_1, x_1, a_1, y_1), \cdots, (w_t, x_t, a_t, y_t) \in \mathbb{R}_+ \times \mathcal{X} \times \mathcal{A} \times \mathbb{R}$. As Foster et al. (2018) mention, the above oracle can often be solved efficiently and is very common in machine learning practice — it is far more reasonable than the cost-sensitive classification oracle (2.1). However, unlike `Regressor Elimination`, the `RegCB` algorithm is not minimax optimal — its worst-case regret could be as large as $\widetilde{\Omega}(T)$. Whether the optimal $\widetilde{O}(\sqrt{KT \log |\mathcal{F}|})$ regret is attainable for an offline-regression-oracle-based algorithm remains unknown in the literature.

More recently, Foster and Rakhlin (2020) propose an algorithm that achieves the optimal regret for contextual bandits using an *online* regression oracle. Their algorithm, called `SquareCB`, builds on the `A`/`BW` algorithm of Abe and Long (1999) (see also the journal version Abe et al. 2003) originally developed for linear contextual bandits — specifically, `SquareCB` replaces the "Widrow-Hofff predictor" used in the `A`/`BW` algorithm by a general online regression predictor, then follows the same probabilistic action selection strategy as the `A`/`BW` algorithm. Foster and Rakhlin (2020) show that by using this simple strategy, contextual bandits can be reduced to online regression in a black-box manner. While the implication that contextual bandits are no harder than online regression is important and insightful, online regression with a general function class itself is a challenging problem. Note that an online regression oracle has to provide robust guarantees for an arbitrary data sequence generated by an

adaptive adversary, which may cause implementation difficulties when the function class $\mathcal{F}$ is complicated — while there is a beautiful theory characterizing the minimax regret rate of online regression with general function classes (Rakhlin and Sridharan 2014), to our knowledge computational efficient algorithms are only known for specific function classes. For example, consider the case of a general finite $\mathcal{F}$, the online algorithm given by Rakhlin and Sridharan (2014) actually requires $\Omega(|\mathcal{F}|)$ computational cost at each round. Therefore, beyond the existing results of Foster and Rakhlin (2020), a more thorough "online-to-offline reduction" from contextual bandits to offline regression is highly desirable.

### 2.1.4 Technical Challenges and Our Approach

Before we proceed to present our results, we would like to illustrate the key technical hurdles of using offline regression oracles to achieve the optimal regret for contextual bandits. We will then briefly explain how our approach overcomes these technical hurdles.

As was pointed out before, three excellent papers Agarwal et al. (2012), Foster et al. (2018), Foster and Rakhlin (2020) have made important progress towards solving contextual bandits via regression approaches. Understanding the gap between the existing results and our desired result is important for understanding the key technical hurdles. Below we discuss three challenges.

**Computational hurdle.** Agarwal et al. (2012) propose a provably optimal but computational inefficient algorithm for contextual bandits with a general finite $\mathcal{F}$. At each round $t$, their algorithm maintains a subset $\mathcal{F}_t \subset \mathcal{F}$ based on successive elimination and solves a complicated optimization problem over $\mathcal{F}_t$. Here, the key difficulty of using an offline regression oracle is that one cannot reformulate the complicated optimization problem over $\mathcal{F}_t$ to a simple optimization problem like least squares regression, as the objective function is far more complicated than a sum of squares. This is also why using a square loss regression oracle is more challenging than using the offline cost-sensitive classification oracle (2.1) — one can understand the latter as a 0-1 loss oracle.

**Statistical hurdle associated with constructing confidence bounds.** Foster et al. (2018) propose a computationally efficient confidence-bounds-based algorithm using an offline weighted least squares oracle. However, their algorithm only has statistical guarantees under some strong distributional assumptions. An important reason is that confidence-bounds-based algorithms typically rely on the ability of constructing *shrinking* confidence intervals on *each*

*context*. While this is possible for a simple $\mathcal{F}$ like a linear class, it is impossible for a general $\mathcal{F}$. Here, the difficulty originates from the fact that all the statistical learning guarantees for offline regression with a general $\mathcal{F}$ require one to take an expectation over contexts. In other words, effective per-context statistical guarantees are generally impossible for an offline regression oracle.

**Statistical hurdle associated with analyzing dependent actions.** Foster and Rakhlin (2020) propose an optimal and efficient contextual bandit algorithm assuming access to an online regression oracle, which is quite different from an offline regression oracle. Statistically, the difference between offline and online regression oracles is that, offline regression oracles only assume statistical guarantees for an i.i.d. data sequence (see §2.3 for our definition of a general offline regression oracle), while online regression oracles assume statistical guarantees for an arbitrary data sequence possibly generated by an adaptive adversary. Evidently, access to an online regression oracle is a stronger assumption than access to an offline regression oracle. As Foster and Rakhlin (2020) mention, their algorithm requires an online regression oracle because "the analysis critically uses that the regret bound (of the online regression oracle) holds when the actions $a_1, \ldots, a_T$ are chosen adaptively, since actions selected in early rounds are used by `SquareCB` to determine the action distribution at later rounds." That is, the technical hurdle of using an offline regression oracle here is that the algorithm's action sequence is not i.i.d. — since offline regression oracles are designed for i.i.d. data, it is unclear how one can deal with dependent actions when one only has access to an offline regression oracle. We note that this hurdle lies at the heart of the "exploration-exploitation trade-off" — essentially, any efficient algorithm's actions must be highly dependent, as they are simultaneously used for exploration and exploitation.

### Our Resolution

We address the three technical hurdles in §2.1.4 in a surprisingly elegant way. Specifically, we derive an algorithm that accesses the offline regression oracle in a mostly "naive" way, without constructing any explicit optimization problems or confidence bounds, thus gets around the first two hurdles simultaneously; further, we overcome the third hurdle by establishing a framework to analyze our algorithm and prove its statistical optimality — in particular, we face the complex dynamics of evolving dependent actions, but analyze them through a different lens (the "dual interpretation" in §2.4), and establish a series of sufficient conditions

for optimal contextual bandit learning under this lens. The final algorithm is simple, but the ideas behind it are highly non-trivial and are supported by novel analysis. The algorithmic details will be presented in §2.2 and §2.3 and the key ideas will be explained in §2.4.

Our approach builds on (and reveals connections between) two lines of research in the contextual bandit literature: (i) a celebrated theory of optimal contextual bandit learning in the agnostic setting using a (seemingly unavoidable) classification oracle, represented by Dudík et al. (2011) (the "monster paper") and Agarwal et al. (2014) ("taming the monster"); (ii) a simple probabilistic selection strategy mapping the predicted rewards of actions to the probabilities of actions, pioneered by Abe and Long (1999) (see also Abe et al. 2003) and extended by Foster and Rakhlin (2020). In particular, we rethink the philosophy behind Dudík et al. (2011) and Agarwal et al. (2014), reform it with our own understanding of the value of realizability, and come up with a new idea of "bypassing" the classification oracle under realizability — our algorithm is essentially a consequence of this new idea; see §2.4.6. Interestingly, our derived algorithm turns out to use essentially the same probabilistic selection strategy as Abe and Long (1999) and Foster and Rakhlin (2020) — this is surprising, as the idea behind the derivation of our algorithm is very different from the ideas behind Abe and Long (1999) and Foster and Rakhlin (2020). This suggests that this simple probabilistic selection strategy might be more intriguing and more essential for bandits than previously understood, and we believe that it is worth further attention from the bandit community. We hope that our work, together with Abe and Long (1999) and Foster and Rakhlin (2020), can provide diverse perspectives on how to understand this strategy.

As a final remark, we emphasize that compared with each line of research that we mention above, our approach has new contributions beyond them which seem necessary for our arguments to hold. We will elaborate on such new contributions in the rest of our article.

### 2.1.5 Organization and Notation

The rest of the paper is organized as follows. For pedagogical reasons, we first present our results in the case of a general finite $\mathcal{F}$ in §2.2, where we introduce our algorithm and state its theoretical guarantees. In §2.3, we extend our results to the general setting and discuss several important consequences. In §2.4, we present our regret analysis and explain the ideas behind our algorithm. We conclude our paper in §2.5. All the proofs of our results are deferred to the appendix.

Throughout the paper, we use $O(\cdot)$ to hide constant factors, and $\widetilde{O}(\cdot)$ to hide polylog$(T)$ factors. Given $\mathcal{D}$, let $\mathcal{D}$ denote the marginal distribution over $\mathcal{X}$. We use $\sigma(Y)$ to denote the $\sigma$-algebra generated by a random variable $Y$, and use $\mathcal{B}(E)$ to denote the Borel $\sigma$-algebra on a set $E$. An *action selection kernel* $p : \mathcal{B}(\mathcal{A}) \times \mathcal{X} \to [0,1]$ is defined as a probability kernel such that $p(a \mid x)$ specifies the probability of selecting action $a \in \mathcal{A}$ given context $x \in \mathcal{X}$; let $\mathcal{P}$ be the space of all action selection kernels. We use $\mathbb{N}$ to denote the set of all positive integers, and $\mathbb{R}_+$ to denote the set of all non-negative real numbers. Without loss of generality, we assume that $|\mathcal{F}| \geq 4$.

## 2.2 Algorithm and Guarantees

Following previous work (Dudík et al. 2011, Agarwal et al. 2012, 2014), we start with the case of a general finite $\mathcal{F}$, as this is the starting point for further studies of an infinite $\mathcal{F}$. For this case, the "gold standard" is an algorithm that achieves $\widetilde{O}(\sqrt{KT \log |\mathcal{F}|})$ regret with the total number of oracle calls being polynomial/sublinear in $T$ (see Agarwal et al. 2012, Foster et al. 2018). As for the oracle, we assume access to the following *least squares regression oracle* that solves

$$\arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t} (f(x_s, a_s) - y_s)^2 \tag{2.3}$$

for any input sequence $(x_1, a_1, y_1), \cdots, (x_t, a_t, y_t) \in \mathcal{X} \times \mathcal{A} \times [0,1]$. Without loss of generality[1], we assume that the oracle (2.3) always returns the same solution for two identical input sequences. Note that the above least squares oracle (2.3) is a concrete optimization oracle and is simpler than the weighted one (2.2) assumed in Foster et al. (2018), as it does not need to consider the weights.

We remark that our reduction is not restricted to this setup — in §2.3, we will extend all our results to the general setting where both $\mathcal{F}$ and the offline regression oracle are generic. Still, the above setup is good for illustrating our results, and allows direct comparisons to the "gold standard."

[1] If the oracle is allowed to return a random solution (when there are multiple optimal solutions), then we can simply incorporate such randomness into the history when we define $\Upsilon_t$ in Appendix A.1.1, and all our proofs will still hold.

### 2.2.1 The Algorithm

We present our algorithm, "FAst Least-squares-regression-oracle CONtextual bandits" (FALCON), in Algorithm 2.1 (a generalized version of this algorithm will be provided in §2.3). The algorithm is very simple and follows the same general template as the A/BW algorithm of Abe and Long (1999) and the SquareCB algorithm of Foster and Rakhlin (2020), with the main difference lying in using a different oracle to generate predictions. We also add a few useful ingredients, including an epoch schedule and a changing learning rate. See the description of the algorithm below.

**Algorithm 2.1** FAst Least-squares-regression-oracle CONtextual bandits (FALCON)

**input** epoch schedule $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$, confidence parameter $\delta$, tuning parameter $c$

**for** epoch $m = 1, 2, \ldots$ **do**
  Let $\gamma_m = c\sqrt{K\tau_{m-1}/\log(|\mathcal{F}|\log(\tau_{m-1})m/\delta)}$ (for epoch 1, $\gamma_1 = 1$).
  Compute $\widehat{f}_m = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^{\tau_{m-1}} (f(x_t, a_t) - r_t(a_t))^2$ via the **offline least squares oracle**.
  **for** round $t = \tau_{m-1} + 1, \cdots, \tau_m$ **do**
    Observe context $x_t \in \mathcal{X}$.
    Compute $\widehat{f}_m(x_t, a)$ for each action $a \in \mathcal{A}$. Let $\widehat{a}_t = \max_{a\in\mathcal{A}} \widehat{f}_m(x_t, a)$. Define

$$p_t(a) = \begin{cases} \frac{1}{K+\gamma_m\left(\widehat{f}_m(x_t,\widehat{a}_t)-\widehat{f}_m(x_t,a)\right)}, & \text{for all } a \neq \widehat{a}_t, \\ 1 - \sum_{a\neq\widehat{a}_t} p_t(a), & \text{for } a = \widehat{a}_t. \end{cases}$$

    Sample $a_t \sim p_t(\cdot)$ and observe reward $r_t(a_t)$.

Our algorithm runs in an epoch schedule to reduce oracle calls, i.e., it only calls the oracle at certain pre-specified rounds $\tau_1, \tau_2, \tau_3, \ldots$. For $m \in \mathbb{N}$, we refer to the rounds from $\tau_{m-1}+1$ to $\tau_m$ as epoch $m$. As a concrete example, consider $\tau_m = 2^m$, then for any (possibly unknown) $T$, our algorithm runs in $O(\log T)$ epochs. As another example, when $T$ is known, consider $\tau_m = \left\lfloor 2T^{1-2^{-m}} \right\rfloor$, then our algorithm runs in $O(\log\log T)$ epochs. We allow very general epoch schedules; in particular, calling the oracle more frequently does not affect the regret analysis.

At the start of each epoch $m$, our algorithm makes two updates. First, it updates a (epoch-varying) learning rate $\gamma_m \simeq \sqrt{K\tau_{m-1}/\log(|\mathcal{F}|/\delta)}$, which aims to strike a balance between exploration and exploitation. Second, it computes a "greedy" predictor $\widehat{f}_m$ from $\mathcal{F}$ that minimizes the empirical square loss $\sum_{t=1}^{\tau_{m-1}} (f(x_t, a_t) - r_t(a_t))^2$. This predictor can be computed via a single call to the offline least squares regression oracle — notably, $\min_{f\in\mathcal{F}} \sum_{t=1}^{\tau_{m-1}} (f(x_t, a_t) - r_t(a_t))^2$ is almost the best way that we can imagine for our oracle

to be called, with no augmented data generated, no weights maintained, and no additional optimization problem constructed.

The decision rule in epoch $m$ is then completely determined by $\gamma_m$ and $\widehat{f}_m$. For each round $t$ in epoch $m$, given a context $x_t$, the algorithm uses $\widehat{f}_m$ to predict each action's reward and finds a greedy action $\widehat{a}_t$ that maximizes the predicted reward. Yet the algorithm does not directly select $\widehat{a}_t$ — instead, it randomizes over all actions according to a probabilistic selection strategy that picks each action other than $\widehat{a}_t$ with probability roughly inversely proportional to how much worse it is predicted to be as compared with $\widehat{a}_t$, as well as roughly inversely proportional to the learning rate $\gamma_m$. The effects of this strategy are twofold. First, *at each round*, by assigning the greedy action the highest probability and each non-greedy action a probability roughly inverse to the predicted reward gap, we ensure that the better an action is predicted to be, the more likely it will be selected. Second, *across different epochs*, by controlling the probabilites of non-greedy actions roughly inverse to the gradually increasing learning rate $\gamma_m$, we ensure that the algorithm "explores more" in the beginning rounds where the learning rate is small, and gradually "exploits more" in later rounds where the learning rate becomes larger — this is why we view our learning rate as a sequential balancer between exploration and exploitation.

**Algorithmic components and comparisons with literature.** `FALCON` is a very simple algorithm, and can be viewed as a combination of three algorithmic components: (i) an epoch schedule, (ii) the greedy use of an offline least squares regression oracle, and (iii) a probabilistic selection strategy that maps reward predictions to action probabilities, controlled by an epoch-varying learning rate. While each component alone is not new in the literature, the combination of the above three components has not been considered in the literature, and it is far from obvious that this particular combination should be effective. In fact, it is quite surprising that such a simple algorithm would work well for general contextual bandits. While there is definitely more to this algorithm than meets the eye (we will explain the essential idea behind `FALCON` in §2.4.5 and §2.4.6), let us first give a few quick comments on component (ii) and (iii), and compare them to existing literature.

We start from component (iii). As we mention before, the idea of mapping the predicted action rewards to action probabilities via an "inverse proportional to the gap" rule is not new: such a probabilistic selection strategy is firstly proposed by Abe and Long (1999) in their study of linear contextual bandits, and recently adopted by Foster and Rakhlin (2020)

in their reduction from contextual bandits to *online* regression. Compared with the existing strategy used in Abe and Long (1999) and Foster and Rakhlin (2020), the strategy that we use here has a notable difference: while the above two papers adopt a constant learning rate $\gamma$ that does not change in the running process of their algorithms, we appeal to an epoch-varying (or time-varying) learning rate $\gamma_m \simeq \sqrt{K\tau_{m-1}/\log(|\mathcal{F}|/\delta)}$ that gradually increases as our algorithm proceeds. This epoch-varying learning rate plays a important role in our statistical analysis, as the proof of our regret guarantee relies on an inductive argument which requires the learning rate to change carefully with respect to epochs and gradually increase over time; see §2.4.4.

*Remark.* While such an epoch-varying learning rate is not necessary when $T$ is known in advance and the oracle calls are "frequent" enough, an epoch-varying learning rate brings certain benefits to the algorithm: first, in the case of unknown $T$, it is required; second, in the case of known $T$, it is necessary whenever one seeks to control the total number of oracle calls within $o(\log T)$ (a fixed learning rate could lead to sub-optimal regret in this setting); third, in our analysis it always leads to tighter regret bounds with better dependence on logarithmic factors. As a result, it seems that an epoch-varying learning rate always dominates a fixed learning rate in our problem.

Component (ii) of our algorithm is particularly interesting. Indeed, our algorithm makes predictions in a surprisingly simple and straightforward way: it always picks the greedy predictor and directly applies it on contexts without any modification — that is, in terms of making predictions, the algorithm is fully greedy. This seems to contradict the conventional idea that greedy-prediction-based algorithms are typically sub-optimal (e.g., Langford and Zhang 2008), and is in sharp contrast to previous elimination-based algorithms (e.g., Dudík et al. 2011, Agarwal et al. 2012) and confidence-bounds-based algorithms (e.g., Abbasi-Yadkori et al. 2011, Chu et al. 2011) ubiquitous in the bandit literature, which spend a lot of efforts and computation resources maintaining complex confidence intervals, version spaces, or distributions over predictors. Even when one thinks about the algorithms of Abe and Long (1999) and Foster and Rakhlin (2020) which are similar to ours, one can find that they appeal to more robust predictors: Abe and Long (1999) appeal to the "Widrow-Hofff predictor" (equivalent to an online gradient descent oracle) and Foster and Rakhlin (2020) appeal to a general online regression oracle. Both of their analysis critically relies on the *online* nature of their oracles, i.e., the oracles can efficiently minimize regret against an adaptive adversary

— essentially, this means that a portion of the heavy lifting regarding the exploration-exploitation trade-off is taken care of by the online oracles, not the algorithms. While seemingly counter-intuitive, we claim that making "naive" greedy predictions is sufficient for optimal contextual bandit learning, which means that our oracle does not care about the exploration-exploitation trade-off at all. This surprising finding suggests that a rigorous analysis of our algorithm should contain some new ideas beyond existing bandit literature. Indeed, we will provide a quite interesting analysis of our algorithm in §2.4, which seem to be conceptually novel.

*Remark.* Readers who are interested in the difference between an offline oracle and an online oracle may compare the regret analysis approach in this paper with the approaches in Abe and Long (1999) and Foster and Rakhlin (2020). The analysis of Abe and Long (1999) and Foster and Rakhlin (2020) is essentially per-round analysis: at each round, the instantaneous bandit regret is upper bounded by the instantaneous online regression regret, with no structure shared across different rounds, so the final regret bound follows from taking a sum over all rounds. By contrast, our analysis has to deal with the shared structure across different rounds, i.e., we have to figure out how the exploration that occurred in early rounds benefits the exploitation in later rounds.

### 2.2.2 Theoretical Guarantees

We show that the simple algorithm `FALCON` enjoys strong performance guarantees.

**Statistical optimality.** Define $m(T) := \min\{m \in \mathbb{N} : T \leq \tau_m\}$, which is the total number of epochs that Algorithm 2.1 executes. The regret guarantee of Algorithm 2.1 is stated in Theorem 2.1. The proof is deferred to Appendix A.1. We will elaborate on the key ideas of the analysis in §2.4.

**Theorem 2.1.** *Consider an epoch schedule such that $\tau_m \leq 2\tau_{m-1}$, $\forall m > 1$ and $\tau_1 \leq 2$. Let $c = 1/30$. For any $T \in \mathbb{N}$, with probability at least $1-\delta$, the regret of Algorithm 2.1 after $T$ rounds is at most*

$$O\left(\sqrt{KT\log(|\mathcal{F}|m(T)/\delta)}\right).$$

When $\tau_m = 2^m$, the above upper bound is $O\left(\sqrt{KT\log(|\mathcal{F}|\log T/\delta)}\right)$, which removes a superfluous $\sqrt{\log T}$ factor in the regret upper bound of Agarwal et al. (2012) (attained by an *inefficient* algorithm), and matches the lower bound proved by Agarwal et al. (2012) up to a

constant or $\sqrt{\log\log T}$ factor. The FALCON algorithm is thus statistically optimal.

**Computational efficiency.** Consider the epoch schedule $\tau_m = 2^m$, $\forall m \in \mathbb{N}$. For any possibly unknown $T$, our algorithm runs in $O(\log T)$ epochs, and in each epoch our algorithm only calls the oracle once. Therefore, our algorithm's computational complexity is $O(\log T)$ calls to a least squares regression oracle across all $T$ rounds (plus $O(K)$ additional cost per round). This leads to potential advantages over existing algorithms. Note that ILOVETOCONBANDITS requires $\widetilde{O}(\sqrt{KT/\log(|\mathcal{F}|/\delta)})$ calls to an offline cost-sensitive classification oracle, and SquareCB requires $O(T)$ calls to an online regression oracle — compared with our algorithm, both of them require considerably more calls to more complicated oracles (as far as a general finite $\mathcal{F}$ is concerned). Also, since a general finite $\mathcal{F}$ is not a convex function class, RegCB requires $O(T^{3/2})$ calls to a weighted least squares regression oracle for this setting — this is also much slower than our algorithm.

When the total number of rounds $T$ is known to the learner, we can make the computational cost of FALCON even lower. For any $T \in \mathbb{N}$, consider an epoch schedule $\tau_m = \left\lfloor 2T^{1-2^{-m}} \right\rfloor$, $\forall m \in \mathbb{N}$ (similar to Cesa-Bianchi et al. 2014). Then FALCON will run in $O(\log\log T)$ epochs, calling the oracle for only $O(\log\log T)$ times over $T$ rounds. In this case, we still have the same regret guarantee (up to a $\log\log T$ factor); see Corollary 2.1 below. The proof can be found in Appendix A.1.6.

**Corollary 2.1.** *For any $T \in \mathbb{N}$, consider an epoch schedule $\tau_m = \left\lfloor 2T^{1-2^{-m}} \right\rfloor$, $\forall m \in \mathbb{N}$ and let $c = 1/30$. With probability at least $1-\delta$, the regret of Algorithm 2.1 after $T$ rounds is at most*

$$O\left(\sqrt{KT\log(|\mathcal{F}|\log T/\delta)}\log\log T\right).$$

## 2.3 General Offline Regression Oracles

We now extend our results to the general setting where $\mathcal{F}$ is generic (possibly infinite). While we can still assume a least squares regression oracle as before (which corresponds to the *empirical risk minimization* (ERM) procedure under square loss in offline supervised learning), for different $\mathcal{F}$, some other types of offline regression procedures (e.g., regularized least squares like Ridge and Lasso, or logistic regression) may be preferred. Moreover, even for a function class where least squares regression is preferred, one may not want to solve the square loss minimization problem exactly, and an oracle that allows optimization error may

be preferred. Therefore, in this section, we state our results in a more general way: we assume access to an arbitrary offline regression oracle with a generic statistical learning guarantee, and design an algorithm that makes calls to this arbitrary oracle and utilizes its statistical learning guarantees. Recall that the goal of this paper is to accomplish an online-to-offline reduction from contextual bandits to offline regression. So ultimately, we want to provide a universal and optimal "offline-to-online converter," such that existing machinery of supervised learning with general function classes can be automatically translated into contextual bandits with general function classes.

In what follows, we introduce the notion of a *general offline regression oracle*.

Given a general function class $\mathcal{F}$, a general offline regression oracle associated with $\mathcal{F}$, denoted by $\texttt{OffReg}_{\mathcal{F}}$, is defined as a procedure that generates a predictor $\widehat{f} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ based on input data[2] and $\mathcal{F}$ (note that $\widehat{f}$ need not be in $\mathcal{F}$). In statistical learning theory, the quality of $\widehat{f}$ is typically measured by its "out-of-sample error," i.e., its expected error on *random* and *unseen* test data. We make the following generic assumption on the statistical learning guarantee of $\texttt{OffReg}_{\mathcal{F}}$.

**Assumption 2.2.** *Let $p$ be an arbitrary action selection kernel (see §2.1.5 for the definition). Given $n$ training samples of the form $(x_i, a_i; r_i(a_i))$ independently and identically drawn according to $(x_i, r_i) \sim \mathcal{D}$, $a_i \sim p(\cdot \mid x_i)$, the offline regression oracle $\texttt{OffReg}_{\mathcal{F}}$ returns a predictor $\widehat{f} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. For any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\mathbb{E}_{x \sim \mathcal{D}, a \sim p(\cdot|x)} \left[ (\widehat{f}(x, a) - f^\star(x, a))^2 \right] \leq \mathcal{E}_{\mathcal{F},\delta}(n).$$

The offline learning guarantee $\mathcal{E}_{\mathcal{F},\delta}(n)$ is a function that decreases with $n$, which bounds the squared $L_2$ distance between $\widehat{f}$ and $f^\star$ on the test data (generated from the same distribution as the training data). Under realizability (i.e., $f^\star \in \mathcal{F}$), this squared distance corresponds to the *estimation error* or *excess risk* of $\widehat{f}$ (under square loss, or more broadly, strongly convex loss[3]). Note that characterizing sharp estimation error / excess risk bounds and designing efficient algorithms to attain such bounds are among the most central tasks in

[2]Without loss of generality, assume that $\texttt{OffReg}_{\mathcal{F}}$ always returns the same predictor for two identical input sequences.

[3]The estimation error / excess risk is defined as $\mathbb{E}[\ell(\widehat{f}(x, a), r(a))] - \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(f(x, a), r(a))]$ for a general loss function $\ell$. When $\ell$ is the square loss, it is equal to $\mathbb{E}[(\widehat{f}(x, a) - f^\star(x, a))^2]$ under realizability. Moreover, any excess risk bound under a strongly convex loss such as the log loss implies an upper bound on $\mathbb{E}[(\widehat{f}(x, a) - f^\star(x, a))^2]$ under realizability.

statistical learning.

The above notion of the offline regression oracle, though being very natural, appears to be new in the contextual bandit literature. In particular, it is not restricted to the least squares oracle (thus finds broader applications), and it is strictly easier to implement than the online regression oracle of Foster and Rakhlin (2020) which has to deal with sequential data generated by an adaptive adversary. Indeed, any oracle satisfying the requirement of Foster and Rakhlin (2020) can be easily converted to an oracle that satisfy our Assumption 2.2.

Reducing contextual bandits to the above general offline regression oracle brings many important advantages, which will be discussed after our reduction is presented; see §2.3.2.

### 2.3.1 Algorithm and Guarantees

We provide an algorithm, called FALCON+, in Algorithm 2.2. The key differences between Algorithm 2.2 and Algorithm 2.1 lie in step 2 and step 3. In step 2, we define a new epoch-varying learning rate based on the offline learning guarantee of $\texttt{OffReg}_{\mathcal{F}}$ — this is a direct generalization of the learning rate defined in Algorithm 2.1. In step 3, instead of feeding all the previous data into the oracle, we only feed the data in epoch $m-1$ into the oracle. We make two comments here. First, while we do not feed all the previous data into the oracle any more, this is still a greedy-type call to the offline oracle, as we do not make any exploration consideration in this step. Second, the strategy of only feeding the data in the last epoch into the oracle is purely due to technical reasons (i.e., Assumption 2.2 requires i.i.d. data), as we want to avoid a more complicated discussion of martingales. Note that as a consequence of this strategy, our algorithm must run in gradually increasing epochs, e.g., $\tau_m = 2^m$ or $\tau_m = \left\lfloor 2T^{1-2^{-m}} \right\rfloor$.

Recall that $m(T)$ is the total number of epochs that Algorithm 2.2 executes. The regret guarantee of Algorithm 2.2 is stated in Theorem 2.2. The proof of Theorem 2 is deferred to Appendix A.1.

**Theorem 2.2.** *Consider an epoch schedule such that $\tau_m \geq 2^m$ for $m \leq m(T)$ and let $c = 1/2$. Without loss of generality, assume that $\gamma_1 \leq \cdots \leq \gamma_{m(T)}$. For any $T \in \mathbb{N}$, with*

**Algorithm 2.2** FAst generaL-offline-regression-oracle CONtextual bandits (FALCON+)

**input** epoch schedule $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$, confidence parameter $\delta$, tuning parameter $c$

**for** epoch $m = 1, 2, \ldots$ **do**

  Let $\gamma_m = c\sqrt{K/\mathcal{E}_{\mathcal{F},\delta/(2m^2)}(\tau_{m-1} - \tau_{m-2})}$ (for epoch 1, $\gamma_1 = 1$).

  Feed (**only**) the data in epoch $m-1$, i.e.,

$$(x_{\tau_{m-2}+1}, a_{\tau_{m-2}+1}; r_{\tau_{m-2}+1}(a_{\tau_{m-2}+1})), \cdots, (x_{\tau_{m-1}}, a_{\tau_{m-1}}; r_{\tau_{m-1}}(a_{\tau_{m-1}}))$$

  into the **offline regression oracle** $\texttt{OffReg}_{\mathcal{F}}$ and obtain $\widehat{f}_m$ (for epoch 1, $\widehat{f}_1 \equiv 0$).

  **for** round $t = \tau_{m-1} + 1, \cdots, \tau_m$ **do**

    Observe context $x_t \in \mathcal{X}$.

    Compute $\widehat{f}_m(x_t, a)$ for each action $a \in \mathcal{A}$. Let $\widehat{a}_t = \max_{a \in \mathcal{A}} \widehat{f}_m(x_t, a)$. Define

$$p_t(a) = \begin{cases} \frac{1}{K + \gamma_m\left(\widehat{f}_m(x_t, \widehat{a}_t) - \widehat{f}_m(x_t, a)\right)}, & \text{for all } a \neq \widehat{a}_t, \\ 1 - \sum_{a \neq \widehat{a}_t} p_t(a), & \text{for } a = \widehat{a}_t. \end{cases}$$

    Sample $a_t \sim p_t(\cdot)$ and observe reward $r_t(a_t)$.

*probability at least $1 - \delta$, the regret of Algorithm 2.2 after $T$ rounds is at most*

$$O\left(\sqrt{K}\sum_{m=2}^{m(T)} \sqrt{\mathcal{E}_{\mathcal{F},\delta/(2m^2)}(\tau_{m-1} - \tau_{m-2})}(\tau_m - \tau_{m-1})\right). \tag{2.4}$$

The above regret bound is general and it typically has the same rate as $O\left(\sqrt{K\mathcal{E}_{\mathcal{F},\delta/\log T}(T)T}\right)$. Therefore, given an arbitrary offline regression oracle with an arbitrary estimation error guarantee $\mathcal{E}_{\mathcal{F},\delta}(\cdot)$, we know that our algorithm's regret is upper bounded by $O\left(\sqrt{K\mathcal{E}_{\mathcal{F},\delta/\log T}(T)T}\right)$.

**Example 2.1** (Statistical Optimality of FALCON+)**.** *Consider a general, potentially nonparametric function class $\mathcal{F}$ whose empirical entropy is $O(\varepsilon^{-p})$, $\forall \varepsilon > 0$ for some constant $p > 0$. Yang and Barron (1999) and Rakhlin et al. (2017) provide several offline regression oracles that achieve the optimal $\mathcal{E}_{\mathcal{F}}(n) = O(n^{-2/(2+p)})$ estimation error rate. By letting $\tau_m = 2^m$ for $m \in \mathbb{N}$, the regret of* FALCON+ *is upper bounded by $O(T^{\frac{1+p}{2+p}} \log T)$ when one ignores the dependence on $K$. Combined with an $\widetilde{\Omega}(T^{\frac{1+p}{2+p}})$ lower bound proved in Foster and Rakhlin (2020), we know that* FALCON+ *is rate-optimal as long as the offline regression oracle is rate-optimal. We thus accomplish a universal and optimal reduction from contextual bandits to offline regression. We note that the above result also helps characterize the minimax regret rate of stochastic contextual bandits with a general, potentially nonparametric $\mathcal{F}$. Note that Foster and Rakhlin (2020) have already provided such a characterization, under*

*a* tensorization *assumption (see their Section 3 for details). We remove this assumption, as the* $O(T^{\frac{1+p}{2+p}}\log T)$ *upper bound implied by our Theorem 2.2 recovers Theorem 3 of Foster and Rakhlin (2020), without assuming tensorization.*

**Example 2.2** (Linear Contextual Bandits)**.** *Consider the linear contextual bandit setting of Chu et al. (2011) with stochastic contexts. This corresponds to setting* $\mathcal{F}$ *to be the linear class*

$$\mathcal{F} = \{(x, a) \mapsto \theta^\top x_a \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1\},$$

*where* $x = (x_a)_{a\in\mathcal{A}}$, $x_a \in \mathbb{R}^d$ *and* $\|x_a\|_2 \leq 1$*. In this case, by using the least squares regression oracle,* FALCON+ *achieves the regret* $O(\sqrt{KT(d + \log T)})$*. Compared with the best known upper bound for this problem,* $\mathrm{poly}(\log\log KT)O(\sqrt{Td\log T\log K})$ *in Li et al. (2019), the regret bound of* FALCON+ *has worse dependence on* $K$ *(which seems to come from the employed sampling strategy), but saves a* $\sqrt{\log T}$ *factor, which means that* FALCON+ *improves the best known regret upper bound for this problem when* $K << T$*. To the best of our knowledge, this is the first time that an algorithm gets over the* $\Omega(\sqrt{Td\log T})$ *barrier for this problem — notably, our new upper bound even "breaks" the* $\Omega(\sqrt{Td\log T\log K})$ *lower bound proved in Li et al. (2019). The caveat here is that Li et al. (2019) study the setting where contexts are chosen by an oblivious adversary, while we are considering the setting where contexts are stochastic. Our finding that the* $\Omega(\sqrt{Td\log T})$ *barrier does not exist for linear contextual bandits with stochastic contexts is quite interesting.*

**Example 2.3** (Contextual Bandits with Neural Networks)**.** *Deriving provable performance guarantees for neural networks is an active area of research. Here we use a recent result of Farrell et al. (2021) to illustrate how estimation error bounds for deep neural networks can be translated into contextual bandits. Specifically, let* $\mathcal{F} = \mathcal{G}^K$*,* $\mathcal{G}$ *be the class of* Multi-Layer Perceptrons *(MLP) as described in Section 2.1 of Farrell et al. (2021), and* $f^\star(x, a) = g^\star_a(x)$ *for* $x \in \mathcal{X}, a \in \mathcal{A}$*. Assume that* $\mathcal{D}$ *is a continuous distribution on* $[-1, 1]^d$ *and* $g^\star_1, \ldots, g^\star_K$ *lie in a Sobolev ball with smoothness* $\beta \in \mathbb{N}$*. By Theorem 1 of Farrell et al. (2021), the deep MLP-ReLU network estimator attains* $\widetilde{O}(n^{-\frac{\beta}{\beta+d}})$ *estimation error. Consequently,* FALCON+ *attains* $\widetilde{O}(T^{\frac{\beta+2d}{2\beta+2d}})$ *regret by using this estimator as the offline regression oracle (we omit the dependence on* $K$ *here). The above result is new, but cannot be directly compared with existing results on "neural contextual bandits" (e.g., Zhou et al. 2020), as the model assumptions are very different.*

In general, one can set $\mathcal{F}$ to be any parametric or nonparametric function class, e.g., high-dimensional parametric class, Lipschitz function class, reproducing kernel Hilbert space, and regression-tree-based or random-forest-based class. For any function class $\mathcal{F}$, we can obtain a practical algorithm achieving the optimal regret for the corresponding contextual bandit problem, as long as we can find a computationally-efficient and statisitcally-optimal offline regression oracle. This usually leads to faster algorithms with improved regret bounds. In particular, our regret upper bounds' dependence on $T$ is usually better than previous upper bounds in the literature, thanks to the fact that we lose very little in terms of dependence on $T$ when we directly convert an offline estimation error bound to a regret bound. Moreover, our results enable people to tackle broader classes of new contextual bandit problems, such as contextual bandits with heavy-tailed rewards, which will be discussed shortly.

### 2.3.2 Discussion

We discuss some interesting observations regarding our Assumption 2.2 and Theorem 2.2, which further demonstrate the generality of our results.

**Exact solutions to ERM are not required.** An important advantage of Assumption 2.2 is that it does not pose any restriction on how the predictor $\widehat{f}$ is generated, thus does not require one to use ERM or exactly solve ERM. This implies that the offline predictor $\widehat{f}$ can be obtained by running iterative optimization algorithms like (stochastic) gradient descent, and its computation can be implemented in an online/streaming fashion on large datasets, which is an important consideration in modern machine learning practice. In other words, $\widehat{f}$ can be computed via various methods, and the optimization error of $\widehat{f}$ is already included in the offline learning guarantee $\mathcal{E}_{\mathcal{F},\delta}(n)$.

**Exact realizability is not required.** Another observation is that some approximation error can also be included in $\mathcal{E}_{\mathcal{F},\delta}(n)$, which enables one to consider some relaxed notions of realizability. Note that the proof of Theorem 2.2 does not rely on the realizability assumption — the proof only relies on Assumption 2.2, which is well-defined even if $f^\star \notin \mathcal{F}$. This means that Algorithm 2.2 and the regret bound (2.4) do not really require $f^\star \in \mathcal{F}$ — all they need is a *known* guarantee $\mathcal{E}_{\mathcal{F},\delta}(n)$ which correctly upper bounds the population $L_2$ distance between $\widehat{f}$ and $f^\star$. As a consequence, our results readily extend to the setting where realizability only holds approximately up to a *known* misspecification error $\epsilon$ (Van Roy and Dong 2019, Lattimore et al. 2020, Foster and Rakhlin 2020). Specifically, suppose that $f^\star \notin \mathcal{F}$ but there

exists a function $\tilde{f} \in \mathcal{F}$ that is close to $f^\star$ in the sense that $\sup_{x,a} |\tilde{f}(x,a) - f^\star(x,a)| \le \epsilon$, then we can deduce that

$$\mathbb{E}[(\widehat{f}(x,a) - f^\star(x,a))^2] \le \epsilon^2 + \underbrace{\mathbb{E}[(\widehat{f}(x,a) - r(a))^2] - \inf_{f \in \mathcal{F}} \mathbb{E}[(f(x,a) - r(a))^2]}_{\text{estimation error}}$$

This means that one can take $\mathcal{E}_{\mathcal{F},\delta}(n)$ to be $\epsilon^2$ plus an upper bound on estimation error which goes to zero with $n$ (note that one can still get sharp $\widetilde{O}(n^{-p})$-type estimation error bounds via offline regression in the misspecified setting; see Rakhlin et al. 2017). Plugging the above choice of $\mathcal{E}_{\mathcal{F},\delta}(n)$ into Algorithm 2.2 and the general regret bound (2.4), one can easily obtain the regret bound in the misspecified setting, which is typically equal to the regret bound in the well-specified setting plus an additive term of $O(\epsilon\sqrt{K}T)$. While this additive term is linear in $T$, it is not surprising and is consistent with existing results in such a setting (e.g., Theorem 5 of Foster and Rakhlin 2020), as the model is misspecified while the regret is still evaluated against the globally optimal policy $\pi_{f^\star}$.

It is worth noting that the misspecification error $\epsilon$ may be *unknown* in practice. The challenge of adapting to an unknown $\epsilon$ is addressed by follow-up work of our paper; see §2.5 for a discussion of follow-up work.

**Rewards can be unbounded/heavy-tailed.** We note that the assumption $r_t \in [0,1]^{\mathcal{A}}$ is not essential to our reduction, if we only want to bound the regret *in expectation* rather than with high probability. Specifically, to obtain the same results on the expected regret, we only need the following condition on the reward distribution:

$$\mathbb{E}_{x \sim \mathcal{D}}[\sup_{a,a' \in \mathcal{A}} (f^\star(x,a) - f^\star(x,a'))] \le \sqrt{K}, \tag{2.5}$$

which is very weak — in the special case of multi-armed bandits, it means "the gap between the *mean rewards* of two actions is no greater than $\sqrt{K}$." Note that (2.5) only concerns the conditional mean $f^\star(x,a)$ rather than the reward distribution, thus allows the "random noise" in the reward to have an *arbitrary* distribution. Moreover, (2.5) allows the scale of $f^\star(x,a)$ to be arbitrarily large and unknown (as it only concerns the gap), thus enables $\mathcal{F}$ to contain unbounded functions. As a consequence, recent advances on "fast rates" for offline unbounded/heavy-tailed regression (see Mendelson 2014 and Section 8 of Xu and Zeevi 2020a) can be translated into contextual bandits. Here, the merit of our reduction

is that different assumptions on the reward distribution only affect our results through the offline learning guarantee $\mathcal{E}_{\mathcal{F},\delta}(n)$ in Assumption 2.2, thus the associated offline regression challenges are "separated" from contextual bandits. Note that while heavy-tailed noise is very well-studied in offline regression, it is rarely studied in contextual bandits, especially with general function classes. Our reduction provides a simple way to obtain such results.

**Robustness to delayed and batched feedback.** In practical applications of contextual bandits (e.g., clinical trials, recommendation systems), the feedback to the learner is typically not immediate and may arrive in batches (Chapelle and Li 2011). In Perchet et al. (2016) and Gao et al. (2019), a "batched bandit" model is developed, where the the learner must split her learning process into a small number of batches due to several practical constraints. Since FALCON / FALCON+ only requires processing the feedback associated with each epoch after this epoch ends, our algorithm naturally handles delayed and batched feedback. In particular, Theorem 2.2 is directly applicable to the batched version of stochastic contextual bandits with general function classes, and implies that $O(\log\log T)$ batches are sufficient for one to achieve the optimal regret rate of $T$, which significantly generalizes existing results on batched bandits. We note that the ability to handle delayed and batched feedback is an important advantage of adopting an epoch schedule and using an offline regression oracle rather than an online regression oracle, as online regression is known to require immediate feedback, and the increase of regret due to delayed rewards is generally much larger in adversarial models than in stochastic models (see Lattimore and Szepesvári 2020).

## 2.4 Regret Analysis

In this section, we elaborate on how our simple algorithm achieves the optimal regret. While we present our analysis based on Algorithm 2.1 and Theorem 2.1, everything is essentially the same for Algorithm 2.2 and Theorem 2.2. We first analyze our algorithm (through an interesting dual interpretation) and provide in §2.4.1 to §2.4.4 a proof sketch of Theorem 2.1. Then, in §2.4.5, we explain the key idea behind our algorithm, and in §2.4.6, we show how this idea leads to the algorithm.

For ease of presentation, in this section we assume that $|\mathcal{X}| < \infty$ but allows $|\mathcal{X}|$ to be arbitrarily large. Focusing on such a setting enables us to highlight important ideas and key insights without the need to invoke measure theoretic arguments (which are necessary for

infinite/uncountable $\mathcal{X}$). We remark that all our results hold for general uncountable $\mathcal{X}$; see Appendix A.1.7 for more details.

Since some notations appearing in Algorithm 2.1 are shorthand and do not explicitly reveal the dependence between different quantities (e.g., $\widehat{a}_t$ and $p_t(\cdot)$ should be written as a function and a conditional distribution explicitly depending on the random context $x_t$), we introduce some new notations which can describe the decision generating process of Algorithm 2.1 in a more systematic way. For each epoch $m \in \mathbb{N}$, given the learning rate $\gamma_m \in \mathbb{R}_+$ and the greedy predictor $\widehat{f}_m : \mathcal{X} \times \mathcal{A} \to [0,1]$ (which are uniquely determined by the data from the first $m-1$ epochs), we can explicitly represent the algorithm's decision rule using $\gamma_m$ and $\widehat{f}_m$. Specifically, for any $x \in \mathcal{X}$, define $\widehat{a}_m(x) := \max_{a \in \mathcal{A}} \widehat{f}_m(x,a)$ and

$$p_m(a \mid x) := \begin{cases} \frac{1}{K + \gamma_m \left(\widehat{f}_m(x, \widehat{a}_m(x)) - \widehat{f}_m(x,a)\right)}, & \text{for all } a \neq \widehat{a}_m(x), \\ 1 - \sum_{a \neq \widehat{a}_m(x)} p_m(a \mid x), & \text{for } a = \widehat{a}_m(x). \end{cases}$$

Then $p_m(\cdot \mid \cdot)$ is a well-defined *action selection kernel* (see §2.1.5) that completely characterizes the algorithm's decision rule in epoch $m$. Specifically, at each round $t$ in epoch $m$, the algorithm first observes a random context $x_t$, then samples its action $a_t$ according to the conditional distribution $p_m(\cdot \mid x_t)$. Note that $p_m(\cdot \mid \cdot)$ depends on all the randomness up to round $\tau_{m-1}$ (including round $\tau_{m-1}$), which means that $p_m(\cdot \mid \cdot)$ depends on $p_1(\cdot \mid \cdot), p_2(\cdot \mid \cdot), \ldots, p_{m-1}(\cdot \mid \cdot)$, and will affect $p_{m+1}(\cdot \mid \cdot), p_{m+2}(\cdot \mid \cdot), \ldots$ in later epochs.

### 2.4.1 A Tale of Two Processes

The conventional way of analyzing our algorithm's behavior at round $t$ in epoch $m$ is to study the following *original* process:

1. Nature generates $x_t \sim \mathcal{D}$.
2. Algorithm samples $a_t \sim p_t(\cdot)$.

The above process is however tricky to analyze, because the algorithm's sampling strategy over actions, $p_t(\cdot) = p_m(\cdot \mid x_t)$, depends on the new random context $x_t$, and cannot be evaluated in advance before observing $x_t$.

A core idea of our analysis is to find a way to examine the algorithm's behavior at round $t$ before observing $x_t$. To this end, we look at the following *virtual* process at round $t$ in

epoch $m$:

1. Algorithm samples $\pi_t \sim Q_m(\cdot)$, where $\pi_t : \mathcal{X} \to \mathcal{A}$ is a *policy*, and $Q_m(\cdot) : \mathcal{A}^{\mathcal{X}} \to [0, 1]$ is a probability distribution over all policies in $\mathcal{A}^{\mathcal{X}}$.

2. Nature generates $x_t \sim \mathcal{D}$.

3. Algorithm selects $a_t = \pi_t(x_t)$ deterministically.

The merit of the above process is that the algorithm's sampling procedure over policies, $Q_m(\cdot)$, is independent of the new context $x_t$. While the algorithm still has to select an action based on $x_t$ in step 3, this is completely deterministic and easier to analyze. Note that at round $t$, $Q_m(\cdot)$ is a stationary distribution which has already been determined at the beginning of epoch $m$.

The second process is however a *virtual* process because it is not how our algorithm directly proceeds. An immediate question is whether we can always find a distribution over policies $Q_m(\cdot)$, such that our algorithm behaves exactly the same as the virtual process in epoch $m$? Recall that the algorithm's decision rule in epoch $m$ is completely characterized by the action selection kernel $p_m(\cdot \mid \cdot)$. In fact, any action selection kernel $p_m(\cdot \mid \cdot)$ can be translated into an "equivalent" distribution over policies $Q_m(\cdot)$, enabling us to study our algorithm's behavior through the virtual process. We complete this translation in §2.4.2.

### 2.4.2 Action Selection Kernel as a Randomized Policy

We define the *universal policy space* as $\Psi := \mathcal{A}^{\mathcal{X}}$, which contains all possible policies. For any $p_m(\cdot \mid \cdot)$, we can construct a (unique) product probability measure $Q_m(\cdot)$ on $\Psi$ such that $Q_m(\pi) = \prod_{x \in \mathcal{X}} p_m(\pi(x) \mid x)$ for all $\pi \in \Psi$ (see Lemma A.3 in the appendix). This $Q_m(\cdot)$ ensures that for every $x \in \mathcal{X}, a \in \mathcal{A}$,

$$p_m(a \mid x) = \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x) = a\} Q_m(\pi). \tag{2.6}$$

That is, for any arbitrary context $x$, the algorithm's action generated by $p_m(\cdot \mid x)$ is probabilistically equivalent to the action generated by $Q_m(\cdot)$ through the virtual process in §2.4.1. Since $Q_m(\cdot)$ is a dense distribution over all *deterministic* policies in the universal policy space, we refer to $Q_m(\cdot)$ as the "equivalent *randomized* policy" induced by $p_m(\cdot \mid \cdot)$.

Since $p_m(\cdot \mid \cdot)$ is completely determined by $\gamma_m$ and $\hat{f}_m$, we know that $Q_m(\cdot)$ is also completely determined by $\gamma_m$ and $\hat{f}_m$.

We emphasize that our algorithm does not compute $Q_m(\cdot)$, but implicitly maintains $Q_m(\cdot)$ through $\gamma_m$ and $\widehat{f}_m$. This is important, as even when $\mathcal{X}$ is known to the learner, computing the product measure $Q_m(\cdot)$ requires $\Omega(|\mathcal{X}|)$ computational cost which is intractable for large $|\mathcal{X}|$. Remember that all of our arguments based on $Q_m(\cdot)$ are only applied for the purpose of statistical analysis and have nothing to do with the algorithm's original implementation.

### 2.4.3 Dual Interpretation in the Universal Policy Space

Through the lens of the virtual process, we find a dual interpretation of our algorithm: *it sequentially maintains a dense distribution $Q_m(\cdot)$ over all the policies in the universal policy space $\Psi$, for epoch $m = 1, 2, 3 \ldots$.* The analysis of the behavior of our algorithm thus could hopefully reduce to the analysis of an evolving sequence $\{Q_m\}_{m\in\mathbb{N}}$ (which is still non-trivial because it still depends on all the interactive data). All our analysis from now on will be based on the above dual interpretation.

As we start to explore how $\{Q_m\}_{m\in\mathbb{N}}$ evolves in the universal policy space, let us first define some *implicit* quantities in this world which are useful for our statistical analysis — they are called "implicit" because our algorithm does not really compute or estimate them at all, yet they are all well-defined and implicitly exist as long as our algorithm proceeds.

Define the "implicit reward" of a policy $\pi \in \Psi$ as

$$\mathcal{R}(\pi) := \mathbb{E}_{x\sim\mathcal{D}}\left[f^\star(x, \pi(x))\right]$$

and define the "implicit regret"[4] of a policy $\pi \in \Psi$ as

$$\mathrm{Reg}(\pi) := \mathcal{R}(\pi_{f^\star}) - \mathcal{R}(\pi).$$

At round $t$ in epoch $m$, given a predictor $\widehat{f}_m$, define the "predicted implicit reward" of a policy $\pi \in \Psi$ as

$$\widehat{\mathcal{R}}_t(\pi) := \mathbb{E}_{x\sim\mathcal{D}}\left[\widehat{f}_m(x, \pi(x))\right]$$

[4]Note that this is an "instantaneous" quantity in $[0, 1]$, not a sum over multiple rounds.

and define the "predicted implicit regret" of a policy $\pi \in \Psi$ as[5]

$$\widehat{\mathrm{Reg}}_t(\pi) := \widehat{\mathcal{R}}_t(\pi_{\widehat{f}_m}) - \widehat{\mathcal{R}}_t(\pi).$$

The idea of defining the above quantities is motivated by the celebrated work of Agarwal et al. (2014), which studies policy-based optimal contextual bandit learning in the agnostic setting (in which setting the above quantities are not implicit but play obvious roles and are directly estimated by their algorithm). There are some differences in the definitions though. For example, Agarwal et al. (2014) define the above quantities for all policies $\pi$ in a given finite policy class $\Pi$, while we define the above quantities for all policies in the universal policy space $\Psi$ (which is strictly larger than $\Pi$). Also, Agarwal et al. (2014) define $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$ based on the inverse propensity scoring estimates, while we define them based on a single predictor. We will revisit these differences later.

After defining the above quantities, we make a simple yet powerful observation, which is an immediate consequence of (2.6): for any epoch $m \in \mathbb{N}$ and any round $t$ in epoch $m$, we have

$$\mathbb{E}_{(x_t,r_t)\sim\mathcal{D},a_t\sim p_m(\cdot|x_t)}\left[r_t(\pi_{f^\star}) - r_t(a_t) \mid \gamma_m, \widehat{f}_m\right] = \sum_{\pi\in\Psi} Q_m(\pi)\mathrm{Reg}(\pi),$$

see Lemma A.4 in the appendix. This means that (under any possible realization of $\gamma_m, \widehat{f}_m$) the expected instantaneous regret incurred by our algorithm is equal to the "implicit regret" of the randomized policy $Q_m$ (as a weighted sum over the implicit regret of every deterministic policy $\pi \in \Psi$). Since $\mathrm{Reg}(\pi)$ is a fixed deterministic quantity for each $\pi \in \Psi$, the above equation indicates that to analyze our algorithm's expected regret in epoch $m$, we only need to analyze the distribution $Q_m(\cdot)$. This property shows the advantage of our dual interpretation: compared with the original process in §2.4.1 where it is hard to evaluate our algorithm without $x_t$, now we can evaluate our algorithm's behavior regardless of $x_t$.

### 2.4.4 Optimal Contextual Bandit Learning in the Universal Policy Space

We proceed to understand how $Q_m(\cdot)$ evolves in the universal policy space. We first state an immediate observation based on the equivalence of $p_m(\cdot \mid \cdot)$ and $Q_m(\cdot)$ given by equation

[5]Note that in §2.1.2 we have defined $\pi_f$ as the reward-maximizing policy induced by a reward function $f$, i.e., $\pi_f(x) = \arg\max_{a\in\mathcal{A}} f(x,a)$ for all $x \in \mathcal{X}$. Also note that *not all* policies in $\Psi$ can be written as $\pi_f$ for some $f \in \mathcal{F}$.

(2.6).

**Observation 2.1.** *For any deterministic policy $\pi \in \Psi$, the quantity $\mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_m(\pi(x)|x)}\right]$ is the expected inverse probability that "the decision generated by the randomized policy $Q_m$ is the same as the decision generated by the deterministic policy $\pi$," over the randomization of context $x$. This quantity can be intuitively understood as a measure of the "decisional divergence" between the randomized policy $Q_m$ and the deterministic policy $\pi$.*

Now let us utilize the closed-form structure of $p_m(\cdot \mid x)$ in our algorithm and point out a most important property of $Q_m(\cdot)$ stated below (see Lemma A.5 and Lemma A.6 in the appendix for details).

**Observation 2.2.** *For any epoch $m \in \mathbb{N}$ and any round $t$ in epoch $m$, for any possible realization of $\gamma_m$ and $\widehat{f}_m$, $Q_m(\cdot)$ is a feasible solution to the following "Implicit Optimization Problem" (*`IOP`*):*

$$\sum_{\pi\in\Psi} Q_m(\pi)\widehat{\mathrm{Reg}}_t(\pi) \le K/\gamma_m, \tag{2.7}$$

$$\forall \pi \in \Psi, \quad \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_m(\pi(x) \mid x)}\right] \le K + \gamma_m\widehat{\mathrm{Reg}}_t(\pi). \tag{2.8}$$

We give some interpretations for the "Implicit Optimization Problem" (`IOP`) defined above. (2.7) says that $Q_m$ controls its predicted implicit regret (as a weighted sum over the predicted implicit regret of every policy $\pi \in \Psi$, based on the predictor $\widehat{f}_m$) within $K/\gamma_m$. This can be understood as an "exploitation constraint" because it require $Q_m$ to put more mass on "good policies" with low predicted implicit regret (as judged by the current predictor $\widehat{f}_m$). (2.8) says that the decisional divergence between $Q_m(\cdot)$ and any policy $\pi \in \Psi$ is controlled by the predicted implicit regret of policy $\pi$ (times a learning rate $\gamma_m$ and plus a constant $K$). This can be understood as an "adaptive exploration constraint," as it requires that $Q_m$ behaves similarly to *every* policy $\pi \in \Psi$ at some level (which means that there should be sufficient exploration), while allowing $Q_m$ to be more similar to "good policies" with low predicted implicit regret and less similar to "bad policies" with high predicted implicit regret (which means that the exploration can be conducted adaptively based on the judgement of the predictor $\widehat{f}_m$). Combining (2.7) and (2.8), we conclude that $Q_m$ elegantly strikes a balance between exploration and exploitation — it is surprising that this is done completely implicitly, as the original algorithm does not explicitly consider these constraints at all.

There are still a few important tasks to complete. The first task is to figure out what exactly the decisional divergence $\mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_m(\pi(x)|x)}\right]$ means. We give an answer in Lemma A.7, which shows that with high probability, for any epoch $m \in \mathbb{N}$ and any round $t$ in epoch $m$, for all $\pi \in \Psi$,

$$|\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)| \leq \frac{\sqrt{K}}{2\gamma_m}\sqrt{\max_{1\leq n\leq m-1} \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_n(\pi(x) \mid x)}\right]}.$$

That is, the prediction error of the implicit reward of every policy $\pi \in \Psi$ can be bounded by the (maximum) decisional divergence between $\pi$ and all previously used randomized policies $Q_1, \ldots, Q_{m-1}$. This is consistent with our intuition, as the more similar a policy is to the previously used randomized policies, the more likely that this policy is implicitly explored in the past, and thus the more accurate our prediction on this policy should be. We emphasize that the above inequality relies on our specification of the learning rate $\gamma_m$: we can bound the prediction error using $1/\gamma_m$ because $1/\gamma_m$ is proportional to $1/\sqrt{\tau_{m-1}}$ and proportional to $\sqrt{\log|\mathcal{F}|}$ — the first quantity $1/\sqrt{\tau_{m-1}}$ is related to the length of the history, and the second quantity $\sqrt{\log|\mathcal{F}|}$ is related to the generalization ability of function class $\mathcal{F}$. This is the first place that our proof requires an epoch-varying learning rate.

The second task is to further bound (the order of) the prediction error of the implicit regret of every policy $\pi$, as the implicit regret is an important quantity that can be directly used to bound our algorithm's expected regret (see §2.4.3). We do this in Lemma A.8, where we show that with high probability, for any epoch $m \in \mathbb{N}$ and any round $t$ in epoch $m$, for all $\pi \in \Psi$,

$$\mathrm{Reg}(\pi) \leq 2\widehat{\mathrm{Reg}}_t(\pi) + 5.15K/\gamma_m,$$

$$\widehat{\mathrm{Reg}}_t(\pi) \leq 2\mathrm{Reg}(\pi) + 5.15K/\gamma_m$$

through an inductive argument. While this is a uniform-convergence-type result, we would like to clarify that this does not mean that there is a uniform convergence of $|\mathrm{Reg}(\pi) - \widehat{\mathrm{Reg}}_t(\pi)|$ for all $\pi \in \Psi$, which is too strong and unlikely to be true. Instead, we use a smart design of $\mathrm{Reg}(\pi) - 2\widehat{\mathrm{Reg}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi) - 2\mathrm{Reg}(\pi)$ (the design is motivated by Lemma 13 in Agarwal et al. 2014), which enables us to characterize the fact that the predicted implicit regret of "good policies" are becoming more and more accurate, while the predicted implicit regret of "bad policies" do not need to be accurate (as their orders directly dominate $K/\gamma_m$). We

emphasize that the above result relies on the fact that our learning rate $\gamma_m$ is gradually increasing from $O(1)$ to $O(\sqrt{T})$, as we use an inductive argument and in order to let the hypothesis hold for initial cases we have to let $\gamma_m$ be very small for small $m$. This is the second place that our proof requires a epoch-varying learning rate.

We have elaborated on how our algorithm implicitly strikes a balance between exploration and exploitation, and how our algorithm implicitly enables some nice uniform-convergence-type results to happen in the universal policy space. This is already enough to guarantee that the dual interpretation of our algorithm achieves optimal contextual bandit learning in the universal policy space. The rest of the proof is standard and can be found in the appendix.

### 2.4.5 Key Idea: Bypassing the Monster

For readers who are familiar with the research line of optimal contextual bandits learning in the agnostic setting using an offline cost-sensitive classification oracle (represented by Dudík et al. 2011, Agarwal et al. 2014), they may find a surprising connection between the `IOP` (2.7) (2.8) that we introduce in Observation 2.2 and the so-called "Optimization Problem" (`OP`) in Dudík et al. (2011) and Agarwal et al. (2014) — in particular, if one takes a look at the `OP` defined in page 4 of Agarwal et al. (2014), one will find that it is almost the same as our `IOP` (2.7) (2.8), except for two fundamental differences:

1. The `OP` of Dudík et al. (2011) and Agarwal et al. (2014) is defined on a given finite policy class $\Pi$, which may have an arbitrary shape. As a result, to get a solution to `OP`, the algorithm must explicitly solve a complicated (non-convex) optimization problem over a possibly complicated policy class — this requires a considerable number of calls to a cost-sensitive classification oracle, and is the major computational burden of Dudík et al. (2011) and Agarwal et al. (2014). Although Agarwal et al. (2014) "tame the monster" and reduce the computational cost by only strategically maintaining a *sparse* distribution over policies in $\Pi$, solving `OP` still requires $\widetilde{O}(\sqrt{KT/\log|\Pi|})$ calls to the classification oracle and is computationally expensive — the monster is still there.

   By contrast, our `IOP` is defined on the universal policy space $\Psi$, which is a nice product space. The `IOP` can thus be viewed as a very "slack" relaxation of `OP` which is extremely easy to solve. In particular, as §2.4 suggests, the solution to `IOP` can have

a completely decomposed form which enables our algorithm to solve it in a complete *implicit* way. This means that our algorithm can implicitly and confidently maintain a *dense* distribution over all policies in $\Psi$, while solving `IOP` in closed forms at no computational cost — there is no monster any more as we simply bypass it.

2. In Dudík et al. (2011) and Agarwal et al. (2014), the quantities $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$ are explicitly calculated based on the model-free inverse propensity scoring estimates. As a result, their regret guarantees do not require the realizability assumption.

   By contrast, in our paper, the quantities $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$ are implicitly calculated based on a single greedy predictor $\widehat{f}$ — we can do this because we have the realizability assumption (or relaxed notions of realizability) which enables us to *learn the reward model* and obtain an $\widehat{f}$ that is close to $f^\star$ (see also Assumption 2.2). As a result, we make a single call to the offline regression oracle here, and this is the main computational cost of our algorithm.

A possible question could then be that, given the fact that the main computational burden of Dudík et al. (2011) and Agarwal et al. (2014) is solving `OP`, why can't they simply relax `OP` as we do in our `IOP`? The answer is that without the realizability assumption, they have to rely on the capacity control of their policy space, i.e., the boundedness of $|\Pi|$, to obtain their statistical guarantees. Indeed, as their $\widetilde{O}(\sqrt{KT\log|\Pi|})$ regret bound suggests, if one let $\Pi = \mathcal{A}^{\mathcal{X}}$, then the regret could be as large as $\Omega(|\mathcal{X}|)$. Specifically, their analysis requires the limited capacity (or complexity) of $\Pi$ in two places: first, a generalization guarantee of the inverse propensity scoring requires limited $|\Pi|$; second, since they have to explicitly compute $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$ without knowing the true context distribution $\mathcal{D}$, they try to approximate it based on the historical data, which also requires limited $|\Pi|$ to enable statistical guarantees.

Our algorithm bypasses the above two requirements simultaneously: first, since we use model-based regression rather than model-free inverse propensity scoring to make predictions, we do not care about the complexity of our policy space in terms of prediction (i.e., the generalization guarantee of our algorithm is governed by the capacity of $\mathcal{F}$ rather than $\Psi$); second, since our algorithm does not require explicit computation of $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$, we do not care about what $\mathcal{D}$ looks like. Essentially, all of these nice properties originate from the realizability assumption. This is how we understand the value of realizability: it does

not only (statistically) give us better predictions, but also (computationally) enables us to remove the restrictions in the policy space , which helps us to bypass the monster.

### 2.4.6 The Birth of FALCON

The idea behind "bypassing the monster," as explained in §2.4.5, is exactly what leads to the derivation of the FALCON algorithm. The derivation is interesting because it reveals deep connections between the celebrated OP studied by Dudík et al. (2011), Agarwal et al. (2014) and the intriguing probabilistic selection strategy studied by Abe and Long (1999) and Foster and Rakhlin (2020). Before we close this section, we describe how FALCON was derived. We hope that this derivation process can provide new perspectives on previous work, and motivate further discovery of new algorithms for other bandit and reinforcement learning problems.

1. We conduct a thought experiment, considering how ILOVETOCONBANDITS (Agarwal et al. 2014) can solve our problem without the realizability assumption, given an induced policy class $\Pi = \{\pi_f \mid f \in \mathcal{F}\}$.

2. ILOVETOCONBANDITS uses an inverse propensity scoring approach to calculate the predicted reward and predicted regret of policies. This can be thought as using a model-free approach (different from our §2.4.3) to calculate $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$ for $\pi \in \Pi$.

3. The computational burden in the above thought experiment is to solve OP over $\Pi$, which requires repeated calls to a cost-sensitive classification oracle.

4. When we have realizability, we can use a regression oracle to obtain a predictor $\widehat{f}_m$ and use it to calculate $\widehat{\mathcal{R}}_t(\pi)$ and $\widehat{\mathrm{Reg}}_t(\pi)$ for $\pi \in \Psi$ (if $\mathcal{D}$ is known). Here, we can operate on the set of all policies rather than only on $\Pi$, as generalization is governed by the capacity of $\mathcal{F}$.

5. An early technical result of Lemma 4.3 in Agarwal et al. (2012) is very interesting. It shows that when one tries to solve contextual bandits using regression approaches, one should try to bound a quantity like "the expected inverse probability of choosing the same action" — note that a very similar quantity also appears in OP in Agarwal et al. (2014). This suggests that an offline-regression-oracle-based algorithm should try to satisfy some requirements similar to OP. (Lemma 4.3 in Agarwal et al. (2012) also

motivates our Lemma A.7. But our Lemma A.7 goes a significant step beyond Lemma 4.3 in Agarwal et al. (2012) by unbinding the relationship between a predictor and a policy and moving forward to the universal policy space.)

6. Motivated by 3, 4, and 5, we relax the domain of `OP` from $\Pi$ to $\Psi$, and obtain the relaxed problem `IOP`. Since the new domain $\Psi = \mathcal{A}^{\mathcal{X}}$ is a product space, we consider the per-context decomposed version of `IOP`, i.e., a problem "conditional on a single $x$":

$$\sum_{\pi(x)\in\mathcal{A}} p_m(\pi(x) \mid x)\gamma_m \left(\widehat{f}_m(\pi_{\widehat{f}_m}(x)) - \widehat{f}_m(\pi(x))\right) \le K,$$
$$\forall \pi(x) \in \mathcal{A}, \quad \frac{1}{p_m(\pi(x) \mid x)} \le K + \gamma_m \left(\widehat{f}_m(\pi_{\widehat{f}_m}(x)) - \widehat{f}_m(\pi(x))\right).$$

Clearly, there is a closed-form solution to the above problem: the conditional probability of selecting an action $\pi(x)$ should be inversely proportional to the predicted reward gap of $\pi(x)$ times $\gamma_m$. This leads to `FALCON`'s decision generating process in epoch $m$.

## 2.5 Concluding Remarks

In this paper, we propose the first provably optimal offline-regression-oracle-based algorithm for general contextual bandits, solving an important open problem in the contextual bandit literature. Our algorithm is surprisingly fast and simple, and our analysis is quite general. We hope that our findings can motivate future research on contextual bandits and reinforcement learning. We discuss some follow-up work and future directions below.

**Follow-up work.** Since the first version of our paper appeared on arXiv (Simchi-Levi and Xu 2020), there have been several developments directly inspired by our work. Here we mention several extensions of our results. Xu and Zeevi (2020b) extend our results to the practical setting of infinite actions. Foster et al. (2020) build on our results to achieve instance-dependent guarantees of contextual bandits, and further extend the results to reinforcement learning. Wei and Luo (2021) extend our results to non-stationary contextual bandits; their approach to deal with non-stationarity is quite general and finds broader applications. Sen et al. (2021) extend our results to a combinatorial action model where one need to select more than one action per round. Krishnamurthy et al. (2021) extend our results to the setting where the model is misspecified and the misspecification error is unknown.

**Future directions.** Going forward, our work motivates many interesting research questions. First, in Example 2.2 (linear contextual bandits), our regret bound has worse dependence on $K$ compared with `LinUCB` (Chu et al. 2011). This seems like a limitation of the employed probabilistic selection strategy, i.e., it does not fully utilize the special properties of some function classes to obtain improved dependence on $K$. Understanding this issue better, and more broadly, understanding how to characterize and achieve the regret's optimal dependence on $K$ for general function classes, is important from both theoretical and practical points of view. Second, our work establishes new connections between policy-based (agnostic) and value-function-based (realizable) contextual bandits. We hope that the techniques and perspectives developed in this paper can find broader applications in reinforcement learning with function approximation. Finally, our work successfully reduces a prominent online decision making problem to a well-studied offline supervised learning problem. Can similar online-to-offline reductions be achieved in other practical learning settings?

# Chapter 3

# Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning

## 3.1 Introduction

How can we adaptively allocate measurements to exploit problem structure in the presence of rich, high-dimensional, and potentially stateful contextual information? In this paper, we investigate this question in the *contextual bandit* problem and its stateful relative, the problem of *reinforcement learning with rich observations*.

The contextual bandit is a fundamental problem in sequential decision making. At each round, the learner receives a *context*, selects an *action*, and receives a *reward*; their goal is to select actions so as to maximize the total long-term reward. This model has been successfully deployed in news article recommendation (Li et al. 2010, Agarwal et al. 2016), where actions represent articles to display and rewards represent clicks, and healthcare (Tewari and Murphy 2017, Bastani and Bayati 2020), where actions represent treatments to prescribe and rewards represent the patient's response. Reinforcement learning with rich observations (Krishnamurthy et al. 2016, Jiang et al. 2017) is a substantially more challenging generalization in which the learner's actions influence the evolution of the contexts, and serves as a stylized model for reinforcement learning with function approximation.

For both settings, our aim is to develop *instance-dependent* algorithms that adapt to gaps between actions in the underlying reward function to obtain improved regret. In the classical

(non-contextual) multi-armed bandit problem, this issue has enjoyed extensive investigation beginning with the work of Lai and Robbins (1985). Here, it is well-understood that when the mean reward function admits a constant gap between the best and second-best action, well-designed algorithms can obtain logarithmic (in $T$, the number of rounds) regret, which offers significant improvement over the worst-case minimax rate of $\sqrt{T}$. Subsequent work has developed a sharp understanding of optimal instance-dependent regret, both asymptotically and with finite samples (Burnetas and Katehakis 1996, Garivier et al. 2016, Kaufmann et al. 2016, Lattimore 2018, Garivier et al. 2019). Beyond the obvious appeal of lower regret, instance-dependent algorithms are particularly compelling for applications such as clinical trials—where excessive randomization may be undesirable or unethical—because they identify and eliminate suboptimal actions more quickly than algorithms that only aim for worst-case optimality.

We take the first step towards developing a similar theory for contextual bandits and reinforcement learning with general function approximation. We focus on the "realizable" or "well-specified" setting in which the learner has access to a class of regression functions $\mathcal{F}$ that is flexible enough to capture the true reward function or value function. Our aim is to develop *learning-theoretic* guarantees for rich, potentially nonparametric function classes that 1) scale only with the statistical capacity of the class, and 2) are efficient in terms of basic computational primitives for the class.

For contextual bandits, instance-dependent regret bounds are not well-understood. Positive results are known for simple classes of functions such as linear classes (Dani et al. 2008, Abbasi-Yadkori et al. 2011, Hao et al. 2019) or nonparametric Lipschitz/Hölder classes (Rigollet and Zeevi 2010, Perchet and Rigollet 2013, Hu et al. 2020). On the other hand, for arbitrary finite function classes, it is known that gap-dependent regret bounds are not possible in general (Foster and Rakhlin 2020). One line of work develops algorithms which attain instance-dependent bounds for general classes under additional structural assumptions or distributional assumptions (Russo and Van Roy 2013, Bietti et al. 2018, Foster et al. 2018), but it is not clear whether these assumptions are fundamental (in particular, they are not required to obtain minimax rates). For reinforcement learning, the situation is more dire: while instance-dependent rates have been explored in the finite state/action setting (Burnetas and Katehakis 1996, Tewari and Bartlett 2008, Ok et al. 2018, Simchowitz and Jamieson 2019), very little is known for the general setting with high dimensional states and

function approximation.

Beyond the basic issue of what instance-dependent rates can be achieved for general function classes, an important question is whether they can be achieved efficiently, using practical algorithms. A recent line of work (Foster et al. 2018, Foster and Rakhlin 2020, Simchi-Levi and Xu 2022, Xu and Zeevi 2020b) develops algorithms that are efficient in terms calls to an oracle for (offline/online) *supervised regression*. A secondary goal in this work is to develop *practical* instance-dependent algorithms based on this primitive.

Altogether, our central questions are:

1. For contextual bandits and reinforcement learning with rich observations, what properties of the function class enable us to adapt to the gap, and what are the fundamental limits?

2. Can we adapt to the gap *efficiently*?

3. More ambitiously, can we get the *best of both worlds*: Adapt to the gap and obtain the minimax rate simultaneously?

### 3.1.1 Main Results

For contextual bandits, we address each of the above research questions. We introduce a family of new complexity measures which are both necessary (in a certain sense) and sufficient to obtain fast gap-dependent regret bounds. We introduce new oracle-efficient algorithms which adapt to the gap and to these complexity measures whenever possible, while also obtaining the minimax rate. Notably, our algorithms only access the hypothesis class through calls to a *weighted least squares regression oracle*, which makes our algorithms highly practical (as they can be combined with any out-of-the-box algorithm for supervised regression for the model of interest). In a large-scale empirical evaluation, we find that our approach often gives superior results for challenging exploration problems.

Moreover, we prove new structural results which—in conjunction with our lower bounds—tie together a number of complexity measures previously proposed in contextual bandits, reinforcement learning, and active learning and provide new insight into their role in determining the optimal instance-dependent regret. See Figure 3-1 for an illustration of the relationship between all the complexity measures (three old, three new) studied in this paper.

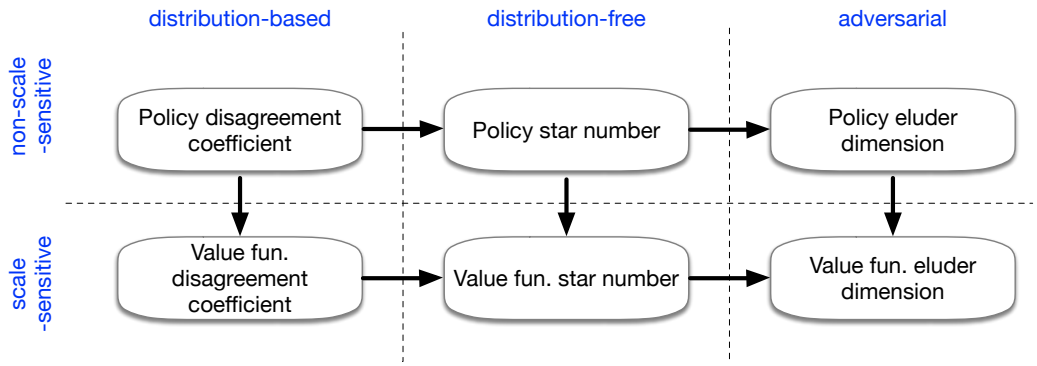


Figure 3-1: Relationship between complexity measures.

Building on our contextual bandit results, we further provide disagreement-based guarantees for episodic reinforcement learning with function approximation in a model called the *block MDP* (Krishnamurthy et al. 2016, Du et al. 2019b), which is an important type of *contextual decision process* (Jiang et al. 2017). Specifically, we extend the *value function disagreement coefficient* that we introduced in the contextual bandit setup to the block MDP setup, and develop a new instance-dependent algorithm that adapts to the gap in the optimal value function to attain improved sample complexity. Our algorithm is oracle-efficient, and attains a tight gap-dependent PAC-RL guarantee whenever the (generalized) value function disagreement coefficient is bounded. Overall, our results for RL are somewhat less complete, but we believe they suggest a number of exciting new directions for future research.

### 3.1.2 Notation

We adopt non-asymptotic big-oh notation: For functions $f, g : \mathcal{X} \to \mathbb{R}_+$, we write $f = O(g)$ (resp. $f = \Omega(g)$) if there exists some constant $C > 0$ such that $f(x) \leq Cg(x)$ (resp. $f(x) \geq Cg(x)$) for all $x \in \mathcal{X}$. We write $f = \widetilde{O}(g)$ if $f = O(g \cdot \mathrm{polylog}(T))$, $f = \widetilde{\Omega}(g)$ if $f = \Omega(g/\mathrm{polylog}(T))$, and $f = \widetilde{\Theta}(g)$ if $f = \widetilde{O}(g)$ and $f = \widetilde{\Omega}(g)$. We use $f \propto g$ as shorthand for $f = \widetilde{\Theta}(g)$. For a vector $x \in \mathbb{R}^d$, we let $\|x\|_2$ denote the euclidean norm and $\|x\|_\infty$ denote the element-wise $\ell_\infty$ norm. For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \dots, n\}$. For a set or a sequence $S$, we let $\mathrm{Unif}(S)$ denote the uniform distribution over all the elements in $S$ (note that a sequence allows identical elements to appear multiple times). For a set $\mathcal{X}$, we let $\Delta(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$. Given a policy $\pi : \mathcal{X} \to \mathcal{A}$, we occasionally overload notation and write $\pi(x, a) = \mathbb{I}\{\pi(x) = a\}$.

### 3.1.3 Organization

Section 3.2 and Section 3.3 give high-level overviews for our contextual bandit and reinforcement learning results, respectively. Due to space constraints, detailed discussion and formal statements for certain results are deferred to Appendix B.1 and Appendix B.2.

## 3.2 Instance-Dependent Complexity of Contextual Bandits

In this section we give a brief overview of our main results for contextual bandits. Please refer to Appendix B.1 for a more thorough tour of the results, including full theorem statements, additional discussion and examples, and a survey of related work.

### 3.2.1 Contextual Bandit Setup

We consider the following stochastic contextual bandit protocol. At each round $t \in [T]$, the learner observes a context $x_t \in \mathcal{X}$, selects an action $a_t \in \mathcal{A}$, then observes a reward $\ell_t(a_t) \in [0,1]$. We assume that contexts are drawn i.i.d. from a fixed but unknown distribution $\mathcal{D}$, and that each reward function $\ell_t : \mathcal{A} \to [0,1]$ is drawn independently from a fixed but unknown *context-dependent* distribution $\mathbb{P}_\ell(\cdot \mid x_t)$. We consider finite actions, with $A := |\mathcal{A}|$.[6]

We assume that the learner has access to a class of value functions $\mathcal{F} \subset (\mathcal{X} \times \mathcal{A} \to [0,1])$ (e.g., regression trees or neural networks) that is flexible enough to model the true reward distribution. In particular, we make the following standard *realizability* assumption (Chu et al. 2011, Agarwal et al. 2012, Foster et al. 2018).

**Assumption 3.1** (Realizability)**.** *There exists a function $f^\star \in \mathcal{F}$ such that $f^\star(x,a) = \mathbb{E}[\ell(a) \mid x]$.*

For each regression function $f \in \mathcal{F}$, let $\pi_f(x) = \arg\max_{a \in \mathcal{A}} f(x,a)$ denote the induced policy (with ties broken arbitrarily, but consistently), and let $\Pi = \{\pi_f \mid f \in \mathcal{F}\}$ be the induced policy class. The goal of the learner is to ensure low *regret* to the optimal policy:

$$\mathrm{Reg} = \sum_{t=1}^{T} \ell_t(\pi^\star(x_t)) - \sum_{t=1}^{T} \ell_t(a_t), \tag{3.1}$$

where $\pi^\star := \pi_{f^\star}$. For simplicity, we assume that $\arg\max_{a \in \mathcal{A}} f^\star(x,a)$ is unique for all $x$, but our results extend when this is not the case.

[6] We refer to each pair $(\mathcal{D}, \mathbb{P}_r)$ for the contextual bandit problem as an *instance*.

**Reward gaps and instance-dependent regret bounds** Consider the simple case where $\mathcal{F}$ is finite. For general finite classes $\mathcal{F}$ under Assumption 3.1, the minimax rate for contextual bandits is $\Theta(\sqrt{AT\log|\mathcal{F}|})$ (Agarwal et al. 2012). The main question we investigate is to what extent this rate can be improved when the instance has a *uniform gap*[7] in the sense that for all $x \in \mathcal{X}$,

$$f^\star(x, \pi^\star(x)) - f^\star(x, a) \geq \Delta \quad \forall a \neq \pi^\star(x). \tag{3.2}$$

For multi-armed bandits, the minimax rate is $\Theta(\sqrt{AT})$, but instance-dependent algorithms can achieve a *logarithmic* regret bound of the form $\mathrm{Reg} \leq O\big(\frac{A\log T}{\Delta}\big)$ when the gap is $\Delta$, and this is optimal (Garivier et al. 2019). Moving to contextual bandits, a natural guess would be that we can achieve

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \frac{A\log|\mathcal{F}|}{\Delta}. \tag{3.3}$$

This is impossible in a fairly strong sense: Foster and Rakhlin (2020) show that exist function classes $\mathcal{F}$ for which any algorithm must have[8]

$$\mathbb{E}[\mathrm{Reg}] = \Omega(1) \cdot \frac{|\mathcal{F}|}{\Delta}.$$

Since $\mathcal{F}$ is exponentially large for most models, polynomial dependence on $|\mathcal{F}|$ is unacceptable. The natural question then, and the one we address, is what structural properties of $\mathcal{F}$ allow for bounds of the form (3.3) that scale only logarithmically with the size of the value function class. We present our main results for the special case of finite classes for simplicity, but our lower bounds and structural results concern infinite classes, and our algorithms make no assumption on the structure.

### 3.2.2 An Efficient, Instance-Dependent Algorithm

Our main algorithm, AdaCB, is presented in Algorithm 3.1. Exploration in AdaCB is based on a probability selection strategy introduced by Abe and Long (1999) (see also Abe et al. (2003)) and extended to contextual bandits with general function classes by Foster and Rakhlin (2020) and Simchi-Levi and Xu (2022) for online and offline regression oracles, respectively. We utilize a general version of the Abe-Long strategy which we refer to by the

[7] This is sometimes referred to as the *Massart noise condition*, which has been widely studied in statistical learning theory in the context of obtaining faster rates for classification.

[8] Foster and Rakhlin (2020) prove this lower bound for adversarial contexts. Our Theorem B.2 implies an analogous lower bound for stochastic contexts.

more descriptive name "inverse gap weighting" (IGW). The strategy is parameterized by a learning rate $\gamma$ and a subset $\mathcal{A}' \subseteq \mathcal{A}$ of actions. Given a context $x$ and reward predictor $\widehat{f} \in \mathcal{F}$, we define a probability distribution $\mathsf{IGW}_{\mathcal{A}',\gamma}(x;\widehat{f}) \in \Delta(\mathcal{A})$ by

$$\left(\mathsf{IGW}_{\mathcal{A}',\gamma}(x;\widehat{f})\right)_a = \begin{cases} \frac{1}{|\mathcal{A}'|+\gamma\left(\widehat{f}(x,\widehat{a})-\widehat{f}(x,a)\right)}, & \text{for all } a \in \mathcal{A}'/\{\widehat{a}\}, \\ 1-\sum_{a\in\mathcal{A}'/\{\widehat{a}\}} p_t(a), & \text{for } a = \widehat{a}, \\ 0, & \text{for } a \notin \mathcal{A}', \end{cases} \tag{3.4}$$

where $\widehat{a} := \arg\max_{a\in\mathcal{A}'} \widehat{f}(x,a)$. Both Foster and Rakhlin (2020) and Simchi-Levi and Xu (2022) apply this strategy with $\mathcal{A}' = \mathcal{A}$, and with the learning rate $\gamma$ selected either constant or following a fixed non-adaptive schedule. Building on this approach, AdaCB follows the same general template as the FALCON algorithm of Simchi-Levi and Xu (2022), but with two key differences. First, rather than applying the IGW scheme to all actions, we restrict only to actions $a$ which are "plausible" in the sense that they are induced by a version space $\mathcal{F}_m$ maintained (implicitly) by the algorithm. Second, we choose the learning rate $\gamma_m$ in a data-driven fashion, i.e., we adaptively update the learning rate based on our empirical estimation of the "hardness" of the underlying problem instance. We refer to Appendix B.1.1 for more a detailed explanation of the design considerations behind AdaCB.

### 3.2.3 Disagreement-Based Guarantees

Our analysis of AdaCB shows that variants of the *disagreement coefficient*, a key parameter in empirical process theory and active learning (Alexander 1987, Hanneke and Yang 2015), play a fundamental role in determining the optimal gap-dependent regret bounds for contextual bandits with rich function classes.

Our most basic results concern a parameter we call the *policy disagreement coefficient*,[9] defined as

$$\boldsymbol{\theta}^{\mathsf{pol}}_{\mathcal{D},\pi^\star}(\Pi,\varepsilon_0) = \sup_{\varepsilon\geq\varepsilon_0} \frac{\mathbb{P}_{\mathcal{D}}(x : \exists \pi \in \Pi_\varepsilon : \pi(x) \neq \pi^\star(x))}{\varepsilon}, \tag{3.5}$$

where $\Pi_\varepsilon := \{\pi \in \Pi : \mathbb{P}_{\mathcal{D}}(\pi(x) \neq \pi^\star(x)) \leq \varepsilon\}$; when $\mathcal{D}$ and $\pi^\star$ are clear from context we abbreviate to $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi,\varepsilon_0)$. This parameter, sometimes called *Alexander's capacity function*,

[9] In fact, for binary actions the policy disagreement coefficient is the same as the usual disagreement coefficient from active learning (Hanneke and Yang 2015); we adopt the name *policy disagreement coefficient* only to distinguish from other parameters we introduce.

**Algorithm 3.1** AdaCB (Adaptive Contextual Bandits)

**input:** Function class $\mathcal{F}$. Number of rounds $T$.
**initialization:**

- $M = \lceil \log_2 T \rceil$. `// Number of epochs.`
- Define $\tau_m = 2^m, t_m = (\tau_m + \tau_{m-1})/2$ and $n_m = \tau_m - \tau_{m-1}$ for $m \in [M]$ `// Epoch schedule.` and $\tau_0 = 0$, $t_0 = 0$, and $n_0 = 1/2$.
- Set $\delta = 1/T$. `// Failure probability.`
- $\beta_m = 16(M - m + 1)\log(2|\mathcal{F}|T^2/\delta)$ for $m \in [M]$ `// Confidence radius.`
- $\mu_m = 64\log(4M/\delta)/n_{m-1}$ for $m \in [M]$ `// Smoothing parameter.`

**notation:**

- $\sum_{t=1}^{0}[\ldots] := 0$ and $\mathbb{E}_{x\sim\mathcal{D}_1}[\ldots] := 1$.
- For $\mathcal{F}' \subset \mathcal{F}$, define

$$\mathcal{A}(x;\mathcal{F}') = \{a \in \mathcal{A} : \pi_f(x) = a \text{ for some } f \in \mathcal{F}'\}, \quad \text{// Candidate action set.}$$
$$w(x;\mathcal{F}') = \mathbb{I}\{|\mathcal{A}(x;\mathcal{F}')| > 1\} \cdot \max_{a\in\mathcal{A}(x;\mathcal{F}')} \sup_{f,f'\in\mathcal{F}'} \left|f(x,a) - f'(x,a)\right|. \quad \text{// Confidence width.}$$

**algorithm:**

**for** epoch $m = 1, 2, \ldots, M$ **do**
  Compute the predictor $\widehat{f}_m = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^{\tau_{m-1}} (f(x_t,a_t) - r_t(a_t))^2$.
  Define

$$\mathcal{F}_m = \left\{ f \in \mathcal{F} \;\middle|\; \sum_{t=1}^{t_{m-1}} (f(x_t,a_t) - r_t(a_t))^2 \le \inf_{f'\in\mathcal{F}} \sum_{t=1}^{t_{m-1}} (f'(x_t,a_t) - r_t(a_t))^2 + \beta_m \right\}.$$

  Compute the *instance-dependent scale factor:* if $m > 1$,

$$\lambda_m = \begin{cases} \frac{\mathbb{E}_{x\sim\mathcal{D}_m}[\mathbb{I}\{|\mathcal{A}(x;\mathcal{F}_m)|>1\}]+\mu_m}{\sqrt{\mathbb{E}_{x\sim\mathcal{D}_{m-1}}[\mathbb{I}\{|\mathcal{A}(x;\mathcal{F}_{m-1})|>1\}]+\mu_{m-1}}}, & \text{Option I: Policy-based exploration,} \\ \mathbb{I}\left\{\mathbb{E}_{x\sim\mathcal{D}_m}\left[w(x;\mathcal{F}_m)\right] \ge \frac{\sqrt{AT\log(|\mathcal{F}|/\delta)}}{n_{m-1}}\right\}, & \text{Option II: Value-based exploration,} \end{cases}$$

  where $\mathcal{D}_m = \text{Unif}(x_{t_{m-1}+1}, \ldots, x_{\tau_{m-1}})$; else, $\lambda_1 = 1$ (Option I) or 0 (Option II).
  Compute the learning rate:

$$\gamma_m = \lambda_m \cdot \sqrt{\frac{An_{m-1}}{\log(2|\mathcal{F}|T^2/\delta)}}.$$

  **for** round $t = \tau_{m-1} + 1, \cdots, \tau_m$ **do**
    Observe context $x_t \in \mathcal{X}$.
    Compute the *candidate action set*

$$\mathcal{A}_t = \mathcal{A}(x_t;\mathcal{F}_m).$$

    Compute $\widehat{f}_m(x_t,a)$ for each action $a \in \mathcal{A}_t$. Let $\widehat{a}_t = \max_{a\in\mathcal{A}_t} \widehat{f}_m(x_t,a)$. Define

$$p_t = \mathsf{IGW}_{\mathcal{A}_t,\gamma_m}(x_t;\widehat{f}_m). \quad \text{// Inverse gap weighting; see Eq. (3.4).}$$

    Sample $a_t \sim p_t$ and observe reward $r_t(a_t)$.

dates back to Alexander (1987), and was rediscovered and termed the disagreement coefficient in the context of active learning by Hanneke (2007, 2011). In empirical process theory and statistical learning, the disagreement coefficient grants control over the fine-grained behavior of VC classes (Giné and Koltchinskii 2006, Raginsky and Rakhlin 2011, Zhivotovskiy and Hanneke 2016), and primarily determines whether certain logarithmic terms can appear in excess risk bounds for empirical risk minimization (ERM) and other algorithms under low-noise conditions. In active learning, the disagreement coefficient plays a more critical role, as it provides a sufficient (and weakly necessary) condition under which one can achieve label complexity logarithmic in the target precision (Hanneke 2007, 2011, Raginsky and Rakhlin 2011, Hanneke 2014, Hanneke and Yang 2015).

Informally, the policy disagreement coefficient measures how likely we are to encounter a context on which *some* near-optimal policy disagrees with $\pi^\star$. Low disagreement coefficient means that all the near-optimal policies deviate from $\pi^\star$ only in a small, shared region of the context space, while large disagreement coefficient means that the points on which disagreement occurs are more prevalent throughout the context space (w.r.t $\mathcal{D}$), so that many samples are required to rule out all of these policies.

We show that AdaCB (Algorithm 3.1) adapts to the gap whenever the policy disagreement coefficient is bounded; see Section 3.2.2 for details. We show that AdaCB achieves the following instance-dependent guarantee stated in terms of $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon)$.

**Theorem B.1.** *For all instances, AdaCB (Algorithm 3.1 with* Option I*) ensures that*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \min_{\varepsilon > 0} \max \left\{ \varepsilon \Delta T, \frac{\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \cdot A \log |\mathcal{F}|}{\Delta} \right\} \tag{3.6}$$

*with no prior knowledge of* $\Delta$ *or* $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon)$.

Theorem B.1 is a best-of-both-worlds guarantee. In the worst case, we have $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \leq 1/\varepsilon$, so that (3.6) becomes $\widetilde{O}(\sqrt{AT \log|\mathcal{F}|})$, the minimax rate. However, if $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) = \mathrm{polylog}(1/\varepsilon)$, then (3.6) ensures that

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \frac{A \log|\mathcal{F}|}{\Delta},$$

so that AdaCB enjoys logarithmic regret. We emphasize that while Theorem B.1 concerns finite classes, this is only a stylistic choice: AdaCB places no assumption on the structure of

$\mathcal{F}$, and the analysis trivially generalizes by replacing $\log|\mathcal{F}|$ with standard learning-theoretic complexity measures such as the pseudodimension.

While this is certainly encouraging, it is not immediately clear whether the rate in (3.6) is fundamental. To this end, we prove that dependence on the disagreement coefficient is qualitatively necessary.

**Theorem B.2.** *For any $A \in \mathbb{N}$, $\Delta > 0$, $\varepsilon > 0$, and functional $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon)$, there exists a function class $\mathcal{F}$ with $A$ actions and a distribution over realizable instances with uniform gap $\Delta$ such that any algorithm has*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{\Omega}(1) \cdot \min_{\varepsilon > 0} \max\left\{ \varepsilon \Delta T, \frac{\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \cdot A \log|\mathcal{F}|}{\Delta} \right\}.$$

Detailed variants of Theorem B.1 and Theorem B.2, as well as examples, can be found in Appendix B.1.2.

### 3.2.4 Scale-Sensitive Guarantees and the Value Function Disagreement Coefficient

Theorem B.2 shows that the regret bound (3.6) attained by AdaCB cannot be improved without further assumptions on $\mathcal{F}$. However, it leaves the possibility of more refined complexity measures that are tighter than $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon)$ for most instances, yet coincide on the construction that realizes the lower bound in Theorem B.2. To this end, we introduce a second complexity measure, the *value function disagreement coefficient*, which can exploit the scale-sensitive nature of the value function class $\mathcal{F}$ to provide tighter bounds. The value function disagreement coefficient is defined as

$$\boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D};f^\star}(\mathcal{F}, \Delta_0, \varepsilon_0) = \\ \sup_{\Delta > \Delta_0, \varepsilon > \varepsilon_0} \sup_{p:\mathcal{X}\to\Delta(\mathcal{D})} \frac{\Delta^2}{\varepsilon^2} \mathbb{P}_{\mathcal{D},p}\Big( \exists f \in \mathcal{F} : |f(x,a) - f^\star(x,a)| > \Delta, \ \ \|f - f^\star\|_{\mathcal{D},p} \le \varepsilon \Big), \quad (3.7)$$

where $\|f\|^2_{\mathcal{D},p} := \mathbb{E}_{x\sim\mathcal{D}, a\sim p(x)}[f^2(x)]$. We abbreviate $\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta_0, \varepsilon_0) \equiv \boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D};f^\star}(\mathcal{F}, \Delta_0, \varepsilon_0)$ when the context is clear. The key difference from the policy disagreement coefficient is that rather than using a binary property ($\pi(x) \neq \pi^\star(x)$) to measure disagreement, we use a more refined scale-sensitive notion: Two functions $f$ and $f^\star$ are said to $\Delta$-disagree on $(x, a)$ if $|f(x,a) - f^\star(x,a)| > \Delta$, and the value function disagreement coefficient simply measures

how likely we are to encounter a context for which a value function that is $\varepsilon$-close to $f^\star$ in $L_2$ distance $\Delta$-disagrees from it (for a worst-case action distribution). This refined view leads to tighter guarantees for common function classes. For example, when $\mathcal{F}$ is a linear function class, i.e. $\mathcal{F} = \left\{(x,a) \mapsto \langle w, \phi(x,a)\rangle \mid w \in \mathbb{R}^d\right\}$ for a fixed feature map $\phi(x,a)$, the policy disagreement coefficient is only bounded for sufficiently regular distributions, whereas the value function disagreement coefficient is always bounded by $d$ (Proposition B.1).

We show that AdaCB, with a slightly different parameter configuration, can adapt to value function disagreement coefficient in a best-of-both-worlds fashion.

**Theorem B.3.** *For all instances, AdaCB with* Option II *ensures that*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \min\left\{\sqrt{AT\log|\mathcal{F}|}, \frac{\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon_T) \cdot A\log|\mathcal{F}|}{\Delta}\right\},$$

*where* $\varepsilon_T \propto \sqrt{\log|\mathcal{F}|/T}$.

In Appendix B.1.3 we show (Theorem B.4) that this dependence on $\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta, \varepsilon)$ is qualitatively necessary, meaning that AdaCB is adapts near-optimally without additional assumptions.

Beyond contextual bandits, our scale-sensitive generalization of the disagreement coefficient is new to both empirical process theory and active learning to our knowledge, and may be of independent interest.

### 3.2.5 Distribution-Free Guarantees and Structural Results

While the distribution-dependent nature of our disagreement-based upper bounds can lead to tight guarantees for benign distributions, it is natural to ask: *For what classes* $\Pi$ *(resp.* $\mathcal{F}$*) can we ensure the policy (resp. value) disagreement coefficient is bounded for any distribution* $\mathcal{D}$*?* Hanneke and Yang (2015) show that the policy disagreement coefficient is always bounded by a combinatorial parameter for $\Pi$ called the (policy) *star number*.[10] An immediate consequence (via Theorem B.1) is that AdaCB enjoys logarithmic regret even in the *distribution-free* setting for classes with bounded policy star number. More interestingly, in Appendix B.1.4 we show (Theorem B.6) that for any class $\Pi$, bounded policy star number is *necessary* to obtain

[10] In fact, the star number exactly coincides with the worst-case value of the disagreement coefficient over all possible distributions and scale parameters.

logarithmic regret in the worst-case (with respect to both $\mathcal{D}$ and the class $\mathcal{F}$ realizing $\Pi$). Thus, we have the following characterization.

**Theorem (informal).** *For any policy class $\Pi$, bounded policy star number is necessary and sufficient to obtain logarithmic regret in the distribution-free setting.*

Compared to our disagreement-based lower bounds, which rely on specially designed function classes, this lower bound holds for *any* policy class.

This characterization motivates us to define a scale-sensitive analogue of the star number called the *value function star number*. The value function star number is a new combinatorial parameter even within the broader literature on active learning and empirical process theory, and we show (Theorem B.7) that it bounds the value function disagreement coefficient for all choices of the context distribution $\mathcal{D}$ and scale parameter $\varepsilon$. We then show (Theorem B.8) that a weak version of the value function disagreement coefficient is *necessary* to obtain logarithmic regret for worst-case context distributions, leading to the following characterization.

**Theorem (informal).** *For any value function class $\mathcal{F}$, bounded value function star number is (weakly) necessary and sufficient to obtain logarithmic regret in the distribution-free setting.*

We refer the reader to Appendix B.1.4 for formal statements for these results.

### 3.2.6 On the Eluder Dimension

The value function star number is closely related to—and in particular always upper bounded by—the (value function) *eluder dimension* of Russo and Van Roy (2013). The eluder dimension was introduced to prove regret bounds for the generalized UCB algorithm and Thompson sampling for contextual bandits with adversarial contexts, and more recently has been used to analyze algorithms for reinforcement learning with function approximation (Osband and Van Roy 2014, Wen and Van Roy 2017, Ayoub et al. 2020, Wang et al. 2020b). An immediate consequence of our (disagreement coefficient) $\leq$ (star number) $\leq$ (eluder dimension) connection is that boundedness of the eluder dimension suffices to obtain logarithmic regret with AdaCB. Unlike the star number though, bounded eluder dimension is not required for the stochastic setting we consider. However, building on our previous lower bounds, we show (Theorem B.9) that a weak version of the eluder dimension is *necessary* to obtain logarithmic regret under adversarial contexts, and give a tighter analysis

of the generalized UCB algorithm to show that it attains this rate (however, this is not a best-of-both-worlds guarantee).

**Theorem (informal).** *For any value function class $\mathcal{F}$, bounded value function eluder dimension is (weakly) necessary and sufficient to obtain logarithmic regret in the adversarial setting.*

This result places the eluder dimension on more solid footing and shows that while it is not required for minimax rates, it plays a fundamental role for instance-dependent rates. See Appendix B.1.5 for a formal statement.

**Instance-Dependent Guarantees: A Comprehensive Picture** The relationship between all of our complexity measures, old and new, is summarized in Figure 3-1. Beyond expanding the scope of settings for which logarithmic regret is achievable, we hope our structural results and lower bounds provide a new lens through which to understand existing algorithms and instance-dependent rates, and provide new clarity.

As a disclaimer, we mention that the primary goal of this work is to understand how *contextual information* shapes the optimal instance-dependent rates for contextual bandits. We believe that this question is challenging and interesting even in the finite-action regime (in fact, even when $A = 2$!) and as such, we do not focus on obtaining optimal dependence on $A$ in our upper or lower bounds, nor do we handle infinite actions. Fully understanding the interplay between contexts and actions is a fascinating open problem, and we hope to see this addressed in future work.

### 3.2.7 Efficiency and Empirical Performance

**Computational efficiency** AdaCB is an *oracle-efficient* algorithm. That is, it accesses the value function class $\mathcal{F}$ only through a *weighted least squares regression oracle* capable of solving problems of the form

$$\mathsf{Oracle}(\mathcal{H}) = \arg\min_{f \in \mathcal{F}} \sum_{(w,x,a,y) \in \mathcal{H}} w(f(x,a) - y)^2 \tag{RO}$$

for a given set $\mathcal{H}$ of examples $(w, x, a, y)$ where $w \in \mathbb{R}_+$ specifies the example weight. This makes the algorithm practical, as it can be combined with any out-of-the-box algorithm for supervised regression for the model of interest. We refer to Section 4 of the full version of our

paper—Foster et al. (2020)—for a detailed description on how AdaCB can be implemented with the above regression oracle.

**Empirical performance** We replicated the large-scale empirical contextual bandit evaluation setup of Bietti et al. (2018), which compares a number of state-of-the art general-purpose contextual bandit algorithms across more than 500 datasets. We found that our new algorithm, AdaCB, typically gives comparable or superior results to existing baselines, particularly on challenging datasets with many actions. We refer to Section 5 of the full version of our paper—Foster et al. (2020)—for the detailed experimental results.

## 3.3 Instance-Dependent Complexity of Reinforcement Learning

We now give a high-level overview of our guarantees for reinforcement learning in the Block MDP model. Detailed results, including full theorem statements and pseudocode for the main algorithm can be found in Appendix B.2.

### 3.3.1 Block MDP Setup

Building on our contextual bandit results, we provide disagreement-based guarantees for episodic reinforcement learning with function approximation in a model called the *block MDP* (Krishnamurthy et al. 2016, Du et al. 2019b), which is an important type of *contextual decision process* (Jiang et al. 2017).

The block MDP may be thought of as a generalization of the contextual bandit problem. Each round of interaction is replaced by an *episode* of length $H$. While the initial context $x_1$ (now referred to as a *state*) in each episode is drawn i.i.d. as in the contextual bandit, the evolution of the subsequent states $x_2, \ldots, x_H$ is influenced by the learner's actions. Now, without further assumptions, this is simply a general MDP, and function approximation provides no benefits in the worst-case. To allow for sample-efficient learning guarantees, the *block MDP* model assumes there is an unobserved latent MDP with $S$ states, and that each observed state $x_h$ is drawn from an *emission distribution* for the current latent state $s_h$. When $H = 1$ and $S = 1$, this recovers the contextual bandit, and in general the goal is to use an appropriate value function class $\mathcal{F}$ to attain sample complexity guarantees that are

polynomial in $S$, but not $|\mathcal{X}|$ (which, as in the contextual bandit, is typically infinite and high-dimensional).

More formally, the block MDP setup we consider is a layered episodic Markov decision process with horizon $H$, state space $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_H$ (with $\mathcal{X}_i \cap \mathcal{X}_j = \varnothing$), and action space $\mathcal{A}$ with $|\mathcal{A}| = A$. We proceed in $K$ episodes. Within each episode we observe rewards and observations through the following protocol, beginning with $x_1 \sim \mu$.

- For $h = 1, \ldots, H$:
  - Choose action $a_h$.
  - Observe reward $r_h$ and next state $x_{h+1} \sim P^\star_h(\cdot \mid x_h, a_h)$.

Note that for this setting we use the subscript $h$ on e.g., $x_h$, to refer to the layer within a fixed episode, whereas for contextual bandits we use the subscript $t$ to refer to the round/episode itself. We always use $h$ for the former setting and $t$ for the latter to distinguish.

As mentioned above, the state space is potentially rich and high-dimensional, and dependence on $|\mathcal{X}|$ is unacceptable. Hence, to enable sample-efficient reinforcement learning guarantees with function approximation, the block MDP model assumes the existence of a *latent state space* $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$, and assumes that each state $x \in \mathcal{X}$ can be uniquely attributed to a latent state $s \in \mathcal{S}$. More precisely, we assume that for each $h$, $P^\star_h$ factorizes, so that we can view $x_{h+1}$ as generated by the process $s_{h+1} \sim P^\star_h(\cdot \mid x_h, a_h)$, $x_{h+1} \sim \psi(s_{h+1})$, where $\psi : \mathcal{S} \to \Delta(\mathcal{X})$ is an (unknown) *emission distribution*, and $s_{h+1}$ is the latent state for layer $h+1$. We make the following standard decodability assumption (Krishnamurthy et al. 2016, Jiang et al. 2017, Du et al. 2019b).

**Assumption 3.2** (Decodability)**.** *For all $s \neq s'$,* $\operatorname{supp}(\psi(s)) \cap \operatorname{supp}(\psi(s')) = \varnothing$.

This assumption implies that the optimal policy $\pi^\star$ depends only on the current context $x_h$. We write the optimal $Q$-function for layer $h$ as $Q^\star_h(x, a)$ and let $V^\star_h(x) = \max_{a \in \mathcal{A}} Q^\star_h(x, a)$ be the optimal value function.

**Function approximation and gaps** As in the contextual bandit setting, take as given a class of functions $\mathcal{F}$ that attempts to model the optimal value function. We let $\mathcal{F}_h \subseteq (\mathcal{X} \times \mathcal{A} \to [0, H])$ be the value function class for layer $h$ (with $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_h$), and we make the following optimistic completeness assumption (Jin et al. 2020, Wang et al. 2019, 2020b).

**Assumption 3.3.** *For all $h$ and all functions $V : \mathcal{X}_{h+1} \to [0, H]$, we have that*

$$(x, a) \mapsto \mathbb{E}[r_h + V(x_{h+1}) \mid x_h = x, a_h = a] \in \mathcal{F}_h.$$

Assumption 3.3 implies that $Q^\star_h \in \mathcal{F}_h$, generalizing the realizability assumption (Assumption 3.1) but it is significantly stronger, as it requires that the function class contains Bellman backups for arbitrary functions.

### 3.3.2 An Efficient, Instance-Dependent Algorithm

We develop a new instance-dependent and oracle-efficient algorithm called RegRL (Algorithm B.1 in Appendix B.2.1) which adapts to the gap in the optimal value function $Q^\star$ to attain improved sample complexity. Define $\Delta(x, a) = V^\star_h(x) - Q^\star_h(x, a)$, and define the worst-case gap as

$$\Delta = \min_s \inf_{x \in \mathrm{supp}(\psi(s))} \min_a \{\Delta(x, a) \mid \Delta(x, a) > 0\}.$$

Algorithm B.1 attains a tight gap-dependent PAC guarantee for reinforcement learning in the Block MDP model whenever an appropriate Block MDP analogue of the value function disagreement coefficient $\boldsymbol{\theta}^{\mathsf{val}}$ introduced in Section 3.2 is bounded (Appendix B.2.2).

**Theorem B.12 (informal).** *For all instances, RegRL (Algorithm B.1) finds an $\varepsilon$-suboptimal policy using* $\mathrm{poly}(S, A, H, \boldsymbol{\theta}^{\mathsf{val}}) \cdot \frac{\log|\mathcal{F}|}{\varepsilon \cdot \Delta}$ *episodes.*

See Appendix B.2.1 and Appendix B.2.2 for pseudocode and a full theorem statement. The theorem has two key features. First, when $\boldsymbol{\theta}^{\mathsf{val}} = \widetilde{O}(1)$, the scaling of $\varepsilon$ and $\Delta$ in the term $\frac{\log|\mathcal{F}|}{\varepsilon \cdot \Delta}$ is optimal even for in the special case of contextual bandits, and improves over the minimax rate, which scales as $\frac{1}{\varepsilon^2}$. Second, and perhaps more importantly, RegRL is computationally efficient, and only requires a regression oracle for the value function class. Previous works require stronger oracles and typically do not attain optimal dependence on $\varepsilon$, but are not fully comparable in terms of statistical assumptions (Krishnamurthy et al. 2016, Jiang et al. 2017, Dann et al. 2018, Du et al. 2019a,b, Misra et al. 2019, Feng et al. 2020, Agarwal et al. 2020a); see Appendix B.2.3 for a detailed comparison. At a conceptual level, the design and analysis of RegRL use several new techniques that leverage our disagreement-based perspective, and we hope that they will find broader use.

## 3.4 Concluding Remarks

We have developed efficient, instance-dependent algorithms for contextual bandits and reinforcement learning with function approximation. We showed that disagreement coefficients and related combinatorial parameters play a fundamental role in determining the optimal instance-dependent rates, and that algorithms that adapt to these parameters can be simple and practically effective. We hope that our techniques will find broader use, particularly for the reinforcement learning setting.

# Chapter 4

# Fundamental Barriers for Offline Reinforcement Learning with Value Function Approximation

## 4.1 Introduction

In offline reinforcement learning, we aim to evaluate or optimize decision making policies using logged transitions and rewards from historical experiments or expert demonstrations. Offline RL has great promise for decision making applications where actively acquiring data is expensive or cumbersome (e.g., robotics (Pinto and Gupta 2016, Levine et al. 2018, Kalashnikov et al. 2018)), or where safety is critical (e.g., autonomous driving (Sallab et al. 2017, Kendall et al. 2019) and healthcare (Gottesman et al. 2018, 2019, Wang et al. 2018, Yu et al. 2019, Nie et al. 2021)). In particular, there is substantial interest in combining offline reinforcement learning with function approximation (e.g., deep neural networks) in order to encode inductive biases and enable generalization across large, potentially continuous state spaces, with recent progress on both model-free and model-based approaches (Ross and Bagnell 2012, Laroche et al. 2019, Fujimoto et al. 2019, Kumar et al. 2019, Agarwal et al. 2020b). However, existing algorithms are extremely data-intensive, and offline RL methods—to date—have seen limited deployment in the aforementioned applications. To enable practical deployment going forward, it is paramount that we develop a strong understanding of the statistical foundations for reliable, sample-efficient offline reinforcement learning with function approximation, as well as an understanding of when and why existing methods

succeed and how to effectively collect data.

Compared to the basic supervised learning problem, offline reinforcement learning with function approximation poses substantial algorithmic challenges due to two issues: *distribution shift* and *credit assignment*. Within the literature on *value* function approximation (or, approximate dynamic programming), all existing methods require both (1) distributional conditions, which assert that the logged data has good coverage (addressing distribution shift), and (2) representational conditions, which assert that the function approximator is flexible enough to represent value functions induced by certain policies (addressing credit assignment). Notably, sample complexity analyses for standard offline RL methods (e.g., fitted Q-iteration) require representation conditions considerably more restrictive than what is required for supervised learning (Munos 2003, 2007, Munos and Szepesvári 2008, Antos et al. 2008), and these methods can diverge when these conditions do not hold (Gordon 1995, Tsitsiklis and Van Roy 1996, 1997, Wang et al. 2021a). Despite substantial research effort, it is not known whether these conditions constitute fundamental limits or whether the algorithms can be improved. Resolving this issue would serve as a stepping stone toward developing a theory for offline reinforcement learning that parallels our understanding of supervised (statistical) learning.

The lack of understanding of fundamental limits in offline reinforcement learning was highlighted by Chen and Jiang (2019), who observed that all existing finite-sample analyses for offline RL algorithms based on *concentrability* (Munos 2003)—the most ubiquitous notion of data coverage—require representation conditions significantly stronger than *realizability*, a standard condition from supervised learning which asserts that the function approximator can represent optimal value functions. Chen and Jiang (2019) conjectured that realizability and concentrability alone do not suffice for sample-efficient offline RL and noted that proving such a result seemed to be out of reach for existing lower bound techniques. Subsequent progress led to positive results for sample-efficient offline RL under coverage conditions stronger than concentrability (Xie and Jiang 2021) and impossibility results under weaker coverage conditions (Wang et al. 2020a, Zanette 2021), but the original conjecture remained open.

**Contributions** We provide information-theoretic lower bounds which show that, in general, concentrability and realizability together are not sufficient for sample efficient offline

reinforcement learning. Our first result concerns the standard offline RL setup, where the data collection distribution is only required to satisfy concentrability, and establishes a sample complexity lower bound scaling polynomially with the size of the state space. This result resolves the conjecture of Chen and Jiang (2019) in the positive. For our second result, we further restrict the data distribution to be induced by a policy (i.e., admissible), and show that any algorithm requires sample complexity either polynomial in the size of the state space or exponential in other problem parameters. Together, our results establish that sample-efficient offline RL in large state spaces is not possible unless more stringent conditions, either distributional or representational, hold.

Our lower bound constructions are qualitatively different from previous approaches and hold even when the number of actions is constant and the value function class has constant size. Our first lower bound highlights the role of a phenomenon we call strong *over-coverage* (first documented by Xie and Jiang (2021)), wherein the data collection distribution is supported over spurious states that are not reachable by any policy. Despite the irrelevance of these states for learning in the online setting, their inclusion in the offline dataset creates significant uncertainty. Our second lower bound discovers a weak variant of over-coverage, wherein the data collection distribution is induced by running an exploratory policy in particular time steps, but many of the states supported by this distribution are not reachable in other time steps, creating spurious correlations. Our work shows that both the strong and weak over-coverage phenomena serve as fundamental, information-theoretic barriers for the design of offline reinforcement learning algorithms.

### 4.1.1 Offline Reinforcement Learning Setting

**Markov decision processes** We consider the infinite-horizon discounted reinforcement learning setting. Formally, a Markov decision process $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ consists of a (potentially large/continuous) state space $\mathcal{S}$, action space $\mathcal{A}$, probability transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$, discount factor $\gamma \in [0, 1)$, and initial state distribution $d_0 \in \Delta(\mathcal{S})$. Each (randomized) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces a distribution over trajectories $(s_0, a_0, r_0), (s_1, a_1, r_1), \ldots$ via the following process. For $h = 0, 1, \ldots$ : $a_h \sim \pi(s_h)$, $r_h = R(s_h, a_h)$, and $s_{h+1} \sim P(s_h, a_h)$, with $s_0 \sim d_0$. We let $\mathbb{E}^{M,\pi}[\cdot]$ and $\mathbb{P}^{M,\pi}(\cdot)$ denote expectation and probability under this process, respectively.

The expected return for policy $\pi$ is defined as $J_M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$, and the value

function and $Q$-function for $\pi$ are given by

$$V_M^\pi(s) := \mathbb{E}^{M,\pi}\left[\sum_{h=0}^\infty \gamma^h r_h \mid s_0 = s\right], \quad \text{and} \quad Q_M^\pi(s,a) := \mathbb{E}^{M,\pi}\left[\sum_{h=0}^\infty \gamma^h r_h \mid s_0 = s, a_0 = a\right].$$

It is well-known that there exists a deterministic policy $\pi_M^\star : \mathcal{S} \to \mathcal{A}$ that maximizes $V_M^\pi(s)$ for all $s \in \mathcal{S}$ simultaneously and thus also maximizes $J_M(\pi)$. Letting $V_M^\star := V_M^{\pi_M^\star}$ and $Q_M^\star := Q_M^{\pi_M^\star}$, we have $\pi_M^\star(s) = \arg\max_{a\in\mathcal{A}} Q_M^\star(s,a)$ for all $s \in \mathcal{S}$. Finally, we define the occupancy measure for policy $\pi$ via $d_M^\pi(s,a) := (1-\gamma)\sum_{h=0}^\infty \gamma^h \mathbb{P}^{M,\pi}(s_h = s, a_h = a)$. We drop the dependence on the model $M$ when it is clear from context.

**Offline policy learning** In the offline policy learning (or, optimization) problem, we do not have direct access to the underlying MDP and instead receive a dataset $D_n$ of tuples $(s,a,r,s')$ with $r = R(s,a)$, $s' \sim P(s,a)$, and $(s,a) \sim \mu$ i.i.d., where $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$ is the *data collection distribution*. The goal of the learner is to use the dataset $D_n$ to learn an $\varepsilon$-optimal policy $\widehat{\pi}$, that is:

$$J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \le \varepsilon,$$

where the expectation $\mathbb{E}[\cdot]$ is over the draw of $D_n$ and any randomness used by the algorithm.

In order to provide sample-efficient learning guarantees that do not depend on the size of the state space, value function approximation methods take advantage of the following conditions.

- **Realizability.** This condition asserts that we have access to a class of candidate value functions $\mathcal{F} \subseteq (\mathcal{S}\times\mathcal{A} \to \mathbb{R})$ (e.g., linear models or neural networks) such that $Q^\star \in \mathcal{F}$. Realizability (that is, a well-specified model) is the most common representation condition in supervised learning and statistical estimation (Bousquet et al. 2004, Wainwright 2019) and is also widely used in contextual bandits (Agarwal et al. 2012, Foster et al. 2018).

- **Concentrability.** Call a distribution $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$ *admissible* for the MDP $M$ if there exists a (potentially stochastic and non-stationary[11]) policy $\pi$ and index $h$ such that $\nu(s,a) = \mathbb{P}^\pi[s_h = s, a_h = a]$. This condition asserts that there exists a constant

[11] A non-stationary policy is a sequence $\{\pi_h\}_{h\ge 0}$, which generates a trajectory via $a_h \sim \pi_h(s_h)$.

$C_{\text{conc}} < \infty$ such that for all admissible $\nu$,

$$\left\| \frac{\nu}{\mu} \right\|_\infty := \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \frac{\nu(s,a)}{\mu(s,a)} \right\} \leq C_{\text{conc}}. \tag{4.1}$$

Concentrability is a simple but fairly strong notion of coverage which demands that the data distribution uniformly covers all reachable states.

Under these conditions, an offline RL algorithm is said to be sample-efficient if it learns an $\varepsilon$-optimal policy with $\text{poly}(\varepsilon^{-1}, (1-\gamma)^{-1}, C_{\text{conc}}, \log|\mathcal{F}|)$ samples. Notably, such a guarantee depends only on the complexity $\log|\mathcal{F}|$ for the value function class, not on the size of the state space.[12]

**Are realizability and concentrability sufficient?** While realizability and concentrability are appealing in their simplicity, these assumptions alone are not known to suffice for sample-efficient offline RL. The most well-known line of research (Munos 2003, 2007, Munos and Szepesvári 2008, Antos et al. 2008, Chen and Jiang 2019) analyzes offline RL methods such as fitted Q-iteration under the stronger representation condition that $\mathcal{F}$ is closed under Bellman updates ("completeness"),[13] and obtains $\text{poly}(\varepsilon^{-1}, (1-\gamma)^{-1}, C_{\text{conc}}, \log|\mathcal{F}|)$ sample complexity. Completeness is a widely used assumption, but it is substantially more restrictive than realizability and can be violated by adding a single function to $\mathcal{F}$. Subsequent years have seen extensive research into algorithmic improvements and alternative representation and coverage conditions, but the question of whether realizability and concentrability alone are sufficient remains open.

### 4.1.2 Main Results

The first of our main results is an information-theoretic lower bound which shows that realizability and concentrability are not sufficient for sample-efficient offline RL.

**Theorem 4.1** (Main theorem)**.** *For all $S \geq 9$ and $\gamma \in (1/2, 1)$, there exists a family of MDPs $\mathcal{M}$ with $|\mathcal{S}| \leq S$ and $|\mathcal{A}| = 2$, a value function class $\mathcal{F}$ with $|\mathcal{F}| = 2$, and a data distribution $\mu$ such that:*

[12] For infinite function classes ($|\mathcal{F}| = \infty$), one can replace $\log|\mathcal{F}|$ with other standard measures of statistical capacity, such as Rademacher complexity or metric entropy. For example, when $\mathcal{F}$ is a class of $d$-dimensional linear functions, $\log|\mathcal{F}|$ can be replaced by the dimension $d$, which is an upper bound on the metric entropy.

[13] Precisely, $\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$, where $\mathcal{T}$ is the *Bellman operator*: $[\mathcal{T}f](s,a) := R(s,a) + \mathbb{E}_{s' \sim P(s,a)}[\max_{a'} f(s',a')]$.

1. *We have* $Q^\pi \in \mathcal{F}$ *for all* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ *(all-policy realizability) and* $C_{\mathrm{conc}} \leq 16$ *(concentrability) for all models in* $\mathcal{M}$*.*

2. *Any algorithm using less than* $c \cdot S^{1/3}$ *samples must have* $J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \geq c'/(1-\gamma)$ *for some instance in* $\mathcal{M}$*, where* $c$ *and* $c'$ *are absolute numerical constants.*

This result shows that even though realizability and concentrability are satisfied, any algorithm requires at least $\Omega(S^{1/3})$ samples to learn a near-optimal policy. Since $S$ can be arbitrarily large, this establishes that sample-efficient offline RL in large state spaces is impossible without stronger representation or coverage conditions and resolves the conjecture of Chen and Jiang (2019).

In fact, the theorem establishes hardness under a substantially stronger representation condition than realizability—*all policy realizability*—which requires that $Q^\pi \in \mathcal{F}$ for *every* policy $\pi$, rather than just for $\pi^\star$. When one has the ability to interact with the MDP starting from the data collection distribution $\mu$ (e.g., via a generative model), it is known that all policy realizability and concentrability suffice for *approximate policy iteration* methods (Antos et al. 2008, Lattimore et al. 2020). However, the offline RL setting does not permit interaction, and so Theorem 4.1 yields a separation between offline RL and online RL with a generative model (and an exploratory distribution). The lower bound construction can also be extended to related settings, including policy evaluation and linear function approximation; see Section 4.2.4 for discussion.

Theorem 4.1 relies on a strong version of the *over-coverage* phenomenon, where the data distribution contains states not visited by any admissible policy.[14] The issue of over-coverage was first noted by Xie and Jiang (2021), who observed that it can lead to pathological behavior in certain algorithms. Our result shows—somewhat surprisingly—that this phenomenon is a fundamental barrier that applies to *any* value approximation method. In particular, we show that over-coverage causes spurious correlations across reachable and unreachable states which leads to significant uncertainty in the dynamics when the number of states is large.

Theorem 4.1 has constant suboptimality gap for $Q^\star$, which rules out gap-dependent regret bounds as a path toward sample-efficient offline RL. We focus on policy optimization and infinite-horizon RL for concreteness, but the lower bound readily extends to the finite-horizon setting (in fact, with $H = 3$), and provides, to our knowledge, the first impossibility result

[14] Note that while the states may not be reachable for a given MDP in the family $\mathcal{M}$, in our construction, all states are reachable for *some* MDP in the family.

for offline RL with constant horizon.

**A lower bound for admissible data distributions** Up to this point, we have considered the most ubiquitous formulation of the offline RL problem, in which $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ is an arbitrary distribution over state-action pairs. Theorem 4.1 exploits this formulation by placing mass on states not reachable by any policy, leading to a strong version of the over-coverage phenomenon. Our next result shows concentrability and realizability are still insufficient for sample-efficient offline RL even when the data distribution $\mu$ is *admissible*, in the sense that it is induced by a policy or mixture of policies. While strong over-coverage is impossible in this setting, the lower bound relies on a weak notion of over-coverage in which $\mu$ places significant mass on low-probability states.

**Theorem 4.2** (Lower bound for admissible data)**.** *For any $S \geq 9$, $\gamma \in (1/2, 1)$, and $C \geq 64$, there exists a family of MDPs $\mathcal{M}$ with $|\mathcal{S}| = S$ and $|\mathcal{A}| = 2$, a value function class $\mathcal{F}$ with $|\mathcal{F}| = 2$, and a data distribution $\mu$ which is a mixture of admissible distributions, such that:*

1. *We have $Q^\pi \in \mathcal{F}$ for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (all-policy realizability) and $C_{\mathrm{conc}} \leq C$ (concentrability) for all models in $\mathcal{M}$.*
2. *Any algorithm using less than $c \cdot \min\{S^{1/3}/(\log S)^2, 2^{C/32}, 2^{1/(1-\gamma)}\}$ samples must have $J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \geq c'$ for some instance in $\mathcal{M}$, where $c$ and $c'$ are absolute numerical constants.*

Compared to Theorem 4.1, which shows that for general data distributions any algorithm must have sample complexity polynomial in the number of states even when concentrability is constant, Theorem 4.2 shows that, for admissible data distributions,[15] any algorithm must have sample complexity that is *either* polynomial in the number of states *or* exponential in concentrability (or the effective horizon $(1-\gamma)^{-1}$). This result is incomparable to Theorem 4.1 since, it is quantitatively slightly weaker from a sample complexity perspective, but stronger in that applies to admissible data distributions. Since admissible distributions are perhaps more natural in practice, Theorem 4.2 serves as a strong impossibility result.

While we cannot rely on strong over-coverage to prove Theorem 4.2, we are still able to create spurious correlations between a set of states that are useful for estimation and the

[15]The fact that the data collection distribution is a mixture, is not critical for the result. It can be weakened to a single admissible distribution with realizability (rather than all-policy realizability).

remaining states, which are less useful. Indeed, our construction embeds a structure used in the proof of Theorem 4.1 in a nested fashion, so that even an admissible data distribution provides insufficient information to disentangle this correlation and learn a near-optimal policy.

### 4.1.3 Related Work

We close this section with a detailed discussion of some of the most relevant related work.

**Lower bounds** While algorithm-specific counterexamples for offline reinforcement learning algorithms have a long history (Gordon 1995, Tsitsiklis and Van Roy 1996, 1997, Wang et al. 2021a), information-theoretic lower bounds are a more recent subject of investigation. Wang et al. (2020a) (see also Amortila et al. (2020)) consider the setting where $\mathcal{F}$ is linear (i.e., $Q^\star(s,a) = \langle \phi(s,a), \theta \rangle$, where $\phi(s,a) \in \mathbb{R}^d$ is a known feature map). They consider a weaker coverage condition tailored to the linear setting, which asserts that $\lambda_{\min}\big(\mathbb{E}_{(s,a)\sim\mu}\big[\phi(s,a)\phi(s,a)^\top\big]\big) \geq \frac{1}{d}$, and they show that this condition and realizability alone are not strong enough for sample-efficient offline RL. The feature coverage condition is strictly weaker than concentrability, so this does not suffice to resolve the conjecture of Chen and Jiang (2019). Instead, the conceptual takeaway is that the feature coverage condition can lead to *under-coverage* and may not be the right assumption for offline RL. This point is further highlighted by Amortila et al. (2020) who show that in the infinite-horizon setting, the feature coverage condition can lead to non-identifiability in MDPs with only two states, meaning one cannot learn an optimal policy even with infinitely many samples. Concentrability places stronger restrictions on the data distribution and underlying dynamics and always implies identifiability when the state and action space are finite. Establishing impossibility of sample-efficient learning under concentrability and realizability requires very new ideas (which we provide in this paper, via the notion of *over-coverage*).

The results of Wang et al. (2020a) and Amortila et al. (2020) are extended by Zanette (2021), who provides a slightly more general lower bound for linear realizability. The results of Zanette (2021) *cannot* resolve the conjecture of Chen and Jiang (2019) either, because for the family of MDPs constructed therein, no data distribution can satisfy concentrability, which means that the failure of algorithms can still be attributed to the failure of concentrability rather than the hardness under concentrability. There is also a parallel line of work providing

lower bounds for *online* reinforcement learning with linear realizability (Du et al. 2020a, Weisz et al. 2021, Wang et al. 2021b), which are based on very different constructions and techniques.

Compared to the offline RL lower bounds above (Wang et al. 2020a, Amortila et al. 2020, Zanette 2021), our lower bounds have a less geometric, more information-theoretic flavor, and share more in common with lower bounds for sparsity and support testing in statistical estimation (Paninski 2008, Verzelen and Villers 2010, Verzelen and Gassiat 2018, Canonne 2020). While previous work considers a relatively small state space but large horizon and feature dimension, we grow the state space, leading to polynomial dependence on $S$ in our lower bounds; the horizon is somewhat immaterial in our construction.

Another interesting feature is that while previous lower bounds (Wang et al. 2020a, Amortila et al. 2020, Zanette 2021) are based on deterministic MDPs, our constructions critically use stochastic dynamics, which is a *necessary* departure from a technical perspective. Indeed, for *any* family of deterministic MDPs, *any* data distribution satisfying concentrability (if such a distribution exists) would enable sample-efficient learning, simply because all MDPs in the family have deterministic dynamics, and the Bellman error minimization algorithm in Chen and Jiang (2019) succeeds under concentrability and realizability when the dynamics are deterministic.[16] Therefore, any construction involving deterministic MDPs (Wang et al. 2020a, Amortila et al. 2020, Zanette 2021) cannot be used to establish impossibility of sample-efficient learning under concentrability and realizability.

**Upper bounds** Classical analyses for offline reinforcement learning algorithms such as FQI (Munos 2003, 2007, Munos and Szepesvári 2008, Antos et al. 2008) provide sample complexity upper bounds in terms of concentrability under the strong representation condition of Bellman completeness. The path-breaking recent work of Xie and Jiang (2021) provides an algorithm which requires only realizability, but uses a stronger coverage condition ("pushforward concentrability") which requires that $P(s' \mid s, a)/\mu(s') \leq C$ for all $(s, a, s')$. Our results imply that this condition cannot be substantially relaxed.

A complementary line of work, primarily focusing on policy evaluation (Uehara et al. 2020, Xie and Jiang 2020, Jiang and Huang 2020, Uehara et al. 2021), provides upper bounds that require only concentrability and realizability, but assume access to an additional

[16] Deterministic dynamics allow one to avoid the well-known *double sampling* problem and in particular cause the conditional variance in Eq. (3) of Chen and Jiang (2019) to vanish.

*weight function class* that is flexible enough to represent various occupancy measures for the underlying MDP. These results scale with the complexity of the weight function class. In general, the complexity of this class may be prohibitively large without prior knowledge; this is witnessed by our lower bound construction.

### 4.1.4 Preliminaries

For any $x \in \mathbb{R}$, let $(x)_+ := \max\{x, 0\}$. For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \dots, n\}$. For a finite set $\mathcal{X}$, $\mathrm{Unif}(\mathcal{X})$ denotes the uniform distribution over $\mathcal{X}$, and $\Delta(\mathcal{X})$ denotes the set of all probability distributions over $\mathcal{X}$. For probability distributions $\mathbb{P}$ and $\mathbb{Q}$ over a measurable space $(\Omega, \mathscr{F})$ with a common dominating measure, we define the total variation distance as $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int |d\mathbb{P} - d\mathbb{Q}|$ and define the $\chi^2$-divergence as $D_{\chi^2}(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\big[\big(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\big)^2\big] = \int \frac{d\mathbb{P}^2}{d\mathbb{Q}} - 1$ when $\mathbb{P} \ll \mathbb{Q}$ and $+\infty$ otherwise.

## 4.2 Fundamental Barriers for Offline Reinforcement Learning

In this section we present the lower bound construction for Theorem 4.1 and prove the result, then discuss consequences. The proof of Theorem 4.2—which can be viewed as a generalization of this result, but is somewhat more involved—is deferred to Appendix C.5, with an overview given in Section 4.3.[17]

### 4.2.1 Construction: MDP Family, Value Functions, and Data Distribution

We first provide our lower bound construction, which entails specifying the MDP family $\mathcal{M}$, the value function class $\mathcal{F}$, and the data distribution $\mu$.

All MDPs in $\mathcal{M}$ belong to a parameterized MDP family with shared transition and reward structure. In what follows, we first describe the structure of the parameterized family (Section 4.2.1) and provide intuition behind why this structure leads to statistical hardness (Section 4.2.1). We then provide a specific collection of parameters that gives rise to the hard family $\mathcal{M}$ (Section 4.2.1) and complete the construction by specifying the value function class $\mathcal{F}$ and data distribution $\mu$ (Section 4.2.1).

[17] Compared to the first version of this paper (arXiv preprint v1), the current version uses a slight modification to the Theorem 4.1 construction. The only purpose of this change is to emphasize similarity to the construction for Theorem 4.2.

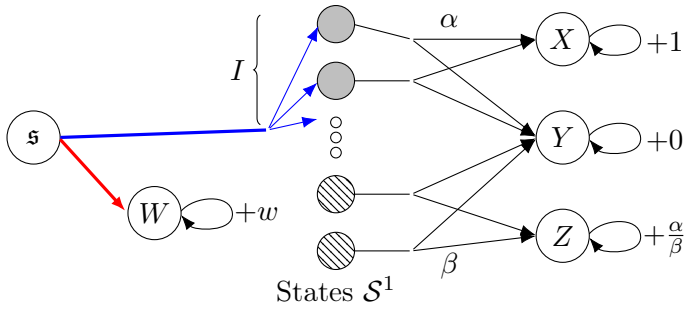


Figure 4-1: The MDPs in $\mathcal{M}$ are parametrized by three scalars $\alpha, \beta, w$ and a subset of states $I$. The state space consists of an *initial state* $\mathfrak{s}$, a large number of *intermediate states* $\mathcal{S}^1$, and four self-looping *terminal states* $\{W, X, Y, Z\}$. From the initial state $\mathfrak{s}$, action 1 (in red) transitions to state $W$, while action 2 (in blue) transitions to a subset of intermediate states $I \subset \mathcal{S}^1$ with equal probability. In all intermediate states and terminal states, actions 1 and 2 have the same effect, with transitions denoted in black. Among the intermediate states, $I \subset \mathcal{S}^1$ (the gray ones) are the *planted states* which transition with probability $\alpha$ to state $X$ and $1-\alpha$ to state $Y$, and the remaining $\mathcal{S}^1 \setminus I$ (the striped ones) are the *unplanted states* which transition with probability $\beta$ to $Z$ and $(1-\beta)$ to $Y$. There are combinatorially many choices for $I$. Only terminal states can generate non-zero rewards: the rewards of the states $W$, $X$, $Y$ and $Z$ are $w$, 1, 0 and $\alpha/\beta$, respectively.

### MDP Parameterization

Let the discount factor $\gamma \in (0,1)$ be fixed, and let $S \in \mathbb{N}$ be given. Assume without loss of generality that $S > 5$ and that $(S-5)/4$ is an integer. We consider the parameterized MDP family illustrated in Figure 4-1. Each MDP takes the form $M_{\alpha,\beta,w,I} = (\mathcal{S}, \mathcal{A}, P_{\alpha,\beta,I}, R_{\alpha,\beta,w}, \gamma, d_0)$, and is parametrized by two probability parameters $\alpha, \beta \in (0,1)$, a reward parameter $w \in [0,1]$, and a subset of states $I$. All MDPs in the family $\{M_{\alpha,\beta,w,I}\}$ share the same state space $\mathcal{S}$, action space $\mathcal{A}$, discount factor $\gamma$, and initial state distribution $d_0$, and differ only in terms of the transition function $P_{\alpha,\beta,I}$ and the reward function $R_{\alpha,\beta,w}$.

**State space** We consider a state space $\mathcal{S} := \{\mathfrak{s}\} \cup \mathcal{S}^1 \cup \{W, X, Y, Z\}$, where $\mathfrak{s}$ is the single *initial* state (occurring at $h = 0$), $W, X, Y, Z$ are four self-looping *terminal* states, and $\mathcal{S}^1$ is a collection of *intermediate* (i.e., neither initial nor terminal) states which may occur between the initial state and the terminal states $\{X, Y, Z\}$. The number of intermediate states is $S_1 := |\mathcal{S}^1| = S - 5$ which ensures $|\mathcal{S}| = S$.

**Action space** Our action space is given by $\mathcal{A} = \{1, 2\}$. For the initial state $\mathfrak{s}$, the two actions have distinct effects, while for all other states in $\mathcal{S} \setminus \{\mathfrak{s}\}$ both actions have identical effects. As a result, the value of a given policy only depends on the action it selects in $\mathfrak{s}$. For

the sake of compactness, we use the symbol $\mathfrak{a}$ as a placeholder to denote either action when taken in $s \in \mathcal{S} \setminus \{\mathfrak{s}\}$, since the choice is immaterial.[18]

**Transition operator** For an MDP $M_{\alpha,\beta,w,I}$, we let $I \subset \mathcal{S}^1$ parameterize a subset of the intermediate states. We call each $s \in I$ a *planted state* and $s \in \bar{I} := \mathcal{S}^1 \setminus I$ an *unplanted state*. The dynamics $P_{\alpha,\beta,I}$ for $M_{\alpha,\beta,w,I}$ are determined by $I$ and the parameters $\alpha, \beta \in (0,1)$ as follows (cf. Figure 4-1):

- *Initial state* $\mathfrak{s}$. We define $P_{\alpha,\beta,I}(\mathfrak{s}, 1) = \text{Unif}(\{W\})$ and $P_{\alpha,\beta,I}(\mathfrak{s}, 2) = \text{Unif}(I)$. That is, from the initial state $\mathfrak{s}$, choosing action 1 makes the MDP transitions to state $W$ deterministically (see the red arrow in Figure 4-1), while choosing 2 makes the MDP transitions to each planted state in $I$ with equal probability (see the blue arrow in Figure 4-1); unplanted states are not reachable.

- *Intermediate states.* Transitions from states in $\mathcal{S}^1$ are defined as follows.

  - For each planted state $s \in I$, define

    $$P_{\alpha,\beta,I}(s, \mathfrak{a}) = \alpha\text{Unif}(\{X\}) + (1-\alpha)\text{Unif}(\{Y\}).$$

  - For each unplanted state $s \in \bar{I}$, define

    $$P_{\alpha,\beta,I}(s, \mathfrak{a}) = \beta\text{Unif}(\{Z\}) + (1-\beta)\text{Unif}(\{Y\}).$$

  That is, the MDP transitions stochastically to either $\{X, Y\}$ or $\{Z, Y\}$, depending on whether the source state $s \in \mathcal{S}^1$ is planted or unplanted; see the black straight arrows in Figure 4-1.

- *Terminal states.* All states in $\{W, X, Y, Z\}$ self-loop indefinitely. That is $P_{\alpha,\beta,I}(s, \mathfrak{a}) = \text{Unif}(\{s\})$ for all $s \in \{W, X, Y, Z\}$.

**Reward function** The initial and intermediate states have no reward, i.e., $R_{\alpha,\beta,w}(s, a) = 0, \forall s \in \{\mathfrak{s}\} \cup \mathcal{S}^1, \forall a \in \mathcal{A}$. Each of the self-looping terminal states $\{W, X, Y, Z\}$ has a fixed

[18] It is conceptually simpler to consider a construction where only a single action is available in $\mathcal{S} \setminus \{\mathfrak{s}\}$ and $\mathcal{S}^1$, but this is notationally more cumbersome.

reward determined by the parameters $\alpha$, $\beta$ and $w$. In particular, we define $R_{\alpha,\beta,w}(W, \mathfrak{a}) = w$, $R_{\alpha,\beta,w}(X, \mathfrak{a}) = 1$, $R_{\alpha,\beta,w}(Y, \mathfrak{a}) = 0$, and $R_{\alpha,\beta,w}(Z, \mathfrak{a}) = \alpha/\beta$.

**Initial state distribution** All MDPs in $M_{\alpha,\beta,w,I}$ start at $\mathfrak{s}$ deterministically (that is, the initial state distribution $d_0$ puts all the probability mass on $\mathfrak{s}$). Note that since $d_0$ does not vary between instances, it may be thought of as *known* to the learning algorithm.

### Intuition Behind the Construction

The family of MDPs $\mathcal{M}$ that witnesses our lower bound is a subset of the collection $\{M_{\alpha,\beta,w,I}\}$. Before specifying the family precisely, we give intuition as to why this MDP structure leads to statistical hardness for offline reinforcement learning.

Evidently, for any MDP $M_{\alpha,\beta,w,I}$, there is only a single effective decision that the learner needs to make: to choose action 1 in $\mathfrak{s}$ (whose value is completely determined by $w$) or to choose action 2 in $\mathfrak{s}$ (whose value is completely determined by $\alpha$). In our construction of the MDP family (to be specified shortly), we keep $w$ fixed over all MDPs (i.e., we make it *known* to the learner), so the only challenge left to the learner is to learn the value of $\alpha$ of the underlying MDP. As we will explain, this seemingly simple task is surprisingly hard, leading to the hardness of offline reinforcement learning. The hardness arises as a result of two general principles, *planted subset structure* and (strong) *over-coverage*.

**Planted Subset Structure** The intermediate states in $\mathcal{S}^1$ are partitioned into planted and unplanted states. Each planted state in $I$ (the *planted subset*) transitions to $X$ and $Y$ with probability $\alpha$ and $1-\alpha$ respectively, while each unplanted state in $\bar{I}$ transitions to $Z$ and $Y$ with probability $\beta$ and $1-\beta$ respectively. We call such a structure the *planted subset structure*, which has two important features:

- The choice of $I \subset \mathcal{S}^1$ is combinatorial in nature (for example, the number of all planted subsets of size $S_1/2$ is $\binom{S_1}{S_1/2}$, which is exponential in $S_1$).
- Planted and unplanted states have the same value, which only depends on $\alpha$. This holds because the rewards of $X, Y, Z$ are $1, 0, \alpha/\beta$ respectively (note that $\alpha \cdot 1 = \beta \cdot (\alpha/\beta)$). As a result, the choice of $I \subset \mathcal{S}^1$ does not affect the value function at all.

The first feature serves as the basis for statistical hardness and leads to the appearance of the state space size in the sample complexity lower bound. For intuition as to why, suppose

we are given a batch dataset of independent examples in which a state in $\mathcal{S}^1$ is selected uniformly at random and we observe a sample from the next state distribution. One can show that basic statistical inference tasks such as estimating the size $|I|$ of the planted subset require poly($S_1$) samples, as this entails detecting the subset based on data generated from a mixture of planted and unplanted states. For example, it is well known that testing if a distribution is uniform on a set $I \subset [N]$ with $|I| = \Theta(N)$ versus uniform on all of $[N]$ requires poly($N$) samples (see e.g., Paninski 2008, Ingster and Suslina 2012, Canonne 2020, Section 5.1).

Building on this hardness, we can show that any algorithm requires at least poly($S_1$) samples to reliably estimate the transition probability parameter $\alpha$ if $\beta$ and $|I|$ are unknown. Intuitively, this arises because the only way to avoid estimating $|I|$ (which is hard) as a means to estimate $\alpha$ is to directly look at the marginal distribution over $\{X, Y, Z\}$. However, the marginal distribution is uninformative for estimating $\alpha$ when there is uncertainty about $\beta$ and $|I|$. For example, the marginal probability of transitioning to $X$ is $\alpha|I|/S_1$, from which $\alpha$ cannot be directly recovered if $|I|$ is unknown.

The takeaway is that while estimating $\alpha$ would be trivial if the dataset only consisted of transitions generated from planted states, estimating this parameter when states are drawn uniformly from $\mathcal{S}^1$ is very difficult because an unknown subset comes from unplanted states. This is relevant because—as we will show—in our construction, any near-optimal policy learning algorithm must have the ability to recover the value of $\alpha$.

The second feature, that all states in $\mathcal{S}^1$ share the same value, is also essential. Since the choice of $I \subset \mathcal{S}^1$ does not affect the value function at all, this feature allows us to consider exponentially many choices of $I$ while ensuring that realizability is satisfied with a value function class $\mathcal{F}$ of constant size. Thus, the poly($|\mathcal{S}|$) factor in our lower bound cannot be attributed to any other problem parameter, such as $\log |\mathcal{F}|$.

**(Strong) Over-coverage** It remains to show that the hardness described above can be embedded in the offline RL setting, since (i) we must ensure concentrability is satisfied, and (ii) the learner observes rewards, not just transitions. Returning to Figure 4-1, we observe that the transitions from the initial state $\mathfrak{s}$ are such that all planted states in $I$ are *reachable*, but the unplanted states in $\mathcal{S}^1 \setminus I$ are *not reachable by any policy*. In particular, since all unplanted states are unreachable, any state that can only be reached from unplanted states

is also unreachable, and hence we can achieve concentrability (4.1) without covering such states. This allows us to choose the data distribution $\mu$ to be (roughly) uniform over all states except for the unreachable state $Z$. This choice satisfies concentrability, but renders all reward observations uninformative (cf. Section 4.2.1). As a consequence, we show that the task described in Section 4.2.1, i.e., detecting $\alpha$ based on transition data, is unavoidable for any algorithm with non-trivial offline RL performance.

The key principle at play here is the over-coverage phenomenon (in particular, the strong version, where $\mu$ is supported over unreachable states). Per the discussion above, we know that if the data distribution $\mu$ were supported only over reachable states for a given MDP, all "time step 1" examples $(s, a, r, s')$ in $D_n$ would have $s \in I$, which would make estimating $\alpha$ trivial. Our construction for $\mu$ is uniform over all states in $\mathcal{S}^1$, and hence satisfies over-coverage, since it is supported over a mix of planted states and spurious (unplanted) states not reachable by any policy. This makes estimating $\alpha$ challenging because—due to correlations between planted and unplanted states—no algorithm can accurately estimate $\alpha$ or recover the planted states until the number of samples scales with the number of states. We emphasize, however, that while strong over-coverage makes the construction for Theorem 4.1 comparatively simple, the weak variant of over-coverage (where all states are reachable, but the offline data distribution creates a spurious correlation by favoring unplanted states) still presents a fundamental barrier and is the mechanism behind Theorem 4.2.

**Specifying the MDP Family**

Using the parameterized MDP family $\{M_{\alpha,\beta,w,I}\}$, we construct the hard family $\mathcal{M}$ for our lower bound by selecting a specific collection of values for the parameters $(\alpha, \beta, w, I)$. Define $\mathcal{I}_\theta := \{I : |I| = \theta S_1\}$ for all $\theta \in (0, 1)$ such that $\theta S_1$ is an integer. We define two sub-families of MDPs,

$$\mathcal{M}_1 := \bigcup_{I \in \mathcal{I}_{\theta_1}} \{M_{\alpha_1,\beta_1,w,I}\}, \quad \text{and} \quad \mathcal{M}_2 := \bigcup_{I \in \mathcal{I}_{\theta_2}} \{M_{\alpha_1,\beta_1,w,I}\},$$

where $w := \gamma(\alpha_1 + \alpha_2)/2$ is fixed for all MDPs, $\mathcal{M}_1$ is specified by $(\theta_1, \alpha_1, \beta_1) = (1/2, 1/4, 3/4)$, and $\mathcal{M}_2$ is specified by $(\theta_2, \alpha_2, \beta_2) = (1/4, 1/2, 1/2)$.[19] Finally, we define the hard family $\mathcal{M}$

[19]Recall that we assume without loss of generality that $S_1/4$ is an integer.

via

$$\mathcal{M} := \mathcal{M}_1 \cup \mathcal{M}_2.$$

Let us discuss some basic properties of the construction that are used to prove the lower bound.

- For all MDPs in $\mathcal{M}$, the rewards of the terminal states $W, X, Y$ are the same. This means there is no uncertainty in the reward function outside of state $Z$, which has $R_{\alpha_1,\beta_1,w}(Z, \mathfrak{a}) = \alpha_1/\beta_1 = 1/3$ when $M \in \mathcal{M}_1$ and $R_{\alpha_2,\beta_2,w}(Z, \mathfrak{a}) = \alpha_2/\beta_2 = 1$ when $M \in \mathcal{M}_2$. As we mentioned, the reward of $Z$ $(= \alpha/\beta)$ is chosen to ensure that all states in $\mathcal{S}^1$ have the same value $(= \gamma\alpha/(1-\gamma))$, given the choice for $(\alpha, \beta)$.

- All MDPs in $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) differ only in the choice of $I \subset \mathcal{S}^1$. This property, along with the aforementioned fact that all states in $\mathcal{S}^1$ have the same value $\gamma\alpha_1/(1-\gamma)$ (resp. $\gamma\alpha_2/(1-\gamma)$), ensures that $Q^\star_M$ is the same for all $M \in \mathcal{M}_1$ (resp. $M \in \mathcal{M}_2$). Furthermore, our choice for $w \in (\gamma\alpha_1, \gamma\alpha_2)$ ensures that the optimal action in $\mathfrak{s}$ is action 1 (resp. action 2) for all MDPs in $\mathcal{M}_1$ (resp. $\mathcal{M}_2$).

- Our choice for $(\theta_1, \alpha_1, \beta_1)$ and $(\theta_2, \alpha_2, \beta_2)$ ensures that the marginal distribution of $s'$ under the process $s \sim \mathrm{Unif}(\mathcal{S}^1), s' \sim P(s, \mathfrak{a})$ is the same for all $M \in \mathcal{M}$. This property is motivated by the hard inference task described in Section 4.2.1, which requires an uninformative marginal distribution.

The exact numerical values for the MDP parameters chosen above are not essential to the result. Any tuple $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; w)$ can be used to establish a result similar to Theorem 4.1, as long as it satisfies certain properties described in Appendix C.1.

**Finishing the Construction: Value Functions and Data Distribution**

We complete our construction by specifying a value function class $\mathcal{F}$ that satisfies (all-policy) realizability and a data distribution $\mu$ that satisfies concentrability (4.1).

**Value function class** Define functions $f_1, f_2 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows; differences are highlighted in blue:

$$f_1(s,a) := \frac{1}{1-\gamma} \cdot \begin{cases} \frac{3}{8}\gamma^2, & s = \mathfrak{s},\ a = 1 \\ \frac{1}{4}\gamma^2, & s = \mathfrak{s},\ a = 2 \\ \frac{1}{4}\gamma, & s \in \mathcal{S}^1 \\ \frac{3}{8}\gamma, & s = W \\ 1, & s = X \\ 0, & s = Y \\ \frac{1}{3}, & s = Z \end{cases} \quad \text{and} \quad f_2(s,a) := \frac{1}{1-\gamma} \cdot \begin{cases} \frac{3}{8}\gamma^2, & s = \mathfrak{s},\ a = 1 \\ \frac{1}{2}\gamma^2, & s = \mathfrak{s},\ a = 2 \\ \frac{1}{2}\gamma, & s \in \mathcal{S}^1 \\ \frac{3}{8}\gamma, & s = W \\ 1, & s = X \\ 0, & s = Y \\ 1, & s = Z \end{cases}. \tag{4.2}$$

The following result is elementary; see Appendix C.2 for a detailed calculation.

**Proposition 4.1.** *For all $M \in \mathcal{M}_1$, we have $Q^\pi_M = f_1$ for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. For all $M \in \mathcal{M}_2$, we have $Q^\pi_M = f_2$ for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.*

It follows that by choosing $\mathcal{F} := \{f_1, f_2\}$, all-policy realizability holds for all $M \in \mathcal{M}$. Note that the all-policy realizability condition ($Q^\pi_M \in \mathcal{F}$ for all $M \in \mathcal{M}$ and for *all* policies $\pi$) is substantially stronger than the standard realizability condition ($Q^\star_M \in \mathcal{F}$ for all $M \in \mathcal{M}$), as it requires $Q^\pi \in \mathcal{F}$ for *every* policy rather than just for $\pi^\star$. Since the conjecture of Chen and Jiang (2019) only asks for a construction that satisfies standard realizability, by considering all-policy realizability, we are proving a *stronger* hardness result. This is possible because in our construction, different actions have identical effects on all states except for the initial state $\mathfrak{s}$; as a result, $Q^\pi$ does not depend on $\pi$ at all (in other words, our construction ensures that $Q^\pi$ is always the same as $Q^\star$).

**Data distribution** Recall that the learner is provided with an i.i.d. dataset $D_n = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ where $(s_i, a_i) \sim \mu$, $s'_i \sim P(\cdot \mid s_i, a_i)$, and $r_i = R(s_i, a_i)$ (here $P$ and $R$ are the transition and reward functions for the underlying MDP). We define the data collection distribution via:

$$\mu := \frac{1}{8}\mathrm{Unif}(\{\mathfrak{s}\} \times \{1,2\}) + \frac{1}{2}\mathrm{Unif}(\mathcal{S}^1 \times \{1,2\}) + \frac{3}{8}\mathrm{Unif}(\{W,X,Y\} \times \{1,2\}).$$

This choice for $\mu$ forces the learner to suffer from the hardness described in Section 4.2.1. Salient properties include: (i) both planted and unplanted states in $\mathcal{S}^1$ are covered, and (ii) the state $Z$ is not covered. Property (i) results in strong over-coverage, which makes estimating the parameters of the underlying MDP from transitions statistically hard, while property (ii) hides the difference between the rewards of $Z$ for the two-subfamilies of MDPs and hence makes all reward observations uninformative.

We now verify the concentrability condition (4.1):

- For time step $h = 0$, for any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the distribution of $(s_0, a_0)$ is $d_0 \times \pi$. It follows that
$$\left\| \frac{d_0 \times \pi}{\mu} \right\|_\infty \leq \frac{1}{\frac{1}{8} \cdot \frac{1}{2}} = 16.$$

- For time step $h = 1$, for any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the distribution of $(s_1, a_1)$ is $\mathrm{Unif}(I) \times \pi$. We conclude that
$$\left\| \frac{\mathrm{Unif}(I) \times \pi}{\mu} \right\|_\infty \leq \frac{\frac{1}{S_1/4}}{\frac{1}{2} \cdot \frac{1}{S_1} \cdot \frac{1}{2}} = 16,$$
where we have used that $|I| \geq S_1/4$.

- For time step $h \geq 2$, for any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the distribution of $(s_h, a_h)$ (denoted by $d_h^\pi$) is supported on $\{W, X, Y\} \times \{1, 2\}$. Therefore, we have
$$\left\| \frac{d_h^\pi}{\mu} \right\|_\infty \leq \frac{1}{\frac{3}{8} \cdot \frac{1}{3} \cdot \frac{1}{2}} = 16.$$

We conclude that the construction satisfies concentrability with $C_{\mathrm{conc}} = 16$.

### 4.2.2 Proof of Theorem 4.1

Having specified the lower bound construction, we proceed to prove Theorem 4.1. For any MDP $M \in \mathcal{M}$, we know from (4.2) that the optimal policy $\pi_M^\star$ has

$$\pi_M^\star(\mathfrak{s}) = \begin{cases} 1, & \text{if } M \in \mathcal{M}_1 \\ 2, & \text{if } M \in \mathcal{M}_2 \end{cases},$$

and that $Q^\star_M$ has a constant gap in value between the optimal and suboptimal actions in the initial state $\mathfrak{s}$:

$$Q^\star_M(\mathfrak{s}, \pi^\star_M(\mathfrak{s})) - Q^\star_M(\mathfrak{s}, a) \geq \frac{1}{8}\frac{\gamma^2}{1-\gamma}, \quad \forall a \neq \pi^\star_M(\mathfrak{s}). \tag{4.3}$$

This implies that any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ with high value must choose action 1 in $\mathfrak{s}$ with high probability when $M \in \mathcal{M}_1$, and must choose action 2 in $\mathfrak{s}$ with high probability when $M \in \mathcal{M}_2$. As a result, any offline RL algorithm with non-trivial performance must reliably distinguish between $M \in \mathcal{M}_1$ and $M \in \mathcal{M}_2$ using the offline dataset $D_n$. In what follows we make this intuition precise.

For each $M \in \mathcal{M}$, let $\mathbb{P}_n^M$ denote the law of the offline dataset $D_n$ when the underlying MDP is $M$, and let $\mathbb{E}_n^M$ be the associated expectation operator. We formalize the idea of distinguishing between $M \in \mathcal{M}_1$ and $M \in \mathcal{M}_2$ using Lemma 4.1, which reduces the task of proving a policy learning lower bound to the task of upper bounding the total variation distance between two *mixture distributions* $\mathbb{P}_n^1 := \frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1}\mathbb{P}_n^M$ and $\mathbb{P}_n^2 := \frac{1}{|\mathcal{M}_2|}\sum_{M\in\mathcal{M}_2}\mathbb{P}_n^M$.

**Lemma 4.1.** *Let $\gamma \in (0,1)$ be fixed. For any offline RL algorithm which takes $D_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ as input and returns a stochastic policy $\widehat{\pi}_{D_n} : \mathcal{S} \to \Delta(\mathcal{A})$, we have*

$$\sup_{M\in\mathcal{M}} \left\{ J_M(\pi^\star_M) - \mathbb{E}_n^M[J_M(\widehat{\pi}_{D_n})] \right\} \geq \frac{\gamma^2}{16(1-\gamma)}\big(1 - D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)\big).$$

Lemma 4.1 implies that if the difference between the average dataset generated by all $M \in \mathcal{M}_1$ and that generated by all $M \in \mathcal{M}_2$ is sufficiently small, no algorithm can reliably distinguish $M \in \mathcal{M}_1$ and $M \in \mathcal{M}_2$ based $D_n$, and hence must have poor performance on some instance. See Appendix C.3 for a proof.

We conclude by bounding $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$. Since directly calculating the total variation distance is difficult, we proceed in two steps. We first design an auxiliary *reference measure* $\mathbb{P}_n^0$, and then bound $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^0\big)$ and $D_{\mathrm{TV}}\big(\mathbb{P}_n^2, \mathbb{P}_n^0\big)$ separately. For the latter step, we move from total variation distance to $\chi^2$-*divergence* and bound $D_{\chi^2}\big(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0\big)$ (resp. $D_{\chi^2}\big(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0\big)$) using a mix of combinatorial arguments and concentration inequalities. This constitutes the most technical portion of the proof, and formalizes the intuition about hardness of estimation under planted subset structure described in Section 4.2.1. Our final bound on the total variation distance (proven in Appendix C.4) is as follows.

**Lemma 4.2.** *For all $n \leq \sqrt[3]{(S-5)}/20$, we have*

$$D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) \leq 1/2.$$

Theorem 4.1 immediately follows by combining Lemma 4.1 and Lemma 4.2. □

### 4.2.3 Discussion

Having proven Theorem 4.1, we briefly interpret the result and discuss some additional consequences. We refer to Section 4.2.4 for extensions and further discussion.

**Separation between online and offline reinforcement learning** In the online reinforcement learning setting, the learner can execute any policy in the underlying MDP and observe the resulting trajectory. Our results show that in general, the separation between the sample complexity of online RL and offline RL can be arbitrarily large, even when concentrability is satisfied. To see this, recall that in the online RL setting, we can evaluate any fixed policy to precision $\varepsilon$ using $\mathrm{poly}((1-\gamma)^{-1}) \cdot \varepsilon^{-2}$ trajectories via Monte-Carlo rollouts. Since the class $\mathcal{M}$ we construct essentially only has two possible choices for the optimal policy and has suboptimality gap $\frac{\gamma^2}{1-\gamma}$, we can learn the optimal policy in the online setting using $\mathrm{poly}((1-\gamma)^{-1})$ trajectories, with no dependence on the number of states. On the other hand, Theorem 4.1 shows that the sample complexity of offline RL for this family can be made arbitrarily large.

**Linear function approximation** The observation above is particularly salient in the context of linear function approximation, where $\mathcal{F} = \big\{(s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^d\big\}$ for a known feature map $\phi(s,a)$. Our lower bound construction for Theorem 4.1 can be viewed as a special case of the linear function approximation setup with $d = 2$ by choosing $\phi(s,a) = (f_1(s,a), f_2(s,a))$. Consequently, our results show that the separation between the complexity of offline RL and online RL with linearly realizable function approximation can be arbitrarily large, even when the dimension is constant. This strengthens one of the results of Zanette (2021), which provides a linearly realizable construction in which the separation between online and offline RL is exponential with respect to dimension.

**Why aren't stronger coverage or representation conditions satisfied?** While our construction satisfies concentrability and realizability, it fails to satisfy stronger coverage and representation conditions for which sample-efficient upper bounds are known. This is to be expected, (or else we would have a contradiction!) but understanding why is instructive. Here we discuss connections to some notable conditions.

*Pushforward concentrability.* The stronger notion of concentrability that $P(s' \mid s, a)/\mu(s') \leq C$ for all $(s, a, s')$, which is used in Xie and Jiang (2021), fails to hold because the state $Z$ is not covered by $\mu$. This presents no issue for standard concentrability because $Z$ is not reachable starting from $\mathfrak{s}$.

*Completeness.* Bellman completeness requires that the value function class $\mathcal{F}$ has $\mathcal{T}_M\mathcal{F} \subseteq \mathcal{F}$ for all $M \in \mathcal{M}$, where $\mathcal{T}_M$ is the Bellman operator for $M$. We show in (4.2) that the set of optimal Q-value functions $\{Q^\star_M\}_{M\in\mathcal{M}}$ is small, but completeness requires that the class remains closed even when we mix and match value functions and Bellman operators from $\mathcal{M}_1$ and $\mathcal{M}_2$, which results in an exponentially large class in our construction. To see why, first note that by Bellman optimality, we must have $\{Q^\star_M\}_{M\in\mathcal{M}} \subseteq \mathcal{F}$ if $\mathcal{F}$ is complete. We therefore also require $\mathcal{T}_{M'}Q^\star_M \in \mathcal{F}$ for $M \in \mathcal{M}_1$ and $M' \in \mathcal{M}_2$. Unlike the optimal Q-functions, which are constant across $\mathcal{S}^1$, the value of $[\mathcal{T}_{M'}Q^\star_M](s, \mathfrak{a})$ for $s \in \mathcal{S}^1$ depends on whether $s \in I$ or $s \in \mathcal{S}^1 \setminus I$, where $I$ is the collection of planted states for $M'$.[20] As a result, there are $\binom{S_1}{|I|}$ possible values for the Bellman backup, which means that the cardinality of $\mathcal{F}$ must be exponential in $S$.

### 4.2.4 Extensions

Theorem 4.1 presents the simplest variant of our lower bound for clarity of exposition. In what follows we sketch some straightforward extensions.

- *Policy evaluation.* Our lower bound immediately extends from policy optimization to policy evaluation. Indeed, letting $\pi^\star_1$ and $\pi^\star_2$ denote the optimal policies for $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, we have $|J_M(\pi^\star_1) - J_M(\pi^\star_2)| \propto \frac{\gamma^2}{1-\gamma}$ for all $M \in \mathcal{M}$, and we know that $J_M(\pi^\star_1)$ is constant across all $M \in \mathcal{M}$. It follows that any algorithm which evaluates policy $\pi^\star_2$ to precision $\varepsilon \cdot \frac{\gamma^2}{1-\gamma}$ with probability at least $1 - \delta$ for sufficiently small

[20] Recall that $f_1$ is the optimal Q-function for any $M \in \mathcal{M}_1$ and consider $\mathcal{T}_{M'}f_1$ where $M' \in \mathcal{M}_2$ has planted set $I$. For $s \in I$, we have $[\mathcal{T}_{M'}f_1](s, \mathfrak{a}) = (1/2 \cdot 1 + 1/2 \cdot 0)\gamma = \gamma/2$ while for $s \in \mathcal{S}^1 \setminus I$, we have $[\mathcal{T}_{M'}f_1](s, \mathfrak{a}) = (1/2 \cdot (1/3) + 1/2 \cdot 0)\gamma = \gamma/6$.

numerical constants $\varepsilon, \delta > 0$ can be used to select the optimal policy with probability $(1-\delta)$, and thus guarantee $J(\pi^\star) - \mathbb{E}[J(\hat{\pi})] \lesssim \delta \frac{\gamma^2}{1-\gamma}$. Hence, such an algorithm must use $n = \Omega(|\mathcal{S}|^{1/3})$ samples by our policy optimization lower bound.

To formally cast this setup in the policy evaluation setting, we take $\Pi = \{\pi_2^\star\}$ as the class of policies to be evaluated, and we require a value function class $\mathcal{F}$ such that $Q_M^\pi \in \mathcal{M}$ for all $\pi \in \Pi$, $M \in \mathcal{M}$. By Proposition 4.1, it suffices to select $\mathcal{F} = \{f_1, f_2\}$.

- *Learning an $\varepsilon$-suboptimal policy.* Theorem 4.1 shows that for any $\gamma \in (1/2, 1)$, $n \gtrsim S^{1/3}$ samples are required to learn a $(1-\gamma)^{-1}$-optimal policy. We can extend the construction to show that more generally, for any $\varepsilon \in (0, 1)$, $n \gtrsim \frac{S^{1/3}}{\varepsilon}$ samples are required to learn an $\varepsilon \cdot (1-\gamma)^{-1}$-optimal policy. We modify the MDP family $M_{\alpha,\beta,w,I}$ by adding a single dummy state $\mathfrak{t}$ with a self-loop and zero reward. The initial state distribution is changed so that $d_0(\mathfrak{t}) = 1 - \varepsilon$ and $d_1(\mathfrak{s}) = \varepsilon$. That is, with probability $1 - \varepsilon$, the agent begins in $\mathfrak{t}$ and stays there forever, collecting no reward, and otherwise the agent begins at $\mathfrak{s}$ and proceeds as in the original construction. Analogously, we replace the original data distribution $\mu$ with $\mu' := (1-\varepsilon)\delta_{\mathfrak{t}} + \varepsilon\mu$, where $\delta_{\mathfrak{t}}$ is a point mass on $\mathfrak{t}$. This preserves the concentrability bound $C_{\mathrm{conc}} \leq 16$. This modification rescales the optimal value functions, and the conclusion of Lemma 4.1 is replaced by

$$\sup_{M \in \mathcal{M}} \left\{ J_M(\pi_M^\star) - \mathbb{E}_n^M[J_M(\widehat{\pi}_{D_n})] \right\} \geq \varepsilon \cdot \frac{\gamma^2}{16(1-\gamma)} \left(1 - D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right)\right).$$

  On the other hand, since samples from the state $\mathfrak{t}$ provide no information about the underlying instance, the effective number of samples is reduced to $\varepsilon n$. One can make this intuition precise and prove that $D_{\mathrm{TV}}\left(\mathbb{P}_n^1, \mathbb{P}_n^2\right) \leq 3/4$ whenever $\varepsilon n \leq c \cdot S^{1/3}$ for a numerical constant $c$. Combining this with the previous bound yields the result.

- *Linear function approximation.* As discussed above, Theorem 4.1 can be viewed as a special case of linear function approximation with $d = 2$ and $\phi(s, a) = (f_1(s, a), f_2(s, a))$. Compared with recent lower bounds in the linear setting (Wang et al. 2020a, Zanette 2021), this result is significantly stronger in that (a) it considers a stronger coverage condition, (b) holds with constant dimension and constant effective horizon, and (c) scales with the number of states, which can be arbitrarily large.

Lastly, it should be clear at this point that our lower bound construction extends to the

finite-horizon setting with $H = 3$ by simply removing the self-loops from the terminal states. The only difference is that the optimal Q-value functions require a new calculation since rewards are no longer discounted.

## 4.3 Proof Overview for Theorem 4.2

In this section we present a high-level overview of the construction and proof for Theorem 4.2. We defer the complete proof, as well as additional discussion, to Appendix C.5. The proof is based on a extension of the construction used in Theorem 4.1. We still use the concept of planted and unplanted states, but since the data collection distribution must be admissible, we cannot rely on strong over-coverage to create spurious correlations. In particular, a naïve adaptation would require that the unplanted states are reachable with sufficient probability, which would necessitate that state $Z$ is supported by $\mu$ with sufficient probability. The resulting construction would not lead to a meaningful lower bound, as the reward information from $Z$ can be used to learn the optimal policy.

To avoid this issue, we modify the construction in Theorem 4.1 to replace $Z$ with another "layer" of states with planted subset structure (see Appendix C.5.1 and Footnote 61 for the precise definition of a layer). By repeating this several times, we obtain a family of MDPs with $L > 1$ layers of planted subset structure, connected in the manner displayed in Figure 4-2 (see Appendix C.5.1 for the details). Specifically, taking action 2 (in blue) from the initial state $\mathfrak{s}$, the $l^{\text{th}}$ layer is selected with probability $\propto 1/2^l$, and we transit uniformly to the states in the $l^{\text{th}}$ layer. In each layer, the planted states behave similarly to the construction for Theorem 4.1, transitioning to terminal states $X$ and $Y$ with specific probabilities that are chosen such that the marginal distribution provides no information. However, except for at the last layer, the unplanted states do not transition directly to the terminal state $Z$, but rather to the planted states of the next layer. Overall, $Z$ can be reached with only $O(1/2^L)$ probability.

Similar to our previous construction, the new multi-layer construction ensures that every MDP in the family differs only in terms of the reward of $Z$ and the transition probabilities for planted and unplanted states. Moreover, while the state $Z$ is no longer unreachable, we know that since all policies only reach $Z$ with exponentially small (in $L$) probability, we can satisfy concentrability with a data collection distribution that places exponentially small

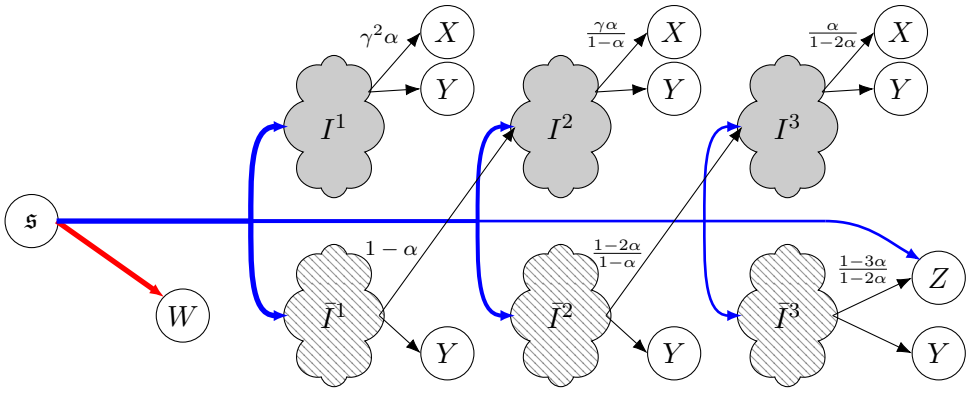


Figure 4-2: Illustration of the MDP family used to prove Theorem 4.2, with $L = 3$ layers of the planted subset structure (note that in general, we take $L > 3$). The rewards of the states $W$, $X$, $Y$ and $Z$ are $w$, 1, 0 and $\alpha/(1 - L\alpha)$, respectively, where $w$ and $\alpha$ are parameters of the MDP family.

mass on $Z$ (which—if it appeared in the dataset—would reveal the optimal policy). As a result, we have that with high probability, all reward observations in the dataset provide no information. Intuitively, this allows us to apply an inductive argument to show that one cannot learn the value function.[21] As the base case, when the reward for $Z$ is unobserved (which happens with high probability), the $L^{\text{th}}$ layer resembles an instance of the construction used to prove Theorem 4.1. Then, going backwards, if one cannot estimate the $(l+1)^{\text{st}}$ layer, we can view the $l^{\text{th}}$ layer as an instance of the previous construction to show that one cannot estimate the value of this layer as well. This induction relies on the delicate design of the data collection distribution $\mu$, which is supported on both planted and unplanted states, but nevertheless exhibits a weak notion of over-coverage resulting in spurious correlations. The argument also requires gradually decreasing the difference in the value function (between the two MDP families) from the $L^{\text{th}}$ layer to the first layer; however, we can ensure that the rate of decrease is very slow, which leads to statistical hardness.

On a technical level, after constructing the MDP family, many of the calculations are similar to those used to prove Theorem 4.1. Analogously to Lemma 4.1, we lower bound the suboptimality of any algorithm by the total variation distance between two mixture distributions. Then we bound this TV distance by constructing auxiliary reference measures and passing to the $\chi^2$-divergence, analogously to Lemma 4.2. Finally, since we rely on a similar planted subset structure, the $\chi^2$-divergence calculation shares many technical elements

[21] The inductive argument is discussed here mainly for providing intuition. The proof in Appendix C.5 is more direct and does not involve a formal inductive argument.

with the proof of Lemma 4.2.

## 4.4 Concluding Remarks

We have proven that concentrability and realizability alone are not sufficient for sample-efficient offline reinforcement learning, resolving the conjecture of Chen and Jiang (2019). Our results establish that sample-efficient offline RL requires coverage or representation conditions beyond what is required for supervised learning and show that over-coverage is a fundamental barrier for offline RL.

For future research, an immediate question is whether it is possible to circumvent our lower bound by considering trajectory-based data rather than $(s, a, r, s')$ tuples. More broadly, while our results elucidate the role of concentrability and realizability, it remains to obtain a sharp, distribution-dependent characterization for the sample complexity of offline RL with general function approximation. Such a characterization would need to recover our result and previous results—both positive and negative—as special cases.

# Chapter 5

# Online Pricing with Offline Data

## 5.1 Introduction

Classical statistical learning theory distinguishes between offline learning and online learning. Offline learning deals with the problem of finding a predictive function based on the entire training data set. The performance of an offline learning algorithm is typically measured by its generalization error (also known as the out-of-sample error) or sample complexity (see, e.g., Hastie et al. 2005). In contrast to the offline learning setting where the entire training data set is directly available before the offline learning algorithm is applied, online learning deals with a setting where data become available in a sequential manner that may depend on the actions taken by the online learning algorithm. The performance of online learning algorithms is typically measured by the regret[22]. While offline learning assumes access to offline data (but not online data) and online learning assumes access to online data (but not offline data), in reality, a broad class of real-world problems incorporate both aspects: there is an offline historical data set (based on historical actions) at the time that the learner starts an online learning process.

Currently, there is no standard framework for the above type of learning problems, as classical offline learning theory and online learning theory have different settings and goals. While establishing a framework that bridges all aspects of offline learning and online learning is generally a very complicated task, in this paper, we propose a framework that bridges

[22]In this paper, when we discuss online learning, we focus more on the literature of stochastic online learning, where the online sequential data arrive in a stochastic manner. There is a vast literature of online learning focusing on the non-stochastic setting where the online sequential data arrive in an adversarial manner (see Cesa-Bianchi and Lugosi 2006), which is not the emphasis of this paper.

the gap between offline learning and online learning in a specific problem setting, which, however, already captures the essence of many dynamic pricing problems that sellers face in practice.

### 5.1.1 The Model: Online Pricing with Offline Data

In this paper, we study the *Online Pricing with Offline Data* (OPOD) problem. Consider a firm selling a single product with an infinite amount of inventory over a selling horizon of $T$ periods. In each period $t = 1, 2, \ldots, T$, the seller chooses a price $p_t$ from a given interval $[l, u] \subset [0, \infty)$ to offer to its customers, and then observes random demand $D_t$. We assume that the demand in each period is a linear function of the price plus some random noise. Specifically, for each $t \geq 1$,

$$D_t = \alpha^\star + \beta^\star p_t + \varepsilon_t, \tag{5.1}$$

where $\alpha^\star$ and $\beta^\star$ are two unknown demand parameters in the known interval $[\alpha_{\min}, \alpha_{\max}] \subseteq (0, \infty)$ and $[\beta_{\min}, \beta_{\max}] \subseteq (-\infty, 0)$ respectively, and $\{\varepsilon_t\}_{t=1}^T$ are *i.i.d.* random variables with zero mean and unknown distribution. We assume that $\varepsilon_t$ is an $R^2$-sub-Gaussian random variable, i.e., there exists a constant $R > 0$ such that $\mathbb{E}[e^{x\varepsilon_t}] \leq e^{\frac{x^2R^2}{2}}$ for any $x \in \mathbb{R}$. For notational convenience, let $\theta^\star := (\alpha^\star, \beta^\star)$ and $\Theta^\dagger := [\alpha_{\min}, \alpha_{\max}] \times [\beta_{\min}, \beta_{\max}]$, and we use $\theta := (\alpha, \beta)$ to denote any possible vector in parameter space $\Theta^\dagger$.

The seller's single-period expected revenue is the price $p$ offered to the customer multiplied by the associated expected demand. To emphasize the dependence on the parameter values, for any $\theta = (\alpha, \beta) \in \Theta^\dagger$, we define the expected revenue function $r(p; \theta)$ as $r(p; \theta) = p(\alpha + \beta p),\ \forall p \in [l, u]$. Let $\psi(\theta)$ be the price that maximizes $r(p; \theta)$ over the interval $[l, u]$, i.e., $\psi(\theta) = \arg\max \big\{r(p; \theta) : p \in [l, u]\big\}$, and use $p^\star$ to denote the true optimal price, i.e., $p^\star = \psi(\theta^\star)$. Let $r^\star(\theta)$ be the optimal expected revenue under demand parameter $\theta$, i.e., $r^\star(\theta) = \psi(\theta)(\alpha + \beta\psi(\theta))$. Without loss of generality,[23] we assume that for any $\theta \in \Theta^\dagger$, the optimal price is an interior point of price range $[l, u]$, and therefore $\psi(\theta) = \frac{\alpha}{-2\beta}$.

**Historical prices and offline data.** In reality, the seller does not know the true demand model, but has to learn such information from the historical data. In this paper, we assume that the seller has some pre-existing offline data before the start of the online learning process.

[23] This is because for any $\theta \in \Theta^\dagger$, $\frac{\alpha}{-2\beta} \in [\frac{\alpha_{\min}}{-2\beta_{\min}}, \frac{\alpha_{\max}}{-2\beta_{\max}}]$, and we can choose $l$ and $u$ such that $[\frac{\alpha_{\min}}{-2\beta_{\min}}, \frac{\alpha_{\max}}{-2\beta_{\max}}] \subset [l, u]$, which guarantees that $\frac{\alpha}{-2\beta}$ is an interior point of interval $[l, u]$.

The offline data set contains $n$ independent samples: $\{(\hat{p}_1, \hat{D}_1), (\hat{p}_2, \hat{D}_2), \ldots, (\hat{p}_n, \hat{D}_n)\}$, where $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ are $n$ fixed prices, and each $\hat{D}_i$ is a demand observation under historical price $\hat{p}_i$, drawn independently according to the underlying linear demand model (5.1). Therefore, for each $1 \leq i \leq n$, $\hat{D}_i = \alpha^\star + \beta^\star \hat{p}_i + \hat{\varepsilon}_i$ for some *i.i.d.* random variables $\{\hat{\varepsilon}_i\}_{i=1}^n$ with the same distribution as that of $\{\varepsilon_t\}_{t=1}^T$.

**Pricing policies and performance metrics.** For each $t \geq 0$, let $H_t$ be the vector of information available at the beginning of period $t+1$, i.e.,

$$H_t = (\hat{p}_1, \hat{D}_1, \ldots, \hat{p}_n, \hat{D}_n, p_1, D_1, \ldots, p_t, D_t).$$

A pricing policy is defined as a sequence of functions $\pi = (\pi_1, \pi_2, \ldots)$, where each $\pi_t$ is a measurable function which maps the realization of $H_t$ (and possibly some external randomness) to a feasible price in $[l, u]$. Let $\Pi$ be the set of all pricing policies. For any policy $\pi \in \Pi$, the *regret* of $\pi$, denoted by $R^\pi_{\theta^\star}(T)$, is defined as the difference between the optimal expected revenue generated by the clairvoyant policy that knows the exact value of $\theta^\star$ and the expected revenue generated by pricing policy $\pi$, i.e.,

$$R^\pi_{\theta^\star}(T) = T \cdot r^\star(\theta^\star) - \mathbb{E}^\pi_{\theta^\star}\Big[\sum_{t=1}^T r(p_t; \theta^\star)\Big],$$

### 5.1.2 Research Question, Observations and Challenges

This paper is inspired by the objective of bridging the gap between offline learning and online learning. The following question naturally arises whenever the offline data are incorporated into the online decision making: how do the offline data affect the *statistical complexity* of online learning? To address this question, the first challenge is to identify the key characteristics of the offline data that intrinsically affect the complexity of the online learning task.

Intuitively, the *size* of the offline data, measured by the number of historical samples $n$, and the *dispersion* of the offline data, measured by the standard deviation of historical prices $\sigma$, i.e., $\sigma = \sqrt{\sum_{i=1}^n (\hat{p}_i - \frac{1}{n}\sum_{j=1}^n \hat{p}_j)^2}$, provide two important metrics that enable quantifying the amount of information collected before the online learning process starts. As $n$ becomes larger, or $\sigma$ increases, the seller can form a better estimation for the unknown demand parameters using offline regression, and the regret may decrease accordingly.

Another crucial, and more intriguing metric of the offline data, is the *location* of the offline data, which is measured by $\delta = |\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i - p^\star|$, i.e., the distance between the average historical price and the optimal price $p^\star$. We refer to $\delta$ as the *generalized distance*, as it intuitively quantifies how far the offline data set is "away" from the (unknown) optimal decision. This is a crucial metric that uniquely appears when offline data are incorporated into the online learning process. Indeed, if there are no offline data available before the start of the online learning process, then there is no $\delta$ at all. Also, if the offline data are only used for estimation or prediction, with no need of online decision making, i.e., the seller is purely interested in estimating the model parameters from the offline data and does not need to make any online pricing decisions, then $\delta$ does not affect her estimation accuracy. Surprisingly, as we prove in this paper, when the offline data are incorporated into online learning, this metric will play a fundamental role.

Besides identifying the above three characteristics of the offline data, a key challenge is to precisely quantify the effects of these offline data characteristics on the online learning task. Specifically, we seek to understand to what extent these three metrics of the offline data influence the behavior and growth rate of the best-achievable regret bound. On the algorithmic side, we also seek to design a simple, intuitive and easy-to-implement pricing policy that exploits the values of both the pre-existing offline data and sequentially-revealed online data, and achieves a tight regret bound with respect to the selling horizon $T$, as well as the three metrics of the offline data, i.e., $n$, $\sigma$, $\delta$. Moreover, since the generalized distance $\delta$ is completely unknown to the seller, the algorithm itself cannot use any information about $\delta$, which implies a more challenging task of designing a learning algorithm whose performance is as good as if $\delta$ were known.

### 5.1.3 Main Results and Technical Highlights

In this paper, we address the above challenges in two settings: (i) *single-historical-price* setting where all the historical prices are identical, i.e., $\sigma = 0$, and (ii) *multiple-historical-price* setting where the historical prices can be different, i.e., $\sigma > 0$. We next summarize our main results and technical highlights. Throughout this paper, we use $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to present upper and lower bounds on the growth rate up to logarithmic factors, and $\widetilde{\Theta}(\cdot)$ to characterize the rate when the upper and lower bounds match (up to logarithmic factors). In addition, we use $A \lesssim B$ and $A \gtrsim B$ to denote $A = \widetilde{\mathcal{O}}(B)$ and $A = \widetilde{\Omega}(B)$ respectively. More formal

definitions of these notations are provided in §5.1.5. For any $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

**Single-historical-price setting.** For the single-historical-price setting, we develop a learning algorithm called *Online and Offline–OFU* (O3FU) algorithm, where OFU refers to the principle of *Optimism in the Face of Uncertainty*, which arises from multi-armed bandits and is widely used in the literature on bandits (see, e.g., Dani et al. 2008, Abbasi-Yadkori et al. 2011). In general, this principle suggests taking actions based on an optimistic guess of the reward associated with each action in each period. We show that the regret of O3FU algorithm has an upper bound $\widetilde{\mathcal{O}}(\sqrt{T} \wedge \frac{T \log T}{(n \wedge T)\delta^2})$. Although this upper bound depends on the unknown quantity $\delta$, the algorithm itself does not require any information about $\delta$. In addition, we prove an information-theoretic lower bound which matches the upper bound, showing that the regret bound cannot be further improved by other algorithms (in a certain sense); we define such an unimprovable regret bound as the *optimal (instance-dependent) regret* for the `OPOD` problem in the single-historical-price setting. We summarize its rate in Table 5.1. In particular, when $n = 0$, or $n = \infty$ and $\delta$ is a constant independent of $T$, the results in the leftmost and rightmost cases with $\delta \gtrsim T^{-1/4}$ in Table 5.1 recover those in Keskin and Zeevi (2014).

**Multiple-historical-price setting.** For the general setting that the historical prices may be different, we modify O3FU algorithm by adding a preliminary step that detects whether a *corner case* $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$ happens or not, and propose the *Modified O3FU* (M-O3FU) algorithm. We prove that M-O3FU algorithm achieves the regret upper bound $\widetilde{\mathcal{O}}(\sqrt{T} \wedge \frac{T \log T}{n\sigma^2 + (n \wedge T)\delta^2})$, except for a corner case where the upper bound becomes $\widetilde{\mathcal{O}}(T\delta^2)$.[24] In addition, we prove an information-theoretic lower bound that matches the upper bound for both cases, showing that our regret bound cannot be further improved (in a certain sense); we define such an unimprovable regret bound as the optimal (instance-dependent) regret for the `OPOD` problem in the multiple-historical-price setting. We summarize its rate in Table 5.2.

**Sufficient condition for self-exploration.** As a byproduct, we provide a sufficient condition for the *myopic* (i.e., greedy) policy to self-explore in the online learning process. Specifically, if the variance of historical prices is sufficiently large, and the average historical price is found to be bounded away from the confidence interval for the optimal price

[24] This corner case rarely happens because it requires the generalized distance $\delta$ to be very small and the price variance $n\sigma^2$ to be very large, such that there is no need of online learning. See the discussion in §5.4.2.

constructed from offline regression, then the myopic policy, the one that always charges the optimal price associated with the least-square estimate obtained in each round, achieves the optimal regret under mild conditions. This result generates additional insights for the performance guarantee of the myopic policy with the help of offline data, and also provides analytical support for the wide use of such policies in practice.

**Methodology contributions.** From a technical perspective, the tight upper and lower bounds that we obtain in this paper are both *instance-dependent* regret bounds, which are much stronger and more challenging to prove than the traditional *worst-case* regret bounds. To prove the instance-dependent upper bound, we conduct a period-by-period trajectory analysis, and develop novel inductive arguments, integrated with the specific property guaranteed by OFU principle, to obtain a sharp characterization on the distance between the algorithm's price and the average historical price. To prove the instance-dependent lower bound, we reduce the `OPOD` problem to a hybrid of estimation and hypothesis testing problems, which requires constructing an instance-dependent prior distribution and an instance-dependent hypothesis set, respectively. To the best of our knowledge, these are the first tight and general instance-dependent regret bounds obtained in (i) the linear-demand online pricing problem, and (ii) a continuous-armed bandit problem where the optimal action may not be an extremal point (in contrast to the extremal-point requirement in Dani et al. 2008 and Abbasi-Yadkori et al. 2011).

### 5.1.4 Key Insights: Phase Transitions and Inverse-Square Law

The characterization of the optimal instance-dependent regret also leads to two important implications on the value of offline data. First, when the offline sample size $n$ changes, the optimal regret rate exhibits significantly different decaying patterns, and we refer to such significant transitions between the regret-decaying patterns as *phase transitions*[25]. For example, when $\sigma = 0$ and $\delta \gtrsim T^{-\frac{1}{4}}$ (see Table 1), the optimal regret rate remains at the level of $\widetilde{\Theta}(\sqrt{T})$ whenever $n \lesssim \frac{\sqrt{T}}{\delta^2}$, and then gradually decays according to $\widetilde{\Theta}(\frac{T}{n\delta^2})$ when $\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim T$, and finally stays at the level of $\widetilde{\Theta}(\frac{\log T}{\delta^2})$ when $n \gtrsim T$. Second, in the regular case, the optimal regret is inversely proportional to the square of the standard deviation $\sigma$ and generalized distance $\delta$, which is referred to as the *inverse-square law*. The optimal

[25] We borrow this terminology from statistical physics; see Domb (2000). See also the discussion of phase transitions in the optimal stopping problem studied by Correa et al. (2022), and in the multi-armed bandit problem studied by Simchi-Levi and Xu (2023).

regret's dependence on $\sigma$ is consistent with our intuition, as more dispersive historical prices indicate more information gained before the online learning process starts, and therefore a smaller regret. The optimal regret's dependence on $\delta$ is more intriguing, as it suggests that the closer the historical prices are to the optimal price, the worse the optimal regret will be. In fact, this is a consequence of the tradeoff between exploration (i.e., experimenting to improve estimates of the unknown demand model) and exploitation (i.e., leveraging current estimates to maximize revenue). Specifically, whenever an algorithm tries to learn the true demand model, it has to make substantial efforts in charging various prices "far away" from the average historical price. Therefore, when $\delta$ is small, such a deviation will also lead to a significant gap with the optimal price, leading to greater revenue loss. These two findings contribute new insights to the fundamental problem of dynamic pricing with demand learning.

Table 5.1: Optimal regret for the single-historical-price setting.

| $\delta \gtrsim T^{-\frac{1}{4}}$ | | | |
|---|---|---|---|
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}}{\delta^2}$ | $\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim T$ | $n \gtrsim T$ |
| optimal regret | $\widetilde{\Theta}(\sqrt{T})$ | $\widetilde{\Theta}(\frac{T}{n\delta^2})$ | $\widetilde{\Theta}(\frac{\log T}{\delta^2})$ |
| $\delta \lesssim T^{-\frac{1}{4}}$ | | | |
| offline sample size | $n \ge 0$ | | |
| optimal regret | $\widetilde{\Theta}(\sqrt{T})$ | | |

Table 5.2: Optimal regret for the multiple-historical-price setting.

| $\delta \gtrsim T^{-\frac{1}{4}}$ and $\sigma \lesssim \delta$ | | | | |
|---|---|---|---|---|
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}}{\delta^2}$ | $\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim T$ | $T \lesssim n \lesssim \frac{T\delta^2}{\sigma^2}$ | $n \gtrsim \frac{T\delta^2}{\sigma^2}$ |
| optimal regret | $\widetilde{\Theta}(\sqrt{T})$ | $\widetilde{\Theta}(\frac{T}{n\delta^2})$ | $\widetilde{\Theta}(\frac{1}{\delta^2})$ | $\widetilde{\Theta}(\frac{T}{n\sigma^2})$ |
| $\delta \gtrsim T^{-\frac{1}{4}}$ and $\sigma \gtrsim \delta$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}}{\sigma^2}$ | | $n \gtrsim \frac{\sqrt{T}}{\sigma^2}$ | |
| optimal regret | $\widetilde{\Theta}(\sqrt{T})$ | | $\widetilde{\Theta}(\frac{T}{n\sigma^2})$ | |
| $\delta \lesssim T^{-\frac{1}{4}}$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}}{\sigma^2}$ | | $\frac{\sqrt{T}}{\sigma^2} \lesssim n \lesssim \frac{1}{\delta^2\sigma^2}$ | $n \gtrsim \frac{1}{\delta^2\sigma^2}$ |
| optimal regret | $\widetilde{\Theta}(\sqrt{T})$ | | $\widetilde{\Theta}(T\delta^2)$ | $\widetilde{\Theta}(\frac{T}{n\sigma^2})$ |

### 5.1.5 Organization and Notation

This paper is organized as follows. In §5.2, we review the relevant literature. In §5.3 and §5.4, we study the OPOD problem in the single-historical-price setting and multiple-historical-price setting respectively. We conduct a numerical study in §5.5, and discuss the self-exploration of the myopic policy in §5.6. In §5.7, we summarize our paper with extensions and future research directions. Most of the technical proofs are deferred to the appendix.

Throughout the paper, all the vectors are column vectors unless otherwise specified. For any $m \in \mathbb{N}$, we use $[m]$ to denote the set $\{1, 2, \ldots, m\}$. For any column vector $x \in \mathbb{R}^n$ and positive semi-definite matrix $A \in \mathbb{R}^{n\times n}$, $||x|| := (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$, and $||x||_A := \sqrt{x^\top Ax}$. The notations $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ are applied to hide constant factors, and $\widetilde{\mathcal{O}}(\cdot)$, $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ are applied to hide both constant and logarithmic factors. That is, $f(T) = \mathcal{O}(g(T))$ means that there exists a constant $C > 0$ such that $f(T) \leq Cg(T)$ for any $T$, and $f(T) = \widetilde{\mathcal{O}}(g(T))$ means that there exist constants $C$ and $\lambda > 0$, such that $f(T) \leq Cg(T)(\log T)^\lambda$ for any $T$. In addition, $f(T) = \Omega(g(T))$ (resp. $f(T) = \widetilde{\Omega}(g(T))$) means $g(T) = \mathcal{O}(f(T))$ (resp. $g(T) = \widetilde{\mathcal{O}}(f(T))$), and $f(T) = \Theta(g(T))$ (resp. $f(T) = \widetilde{\Theta}(g(T))$) means $f(T) = \mathcal{O}(g(T))$ and $f(T) = \Omega(g(T))$ (resp. $f(T) = \widetilde{\mathcal{O}}(g(T))$ and $f(T) = \widetilde{\Omega}(g(T))$).

## 5.2 Related Literature

### 5.2.1 Dynamic Pricing with Online Learning

When there are no offline data, the OPOD problem becomes a pure online learning problem, i.e. dynamic pricing with an unknown linear demand model, and belongs to a broad category referred to as the *online pricing* problems. Online pricing problems have generated great interest in recent years in the operations research and management science (OR/MS) community; see den Boer (2015) for a comprehensive survey. In particular, there is a vast literature (e.g., den Boer and Zwart 2013, den Boer 2014, Keskin and Zeevi 2014, Wang et al. 2014, Keskin and Zeevi 2016, Qiang and Bayati 2016, Nambiar et al. 2019, Ban and Keskin 2021, den Boer and Keskin 2022) studying dynamic pricing problems with an unknown linear (or generalized linear) demand model, which is arguably one of the most fundamental demand models for pricing. All of the existing papers purely focus on online learning. In this paper, we take the fundamental problem of dynamic pricing with a linear demand model as our baseline, but significantly extend it by incorporating offline data into online decision

making.

Keskin and Zeevi (2014) is the most relevant paper to this work. The authors consider dynamic pricing with an unknown linear demand model, studying an important question of how knowing an *exact* point at the demand curve (i.e., the exact expected demand under a single price) in advance helps reduce the optimal regret. Depending on whether the seller knows this exact point or not, they prove that the best achievable regret is $\Theta(\log T)$ and $\widetilde{\Theta}(\sqrt{T})$ respectively. Compared with their work, the `OPOD` problem studied in this paper seems more relevant to practice, and is more general in theory. Practically, while firms will never know the true expected demand under a given price exactly (which requires infinitely many demand observations), they usually have some pre-existing offline data (which are finitely many) prior to the online learning process. Theoretically, the results in Keskin and Zeevi (2014) (for the single-product setting) can be viewed as two special cases of our results when (i) $n = 0$; and (ii) $\sigma = 0$, $n = \infty$, and $\delta = \Theta(1)$, with an additional assumption that $\delta$ is lower bounded by a *known* constant (as their algorithms for case (ii) rely on this knowledge). Since $\delta$ is completely unknown and can be small in our setting (and in reality), their algorithms and analysis do not apply here. In fact, the principle of our algorithm design and the approach of our regret analysis are very different from theirs.

There is also a stream of literature in Bayesian learning, where the decision maker is assumed to have a known prior distribution for the unknown parameter, and can update her belief on the prior distribution from online observations. For recent works on dynamic pricing with Bayesian learning, we refer the interested readers to Harrison et al. (2012) and Agrawal et al. (2017) that focus on the worst-case regret, and to Ferreira et al. (2018) and Miao and Chao (2020) that focus on the Bayesian regret. While the prior distribution in Bayesian learning can be estimated using offline data, the modeling approach and results of these papers are very different from this work. First, in Bayesian learning, it is usually assumed that the decision maker knows the exact prior distribution, which typically belongs to some specific parametric family. By contrast, in this work, we do not assume any prior distribution or impose any parametric assumption on the distribution of demand parameter, but directly incorporate offline data into online learning. Second, as a main contribution of this paper, we characterize the effects of the size, dispersion and location of the offline data on the statistical complexity of online learning, which are not discussed in and not the focus of the current literature on Bayesian learning.

### 5.2.2 Multi-Armed Bandits

Our paper is also related to the literature of multi-armed bandits (MAB). In the classical $K$-armed bandit problem, the decision maker chooses one of the $K$ arms in each round and observes a random reward generated from some unknown distribution associated with the arm being played, with the goal of minimizing the regret; see Lattimore and Szepesvári (2020) for more references on this topic. In most of the literature on bandit problems (see, e.g., Auer et al. 2002a, Dani et al. 2008, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori et al. 2011, Filippi et al. 2010), the decision maker has to start from scratch (i.e., with no historical information). By contrast, a few papers study bandit problems in settings where the algorithms may utilize different types of historical information; see, e.g., Shivaswamy and Joachims (2012), Bouneffouf et al. (2019), Hsu et al. (2019), Ye et al. (2020), Bastani et al. (2022), Gur and Momeni (2022), of which Shivaswamy and Joachims (2012) and Gur and Momeni (2022) are the most relevant to this paper.

Shivaswamy and Joachims (2012) study the MAB problem with offline observations of rewards collected before the online learning algorithm starts; we refer to their problem as "MAB with offline data". While our idea of incorporating offline data into an online learning problem is similar to theirs, there are significant differences between the two papers in terms of model settings, main results and analytical techniques. First, Shivaswamy and Joachims (2012) study the MAB problem with discrete and finitely many arms, while our model builds on the literature of online pricing problems (see §5.2.1 for references), where the prices are continuous and infinitely many, and the rewards are nonlinear with respect to prices. The properties and results for these two classes of problems are very different. Second, under the *well-separated condition*, Shivaswamy and Joachims (2012) prove some regret upper bounds that change from $\mathcal{O}(\log T)$ to $\mathcal{O}(1)$ when the amount of offline observations of rewards for *each* arm exceeds $\Omega(\log T)$, with no regret lower bound proven and hence no discussion of phase transitions. In comparison, we characterize the optimal regret via matching upper and lower bounds, and figure out surprising phase transitions of the optimal regret rate as the offline sample size changes. Moreover, we also discover the inverse square law, which does not appear in the previous literature. Third, while Shivaswamy and Joachims (2012) use a conventional approach in bandit literature to upper-bound the regret via the so-called *sub-optimality gap*, since we are bounding the regret via $\sigma$ and $\delta$, we present different regret

analysis that may be of independent interest.

In a recent paper by Gur and Momeni (2022), a generalized MAB formulation is studied, where some additional information may become available before each online decision is made. Their formulation includes "MAB with offline data" as a special case. Under the general formulation, the authors characterize the optimal regret as a function of the information arrival process, and study the effect of the characteristics of this process on the algorithm design and the best achievable regret bound. Since their formulation and analysis crucially rely on the model setting of discrete and finitely many arms, the results, techniques, and insights of our paper and their paper are significantly different.

Interestingly, by specializing the regret lower bound in Gur and Momeni (2022) to the problem of "MAB with offline data", and combining this lower bound with the regret upper bound in Shivaswamy and Joachims (2012), one can obtain a characterization of the optimal regret for this problem under some mild conditions, which also leads to phase transitions not discussed before. The phase transitions are very different from the phase transitions that we discover in `OPOD`. We provide more discussions on this observation and the differences in Appendix D.7.

## 5.3 Single Historical Price

In this section, we study the single-historical-price setting: where all the $n$ historical prices are identical to $\hat{p}$. As pointed out in Harrison et al. (2012) and Keskin and Zeevi (2014), in finance industry, for many consumer lending products, banks often keep a fixed interest rate over some periods of time before they conduct price experimentation. Similarly, in the retail industry, there are many scenarios where the seller charges a fixed price based on the manufacturer's suggestion, branding or competitors' price before using a dynamic pricing strategy. Thus, we start with this simple but important single-historical-price setting in this section. We first design a learning algorithm with a per-instance regret upper bound in §5.3.1, and then characterize the regret lower bound in §5.3.2. Some important implications are discussed in §5.3.3.

### 5.3.1 O3FU Algorithm and Regret Upper Bound

Our proposed algorithm *Online and Offline–Optimism in the Face of Uncertainty* (O3FU) is constructed based on the celebrated *Optimism in the Face of Uncertainty* (OFU) principle, which effectively addresses the exploration-exploitation dilemma inherent in many online learning problems (see, e.g., §7.1 of Lattimore and Szepesvári 2020 for a reference). For any $t \geq 0$, we define a confidence radius $w_t$ that will be used to construct a confidence ellipsoid for the demand parameter at the end of period $t$, and the expression of $w_t$ is as follows:

$$w_t = R\sqrt{2\log\Big(\frac{1}{\epsilon}\big(1+(1+u^2)(t+n)/\lambda\big)\Big)} + \sqrt{\lambda(\alpha_{\max}^2+\beta_{\min}^2)}, \tag{5.2}$$

where $\epsilon$ and $\lambda$ will be specified in the description of the algorithm. The choice of $w_t$ is based on the high-probability confidence bound developed in Theorem 2 of Abbasi-Yadkori et al. (2011), which will also be used throughout our regret analysis. The pseudo-code of O3FU algorithm is provided in Algorithm 5.1.

**Algorithm 5.1** O3FU Algorithm

**Input**: historical price $\hat{p}$, offline demand data $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_n$, support of unknown parameters $\Theta^\dagger$, price range $[l,u]$, regularization parameter $\lambda = 1+u^2$, $\{w_t\}_{t\geq 1}$ defined in (5.2) with $\epsilon = \frac{1}{T^2}$

**Initialization**: $V_{0,n} = \lambda I + n[1\ \hat{p}]^\intercal[1\ \hat{p}]$, $Y_{0,n} = (\sum_{i=1}^n \hat{D}_i)[1\ \hat{p}]^\intercal$

**for** $t \in [T]$ **do**
  **if** $t = 1$ **then**
    Charge price $p_1 = l\cdot\mathbb{I}\{\hat{p} > \frac{u+l}{2}\} + u\cdot\mathbb{I}\{\hat{p}\leq\frac{u+l}{2}\}$, and observe demand realization $D_1$
    Compute $V_{1,n} = V_{0,n} + [1\ p_1]^\intercal[1\ p_1]$, $Y_{1,n} = Y_{0,n} + D_1[1\ p_1]^\intercal$, $\hat{\theta}_1 = V_{1,n}^{-1}Y_{1,n}$
    Compute confidence ellipsoid $\mathcal{C}_1 = \big\{\theta\in\mathbb{R}^2 : ||\theta-\hat{\theta}_1||_{V_{1,n}} \leq w_1\big\}$
  **else**
    If $\mathcal{C}_{t-1}\cap\Theta^\dagger \neq \varnothing$, let $(p_t,\tilde{\theta}_t) \in \arg\max_{p\in[l,u],\theta\in\mathcal{C}_{t-1}\cap\Theta^\dagger} p(\alpha+\beta p)$; otherwise, let $p_t = p_1$
    Charge price $p_t$, and observe demand realization $D_t$
    Update $V_{t,n} = V_{t-1,n} + [1\ p_t]^\intercal[1\ p_t]$, $Y_{t,n} = Y_{t-1,n} + D_t[1\ p_t]^\intercal$, $\hat{\theta}_t = V_{t,n}^{-1}Y_{t,n}$
    Update confidence ellipsoid $\mathcal{C}_t = \big\{\theta\in\mathbb{R}^2 : ||\theta-\hat{\theta}_t||_{V_{t,n}} \leq w_t\big\}$.

In O3FU algorithm, when $t = 1$, the price is chosen from boundary points $\{l,u\}$, depending on which one has a larger distance from historical price $\hat{p}$. The choice of such an initial price is not unique, and any price that is bounded away from $\hat{p}$ by a constant distance will also work. For each $t \geq 2$, we first maintain a confidence ellipsoid $\mathcal{C}_{t-1}$

for the unknown parameter $\theta^\star$, and then O3FU algorithm selects an optimistic estimator $\tilde{\theta}_t \in \arg\max_{\theta \in \mathcal{C}_{t-1} \cap \Theta^\dagger} \max_{p \in [l,u]} p(\alpha + \beta p)$, and charges price $p_t = \arg\max_{p \in [l,u]} p(\tilde{\alpha}_t + \tilde{\beta}_t p)$, which is optimal with respect to estimator $\tilde{\theta}_t$. Note that when $\max_{p \in [l,u]} p(\alpha + \beta p)$, as a function of $\theta \in \mathcal{C}_{t-1} \cap \Theta^\dagger$, has multiple maximizers, $\tilde{\theta}_t$ can be set as any maximizer. Figure 5-1 shows how O3FU algorithm works, where the three blue curves depict the expected revenues with three different parameters belonging to set $\mathcal{C}_{t-1} \cap \Theta^\dagger$ (we only draw three curves for illustration), and the red curve is the upper envelope of all the possible candidate revenue curves, which is also the revenue function associated with the demand parameter $\tilde{\theta}_t$, i.e., $r(p; \tilde{\theta}_t)$.

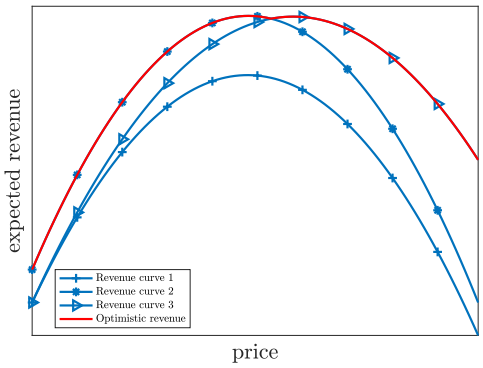


Figure 5-1: Revenue curves under three different parameters (blue), and the optimistic revenue (red).

Intuitively, if we knew that generalized distance $\delta$ would be large, then trying prices far away from $\hat{p}$ is beneficial for both exploration and exploitation. By contrast, if we knew that $\delta$ would be small, then striking a balance between exploration and exploitation would be very important, because choosing prices close to $\hat{p}$ is only effective for exploitation but not for exploration. Of course, the seller does not know the true value of $\delta$, which makes designing a learning algorithm that achieves the right balance between exploration and exploitation a more challenging task. O3FU algorithm achieves this objective by maximizing the *optimistic revenue*, which is defined as $\max_{\theta \in \mathcal{C}_{t-1} \cap \Theta^\dagger} p(\alpha + \beta p)$, and can be treated as the estimated revenue plus a "bonus" of exploration. In fact, as implied from equation (19.8) in Lattimore and Szepesvári (2020), when $\mathcal{C}_{t-1} \subseteq \Theta^\dagger$, we have $\max_{\theta \in \mathcal{C}_{t-1} \cap \Theta^\dagger} p(\alpha + \beta p) = p(\hat{\alpha}_{t-1} + \hat{\beta}_{t-1} p) + w_{t-1} \sqrt{[1\ p] V_{t-1,n}^{-1} [1\ p]^\top}$. Therefore, exploitation and exploration are both incorporated into the objective function through the first term and the second term, respectively.

It's worth noting that our O3FU algorithm is *parameter-free* in the sense that it does not need to use any information about $\delta$. In addition, while O3FU algorithm takes $T$ as input, one can easily extend the current algorithm to work with unknown $T$ using the standard *doubling trick* (see, e.g., Lattimore and Szepesvári 2020) and construct an *anytime* algorithm that does not need to know $T$.

We now provide an upper bound on the regret of O3FU algorithm.

**Theorem 5.1.** *Let $\pi$ be O3FU algorithm. Then there exists a finite constant $K_1 > 0$ such that for any $T \geq 1$, $n \geq 0$ and $\hat{p} \in [l, u]$, and for any possible value of $\theta^\star \in \Theta^\dagger$,*

$$R^{\pi}_{\theta^\star}(T) \leq K_1 \cdot \left(\sqrt{T} \wedge \frac{T \log T}{(n \wedge T)\delta^2}\right) \cdot \log T.$$

Theorem 5.1 provides a regret upper bound $\widetilde{\mathcal{O}}(\sqrt{T} \wedge \frac{T \log T}{(n \wedge T)\delta^2})$ that depends on the problem instance through the value of $\delta$, which is therefore called the *instance-dependent* upper bound. If $\delta$ is a constant, when $n = 0$ or $n = \infty$, i.e., there are no offline data or infinitely many offline data under price $\hat{p}$, the upper bound reduces to $\widetilde{\mathcal{O}}(\sqrt{T})$ and $\widetilde{\mathcal{O}}(\log T)$ respectively. If $\delta$ is not a constant, with an order shrinking to zero as $T$ grows, the regret upper bound is then inversely proportional to $\delta^2$. We summarize the regret upper bound under different $(n, \delta)$ combinations in Table D.1 of Appendix D.8.

We next outline the key ideas to prove Theorem 5.1 and leave the detailed analysis to Appendix D.1.1. From the statement of Theorem 5.1, it suffices to show an *instance-independent* upper bound $\widetilde{\mathcal{O}}(\sqrt{T})$ and an *instance-dependent* upper bound $\widetilde{\mathcal{O}}(\frac{T}{(n \wedge T)\delta^2})$. The instance-independent bound can be proved using similar arguments from stochastic linear bandits, e.g., Abbasi-Yadkori et al. (2011), by noting that the expected revenue is the inner product of the unknown parameter $[\alpha\ \beta]$ and the action vector $[p\ p^2]$. Showing the instance-dependent bound is the novel part in our proof, which relies on the following crucial lemma.

**Lemma 5.1.** *Suppose $T \geq T_0$, $\delta \geq \frac{\sqrt{2(\alpha^2_{\max} + \beta^2_{\max})} w_T}{\beta^2_{\max} n^{1/4}}$, and $\theta^\star \in \mathcal{C}_t$ for each $t \in [T]$, then two sequences of events $\{U_{t,1}\}_{t=1}^T$ and $\{U_{t,2}\}_{t=2}^T$ also hold, where*

$$U_{t,1} = \left\{|p_t - \hat{p}| \geq \min\left\{1 - \frac{\sqrt{2}}{2}, \frac{C_0}{2}\right\} \cdot \delta\right\},$$
$$U_{t,2} = \left\{||\tilde{\theta}_t - \theta^\star||^2 \leq C_2 \cdot \frac{w^2_{t-1}}{(n \wedge (t-1))\delta^2}\right\},$$

*and* $C_0 = \frac{l|\beta_{\max}|}{u|\beta_{\min}|}$, $C_1 = 4(C_0+1)^2 C_0^{-2}\big(1+(4u+1)^2\big)$, $C_2 = \max\Big\{4(u-l)^2,\, 2C_1,\, 4((4u+1)^2+1)(\min\{\frac{C_0^2}{4}, (1-\frac{\sqrt{2}}{2})^2\})^{-1}\Big\}$, $T_0 = \min\Big\{t\in\mathbb{N} : w_t \geq \sqrt{C_1}\beta_{\max}^2(2(\alpha_{\max}^2+\beta_{\max}^2))^{-1/2}\Big\}$.

Lemma 5.1 is interpreted as follows. When the optimal price has a certain distance from historical price $\hat{p}$, i.e., $\delta \geq \frac{\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}w_T}{\beta_{\max}^2 n^{1/4}}$, given that the demand parameter $\theta^\star$ belongs to the confidence ellipsoid $\mathcal{C}_t$ in each period $t$, the algorithm's pricing sequence $\{p_t\}_{t=1}^T$ is also uniformly bounded away from $\hat{p}$ proportional to the unknown quantity $\delta$ (as implied by events $\{U_{t,1}\}_{t=1}^T$), and will gradually approach the true optimal price in a rate of $\mathcal{O}(\frac{w_t^2}{(n\wedge t)\delta^2})$ (as implied by events $\{U_{t,2}\}_{t=2}^T$). This implies that the algorithm can "adaptively" explore to a suitable degree, to create an efficient "collaboration" between the online prices and the historical price, while concurrently approaching the unknown optimal price. This property is nontrivial and cannot be implied from the existing analysis of the OFU-type algorithms. To prove this lemma, we conduct a period-by-period trajectory analysis of the random pricing sequence generated by our algorithm. Specifically, we find that the occurrence of $U_{t,2}$ relies on the joint occurrence of $U_{1,1},\ldots,U_{t-1,1}$, while the occurrence of $U_{t,2}$ (combined with the specific structure of the optimistic revenue curve) in turn leads to the occurrence of $U_{t,1}$. We thus introduce novel induction-based arguments to prove Lemma 5.1; see details in Appendix D.1.2. The induction-based arguments also explain why we set the initial price in the algorithm to be a boundary point (or any price that has a constant distance from $\hat{p}$), since this enables $U_{1,1}$ to occur.

We remark that for the stochastic linear bandit problem with a polytope action set, Abbasi-Yadkori et al. (2011) prove an instance-dependent upper bound of $\mathcal{O}(\frac{\log T}{\Delta})$, where $\Delta$ is defined as the sub-optimality gap between the rewards of the best and second best extremal points of the action set. We emphasize that their result and analysis cannot be applied to prove our instance-dependent upper bound due to the following reasons. First, the instance-dependent upper bound in our problem is developed to capture the effect of the generalized distance $\delta$ on the regret bound, which does not exist in the stochastic linear bandit problem. Second, the instance-dependent upper bound in Abbasi-Yadkori et al. (2011) relies on two strong conditions: (i) their algorithm only selects actions among the extremal points of the action set, and (ii) every sub-optimal action taken by their algorithm is bounded away from the optimal action by a reward gap $\Delta$. Such conditions only hold under their setting and assumptions. Our problem, however, has a quadratic objective function, with the optimal price being an interior point of the interval $[l,u]$, which requires the algorithm's

actions to converge to the optimal action. As a result, the sub-optimality gap $\Delta$ becomes zero, and standard arguments based on $\Delta$ do not work.

### 5.3.2 Lower Bound on Regret

In this subsection, we establish a lower bound on the performance of any algorithm for the `OPOD` problem with a single historical price. We first introduce the following set of *admissible policies* denoted by $\Pi^\circ$, which includes all the policies whose regret is guaranteed to be $\widetilde{\mathcal{O}}(\sqrt{T})$ for any possible value of demand parameter $\theta^\star$, i.e.,

$$\Pi^\circ = \Big\{\pi \in \Pi : \sup_{\theta^\star \in \Theta^\dagger} R^\pi_{\theta^\star}(T) \le K_0 \sqrt{T} (\log T)^{\lambda_0}\Big\}, \tag{5.3}$$

where $K_0 > 0$ and $\lambda_0 \ge 0$ are arbitrary constants. Intuitively, $\Pi^\circ$ excludes those "bad" policies that are not robust and suffer from large worst-case regret, e.g., a policy that never explores and always chooses $\hat{p}$, incurring zero regret when $\delta = 0$ but linear regret when $\delta = \Theta(1)$. Restricting our attention to $\Pi^\circ$ (which O3FU and many existing algorithms obviously belong to) ensures that the considered policies are reasonable enough. Note that $\Pi^\circ$ is specified by a pair of $(K_0, \lambda_0)$, but for simplicity, when there is no ambiguity, we drop the dependence on $(K_0, \lambda_0)$ in the notation. To facilitate our discussion, let $R^\pi_\theta(T, n, \delta)$ be defined as the regret for admissible policy $\pi \in \Pi^\circ$ when the demand parameter is $\theta = (\alpha, \beta)$, i.e., $R^\pi_\theta(T, n, \delta) = T \cdot r^\star(\theta) - \mathbb{E}^\pi_\theta[\sum_{t=1}^T r(p_t; \theta)]$. We also denote $\mathcal{D}$ as the generic distribution of $\{\hat{\varepsilon}_i\}_{i=1}^n$ and $\{\varepsilon_t\}_{t=1}^T$, and $\mathcal{E}(R)$ as the class of sub-Gaussian distributions with parameter $R$.

The following theorem provides a regret lower bound for any admissible policy in terms of the generalized distance $\delta$. For any generalized distance $\delta$, we define an *instance-dependent* environment class $\{\theta \in \Theta^\dagger : |\hat{p} - \psi(\theta)| \in [(1-\xi)\delta, (1+\xi)\delta]\}$, which is the set of all possible values of the demand parameter whose associated optimal prices are $\Theta(\delta)$-distance away from $\hat{p}$ (here $\xi$ can be any fixed constant in $(0, 1)$). This environment class highlights the role of $\delta$ as a key instance-dependent quantity, and enables us to establish an instance-dependent regret lower bound that holds for all possible values of $\delta$; see Theorem 5.2 (note that the environment class appears under the sup operator in the LHS of (5.4)).

**Theorem 5.2.** *There exists a positive constant $K_2$ such that for any admissible policy*

$\pi \in \Pi^\circ$, *for any* $\xi \in (0,1)$, $T \geq 2$ *and* $n \geq 0$, *and for any* $\delta \in [0, u-l]$,

$$\sup_{\substack{\mathcal{D}\in\mathcal{E}(R);\\ \theta\in\Theta^\dagger:|\hat{p}-\psi(\theta)|\in[(1-\xi)\delta,(1+\xi)\delta]}} R_\theta^\pi(T,n,\delta) \geq \begin{cases} K_2 \cdot \left( (\sqrt{T} \wedge \frac{T}{(n\wedge T)\delta^2}) \vee \log T \right), & \text{if } \delta \gtrsim T^{-\frac{1}{4}}; \\ K_2 \cdot \left( (T\delta^2) \vee \frac{\sqrt{T}}{(\log T)^{\lambda_0}} \right), & \text{if } \delta \lesssim T^{-\frac{1}{4}}. \end{cases} \tag{5.4}$$

**Remark 5.1.** *We emphasize that finding a "right" definition of the instance-dependent environment class is important for capturing the true role of* $\delta$ *in determining the instance-dependent regret. While there may be other ways to specify the environment class, they may fail to accurately reflect the instance-dependent complexity of the* OPOD *problem. For example, if one sets the environment class to be the entire parameter space* $\Theta^\dagger$*, then one can obtain a single lower bound for the worst-case regret (independent of* $\delta$*); however, such a definition is too conservative and does not fully capture the value of offline data. Another seemingly natural way to specify the environment class is to consider* $\{\theta \in \Theta^\dagger : |\psi(\theta) - \hat{p}| = \delta\}$*, which is the set of all possible values of the demand parameter whose associated optimal price has a distance from* $\hat{p}$ *exactly equal to* $\delta$*. However, this definition cannot preclude certain speculative behavior of algorithms, and would result in an unrealistic regret bound that cannot be attained by any single algorithm. We refer to Appendix D.4 for more details regarding the limitations of the above two definitions of the environment class.*

We explain Theorem 5.2 as follows. First, when $\delta \gtrsim T^{-\frac{1}{4}}$, the regret lower bound is $\Omega((\sqrt{T} \wedge \frac{T}{(n\wedge T)\delta^2}) \vee \log T)$, and in particular, if $\delta$ is a constant and $n = \infty$, the regret lower bound reduces to $\Omega(\log T)$, which recovers Theorem 3 in Keskin and Zeevi (2014) for their incumbent-price setting. Second, when $\delta \lesssim T^{-\frac{1}{4}}$, the regret lower bound is always $\widetilde{\Omega}(\sqrt{T})$, regardless of offline sample size $n$. The intuition is as follows. When restricting attention to $\Pi^\circ$, we exclude those "unreasonable" policies that seldom explore but make pricing decisions in a naive way, e.g., the one that always chooses price $\hat{p} + \delta$, because the regret of such policies cannot always be upper bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$ for any possible value of $\theta^\star$. In this case, any admissible policy $\pi \in \Pi^\circ$ should be able to make sufficient exploration to distinguish between different demand curves. However, to achieve this, the policy must deviate from $\hat{p}$, which is *less informative* since the seller already has collected some data under this price, to gain more information about the true demand curve. When $\delta$ is very small, charging prices away from $\hat{p}$ leads to a significant gap relative to the optimal price, and therefore a large

regret bound. We summarize the regret lower bound under different $(n, \delta)$ combinations in Table D.2 of Appendix D.8.

We next highlight the key steps in proving Theorem 5.2 and leave the detailed analysis to Appendix D.1.3. The proof idea is to reduce the `OPOD` problem to a hybrid of an estimation problem (see Step 1) and a hypothesis testing problem (see Step 2).

**Step 1.** In this step, we prove that when $\varepsilon$ follows a normal distribution, for any pricing policy $\pi \in \Pi$ (not necessarily in $\Pi^\circ$),

$$\sup_{\theta \in \Theta^\dagger : |\hat{p} - \psi(\theta)| \in [(1-\xi)\delta, (1+\xi)\delta]} R_\theta^\pi(T, n, \delta) = \Omega\Big( \big(\sqrt{T} \wedge \big(\frac{T}{\delta^{-2} + (n \wedge T)\delta^2}\big)\big) \vee \log(1 + T\delta^2) \Big). \tag{5.5}$$

To prove (5.5), we consider an "auxiliary" estimation problem for the optimal price $\psi(\theta)$, and appeal to the multivariate van Trees inequality (cf. Gill and Levit 2001) to construct a lower bound for the Bayesian regret. In particular, when applying the van Trees inequality, we need to carefully choose a suitable instance-dependent prior distribution $q(\cdot)$ whose Fisher information grows at an appropriate rate with respect to $\delta$, and upper-bound the resulting Fisher information of the sequential estimators $\{p_t\}_{t=1}^T$ in different cases. Then we can rightly control the growth rate of the Bayesian regret.

**Step 2.** In the second step, we show that when $\varepsilon$ follows a normal distribution and $\delta \lesssim T^{-\frac{1}{4}} (\log T)^{-\frac{1}{2}\lambda_0}$, for any admissible policy $\pi \in \Pi^\circ$, there exists $\theta \in \Theta^\dagger$ satisfying $|\psi(\theta) - \hat{p}| \in [(1-\xi)\delta, (1+\xi)\delta]$ such that

$$R_\theta^\pi(T, n, \delta) = \Omega\Big( \frac{\sqrt{T}}{(\log T)^{\lambda_0}} \Big). \tag{5.6}$$

The proof of (5.6) is based on arguments using Kullback-Leibler divergence and Bretagnolle-Huber inequality (Theorem 2.2 in Tsybakov 2009), whose key idea is as follows. We construct two problem instances with parameters $\theta_1$ and $\theta_2$ such that (i) the two demand curves under $\theta_1$ and $\theta_2$ intersect at price $\hat{p}$; (ii) the optimal prices under $\theta_1$ and $\theta_2$ are $\hat{p} + \delta$ and $\hat{p} + \delta + \Delta$ respectively, with $\Delta = \Theta(T^{-\frac{1}{4}} (\log T)^{\frac{1}{2}\lambda_0})$. For any pricing policy $\pi \in \Pi^\circ$, it has to perform well under both constructed problem instances, i.e., the regret upper bound is $\widetilde{\mathcal{O}}(\sqrt{T})$ under either instance, and therefore should be able to distinguish between the demand environments under $\theta_1$ and $\theta_2$. Moreover, any policy with the goal of separating $\theta_1$ and $\theta_2$ should charge

prices significantly different from the intersected price $\hat{p}$, i.e., the KL-divergence between the two probability measures under $\theta_1$ and $\theta_2$ induced by policy $\pi$ is large. Nevertheless, since the optimal price associated with $\theta_1$ is $\hat{p}+\delta$, which is extremely close to $\hat{p}$, the policy will incur large regret when the underlying parameter is in fact $\theta_1$. Therefore, the regret under $\theta_1$ is always lower bounded by $\Omega(\frac{\sqrt{T}}{(\log T)^{\lambda_0}})$ no matter how large $n$ is.

### 5.3.3 Phase Transitions and Inverse-Square Law

In this subsection, we discuss two important implications. By comparing Theorems 5.1 and 5.2, one can easily verify that the regret upper bound $\widetilde{\mathcal{O}}(\sqrt{T} \wedge \frac{T \log T}{(n \wedge T)\delta^2})$ achieved by O3FU algorithm, after ignoring the logarithm factor, is unimprovable within the class of all admissible policies under the instance-dependent environment class considered in Theorem 5.2. Motivated by this result, for $\Pi^\circ$ with $\lambda_0 \geq 1$, we define the *optimal (instance-dependent) regret* $R^\star(T, n, \delta)$ as

$$R^\star(T, n, \delta) = \inf_{\pi \in \Pi^\circ} \sup_{\substack{\mathcal{D} \in \mathcal{E}(R); \\ \theta \in \Theta^\dagger : |\psi(\theta) - \hat{p}| \in [(1-\xi)\delta, (1+\xi)\delta]}} R_\theta^\pi(T). \tag{5.7}$$

Thus, $R^\star(T, n, \delta)$ characterizes the statistical complexity of the `OPOD` problem in the sense that no algorithm in the admissible policy class can perform better than this rate when the true optimal price is allowed to center around $\hat{p}$ within $\Theta(\delta)$. We state this result in the following corollary.

**Corollary 5.1.** *The optimal regret defined in* (5.7) *for the single-historical-price setting is*

$$R^\star(T, n, \delta) = \widetilde{\Theta}\left(\sqrt{T} \wedge \left(\frac{T}{n\delta^2} \vee \frac{\log T}{\delta^2}\right)\right).$$

The characterization of the optimal regret leads to two important implications. First, the decaying patterns of the optimal regret rate are different when offline sample size $n$ belongs to different ranges. To better illustrate this phenomenon, we first consider the *well-separated* case where $\delta$ is a constant independent of $T$. This case frequently happens in reality as it suggests that the seller's historical price is suboptimal and quite different from the true optimal price. In this case, as $n$ increases, the optimal regret rate first remains at the level of $\widetilde{\Theta}(\sqrt{T})$ when $n \lesssim \sqrt{T}$, then gradually decays according to $\widetilde{\Theta}(\frac{T}{n})$ when $\sqrt{T} \lesssim n \lesssim T$, and finally reaches $\widetilde{\Theta}(\log T)$ when $n \gtrsim T$. This is depicted in Figure 5-2, from which we can

clearly see that there are three ranges of $n$, i.e., $0 < n \lesssim \sqrt{T}$, $\sqrt{T} \lesssim n \lesssim T$ and $n \gtrsim T$, referred to as the first, second and third *phase* respectively, and the optimal regret shows different properties in different phases. We refer to the significant transitions between the regret-decaying patterns of different phases as *phase transitions*.

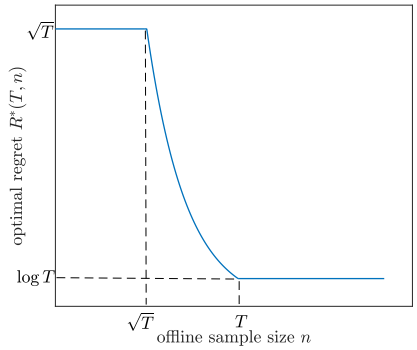


Figure 5-2: Phase transitions for the single-historical-price setting with constant $\delta$.

In contrast to the well-separated case where the phase transitions do not depend on the value of $\delta$, in the general case where $\delta$ may be very small, we cannot simply ignore the effect of $\delta$ in the optimal regret, and as a result, the number of phases and the thresholds of the offline sample size that define different phases are closely related to the magnitude of $\delta$. As illustrated in Figure 5-3, when $\delta = \widetilde{\Omega}(T^{-\frac{1}{4}})$, similar to the well-separated case, there are three phases defined by two thresholds of $n$: the optimal regret remains at the level of $\widetilde{\Theta}(\sqrt{T})$ in the first phase, and gradually decays according to $\widetilde{\Theta}(\frac{T}{n\delta^2})$ in the second phase, and stays at the level of $\widetilde{\Theta}(\frac{\log T}{\delta^2})$ in the third phase. When $\delta = \widetilde{\mathcal{O}}(T^{-\frac{1}{4}})$, there is no phase transition, and the optimal regret rate is always $\widetilde{\Theta}(\sqrt{T})$.

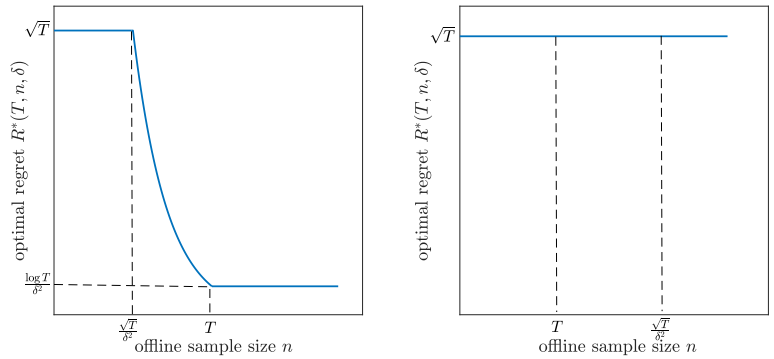


Figure 5-3: Phase transitions for the single-historical-price setting with general $\delta$. Left figure: $\delta \gtrsim T^{-\frac{1}{4}}$; right figure: $\delta \lesssim T^{-\frac{1}{4}}$.

Second, Corollary 5.1 also characterizes the impact of the location of offline data relative to the optimal price on the optimal regret, which can be stated in the following *inverse-square law*: whenever offline data take effect, i.e., $\delta = \widetilde{\Omega}(T^{-\frac{1}{4}})$, and $n$ is in the second phase or the third phase, the optimal regret is inversely proportional to the square of generalized distance $\delta$. Therefore, the factor $\delta^{-2}$ is intrinsic in the regret bound. Seemingly counter-intuitive, the inverse-square law indicates that the closer the historical price is to the optimal price, the more difficult it is to learn the demand parameter, and the larger the optimal regret will be. In fact, this is a consequence of the exploration-exploitation trade-off. In the presence of offline data, a "good" learning algorithm needs to deviate from historical price $\hat{p}$ to conduct price experimentation. However, when $\delta$ is extremely small, such a deviation will also lead to a significant gap with the optimal price, and therefore incurs greater revenue loss to the seller. In an extreme case when the historical price happens to be the optimal price, i.e., $\delta = 0$, even if $n = \infty$, the optimal regret is always $\widetilde{\Theta}(\sqrt{T})$.

## 5.4 Multiple Historical Prices

In this section, we consider the multiple-historical-price setting, where the $n$ historical prices can be different. In this case, $\sigma$ can be strictly positive and will play an important role to further reducing the complexity of the online learning task.

### 5.4.1 M-O3FU Algorithm and Regret Upper Bound

In this subsection, we develop a learning algorithm for the multiple-historical-price setting. We first make the following observations.

(i) If $n\sigma^2 \gtrsim \sqrt{T}$ and $\delta \lesssim T^{-\frac{1}{4}}$, then the offline data provide so much information that there is no need for online learning. In fact, by simply running linear regression on the offline data, we can obtain the estimate $\hat{\theta}_0$ for the true demand parameter with the squared estimation error of $\mathcal{O}(\frac{1}{n\sigma^2})$, i.e., $\mathbb{E}[||\hat{\theta}_0 - \theta^\star||^2] = \mathcal{O}(\frac{1}{n\sigma^2})$, which means that by simply charging price $\psi(\hat{\theta}_0)$ in each online period, we achieve the regret of $\mathcal{O}(\frac{T}{n\sigma^2})$. Note that this $\mathcal{O}(\frac{1}{n\sigma^2})$-type estimation error cannot be further improved in the online

process by policies within $\Pi^\circ$, since when $T\delta^2 \lesssim \sqrt{T} \lesssim n\sigma^2$, we have

$$\begin{aligned}\mathbb{E}[J(\hat{p}_1,\ldots,\hat{p}_n,p_1,\ldots,p_T)] &\le 2\Big(\mathbb{E}[J(\hat{p}_1,\ldots,\hat{p}_n)] + \sum_{t=1}^{T}\mathbb{E}[(p_t-p^\star)^2] + T(\bar{p}_{1:n}-p^\star)^2\Big)\\ &\lesssim n\sigma^2, \end{aligned} \tag{5.8}$$

where $J(x_1,x_2,\ldots,x_k) := \sum_{i=1}^k (x_i - \frac{1}{k}\sum_{j=1}^k x_j)^2$ for any sequence $\{x_i\}_{i=1}^k$ and $k \ge 1$. This suggests that in the online process, exploration is "useless" in the sense that it cannot bring any theoretical improvement (in terms of reducing the *order* of estimation error) beyond offline regression. Therefore, if the algorithm knew that conditions $n\sigma^2 \gtrsim \sqrt{T}$ and $\delta \lesssim T^{-\frac{1}{4}}$ hold, then there is no exploration-exploitation trade-off at all.

(ii) If in addition to the conditions in (i), a further extreme condition $\delta^2 \lesssim \frac{1}{n\sigma^2}$ occurs, then even the above offline-regression-based approach may still be conservative: if an algorithm knew that $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$, then by simply charging $\bar{p}_{1:n}$ in every online period, it achieves the regret of $\mathcal{O}(T\delta^2)$, which is even better than $\mathcal{O}(\frac{T}{n\sigma^2})$. We refer to $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$ as the *corner case*, and its complement as the *regular case*.

(iii) However, since the algorithm does not know the value of $\delta$ in advance, it does not know whether it is in the corner case (i.e., whether $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$ is true) in advance. If the conditions in (i) do not hold, then the algorithm still needs online exploration; if the condition in (ii) does not hold, then the algorithm still needs offline regression.

Motivated by the above observations, we design the following *Modified O3FU* (M-O3FU) algorithm. With an abuse of terminology, we refer to O3FU algorithm in this section as the one proposed in §5.3.1 after natural modification to the multiple-historical-price setting by letting $V_{0,n} = \lambda I + \sum_{i=1}^n [1\ \hat{p}_i]^\top [1\ \hat{p}_i]$, $Y_{0,n} = \sum_{i=1}^n \hat{D}_i [1\ \hat{p}_i]^\top$, and $p_1 = l \cdot \mathbb{I}\{\bar{p}_{1:n} > \frac{u+l}{2}\} + u \cdot \mathbb{I}\{\bar{p}_{1:n} \le \frac{u+l}{2}\}$.

We next make several highlights about M-O3FU algorithm. First, in comparison with O3FU algorithm, before the start of the online learning process, M-O3FU algorithm takes a preliminary step that tests whether the distance between $\bar{p}_{1:n}$ and interval $\{\psi(\theta) : \theta \in \mathcal{C}_0\}$ is smaller than a constant times the length of interval $\{\psi(\theta) : \theta \in \mathcal{C}_0\}$. The goal of this step is to test whether condition $\delta^2 \lesssim \frac{1}{n\sigma^2}$ holds or not. If this condition is inferred to hold based on the empirical observation, and in addition, $n\sigma^2 \ge \sqrt{T}$, the algorithm keeps using

**Algorithm 5.2** M-O3FU Algorithm

**Input**: offline data $\{(\hat{p}_i, \hat{D}_i)\}_{i=1}^n$, support of demand parameters $\Theta^\dagger$, price range $[l, u]$, regularization parameter $\lambda = 1 + u^2$, $\{w_t\}_{t\geq 0}$ defined in (5.2) with $\epsilon = \frac{1}{T^2} \wedge \frac{1}{n\sigma^2}$, parameter $K > 1$

**Initialization**: $V_{0,n} = \lambda I + \sum_{i=1}^n [1\ \hat{p}_i]^\mathsf{T}[1\ \hat{p}_i]$, $Y_{0,n} = \sum_{i=1}^n \hat{D}_i[1\ \hat{p}_i]^\mathsf{T}$, $\hat{\theta}_0 = V_{0,n}^{-1} Y_{0,n}$, $\mathcal{C}_0 = \{\theta \in \Theta^\dagger : ||\theta - \hat{\theta}_0||_{V_{0,n}} \leq w_0\}$

1: **if** $\frac{\min_{\theta\in\mathcal{C}_0} |\bar{p}_{1:n} - \psi(\theta)|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0} |\psi(\theta_1) - \psi(\theta_2)|} \leq K$, and $n\sigma^2 \geq \sqrt{T}$ **then**
2: Charge price $p_t = \bar{p}_{1:n}$ for each $t \in [T]$
3: **else**
4: Run O3FU Algorithm.

$\bar{p}_{1:n}$ for each online period. Otherwise, the algorithm simply runs O3FU algorithm. Second, parameter $\epsilon$ defined in $w_t$ is modified from $\frac{1}{T^2}$ (used in O3FU) to $\frac{1}{T^2} \wedge \frac{1}{n\sigma^2}$, which guarantees that $\theta^\star$ belongs to each confidence ellipsoid $\mathcal{C}_t$ with sufficiently high probability, and the revenue loss incurred when $\theta^\star$ does not belong to some confidence ellipsoid can be bounded by $\mathcal{O}(\frac{T}{n\sigma^2} \wedge \frac{1}{T})$.

The following theorem provides an upper bound on the regret of M-O3FU algorithm.

**Theorem 5.3.** *Let $\pi$ be M-O3FU algorithm. Then there exists a finite constant $K_3 > 0$ such that for any $T \geq 1$, $n \geq 0$, $\sigma \geq 0$ and $\bar{p}_{1:n} \in [l, u]$, and for any possible value of $\theta^\star \in \Theta^\dagger$, we have*

$$R^\pi_{\theta^\star}(T) \leq \begin{cases} K_3 \cdot \big(T\delta^2 + 1\big), & \text{if } \delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}; \\ K_3 \cdot \Big(\big(\sqrt{T}\log T\big) \wedge \frac{T(\log T)^2}{n\sigma^2 + (n\wedge T)\delta^2} + 1\Big), & \text{otherwise.} \end{cases}$$

Theorem 5.3 shows that the regret upper bound has different forms in two different cases. When $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$, M-O3FU algorithm achieves the regret upper bound $\mathcal{O}(T\delta^2 + 1)$, which matches the ideal regret bound in the above item (ii) discussed at the beginning of this subsection. Otherwise, the regret upper bound becomes $\widetilde{\mathcal{O}}(\sqrt{T} \wedge \frac{T}{n\sigma^2 + (n\wedge T)\delta^2} + 1)$. Compared with the upper bound in Theorem 5.1, there is an additional term $n\sigma^2$ in the denominator capturing the effect of the dispersion of offline data. We summarize the regret upper bound under different $(n, \sigma, \delta)$ combinations in Table D.3 of Appendix D.8.

The proof of Theorem 5.3 can be found in Appendix D.2.1. Similar to the proof of Theorem 5.1, we also need an important technical lemma stated as follows.

**Lemma 5.2.** *Suppose we run O3FU algorithm from $t = 1$ with given input offline data*

$\{(\hat{p}_i, \hat{D}_i)\}_{i=1}^n$, $\sigma \le \delta$, $\delta \ge \max\{\frac{\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}}{\beta_{\max}^2}\frac{T^{1/4}w_T}{n^{1/2}}, \sqrt{C_1}T^{-1/4}\}$, *and* $\theta^\star \in \mathcal{C}_t$ *for each* $t \in [T]$, *then two sequences of events* $\{U_{t,3}\}_{t=1}^T$ *and* $\{U_{t,4}\}_{t=2}^T$ *also hold, where*

$$U_{t,3} = \Big\{|p_t - \bar{p}_{1:n}| \ge \min\big\{1 - \frac{\sqrt{2}}{2}, \frac{C_0}{2}\big\} \cdot \delta\Big\},$$
$$U_{t,4} = \Big\{||\tilde{\theta}_t - \theta^\star||^2 \le C_3 \frac{w_{t-1}^2}{n\sigma^2 + (n \wedge (t-1))\delta^2}\Big\},$$

*and* $C_0$ *and* $C_1$ *are defined in Lemma 5.1, and* $C_3 = \max\Big\{8(u-l)^2, 4C_1, 2\max\{2(\sqrt{2}+1)^2, \frac{4}{C_0^2}\} \cdot ((4u+1)^2+1)\Big\}$.

Similar to Lemma 5.1, Lemma 5.2 is also proved based on induction arguments. Besides, we need to use the following lower bound on the sum of squared price deviations:

$$J(\hat{p}_1, \ldots, \hat{p}_n, p_1, \ldots, p_t) \ge J(\hat{p}_1, \ldots, \hat{p}_n) + \frac{n}{n+t}\sum_{s=1}^t (p_s - \bar{p}_{1:n})^2, \tag{5.9}$$

where $J(x_1, x_2, \ldots, x_k) := \sum_{i=1}^k (x_i - \frac{1}{k}\sum_{j=1}^k x_j)^2$ for any sequence $\{x_i\}_{i=1}^k$ and $k \ge 1$. We can interpret $J(x_1, x_2, \ldots, x_k)$ as the information metric capturing the variation for a sequence $\{x_i\}_{i=1}^k$. Then inequality (5.9) bounds the information accumulated up to period $t$ from below, through the pre-existing offline information, plus the information due to the deviation of the algorithm's prices from the average historical price. The proof of Lemma 5.2 is provided in Appendix D.2.2.

### 5.4.2 Lower Bound on Regret

In this subsection, we establish a lower bound on the best-achievable regret for the `OPOD` problem among the class of admissible policies $\Pi^\circ$ defined in a similar way to (5.3). Again, we denote $R_\theta^\pi(T, n, \delta, \sigma)$ as the regret incurred by policy $\pi \in \Pi^\circ$ under the demand parameter $\theta$.

**Theorem 5.4.** *There exists a positive constant* $K_4$ *such that for any admissible policy* $\pi \in \Pi^\circ$, *for any* $\xi \in (0,1)$, $T \ge 2$, $n \ge 0$ *and* $\sigma \ge 0$, *and for any* $\delta \in [0, u-l]$,

$$\sup_{\substack{\mathcal{D}\in\mathcal{E}(R);\\ \theta\in\Theta^\dagger:|\bar{p}_{1:n}-\psi(\theta)|\in[(1-\xi)\delta,(1+\xi)\delta]}} R_\theta^\pi(T,n,\delta,\sigma) \ge \begin{cases} K_4 \cdot T\delta^2, & \text{if } \delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}; \\ K_4 \cdot \Big(\frac{\sqrt{T}}{(\log T)^{\lambda_0}} \wedge \frac{T}{n\sigma^2+(n\wedge T)\delta^2}\Big), & \text{otherwise.} \end{cases}$$

Similar to Theorem 5.2, the instance-dependent environment class is defined as the set of instances whose associated optimal prices are away from $\bar{p}_{1:n}$ by a distance $\Theta(\delta)$. Since M-O3FU algorithm achieves the regret upper bound $\mathcal{O}(\sqrt{T}\log T)$ for any value of $\theta^\star \in \Theta^\dagger$ (thus belongs to the admissible policy class $\Pi^\circ$ with $\lambda_0 \geq 1$), Theorem 5.4 demonstrates that for both the corner and regular cases, the regret rate achieved by M-O3FU algorithm in Theorem 5.3 cannot be further improved by any policy in $\Pi^\circ$. The proof of Theorem 5.4 is provided in Appendix D.2.4, which is a generalization to that of Theorem 5.2. We also summarize the regret lower bound under different $(n, \sigma, \delta)$ combinations in Table D.4 of Appendix D.8.

### 5.4.3 Phase Transitions and Generalized Inverse-Square Law

Motivated from the matching upper and lower bounds (after ignoring logarithm factors) in Theorems 5.3 and 5.4 respectively, we define the optimal instance-dependent regret for the `OPOD` problem in the multiple-historical-price setting as follows:

$$R^\star(T, n, \delta, \sigma) = \inf_{\pi \in \Pi^\circ} \sup_{\substack{\mathcal{D} \in \mathcal{E}(R); \\ \theta \in \Theta^\dagger : |\psi(\theta) - \bar{p}_{1:n}| \in [(1-\xi)\delta, (1+\xi)\delta]}} R^\pi_\theta(T, n, \delta, \sigma), \tag{5.10}$$

where a slight difference compared with (5.7) is the modification from the single historical price $\hat{p}$ to the average historical price $\bar{p}_{1:n}$.

Combining Theorem 5.3 and Theorem 5.4, we are able to characterize the optimal regret of the `OPOD` problem for the multiple-historical-price setting.

**Corollary 5.2.** *The optimal regret defined in* (5.10) *for the multiple-historical-price setting is*

$$R^\star(T, n, \delta, \sigma) = \begin{cases} \widetilde{\Theta}\Big(\sqrt{T} \wedge \frac{T}{n\sigma^2 + (n \wedge T)\delta^2}\Big), & \textit{for the regular case;} \\ \widetilde{\Theta}\big(T\delta^2\big), & \textit{for the corner case.} \end{cases}$$

Recall that in the single-historical-price setting, the threshold $\widetilde{\Theta}(T^{-\frac{1}{4}})$ of $\delta$ plays an important role in characterizing the behavior of the optimal regret rate. This threshold $\widetilde{\Theta}(T^{-\frac{1}{4}})$ also plays a role in the optimal regret rate of the multiple-historical-price setting.

When $\delta \gtrsim T^{-\frac{1}{4}}$, there are significant differences for the behaviors of the optimal regret rate, depending on whether $\sigma$ is less than, equal to or greater than $\delta$. This is illustrated in

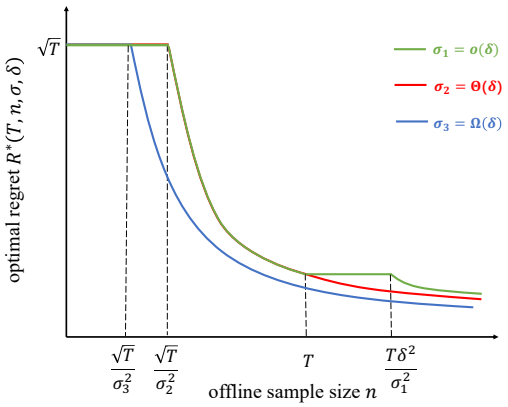


Figure 5-4: Multiple-historical-price setting with $\delta \gtrsim T^{-\frac{1}{4}}$ and different $\sigma$.

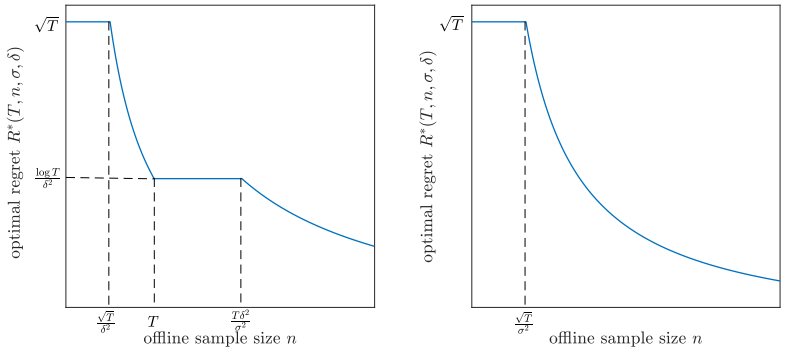


Figure 5-5: Phase transitions for the multiple-historical-price setting with $\delta \gtrsim T^{-\frac{1}{4}}$. Left figure: $\sigma = o(\delta)$; right figure: $\sigma = \Omega(\delta)$.

Figure 5-4, where the green, red and blue curves depict the above three cases respectively. If $\sigma = o(\delta)$, as shown in the green curve, the optimal regret rate exhibits four decaying patterns as $n$ changes between different ranges. Specifically, the optimal regret rate first remains at $\widetilde{\Theta}(\sqrt{T})$ when $n \lesssim \frac{\sqrt{T}}{\delta^2}$, and then decreases according to $\widetilde{\Theta}(\frac{T}{n\delta^2})$ when $\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim T$. After that, the optimal regret rate stays at $\widetilde{\Theta}(\frac{\log T}{\delta^2})$ when $T \lesssim n \lesssim \frac{T\delta^2}{\sigma^2}$, and finally, it decreases according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$ when $n \gtrsim \frac{T\delta^2}{\sigma^2}$. If $\sigma = \Theta(\delta)$ or $\sigma = \Omega(\delta)$ as shown in the red or blue curve, the optimal regret rate exhibits two phases: it remains at the level of $\widetilde{\Theta}(\sqrt{T})$ when $n \lesssim \frac{\sqrt{T}}{\sigma^2}$, and decays according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$ when $n \gtrsim \frac{\sqrt{T}}{\sigma^2}$. Therefore, when $\sigma$ gradually increases, depending on its magnitude compared with $\delta$, the number of phases of the optimal regret rate also experiences the change from four phases to two phases, and the entire patterns of the phase transitions of the optimal regret rate also change accordingly.

Corollary 5.2 also reveals a generalized inverse-square law. Specifically, the optimal

regret is inversely proportional to the square of both $\delta$ and $\sigma$, which quantifies the effect of the location and dispersion of the offline data on the optimal regret. The intuition for the dependence of the optimal regret on $\delta$ is similar to the single-historical-price setting. For the dependence of the optimal regret on $\sigma$, as the historical prices become more dispersive, i.e., $\sigma$ increases, the seller can obtain a more accurate estimate for the unknown demand parameter from offline regression, which helps to further reduce the optimal regret of the online learning process.

It's also worth noting that the thresholds of the offline sample size that define different phases of the optimal regret depend on both $\delta$ and $\sigma$. When $\sigma = \mathcal{O}(\delta)$ and $\delta = \widetilde{\Omega}(T^{-\frac{1}{4}})$, the first threshold of $n$ that defines the first and second phases, i.e., $\widetilde{\Theta}(\frac{\sqrt{T}}{\delta^2})$, decreases in $\delta$. When $\sigma = \Omega(\delta)$ and $\delta = \widetilde{\Omega}(T^{-\frac{1}{4}})$, the threshold of $n$ that defines the first and second phases, i.e., $\widetilde{\Theta}(\frac{\sqrt{T}}{\sigma^2})$, decreases in the standard deviation $\sigma$. This implies that more offline data will be required to overcome the challenges caused by a shorter generalized distance $\delta$ or a smaller standard deviation $\sigma$.

When $\delta \lesssim T^{-\frac{1}{4}}$, Corollary 5.2 indicates that there are three phases of the optimal regret rate as $n$ changes. When $n \lesssim \frac{\sqrt{T}}{\delta^2}$, the optimal regret remains at $\widetilde{\Theta}(\sqrt{T})$. When $\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim \frac{1}{\delta^2\sigma^2}$, the optimal regret experiences a *sudden* drop from $\widetilde{\Theta}(\sqrt{T})$ to $\widetilde{\Theta}(T\delta^2)$. When $n \gtrsim \frac{1}{\delta^2\sigma^2}$, the optimal regret decays according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$. Such transitions of the optimal regret with different $n$ are illustrated in Figure 5-6. In particular, the second phase corresponds to the corner case defined in §5.4.1. In this case, smaller $\delta$ leads to lower optimal regret, which is in contrast to the inverse-square law in the regular case. This is because in the corner case, as discussed in §5.4.1, there is no need for online learning and therefore no exploration-exploitation trade-off, and the policy that always charges $\bar{p}_{1:n}$ incurs very small regret. In this case, the closer the average historical price is to the optimal price, the smaller the optimal regret will be. By contrast, the inverse-square law in the regular case is a consequence of the exploration-exploitation trade-off.

## 5.5 Numerical Experiments

In this section, we test the performance of our algorithm on synthetic data sets. We define the relative regret for a given learning algorithm $\pi$ as $\frac{Tp^\star\cdot(\alpha^\star+\beta^\star p^\star)-\sum_{t=1}^{T}\mathbb{E}^{\pi}_{\theta^\star}[p_t(\alpha^\star+\beta^\star p_t)]}{Tp^\star\cdot(\alpha^\star+\beta^\star p^\star)} \times 100\%$, and the following three problem instances are tested:

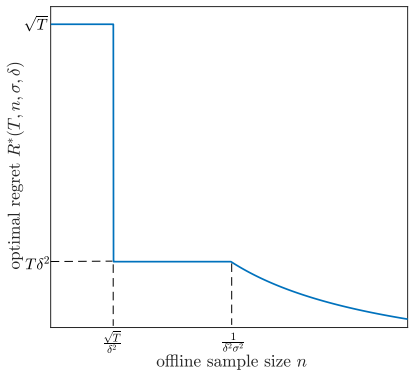


Figure 5-6: Phase transitions for the multiple-historical-price setting with $\delta \lesssim T^{-\frac{1}{4}}$.

(1) $\theta^\star = [2.6, -1.8]$, $[\alpha_{\min}, \alpha_{\max}] = [2.5, 3.5]$, $[\beta_{\min}, \beta_{\max}] = [-2, -1.3]$, $[l, u] = [0.1, 2]$, $R = 2.2$;

(2) $\theta^\star = [3.7, -3.15]$, $[\alpha_{\min}, \alpha_{\max}] = [3.5, 5]$, $[\beta_{\min}, \beta_{\max}] = [-3.2, -2.5]$, $[l, u] = [0.5, 1.3]$, $R = 2.5$;

(3) $\theta^\star = [2.9, -2.6]$, $[\alpha_{\min}, \alpha_{\max}] = [2.8, 3.5]$, $[\beta_{\min}, \beta_{\max}] = [-2.8, -1]$, $[l, u] = [0.2, 2]$, $R = 1.8$.

and $\varepsilon$ follows a normal distribution with standard deviation $R$. For each of the above instance, we repeat the experiments for 500 times, and the results are computed after averaging over the 500 experiments. Under the multiple-historical-price setting, we test a simplified version of M-O3FU algorithm by directly running O3FU, without checking the preliminary condition. Thus, throughout this section, we simply call our algorithm "O3FU algorithm."

First, we compare our O3FU algorithm with the modified Constrained Iterated Least Squares (CILS) algorithm. When there are no offline data, we adopt CILS algorithm directly from Keskin and Zeevi (2014). When there are offline data, no existing learning algorithm in prior literature is directly suitable for this setting, so we make a natural modification to the original CILS by incorporating offline data into the least-square estimation. In both cases, we set the tuning parameter $\kappa$ in CILS to be 0.1 following Keskin and Zeevi (2014), and also 0.5 which seems to lead to the best performance of CILS. Figures 5-7 and 5-8 show the performances of O3FU and CILS algorithms for the settings when there are no offline data, and when there are $n = 1000$ offline demand data under a single historical price (specifically, we set $\hat{p} = 1.8, 0.9, 1$ for instances (1)-(3) respectively). As seen from Figure 5-7, without

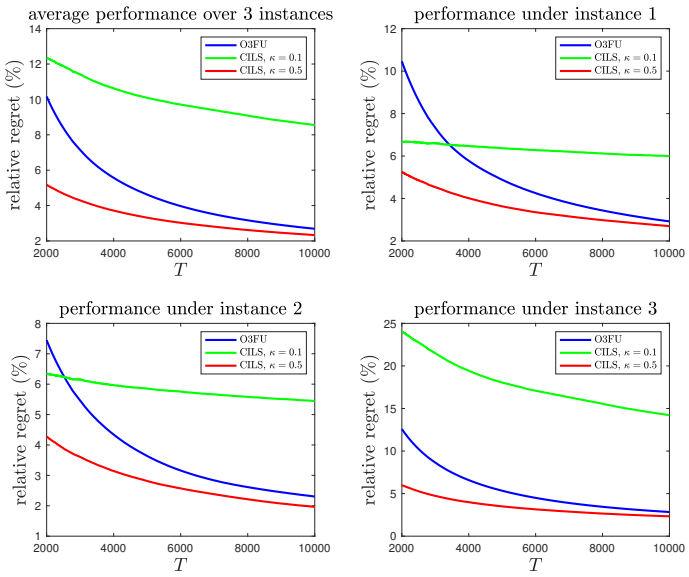


Figure 5-7: Comparison between O3FU and CILS when there are no offline data.

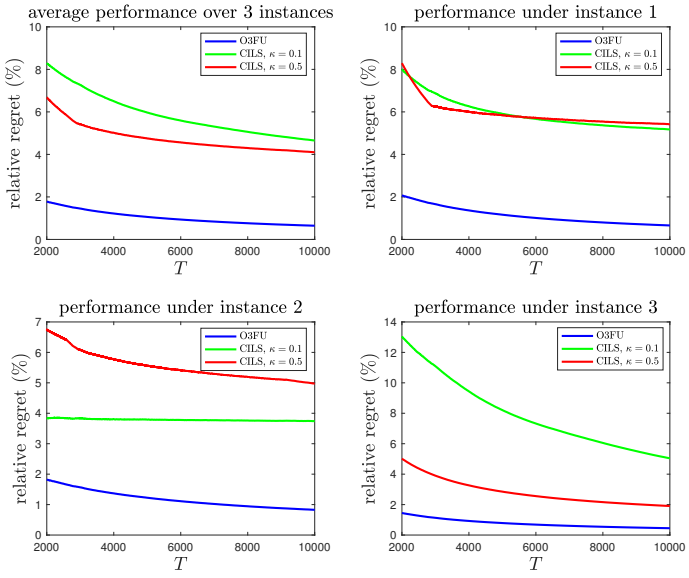


Figure 5-8: Comparison between O3FU and CILS when there are $n = 1000$ offline demand data.

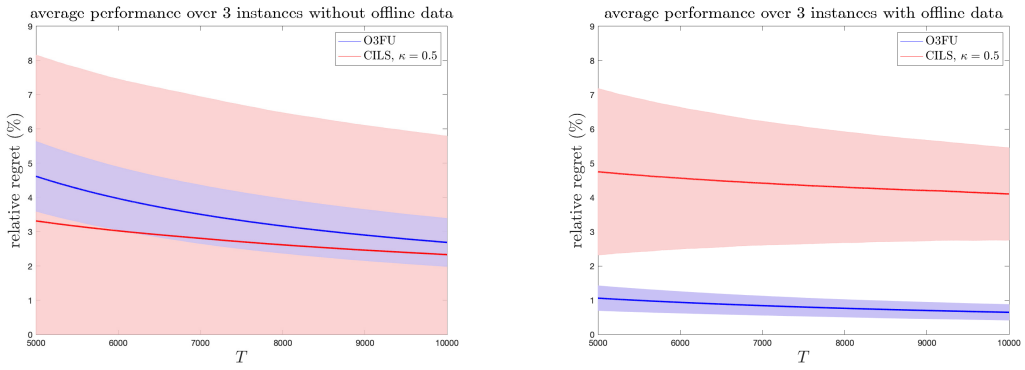


Figure 5-9: 95% confidence-region comparison between O3FU and CILS with $\kappa = 0.5$.

offline data, O3FU performs better than CILS with $\kappa = 0.1$ and comparably to CILS with $\kappa = 0.5$ as $T$ becomes larger. Figure 5-8 reveals that with the help of offline data, the regret of O3FU algorithm is significantly reduced for all $T$ under all instances. By contrast, for CILS algorithms, the impact of offline data on the empirical regret is not obvious and heavily relies on the tuning parameter and specific problem instance. For CILS with $\kappa = 0.1$, the improvement of the relative regret is clear under instance (3), but rather minimal under instances (1) and (2). For CILS with $\kappa = 0.5$, the regret only decreases a little under instance (3), and even becomes larger under instances (1) and (2). Therefore, compared with CILS algorithms, O3FU algorithm better exploits the value of offline data and is more robust to different problem instances.

Second, Figure 5-9 plots the 95% confidence region of O3FU algorithm and CILS algorithm with $\kappa = 0.5$, for both cases when there are no offline data and when there are $n = 1000$ offline data. The left figure shows that while CILS with properly tuned parameter performs slightly better than O3FU on average when there are no offline data, the standard deviation of CILS among the 500 simulations is much larger than O3FU. This implies that O3FU is more stable than CILS. The right figure shows that with offline data, O3FU always outperforms CILS, in terms of both the average regret and standard deviation. Since O3FU algorithm has highly stable performance, we believe that it should be preferable in many real-life business settings.

Third, we investigate the effect of offline sample size $n$ on the empirical regret of O3FU algorithm. In Figure 5-10, we plot the relative regret of O3FU algorithm given different amount of offline data (with $n$ ranging from 20 to 12000), under the single-historical-price setting (with $\hat{p} = 1.8, 0.9, 1$ for instances (1)-(3) respectively). The x-axis is depicted on a log

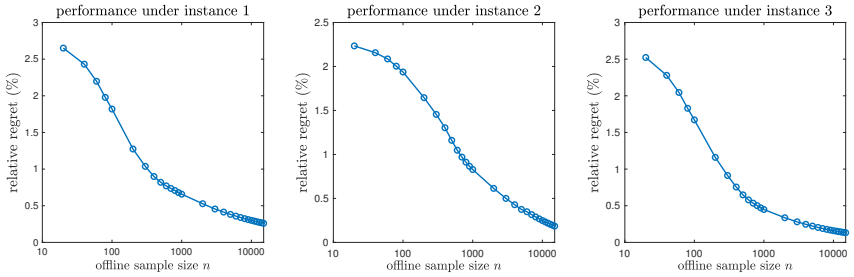


Figure 5-10: $T = 10^4$-period relative regret for the single-historical-price setting with different $n$.

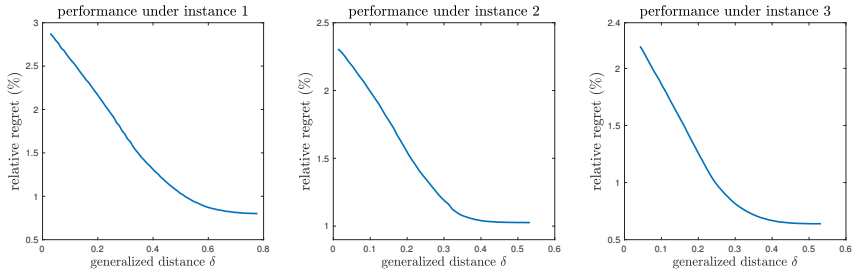


Figure 5-11: $T = 10^4$-period relative regret for the single-historical-price setting with different $\delta$.

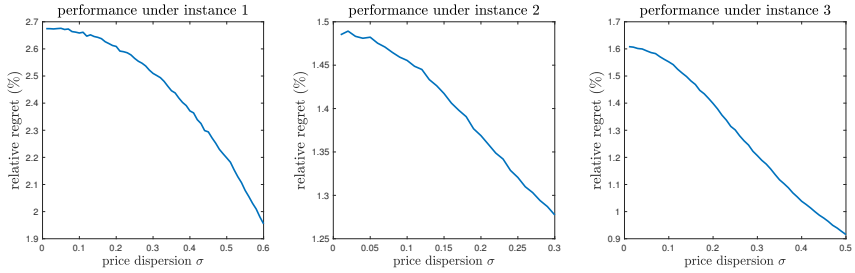


Figure 5-12: $T = 10^4$-period relative regret for the multiple-historical-price setting with different $\sigma$.

scale. We can see clearly that for each problem instance, as the offline sample size increases, the relative regret decreases, which is consistent with the phase transitions implied from our theoretical results.

Finally, we investigate the effects of generalized distance $\delta$ and price dispersion $\sigma$ on the empirical regret of our algorithm. Figure 5-11 shows the relative regret of O3FU algorithm given $n = 500$ offline demand observations under historical price $p^\star + \delta$ with different $\delta$, and Figure 5-12 shows the relative regret of O3FU algorithm given 250 offline demand observations under historical price $\bar{p}_{1:n} - \sigma$, and 250 offline demand observations under historical price $\bar{p}_{1:n} + \sigma$ with different $\sigma$, where $\bar{p}_{1:n} = 0.8, 0.8, 0.7$ for instances (1)-(3) respectively. As seen from Figure 5-11 and 5-12, when $\delta$ or $\sigma$ increases, the empirical regret of our algorithm decays, which also matches the inverse-square law.

**Remark 5.2.** *We remark that the empirical evidence for the phase transitions and inverse-square law is not always observed under every problem instance. This is because according to its definition through the supremum over some instance-dependent environment class, the optimal regret should be attained at some "hard" instances, and so do its implications of the phase transitions and inverse-square law. Besides, when discussing the optimal regret rate and its implications, we require $T$ to be sufficiently large, and ignore all the constant factors. Our choices of instances (1)-(3) capture the aforementioned hard instances, and also avoid that the problem falls into the regimes where constant factors significantly affect the overall regret rate.*

## 5.6 Further Discussion: Offline Data and Self-Exploration

In M-O3FU algorithm proposed in §5.4.1, there is a preliminary step testing whether $\frac{\min_{\theta\in\mathcal{C}_0}|\bar{p}_{1:n}-\psi(\theta)|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} \leq K$ holds or not. We find that this step also has an important implication in practice: $\frac{\min_{\theta\in\mathcal{C}_0}|\bar{p}_{1:n}-\psi(\theta)|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} > K$ is actually a sufficient condition for *self-exploration* in our `OPOD` problem. That is, with high probability, when this condition holds, the myopic (i.e., greedy) policy can achieve the optimal regret without any active exploration.

The myopic policy is defined as follows. Let $\mathcal{C}_0 = \{\theta \in \Theta^\dagger : ||\theta - \theta_0^{\mathrm{LS}}||^2_{V_{0,n}} \leq w_0^2\}$, where $V_{0,n} = \lambda I + \sum_{i=1}^n [1\ \hat{p}_i]^\top [1\ \hat{p}_i]$, $\theta_0^{\mathrm{LS}} = \arg\min_{\theta\in\Theta^\dagger} \sum_{i=1}^n ((\hat{D}_i - \alpha - \beta\hat{p}_i)^2 + \lambda(\alpha^2+\beta^2))$, and $w_0 = R\sqrt{2\log((T^2 \vee n\sigma^2)(1+(1+u^2)n/\lambda))} + \sqrt{\lambda(\alpha^2_{\max}+\beta^2_{\min})}$. Let $\{p_t^{\mathrm{myopic}}\}_{t\geq 1}$ be the

sequence of prices charged by the myopic policy. For $t = 1$, $p_t^{\text{myopic}} = \psi(\theta_0^{\text{LS}})$, and for each $t \geq 2$, we first compute the least-square estimator $\theta_{t-1}^{\text{LS}}$ based on offline data and all the available online data within confidence ellipsoid $\mathcal{C}_0$:

$$\theta_{t-1}^{\text{LS}} = \arg\min_{\theta \in \mathcal{C}_0} \Big( \sum_{i=1}^{n} (\hat{D}_i - \alpha - \beta \hat{p}_i)^2 + \sum_{s=1}^{t-1} (D_s - \alpha - \beta p_s)^2 + \lambda(\alpha^2 + \beta^2) \Big),$$

and then let $p_t^{\text{myopic}} = \psi(\theta_{t-1}^{\text{LS}})$. The next proposition shows that the myopic policy is guaranteed to be optimal if certain condition holds.

**Proposition 5.1.** *Suppose $n\sigma^2 \geq \sqrt{T}$. Then with probability at least $1 - \frac{1}{T^2} \wedge \frac{1}{n\sigma^2}$, the following event holds: if $\frac{\min_{\theta \in \mathcal{C}_0} |\bar{p}_{1:n} - \psi(\theta)|}{\max_{\theta_1, \theta_2 \in \mathcal{C}_0} |\psi(\theta_1) - \psi(\theta_2)|} > K$ for some $K > 1$, then the myopic policy ensures that the regret is $\widetilde{\mathcal{O}}(\frac{T}{n\sigma^2 + (n \wedge T)\delta^2})$.*

The intuition of Proposition 5.1 is as follows. Note that the key step to prove the instance-dependent upper bound $\widetilde{\mathcal{O}}(\frac{T}{n\sigma^2 + (n \wedge T)\delta^2})$ in Theorem 5.3 is to show events $\{U_{t,3}\}_{t=1}^T$ and $\{U_{t,4}\}_{t=2}^T$ in Lemma 5.2 hold. Since the myopic policy charges prices based on estimator $\theta_{t-1}^{\text{LS}} \in \mathcal{C}_0$ in each period $t$, and $\mathcal{C}_0$ contains $\theta$ with high probability, under the condition in Proposition 5.1, we can easily verify that the myopic price $p_t^{\text{myopic}}$ is bounded away from $\bar{p}_{1:n}$ by a distance proportional to $\delta$. In other words, event $U_{t,3}$ in Lemma 5.2 is automatically satisfied for each period $t$, and in this case, when $\theta^\star \in \mathcal{C}_t = \{\theta \in \Theta^\dagger : ||\theta - \theta_t^{\text{LS}}||_{V_{t,n}}^2 \leq w_t^2\}$, we can further show that event $U_{t,4}$ also holds. Therefore, the myopic policy ensures that the regret is $\widetilde{\mathcal{O}}(\frac{T}{n\sigma^2 + (n \wedge T)\delta^2})$.

We also make several remarks about Proposition 5.1. First, the interpretation of probability "$1 - \frac{1}{T^2} \wedge \frac{1}{n\sigma^2}$" is similar to the interpretation of "95%" in a 95% confidence interval. Such a probabilistic statement is common in frequentist statistics, when one wants to make some inference (e.g., myopic policy is optimal or not) based on some empirical observations (e.g., $\frac{\min_{\theta \in \mathcal{C}_0} |\psi(\theta) - \bar{p}_{1:n}|}{\max_{\theta_1, \theta_2 \in \mathcal{C}_0} |\psi(\theta_1) - \psi(\theta_2)|} > K$). Second, if the regular case happens, i.e., $n\sigma^2 \gtrsim \sqrt{T}$ and $\delta^2 \gtrsim \frac{1}{n\sigma^2}$, one can easily verify that the empirical condition described in Proposition 5.1 always holds. In this case, the myopic algorithm always ensures $\widetilde{\mathcal{O}}(\frac{T}{n\sigma^2 + (n \wedge T)\delta^2})$ regret. Nevertheless, verifying the condition $\delta^2 \gtrsim \frac{1}{n\sigma^2}$ requires knowing the true parameter $\theta^\star$ in advance, which is not practical in reality. Thus, we make a probabilistic statement in Proposition 5.1 about the regret bound under an empirical condition that can be directly verified by the algorithm. Third, the choice of $1 - \frac{1}{T^2} \wedge \frac{1}{n\sigma^2}$ is not essential in Proposition 5.1. In fact, one

can achieve any higher probability bound that is arbitrarily close to 1 by defining a larger confidence ellipsoid $\mathcal{C}_0$, although in that case, the condition $\frac{\min_{\theta\in\mathcal{C}_0}|\psi(\theta)-\bar{p}_{1:n}|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} > K$ will be more difficult to be satisfied.

In reality, myopic policies are commonly adopted in many industries, since they are quite easy to explain to managers, and relatively simple to implement in practice. See the discussion of myopic policies in, e.g. Harrison et al. (2012), Keskin and Zeevi (2014), Qiang and Bayati (2016). However, due to the lack of active exploration, myopic policies typically suffer from *incomplete learning*, thus usually have poor theoretical performance in dynamic pricing. Proposition 5.1 shows how offline data may help myopic policies to achieve self-exploration in dynamic pricing: when there are enough dispersive offline data, then with high probability, as long as $\bar{p}_{1:n}$ is bounded away from offline confidence interval of $p^\star$, the issue of incomplete learning could be resolved, and the myopic policy could achieve self-exploration.

## 5.7 Concluding Remarks

In this paper, we investigate the impact of offline data on online learning in the context of dynamic pricing. In contrast to previous literature that involves only offline data or only online data, we consider a more practical problem involving both offline data and online data, aiming to understand whether and how the pre-existence of offline data would benefit the online learning process. For both single-historical-price and multiple-historical-price settings, we design a learning algorithm based on the OFU principle with a provable instance-dependent regret upper bound, and establish a regret lower bound that matches the upper bound up to logarithmic factors. Two important and nontrivial implications implied by our results are *phase transitions* and the *inverse-square law*, characterizing the joint effect of the size, location, and dispersion of the offline data on the optimal regret. The numerical experiments demonstrate the effectiveness, robustness and stability of our algorithm, and reveal the empirical evidence for phase transitions and the inverse-square law. Besides, we also develop a sufficient condition for the myopic policy to achieve the optimal regret in the regular case.

We discuss two extensions of this paper. First, while we focus on the linear demand model in this paper, the regret upper bounds developed in Theorems 5.1 and 5.3 can be

extended to the generalized linear model $D_t = g(\alpha^\star + \beta^\star p_t) + \varepsilon_t$ for some link function $g(\cdot)$, under certain smoothness conditions. In particular, these conditions guarantee that the regret in each single period $t$ for any given policy $\pi$ is of the same order as the quadratic estimation error $\mathbb{E}^\pi_{\theta^\star}[(\psi(\theta^\star) - p^\pi_t)^2]$, and that Lemmas 5.1 and 5.2 continue to hold. We refer the interested readers to Appendix D.5 for more details. Second, we assume that historical prices are fixed constants in this paper. In reality, offline pricing decisions can also be made based on the previous prices and sales observations according to some offline pricing policy, in which case offline data will be generated in an adaptive way. By modifying the performance metric to the expected regret conditioned on the observed offline price trajectory, we can extend our results to the setting with adaptive offline data. This extension is discussed in Appendix D.6.

This paper also suggests various directions for future research. First, with the development of information technology, firms have access to more detailed data that record customer information and product characteristics. It will be interesting to incorporate such contextual information into the model, and study context-based dynamic pricing with online learning and offline data. In this case, it's important to understand how the definition of the location metric of offline data should be modified accordingly. Second, we believe that the framework of online learning with offline data is quite general and widely applicable, and it will be also interesting to explore how to extend such a framework to derive new results and insights for other data-driven operational problems, e.g., pricing under substitutable products, bandit with knapsack constraints, inventory control with demand learning, etc. Third, by leveraging the location metric of offline data, this paper develops the instance-dependent regret bound, which goes beyond the traditional worst-case regret and is new to the literature on dynamic pricing with demand learning. It will be valuable to explore whether other types of instance-dependent bounds can be developed for dynamic pricing and revenue management problems by utilizing certain historical information.

# Chapter 6

# Bandits with Switching Constraints

## 6.1 Introduction

The multi-armed bandit (MAB) problem is one of the most fundamental problems in online machine learning, with diverse applications ranging from pricing and online advertising to clinical trails. Over the past several decades, it has been a very active research area spanning different disciplines, including computer science, economics, operations research, and statistics.

In a traditional multi-armed bandit problem, the learner (i.e., decision-maker) is allowed to switch freely between actions, and an effective learning policy may incur frequent switching — indeed, the learner's task is to balance the exploration-exploitation trade-off, and both exploration (i.e., acquiring new information) and exploitation (i.e., optimizing decisions based on up-to-date information) require switching. However, in many real-world scenarios, it is costly to switch between different alternatives, and a learning policy with limited switching behavior is preferred. The learner thus has to consider the cost of switching in her learning task.

In this paper, we introduce the *Bandits with Switching Constraints* (BwSC) problem. We note that most previous research in multi-armed bandits has modeled the switching cost as a penalty in the learner's objective, and hence the learner's switching behavior is a complete output of the learning algorithm. However, in many real-world applications, there are strict limits on the learner's switching behavior, which should be modeled as a *hard constraint*, and hence the learner's allowable level of switching is an input to the algorithm. In addition, while most prior research assumes specific structures on switching costs (e.g.,

unit or homogeneous costs), in reality, switching between different pairs of actions may incur heterogeneous costs that do not follow any parametric form. These gaps motivate us to propose the BwSC framework, which includes a hard constraint imposed on the total switching cost.

In addition to its strong modeling power and practical significance, the BwSC problem is theoretically important, as it is a natural framework to study the fundamental trade-off between the best achievable regret rate and the maximum incurred switching cost in the classical multi-armed bandit problem. In particular, it enables characterizing important switching patterns associated with any effective exploration-exploitation policies. Thus, the study of the BwSC problem leads to a series of new results for the classical multi-armed bandit problem.

### 6.1.1 Motivating Examples

The BwSC framework has numerous applications, including dynamic pricing, online assortment optimization, online marketplaces, clinical trails, labor markets, supply chain management, etc. We describe some representative examples below.

**Dynamic pricing with demand learning.** Dynamic pricing with demand learning has proven its effectiveness in revenue management (den Boer 2015). However, it is well known that in practice, sellers often face business constraints that prevent them from conducting extensive price experimentation and making frequent price changes; see Cheung et al. (2017), Chen and Chao (2019) and Chen et al. (2020) for discussions of multiple practical reasons. The seller's sequential decision-making problem can be modeled as a BwSC problem, where changing from each price to another price incurs some cost, and there is a limit on the total cost incurred by price changes. In particular, under different switching cost structures (which can be flexibly specified in BwSC), the total cost to be limited can have various practical interpretations: e.g., the total number of price changes, the total distance of price movements (Koren et al. 2017), or the total number of price increases (which is relevant to sellers who prefer *markdown pricing* (Jia et al. 2021)).

**Promotion and assortment strategies in retail and financial services.** Similar to the example of dynamic pricing, many retailers and financial service providers have started to use online learning techniques to dynamically adjust their promotion strategies (e.g., deals, referral programs, sign-up bonus offers) or product assortments based on sequentially

collected data. In many scenarios, frequent changes of public offerings not only increase operational and marketing costs (e.g., inventory and advertising costs) but also lead to customer dissatisfaction and negative public image (Simchi-Levi et al. 2008). The sequential promotion planning problem, and the sequential assortment planning problem (with a *fixed* number of assortment candidates[26]), can both be modeled as BwSC. The application of assortment planning also provides an example of general switching costs: if the seller wants to limit the total number of *product changes* rather than assortment changes, then a fine-grained definition of the switching costs between assortments is required (Dong et al. 2020).

**Sequential experiments in online marketplaces.** Consider an online e-commerce platform (e.g., Uber, Airbnb) choosing a mechanism (e.g., a surge pricing algorithm, a listing ranking rule) among several alternatives. It is common practice for platforms to conduct sequential experiments of mechanisms using bandit approaches to optimize long-term revenue. However, frequent changes of mechanisms may be highly undesirable for a marketplace, because not only the platform (e.g., Uber) but also the *market participants* (e.g., drivers and riders) may suffer from switching costs: each time the platform announces a new mechanism, the market participants will make efforts to adapt to the new mechanism (e.g., if Uber announces that trips completed during select hours each day earn extra rewards, then a driver may be incentivized to change his work schedule (Scheiber 2017)); as a result, market participants will get annoyed when they find that the mechanism changes frequently (which means that they have to re-develop their business strategies frequently (Kerr 2015)). Therefore, platforms usually have to limit their number of mechanism changes in sequential experiments.

### 6.1.2 Problem Formulation

We now introduce our model. Consider a (stochastic) $K$-armed bandit problem where a learner chooses actions from a fixed set $[K] = \{1, \dots, K\}$. There is a total of $T$ rounds ($T \geq K$). In each round $t \in [T]$, the learner first chooses an action $a_t \in [K]$, then observes and collects a reward $X^t(a_t) \in \mathbb{R}$. For each action $k \in [K]$, the reward of action $k$ is i.i.d. drawn from an (unknown) distribution $\mathcal{D}_k$ with (unknown) expected value $\mu_k$. We

[26]Here we refer to the setting that a retailer chooses one assortment from a few assortment candidates (which captures many retailers' practice under complicated business constraints). A different setting in literature allows one to consider exponentially many assortment candidates under an MNL choice model, see Agrawal et al. (2019), Dong et al. (2020).

assume that the distributions $\mathcal{D}_1, \ldots, \mathcal{D}_K$ are standardized sub-Gaussian.[27] Without loss of generality, we assume $\sup_{i,j\in[K]} |\mu_i - \mu_j| \in [0, 1]$.

In the BwSC problem, the learner incurs a switching cost $c_{i,j} \geq 0$ each time she switches from action $i$ to action $j$ $(i, j \in [K])$.[28] In particular, $c_{i,i} = 0$ for $i \in [K]$. There is a pre-specified *switching budget* $S \geq 0$ representing the maximum amount of switching costs that the learner can incur in total. Once the total switching cost exceeds the switching budget $S$, the learner cannot switch her actions any more. The learner's goal is to maximize the expected total reward over $T$ rounds.

### Admissible Policies

Let $\pi$ denote the learner's (non-anticipating) learning policy; specifically, $\pi$ is a sequence $(\pi_1, \ldots, \pi_T)$, where $\pi_t$ establishes a probability kernel acting from the (measurable) space of historical actions and observations before round $t$ to the (measurable) space of actions at round $t$. Let $a_t$ denote the (random) action selected by policy $\pi$ at round $t$, and $X^t(a_t)$ denote the (random) reward observed by policy $\pi$ at round $t$ (note that both $a_t$ and $X^t(a_t)$ depend on the underlying distributions $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_K)$). Let $\mathbb{P}^\pi_{\mathcal{D}}$ denote the law of the random variables $\big(a_1, X^t(a_1)\big), \ldots, \big(a_T, X^T(a_T)\big)$, and let $\mathbb{E}^\pi_{\mathcal{D}}[\cdot]$ be the associated expectation operator.

According to our model, we only need to restrict our attention to the *$S$-switching-budget* policies, which take $S$, $K$ and $T$ as input and are defined below.[29]

**Definition 6.1.** *A policy $\pi$ is said to be an $S$-switching-budget policy if for all $\mathcal{D}$,*

$$\mathbb{P}^\pi_{\mathcal{D}}\left[\sum_{t=1}^{T-1} c_{a_t,a_{t+1}} \leq S\right] = 1.$$

Let $\Pi_S$ denote the set of all $S$-switching-budget policies, which is also the admissible policy class of the BwSC problem.

[27] This is a standard assumption in the stochastic bandit literature. Note that the class of sub-Gaussian distributions is sufficiently wide as it contains Gaussian, Bernoulli and all bounded distributions.

[28] We allow $c_{i,j} = \infty$, which means that switching from $i$ to $j$ is prohibited. We also allow $c_{i,j} \neq c_{j,i}$, which means that the switching costs are asymmetric.

[29] Even if the learner does not intentionally pick an $S$-switching-budget policy at the beginning, the switching constraint will force the learner's policy to be an $S$-switching-budget policy.

**Regret and Optimal Regret**

The performance of a learning policy is measured against a clairvoyant policy that maximizes the expected total reward given foreknowledge of the *environment* (i.e., underlying distributions) $\mathcal{D}$. Let $k^\star = \arg\max_{k\in[K]} \mu_k$ and $\mu^\star = \max_{k\in[K]} \mu_k$. If a clairvoyant knows $\mathcal{D}$ in advance, then she would choose the "optimal" action $k^\star$ for every round and her expected total reward would be $T\mu^\star$. We define the *regret* of policy $\pi$ as the worst-case difference between the expected performance of the optimal clairvoyant policy and the expected performance of policy $\pi$:

$$R^\pi(T) := \sup_{\mathcal{D}} \left\{ T\mu^\star - \mathbb{E}^\pi_{\mathcal{D}} \left[ \sum_{t=1}^{T} X^t(a_t) \right] \right\} = \sup_{\mathcal{D}} \left\{ T\mu^\star - \mathbb{E}^\pi_{\mathcal{D}} \left[ \sum_{t=1}^{T} \mu_{a_t} \right] \right\},$$

which is a non-negative function of the policy $\pi$, the number of actions $K$, and the *horizon* (i.e., number of rounds) $T$; occasionally, we will use the notation $R^\pi(K,T)$ to highlight its dependence on $K$. Furthermore, the *optimal* (i.e., minimax) regret of BwSC is defined as

$$R^\star_S(T) := \inf_{\pi\in\Pi_S} R^\pi(T),$$

which is a non-negative function of the switching budget $S$, the number of actions $K$, and the horizon $T$; occasionally, we will use the notation $R^\star_S(K,T)$ to highlight its dependence on $K$. Note that the optimal regret is an intrinsic quantity that can help us to characterize the statistical complexity of the BwSC problem.

*Remark.* There are two notions of regret in the stochastic bandit literature. The $R^\pi(T)$ regret that we consider is called the *distribution-independent* (or *worst-case*) regret, as it does not depend on $\mathcal{D}$. On the other hand, one can also define the *distribution-dependent* (or *instance-dependent*) regret $R^\pi_{\mathcal{D}}(T) = T\mu^\star - \mathbb{E}^\pi_{\mathcal{D}} \left[ \sum_{t=1}^{T} \mu_{a_t} \right]$ that depends on $\mathcal{D}$. Unlike the classical MAB problem where there are policies that simultaneously achieve near-optimal bounds under both regret notions, in the BwSC problem, due to the limited switching budget, finding a policy that simultaneously achieves near-optimal bounds under both regret notions is usually impossible. Thus in the main body of the paper, we focus on the distribution-independent regret. However, in Appendix E.1, we extend our results to the distribution-dependent regret.

**Research Questions**

Two fundamental tasks in the study of bandits are: (i) to understand the *growth rate* of the optimal regret (i.e., "optimal regret rate") as $T$ grows, or as both $K$ and $T$ grow, and (ii) to design efficient algorithms that attain near-optimal regret. In this paper, we seek to address both of these challenges for BwSC. Moreover, motivated by the relationship between BwSC and MAB,[30] we seek to understand how the switching constraint fundamentally affects the statistical nature of bandits. Altogether, our central questions are:

1. What is the statistical complexity (i.e., optimal regret rate) of BwSC?
2. Can we design practical algorithms to attain the optimal regret rate?
3. How does the optimal regret rate of BwSC changes with respect to the switching budget $S$, and how is it affected by the structure of switching costs $(c_{i,j})$?

### 6.1.3 Main Results and Technical Highlights

The main contributions of this paper lie in fully addressing the above three research questions for BwSC under the unit switching cost structure, partially addressing the three questions for BwSC under the general switching cost structure, and discovering surprising "phase transition" behavior of the optimal regret (under both unit and general switching cost structures). We devise a series of efficient algorithms which attain sharp regret upper bounds, and introduce a highly non-trivial five-step method which provides matching (or nearly matching) lower bounds. As a by-product, we develop a new information-theoretic inequality, namely the *Generalized Reverse Fano-Type* inequality, which plays a critical role in our five-step method.

We summarize our main results and technical contributions as follows.

**Effective algorithms for the U-BwSC problem.** We first study the BwSC problem under the most fundamental switching cost structure — the unit switching cost structure: $c_{i,j} = 1$ for all $i \neq j$. The problem is referred to as the *unit-switching-cost* BwSC problem (or U-BwSC for short), and can be interpreted as "MAB with limited number of switches." As a preliminary attempt, we present a simple and intuitive algorithm, called LS-SE, which builds on the "batched elimination" framework recently developed by Perchet et al. (2016) and Gao

[30] Note that BwSC and MAB share the same definition of $R^{\pi}(T)$, and the only difference between BwSC and MAB is the existence of a switching constraint $\pi \in \Pi_S$, determined by $(c_{i,j}) \in \overline{\mathbb{R}}_{\geq 0}^{K \times K}$ and $S \in \overline{\mathbb{R}}_{\geq 0}$ (when $S = \infty$, BwSC degenerates to MAB).

et al. (2019), and ensures the following regret:

$$\widetilde{\mathcal{O}}(1) \cdot K^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}}, \tag{6.1}$$

where $q(S,K) := \left\lfloor \frac{S-1}{K-1} \right\rfloor$ is the *quotient* of the *Euclidian division* of $(S-1)$ by $(K-1)$.[31]

The LS-SE algorithm, though being very simple, has several drawbacks, including a potentially large waste of the switching budget, and overly low *adaptivity* (i.e., it learns from data in an overly infrequent manner; see Section 6.3.2). To overcome these drawbacks, we design a new algorithm, AdaLS, which builds on and improves upon LS-SE by (i) adopting a novel hybrid and randomized exploration strategy and (ii) deciding when to make switches in a more data-driven fashion. These two features enable AdaLS to make better use of the switching budget and enjoy higher adaptivity.

We show that AdaLS attains an improved regret bound of

$$\widetilde{\mathcal{O}}(1) \cdot \max\left\{ \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}},\ K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}} \right\}, \tag{6.2}$$

where $r(S,K) := (S-1)\%(K-1)$ is the *remainder* of the *Euclidean division* of $(S-1)$ by $(K-1)$.[32] Since $0 \le r(S,K) \le K-2$, the rate of (6.2) is at most $K^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}}$, which is the same as (6.1), and at least $\max\left\{ K^{-1} T^{\frac{1}{2-2^{-q(S,K)}}}, K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}} \right\}$, which can be much smaller than (6.1) when $K$ is large, as the first term in "max" has a sharp $K^{-1}$ factor and the second term in "max" has a smaller order of $T$. This implies that AdaLS reduces the regret of LS-SE by a multiplicative factor of at least $\Omega(1)$ and at most $\mathcal{O}(K^{3/2})$. See Table 6.1 for detailed illustrations and comparisons of regret bounds (6.1) and (6.2), and Section 6.3.2 for more explanations.

**Tight lower bound for the U-BwSC problem.** The AdaLS algorithm, though being a significantly refined version of the LS-SE algorithm, still seems to leave plenty of room for improvement. In particular, it only improves the regret rate when $K$ is permitted to grow with $T$, failing to directly improve the regret's dependence on the most important parameter $T$ when $K = \widetilde{\mathcal{O}}(1)$. Several challenging questions remain open: Is it possible to directly improve the regret in terms of $T$? Can the dependence on $K$ be further improved? What is

[31] See https://en.wikipedia.org/wiki/Euclidean_division for a definition of the Euclidian division. We use $\lfloor \cdot \rfloor$ to denote the floor function; see Section 6.1.4 for details.

[32] We use % to denote the modulo operation; see Section 6.1.4 for details.

the fundamental limit of the U-BwSC problem? We settle these questions by establishing a strong (and quite surprising) information-theoretic lower bound that directly match the upper bound (6.2) for any $S$, any $K$, and any $T$ — specifically, we show that *no* admissible policy can avoid a regret lower bound of

$$\widetilde{\Omega}(1)\cdot\max\left\{\frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K}T^{\frac{1}{2-2^{-q(S,K)}}},\ K^{1-\frac{1}{2-2^{-q(S,K)-1}}}T^{\frac{1}{2-2^{-q(S,K)-1}}}\right\},\quad(6.3)$$

which implies that AdaLS is optimal up to logarithmic factors. The proof of the lower bound is highly non-trivial. The methodological contributions in the lower bound proof will be elaborated shortly.

The tight lower bound proved for U-BwSC also motivate new insights about the classical MAB. In particular, it implies that $\Omega(K\log\log T)$ switches are *necessary* to achieve $\widetilde{\mathcal{O}}(\sqrt{KT})$ (near-optimal) regret in MAB, which appears to be a fundamental yet new result.

**Phase transitions associated with the optimal regret.** Combining (6.2) and (6.3), we completely characterize the optimal regret of the U-BwSC problem. The characterization reveals surprising findings: when $K=\widetilde{\mathcal{O}}(1)$, the quotient function $q(S,K)$ (as a floor function) uniquely determines the optimal regret rate; when $K$ grows with $T$ in a non-negligible way, the remainder function $r(S,K)$ also affects the optimal regret rate through the dependence on $K$, but only when $r(S,K)$ is large enough such that $r(S,K)=K-o(K)$. To the best of our knowledge, this is the first example of an online learning setting where (i) a floor function naturally arises in the exponent of $T$ in the optimal regret, and (ii) the optimal regret exhibits a non-conventional growth rate which is surprisingly characterized by an Euclidean division.

As a consequence of these findings, we discover surprising *phase transitions* regarding how the optimal regret rate changes with respect to the switching budget $S$, which can be summarized by the following two cases: when $K$ is fixed (and $T$ grows), there are equal-length phases defined by $S$, where the optimal regret rate remains the same (up to logarithmic factors) within each phase and exhibits abrupt changes between phases; when $K$ grows with $T$ in a non-negligible manner, such abrupt changes become subtler and may disappear, but a generalized form of phase transitions involving the "budget-to-arm ratio" (BAR) still exist. We will provide a rigorous and detailed treatments of phase transitions in Section 6.3.4.

**Extensions to general switching cost structures.** We extend the results obtained in

the U-BwSC problem to the general BwSC problem. Specifically, we study two types of general switching cost structures (with one being symmetric and the other being asymmetric): (i) the *general symmetric switching cost* structure (corresponding to the G-BwSC problem), where $c_{i,j} = c_{j,i}$ $(i \neq j)$ can be any non-negative real number; and (ii) the *departure cost* structure (corresponding to the D-BwSC problem), where $c_{i,j} = c_i$ for all $j \neq i$ and $c_i$ can be any non-negative real number, i.e., the switching cost between any pair of actions only depends on the action that the learner departs from. For both G-BwSC and D-BwSC, we design efficient algorithms, and prove corresponding lower bounds on regret. Under the condition of $K = \widetilde{\mathcal{O}}(1)$, we show that our regret upper and lower bounds almost match for G-BwSC, and exactly match for D-BwSC (both in terms of the dependence on $T$); the optimal regret again exhibits *phase transitions*. Our results in this part make conceptual contributions by revealing an interesting connection between bandit problems and *graph traversal* problems.

**Methodological contributions in the lower bound analysis.** As we mentioned, the proof of the lower bound (6.3) requires significant technical effort, and contains several technical highlights of this paper. In particular, to show that the quotient function $q(S, K)$ *necessarily* appears in the exponent of $T$ and the remainder function $r(S, K)$ *necessarily* affects the order of $K$ through the $(K - r(S, K))$ term, we develop a host of new techniques, largely from first principles, to characterize how the switching constraint affects the learning dynamics of an *arbitrary* admissible policy, and how concrete classes of certain learning dynamics (represented by *risky events*) lead to fundamental performance limits. These techniques are integrated in a five-step proof program called RECAP.

As an aside, we establish a new information-theoretic inequality, namely the *Generalized Reverse Fano-Type* (GRF) inequality, which is of independent interest. We believe that the GRF inequality, together with the ideas and techniques arising in the RECAP method, can find broader applications in learning theory and statistics. We refer the interested reader to Appendix E.4 for a detailed introduction of the GRF inequality and the RECAP method.

### Comparison with Results on "Batched Bandits"

The U-BwSC problem is closely related to the "batched (multi-armed) bandit" problem (Perchet et al. 2016, Gao et al. 2019). Here, we explain the major differences between our results and the results on batched bandits.

The $M$-batched bandit problem is defined as follows: given a classical MAB, assumes that

the learner must split her learning process into $M$ batches and is only able to observe data (i.e., realized rewards) from a given batch after the entire batch is completed. This implies that all actions within a batch are determined at the beginning of this batch. Here $M$ can be viewed as a quantity measuring the learner's *adaptivity*, i.e., her ability to learn from her data and adapt to the environment. An $M$-batch policy is defined as a policy that only observes the available data for $M-1$ times through the entire horizon. Perchet et al. (2016) study the above problem in the two-armed case. They propose an $M$-batch policy with $\widetilde{\Theta}\Big(T^{\frac{1}{1-2^{1-M}}}\Big)$ regret, and show that no $M$-batch policy can attain a better regret rate under the "static grid" restriction which requires the policy to *pre-determine* the batch sizes before the learning process. Gao et al. (2019) extend their algorithm and results to the general $K$-armed case, and show that even without the "static grid" restriction (i.e., even when the batch sizes can be *adaptively chosen* in a batch-by-batch manner), no $M$-batch policy can attain a better regret rate. The statistical complexity of the batched bandit problem is thus completely characterized.

The batched bandit problem and the U-BwSC problem are two closely related but fundamentally different problems: while the batched bandit problem explicitly limits the number of times of making observations (i.e., adaptivity), the U-BwSC problem only limits the number of times of action changes, and (importantly!) *allows unlimited number of times of making observations*. As we shall see in Section 6.3.5, the U-BwSC model can be seen as a strict relaxation of the batched bandit model, in the sense that the admissible policy class of U-BwSC is much richer and contains more efficient policies. This difference allows U-BwSC to enjoy a fundamentally smaller optimal regret rate when $K$ is large; see Section 6.3.5 for a detailed discussion on the relationship and difference of the two problems.

We note that existing results and techniques of batched bandits cannot provide a satisfying solution to the U-BwSC problem, neither in terms of designing rate-optimal algorithms nor in terms of establishing fundamental limits. While it is relatively easier to obtain an $S$-switching-budget policy by modifying a $(q(S,K)+1)$-batch policy (see Section 6.3.1 and the LS-SE algorithm), such approach suffers from several drawbacks and is generally sub-optimal when $K$ is large, as an $S$-switching-budget policy can utilize data more frequently than a $(q(S,K)+1)$-batch policy and achieve better regret (see Section 6.3.2). As a result, we have to develop new algorithmic ideas to design a more advanced algorithm (the AdaLS algorithm) to achieve the optimal regret rate of the U-BwSC problem.

More importantly, since U-BwSC is more relaxed than the batched bandit problem, a regret lower bound for the batched bandit problem cannot imply a regret lower bound for U-BwSC. Therefore, we need to establish new lower bounds for U-BwSC, which also imply lower bounds for the batched bandit problem. In fact, from an information-theoretic perspective, when we aim for lower bounds, dealing with a switching constraint (which does not impose any constraint on the number of queries of information) is considerably more challenging than dealing with a batch constraint (which directly restricts the ability to inquire information): the challenge is that an $S$-switching-budget policy can continuously gain new information at every round; as a result, it may be difficult to establish sharp impossibility results for such a policy via standard information-theoretic arguments. We address this challenge by establishing the RECAP method from first principles. We remark that the lower bound results in our paper are strong. Indeed, existing lower bounds of Perchet et al. (2016) and Gao et al. (2019) can be seen as corollaries of our lower bound (6.3); see Section 6.3.5.

### 6.1.4 Organization and Notation

The rest of the paper is organized as follows. In Section 6.2, we review other related literature. In Section 6.3, we discuss the unit-switching-cost model. In Section 6.4, we discuss two general-switching-cost models. In Section 6.5, we numerically test our algorithms. We conclude the paper in Section 6.6.

Let $\mathbb{N}$ (resp. $\mathbb{N}_{>0}$) be the set of all non-negative (resp. positive) integers. For all $n_1, n_2 \in \mathbb{N}$ such that $n_1 \le n_2$, we use $[n_1]$ to denote the set $\{1, \ldots, n_1\}$, and use $[n_1 : n_2]$ (resp. $(n_1 : n_2]$) to denote the set $\{n_1, n_1+1, \ldots, n_2\}$ (resp. $\{n_1+1, \ldots, n_2\}$). For all $x \ge 0$, we use $\lfloor x \rfloor$ to denote the largest integer less than or equal to $x$. For ease of presentation, we define $\lfloor x \rfloor = 0$ for all $x < 0$. For all $m \in \mathbb{N}, n \in \mathbb{N}_{>0}$, we define $m \% n := m - n\lfloor m/n \rfloor$ (i.e., the remainder of the Euclidean division of $m$ by $n$). For all $m, n \in \mathbb{R}$, let $m \vee n := \max\{m, n\}$ and $m \wedge n := \min\{m, n\}$. Throughout the paper, we adopt non-asymptotic big-oh notation: for functions $f, g : \mathcal{X} \to \mathbb{R}_{\ge 0}$, we write $f = \mathcal{O}(g)$ (resp. $f = \Omega(g)$) if there exists some constant $C > 0$ such that $f(x) \le Cg(x)$ (resp. $f(x) \ge Cg(x)$) for all $x \in \mathcal{X}$. We write $f = \widetilde{\mathcal{O}}(g)$ if $f = \mathcal{O}(g \cdot \mathrm{polylog}(T))$, $f = \widetilde{\Omega}(g)$ if $f = \Omega(g/\mathrm{polylog}(T))$, and $f = \widetilde{\Theta}(g)$ if $f = \widetilde{\mathcal{O}}(g)$ and $f = \widetilde{\Omega}(g)$. We use $f \asymp g$ as shorthand for $f = \widetilde{\Theta}(g)$. We write $f = o(g)$ if $\lim_{x\to\infty} f(x)/g(x) = 0$.

## 6.2 Related Literature

### 6.2.1 Stochastic MAB with Switching Costs

The stochastic MAB problem has been extensively studied for more than fifty years. It is well known that the optimal distribution-dependent regret is $\Theta(K \log T)$ (Lai and Robbins 1985) and the optimal distribution-independent regret is $\Theta(\sqrt{KT})$ (Auer et al. 2002a). We point out two excellent surveys Lattimore and Szepesvári (2020) and Slivkins (2019) for more reference about this topic.

There is rich literature focusing on stochastic MAB with switching costs.[33] Most of the papers model the switching cost as a penalty in the learner's objective, i.e., they measure a policy's regret and incurred switching cost using the same metric and the objective is to minimize the sum of these two terms[34] (e.g., Agrawal et al. 1988, 1990, Brezzi and Lai 2002, Cesa-Bianchi et al. 2013; there are other variations with discounted rewards: Banks and Sundaram 1994, Asawa and Teneketzis 1996, Bergemann and Välimäki 2001, see Jun 2004 for a survey). Though this conventional "switching penalty" model has attracted significant research interest in the past, it has two limitations. First, under this model, the learner's total switching cost is an output determined by the algorithm. However, in many real-world applications, there are strict limits on the learner's total switching cost, which should be modeled as a *hard constraint*, and hence the learner's switching budget should be an input that helps determine the algorithm. In particular, while the algorithm in Cesa-Bianchi et al. (2013) developed for the "switching penalty" model can achieve $\widetilde{\mathcal{O}}(\sqrt{KT})$ (near-optimal) regret with $\mathcal{O}(K \log \log T)$ switches, if the learner wants a policy that always incurs finite switching cost independent of $T$, then prior literature does not provide an answer. Second, the "switching penalty" model has fundamental weakness in studying the trade-off between the regret rate and the incurred switching cost in stochastic MAB — since the $\log \log T$-type bound on the incurred switching cost of a policy is negligible compared with the $\sqrt{T}$-type bound on its best achievable regret, when adding the two terms up, the term associated with incurred switching cost is always dominated by the regret (in terms of the growth rate), thus no trade-off can be identified. As a result, to the best of our knowledge, prior literature

[33] It is worth noting that there is also a vast literature on *adversarial* MAB with switching costs. In particular, Dekel et al. (2014) prove a striking $\widetilde{\Omega}(K^{1/3}T^{2/3})$ lower bound for this problem, indicating a fundamental difference between the roles of switching costs in stochastic MAB and in adversarial MAB.

[34] If the "total regret plus total switching cost" is guaranteed to be small, then the "total reward minus total switching cost" is guaranteed to be large.

has not characterized the fundamental trade-off between the regret rate and the incurred switching cost in stochastic MAB.

The BwSC framework addresses the issues associated with the "switching penalty" model in several ways. First, it introduces a hard constraint on the total switching cost, enabling us to design good policies that guarantee limited switching cost. While $\mathcal{O}(K \log\log T)$ switches has proven to be sufficient for a learning policy to achieve near-optimal regret in MAB, in BwSC, we are mostly interested in the setting of horizon-independent or $o(K \log\log T)$ switching budget, which is highly relevant in practice. Second, by focusing on rewards in the objective function and incurred switching cost in the switching constraint, the BwSC framework enables the characterization of the fundamental trade-off between regret and maximum incurred switching cost in MAB. Third, while most prior research assumes specific structures on switching costs (e.g., unit or homogeneous costs), BwSC allows general switching costs, which makes it a powerful modeling framework.

### 6.2.2 Online Learning with Limited Switches.

This paper is not the first one to study online learning problems with limited switches.[35] In Indeed, a few authors have realized the practical significance of limited switching budget. For example, Cheung et al. (2017) consider a dynamic pricing model where the demand function is unknown but belongs to a known finite set, and a pricing policy is allowed to make at most $m$ price changes. Their constraint on the total number of price changes is motivated by collaboration with Groupon, a major e-commerce marketplace in North America. In such an environment, Groupon limits the number of price changes, either because of implementation constraints, or for fear of confusing customers and receiving negative customer feedback. They propose a pricing policy that guarantees $\mathcal{O}(\log^{(m)} T)$ (or $m$ iterations of the logarithm) regret with at most $m$ price changes, and report that in a field experiment, this pricing policy with a single price change increases revenue and market share significantly. Chen and Chao (2019) study the joint pricing and inventory control problem with unknown demand and limited price changes. Assuming that the demand function is drawn from a parametric

[35] Here, "online learning with limited switches" refers to online learning problems with constraints on the *learner*'s number of changes of *decisions*. There is another line of research studying (non-stationary) online learning problems with constraints on *nature*'s number of changes of *environments* (e.g., Herbster and Warmuth 1998, Jun et al. 2017). Though this line of research focuses on completely different learning challenges (i.e., non-stationarity), it is conceptually relevant as it shares the same flavor of characterizing the regret using a "budget for making changes."

class of functions, they develop a finite-price-change policy based on maximum likelihood estimation (MLE) that achieves the optimal regret rate. Chen et al. (2020) also study the dynamic pricing and inventory control problem with limited price changes, but in a more challenging setting with *censored* demand. They prove matching upper and lower bounds on the optimal regret, and devise an MLE-based policy to achieve the optimal regret rate.

We note that all of Cheung et al. (2017), Chen and Chao (2019), Chen et al. (2020) focus on specific revenue management problems, and their results rely on certain assumptions that are specialized to their models. The BwSC model in our paper has a different flavor, in the sense that it is very generic and relies on very few assumptions, but can only handle discrete actions. The results and techniques of our paper are thus very different from the above papers. Also, the switching constraint in the BwSC problem is more general than the price-change constraints in previous models.

In the Bayesian bandit setting, Guha and Munagala (2013) (see also Guha and Munagala 2009) study the "bandits with metric switching costs" problem that allows a constraint involving metric switching costs. Using competitive ratio as the performance metric and assuming Bayesian priors, they develop a 4-approximation algorithm for the problem. The competitive ratio is measured against an optimal online policy that does not know the true distributions. As pointed out by the authors, the optimal online policy can be directly determined by a dynamic program, so the main challenge in their model is a computational one. Our work is different, as we are using regret as our performance metric, and we are competing with an optimal clairvoyant policy that knows the true distributions — a much stronger benchmark. Our problem thus involves both statistical and computational challenges. In fact, the algorithm in Guha and Munagala (2013) cannot avoid a linear regret when applied to the BwSC problem.

In the adversarial bandit setting, Altschuler and Talwar (2021) study the adversarial MAB problem with limited number of switches, which can be viewed as an adversarial counterpart of the U-BwSC problem. For any policy that makes no more than $S \leq T$ switches, they prove that the optimal regret is $\widetilde{\Theta}(T\sqrt{K}/\sqrt{S})$. Since we are considering a different setting from them (our problem is stochastic while their problem is adversarial), the results and techniques in our paper are fundamentally different from their paper. In particular, while any fixed-switching-budget policy cannot avoid linear regret in the adversarial setting, in the stochastic setting, a fixed number of switches may already guarantee sublinear regret

(assuming $K$ is fixed). Moreover, while the optimal regret rate in Altschuler and Talwar (2021) decreases smoothly as $S$ increases from 0 to $T$, in the stochastic setting, we identify surprising behavior of the optimal regret rate as $S$ increases from 0 to $\Theta(K \log\log T)$, which, to the best of our knowledge, has not been identified in the bandit literature before.

## 6.3 Unit Switching Costs

In this section, we consider the unit-switching-cost BwSC problem (abbreviated as U-BwSC), where $c_{i,j} = 1$ for all $i \neq j$. In this case, since every switch incurs a unit cost, the switching budget $S$ can be interpreted as the maximum number of switches that the learner can make in total. Without loss of generality, in this section we assume that $S$ is a non-negative integer, and refer to an $S$-switching-budget policy as an $S$-switch policy. Note that the U-BwSC problem can be simply interpreted as "MAB with limited number of switches."

The section is organized as follows. In Section 6.3.1, we present a simple and intuitive algorithm and an initial upper bound on regret. In Section 6.3.2, we propose a refined algorithm that attains an improved upper bound on regret. In Section 6.3.3, we establish a matching lower bound on regret, indicating that the algorithm in Section 6.3.2 is rate-optimal. In Section 6.3.4, we discuss several surprising findings in U-BwSC, namely "phase transitions" of the optimal regret. In Section 6.3.5, we discuss the relationship between limited switches and limited adaptivity in bandit problems.

**Algorithmic notation.** We adopt the following notation to facilitate the descriptions of our algorithms. For any execution of an algorithm, for any action $i \in [K]$, for any round $t \in [T]$, let $N_i(t)$ denote the number of plays of action $i$ up to round $t$ (inclusive), $\bar{\mu}_i(t)$ denote the average *observed* reward of action $i$ up to round $t$ (for notational convenience, we define $\bar{\mu}_i(0) = -\infty$), and

$$\mathsf{UCB}_i(t) := \bar{\mu}_i(t) + \sqrt{\frac{6 \log T}{N_i(t)}}, \qquad \mathsf{LCB}_i(t) := \bar{\mu}_i(t) - \sqrt{\frac{6 \log T}{N_i(t)}} \tag{6.4}$$

denote the *upper confidence bound* and *lower confidence bound* of action $i$ up to round $t$, respectively.

### 6.3.1 The LS-SE Algorithm

As a preliminary attempt, we provide a simple and intuitive algorithm for U-BwSC, namely the *Limited-Switch Successive Elimination* (LS-SE) algorithm; see Algorithm 6.1 for details.[36] The algorithm builds on the "batched elimination" framework recently developed by Perchet et al. (2016) and Gao et al. (2019) for the batched bandit problem, which splits the $T$ rounds into a given number of pre-determined batches and successively eliminates "poorly performing" actions (based on confidence bounds) in a batch-by-batch manner; the key ingredient of this framework is a delicate batch schedule (i.e., splitting rule) that strikes a balance between exploration and exploitation given a limited number of batches (cf. Section 4 of Perchet et al. 2016). Since we are studying a different problem, directly applying a batched bandit algorithm to the U-BwSC problem may not work — in batched bandits, the number of batches is a given constraint, while in U-BwSC, the switching budget is the given constraint. We thus add two ingredients into the LS-SE algorithm: (i) an index $q(S,K)$ suggesting how many batches should be used to split the entire horizon, and (ii) a switching rule ensuring that the total number of switches across all $K$ actions cannot exceed the budget $S$.

**Intuition about LS-SE.** The algorithm divides the $T$ rounds into $q(S,K)+1 = \left\lfloor \frac{S-1}{K-1} \right\rfloor + 1$ epochs in advance, where an epoch corresponds to a batch in batched bandits. Note that there is no adaptivity *within* each epoch: decisions are determined at the beginning of the epoch and do not depend on the rewards observed in this epoch. The epoch schedule follows the batch schedules given by Perchet et al. (2016) and Gao et al. (2019), with slight differences in the dependence on $K$.[37] Combined with the celebrated *successive elimination* strategy (see Line 9) in bandits, this schedule ensures that exploration and exploitation are balanced and the (worst-case) regret incurred during each epoch is at the same level (more specifically, the schedule ensures that $t_1 \asymp \frac{t_2}{\sqrt{t_1/K}} \asymp \cdots \asymp \frac{T}{\sqrt{t_{q(S,K)}/K}}$, where $\frac{t_l}{\sqrt{t_{l-1}/K}}$ approximately controls the worst-case regret incurred during epoch $l$). In addition, our two new ingredients (the index and the switching rule) guarantee the following properties:

[36] Note that in Line 4 and Line 7 of Algorithm 6.1, $\frac{t_l - t_{l-1}}{|A_l|}$ might be fractional. For ease of presentation, we defer the rigorous treatment of such (minor) rounding issues to Appendix E.2. The same principle applies to Algorithm 6.3.

[37] The batch schedule of Perchet et al. (2016) does not involve $K$ because they only study the two-armed case. The batch schedule of Gao et al. (2019) does not involve $K$ because they allow $\sup_{i,j\in[K]}|\mu_i - \mu_j| \in [0, \sqrt{K}]$. Our epoch schedule is optimized for the (usual) setting of $\sup_{i,j\in[K]}|\mu_i - \mu_j| \in [0,1]$ and leads to better regret in this setting.

**Algorithm 6.1** Limited-Switch Successive Elimination (LS-SE)

**Input:** Switching budget $S$, number of actions $K$, horizon $T$.

**Initialization:** Compute $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$. Divide the entire time horizon $T$ into $q(S,K)+1$ epochs: $(t_0 : t_1], (t_1 : t_2], \ldots, (t_{q(S,K)} : t_{q(S,K)+1}]$, where the endpoints are defined by $t_0 = 0$ and

$$t_j = \left\lfloor K^{1-\frac{2-2^{-(j-1)}}{2-2^{-q(S,K)}}} T^{\frac{2-2^{-(j-1)}}{2-2^{-q(S,K)}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+1.$$

Let $A_1 = [K]$. Let $a_0$ be a random action in $[K]$.

**Policy:**

**for** $l = 1, \ldots, q(S,K)$ **do**
  **if** $a_{t_{l-1}} \in A_l$ **then**
    **for** $i = a_{t_{l-1}}$ and then $i \in A_l \setminus \{a_{t_{l-1}}\}$ **do** ▷ starting from $i = a_{t_{l-1}}$ is critical
      Choose action $i$ for $\frac{t_l - t_{l-1}}{|A_l|}$ *consecutive* rounds.
  **else**
    **for** $i \in A_l$ **do**
      Choose action $i$ for $\frac{t_l - t_{l-1}}{|A_l|}$ *consecutive* rounds.
  Mark the last chosen action as $a_{t_l}$.
  Elimination: compute $\mathsf{UCB}_i(t_l)$ and $\mathsf{LCB}_i(t_l)$ for all $i \in A_l$ and let ▷ learn from data

$$A_{l+1} = \left\{ i \in A_l \mid \mathsf{UCB}_i(t_l) \geq \max_{j \in A_l} \mathsf{LCB}_j(t_l) \right\}.$$

For $l = q(S,K)+1$, find an action $i \in A_l$ that maximizes $\bar{\mu}_i(t_{l-1})$. Keep choosing this action until round $T$.

- Limited switches within each epoch: In epoch $l$, only $|A_l| - 1 \leq K - 1$ switches happen.
- At most one switch between two consecutive epochs: If the last action chosen in epoch $l$ remains in $A_{l+1}$ ($l < q(S,K)$), then it will be the first action chosen in epoch $l+1$, and no switch occurs between these two epochs. If the last action chosen in epoch $l$ is eliminated from $A_{l+1}$, then epoch $l+1$ starts from another action in $A_{l+1}$, and one switch occurs between these two epochs.
- No switch within the last epoch: In the last epoch, only the empirical best action is chosen.
- At most $S$ switches in $T$ rounds: by combining the above three properties with $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$, one can show that the total number of switches is at most $q(S,K)(K-1)+1 \leq S$.

We show that LS-SE is indeed an $S$-switch policy, and ensures the following upper bound on regret. The proof is standard[38] and deferred to Appendix E.7.

[38] The regret analysis of LS-SE is similar to the analysis of Gao et al. (2019); we present it for completeness. A difference is that we obtain slightly better dependence on $K$ under the condition $\sup_{i,j \in [K]} |\mu_i - \mu_j| \in [0,1]$.

**Proposition 6.1.** *Let $\pi$ be the LS-SE policy, then $\pi \in \Pi_S$. There exists an absolute constant $C \geq 0$ such that for all $K > 1$, $S \geq 0$ and $T \geq K$,*

$$R^{\pi}(K,T) \leq C(\log K \log T) K^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}},$$

*where $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$.*

*Remark.* Proposition 6.1 implies that for the classical MAB problem, $\mathcal{O}(K \log \log T)$ times of switches are sufficient for a learner to achieve the optimal $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret, which recovers a well-known result of Cesa-Bianchi et al. (2013) (see also Perchet et al. 2016, Gao et al. 2019).

### 6.3.2 The AdaLS Algorithm

The LS-SE algorithm, though being very simple, has several drawbacks that may degrade its performance. Specifically:

- The LS-SE policy does not make full use of its switching budget. Consider the case of $S = 2K - 2$. Since $q(2K-2, K) = \left\lfloor \frac{2K-3}{K-1} \right\rfloor = 1 = q(K,K)$, the LS-SE policy will just run as if it could only make $K$ switches, despite the fact that it can actually make $2K-2$ switches — in this case, nearly half of the switching budget will never be used. Intuitively, an effective learning policy should make full use of its switching budget. It seems that by tracking and allocating the switching budget in a more careful way, one can achieve lower regret.

- The LS-SE policy has (unnecessarily) low adaptivity. Note that the LS-SE policy is a batched policy that utilizes data in a very restrictive way: it only learns from data at the end of each epoch, for at most $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$ times. For example, consider the case of $S = 2K - 2$. The LS-SE policy will observe the data only once throughout the entire horizon. This is a waste of a policy's information acquisition ability in BwSC, where the learner is more flexible than in batched bandits, and can observe data at *every* round. Intuitively, data should be utilized to save switches and reduce regret, and one would expect that an effective policy will have a higher degree of *adaptivity*, that is, it should learn from the available data and adapt to the environment more frequently than LS-SE.

To overcome the above drawbacks, we design a new algorithm, namely the *Adaptive Limited-Switch* (AdaLS) algorithm; see Algorithm 6.2 for details. AdaLS builds on and improves upon LS-SE by (i) adopting a novel hybrid and randomized exploration strategy and (ii) deciding when to make switches in a more data-driven fashion. These two features enable AdaLS to make better use of its switching budget and enjoy higher adaptivity (i.e., learn from data more frequently). We explain the key ideas of AdaLS below.

**Two key indices.** At initialization, AdaLS performs the Euclidian division of $(S-1)$ by $(K-1)$ and obtains two key indices: the quotient $q(S,K)$ and the remainder $r(S,K) \in \{0,\cdots,K-2\}$. While the quotient $q(S,K)$ is already used by LS-SE to determine the number of epochs (which ensures that LS-SE makes at most $q(S,K)(K-1)+1$ switches), the remainder $r(S,K) = S-1-q(S,K)(K-1)$ is a new index that reflects the "abandoned switching budget" of LS-SE, i.e., the amount of switching budget that will never be used by LS-SE. Intuitively, the larger $r(S,K)$ is, the more AdaLS can (hopefully) improve upon LS-SE by making better use of the switching budget.

**Random partition of the action set.** An important goal of AdaLS is to make better use of the switching budget when $r(S,K)$ is large. This is however a non-trivial task: since $r(S,K) \le K-2$, the additional switching budget does not allow for visiting all actions in $[K]$; that is, we may only conduct additional exploration for a *subset* of actions in $[K]$, and the selection of this subset requires careful consideration. We address this issue by utilizing the idea of *randomization*: at initialization, we randomly split $[K]$ into two subsets $A_1^{(1)}$ and $A_1^{(2)}$, with $|A_1^{(1)}| = K - \widehat{r}(S,K)$ and $|A_1^{(2)}| = \widehat{r}(S,K)$ (we will explain the configuration of $\widehat{r}(S,K)$ shortly). Then, in the execution of the policy, we treat the actions in $A_1^{(1)}$ and $A_1^{(2)}$ differently, allowing the actions in $A_1^{(2)}$ to be explored more frequently than the actions in $A_1^{(1)}$. Specifically, before the last switch (where we commit to a single action; see Line 16), we allow AdaLS to switch to each action in $A_1^{(1)}$ for at most $q(S,K)$ times, while allowing it to switch to each action in $A_1^{(2)}$ for at most $q(S,K)+1$ times (as a comparison, LS-SE switches to every action in $[K]$ for at most $q(S,K)$ times before the last switch). By letting $\widehat{r}(S,K) = \max\{r(S,K)+1-q(S,K),0\}$, we enable AdaLS to make good use of the switching budget while never going over it: if $\widehat{r}(S,K) > 0$ (i.e., $r(S,K)$ is large enough), then AdaLS will make up to $q(S,K)(K-\widehat{r}(S,K)) + (q(S,K)+1)\widehat{r}(S,K) - 1 + 1 = S$ switches; otherwise, AdaLS will behave similar to LS-SE and make up to $q(S,K)(K-1)+1 < S$ switches. See Appendix E.3 for an illustration of how AdaLS makes switches. We remark that it is crucial

**Algorithm 6.2** Adaptive Limited-Switch Policy (AdaLS)

**Input:** Switching budget $S$, number of actions $K$, horizon $T$, tuning parameter $\lambda = 1/2$.

**Initialization:** Compute $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$ and $r(S,K) = (S-1)\%(K-1)$. Define $\widehat{r}(S,K) = \max\{r(S,K)+1-q(S,K), 0\}$. Define $T_0^{(1)} = T_0^{(2)} = 0$, $t_0^{(1)} = t_0^{(2)} = 0$ and

$$t_j^{(1)} = \left\lfloor (K - \widehat{r}(S,K))^{1-\frac{2-2^{1-j}}{2-2^{-q(S,K)}}} T^{\frac{2-2^{1-j}}{2-2^{-q(S,K)}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+1,$$

$$t_j^{(2)} = \left\lfloor K^{1-\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} T^{\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+2.$$

Let $A_1 = [K]$. Let $A_1^{(2)}$ be a subset of $A_1$ obtained by uniformly sampling $\widehat{r}(S,K)$ actions from $A_1$ *without replacement* (thus $|A_1^{(2)}| = \widehat{r}(S,K)$). Let $A_1^{(1)} = A_1 \setminus A_1^{(2)}$. Let $a_0$ be a random action in $A_1^{(1)}$.

**Policy:**

1: **for** $l = 1, \ldots, q(S,K)$ **do**

2: Starting from an arbitrary action in $A_l^{(2)}$, choose each action in $A_l^{(2)}$ for $n_l^{(2)} = \lfloor \frac{t_l^{(2)} - T_{l-1}^{(1)}}{|A_l|} \rfloor$ *consecutive* rounds. Mark the last round as $T_l^{(2)}$.

3: **if** $a_{T_{l-1}^{(1)}} \in A_l^{(1)}$ **then**

4: **for** $i = a_{T_{l-1}^{(1)}}$ and then $i \in A_l^{(1)} \setminus \{a_{T_{l-1}^{(1)}}\}$ **do** ▷ starting from $i = a_{T_{l-1}^{(1)}}$ is critical

5: Choose action $i$ for $n_l^{(2)}$ *consecutive* rounds. Mark the last round as $T_{l,i}^{(1)}$.

6: **if** $\mathtt{UCB}_i(T_{l,i}^{(1)}) \geq \max_{j \in A_l} \mathtt{LCB}_j(T_{l,i}^{(1)})$ **then** ▷ learn from data

7: Choose action $i$ for additional $\max\left\{ \lfloor \frac{\lambda t_l^{(1)} - T_l^{(2)}}{|A_l^{(1)}|} \rfloor - n_l^{(2)}, 0 \right\}$ *consecutive* rounds.

8: **else**

9: **for** $i \in A_l^{(1)}$ **do**

10: The same steps as Lines 5 to 7. ▷ learn from data

11: Mark the last round as $T_l^{(1)}$, and mark the last chosen action as $a_{T_l^{(1)}}$.

12: Elimination: compute $\mathtt{UCB}_i(T_l^{(1)})$ and $\mathtt{LCB}_i(T_l^{(1)})$ for all $i \in A_l$, and let ▷ learn from data

$$A_{l+1}^{(1)} = \left\{ i \in A_l^{(1)} \mid \mathtt{UCB}_i(T_l^{(1)}) \geq \max_{j \in A_l} \mathtt{LCB}_j(T_l^{(1)}), \mathtt{UCB}_i(T_{l,i}^{(1)}) \geq \max_{j \in A_l} \mathtt{LCB}_j(T_{l,i}^{(1)}) \right\},$$

$$A_{l+1}^{(2)} = \left\{ i \in A_l^{(2)} \mid \mathtt{UCB}_i(T_l^{(1)}) \geq \max_{j \in A_l} \mathtt{LCB}_j(T_l^{(1)}) \right\},$$

and $A_{l+1} = A_{l+1}^{(1)} \cup A_{l+1}^{(2)}$.

13: **for** $l = q(S,K)+1$ **do**

14: Starting from an arbitrary action in $A_l^{(2)}$, choose each action in $A_l^{(2)}$ for $n_l^{(2)} = \lfloor \frac{t_l^{(2)} - T_{l-1}^{(1)}}{|A_l|} \rfloor$ *consecutive* rounds. Mark the last round as $T_l^{(2)}$.

15: Elimination: compute $\mathtt{UCB}_i(T_l^{(2)})$ and $\mathtt{LCB}_i(T_l^{(2)})$ for all $i \in A_l$, and let ▷ learn from data

$$A_{l+1}^{(2)} = \left\{ i \in A_l^{(2)} \mid \mathtt{UCB}_i(T_l^{(2)}) \geq \max_{j \in A_l} \mathtt{LCB}_j(T_l^{(2)}) \right\}.$$

If $A_{l+1}^{(2)}$ is non-empty, let $A_{l+1} = A_{l+1}^{(2)}$; otherwise, let $A_{l+1} = A_l^{(1)}$.

16: Find an action $i \in A_{l+1}$ that maximizes $\bar{\mu}_i(T_l^{(2)})$. Keep choosing this action until round $T$.

to determine $A_1^{(1)}$ and $A_1^{(2)}$ *randomly* rather than deterministically, as randomization enables better worst-case performance.

**Hybrid exploration scheme.** At initialization, we define two series of time points $(t_j^{(1)})_{j=1}^{q(S,K)+1}$ and $(t_j^{(2)})_{j=1}^{q(S,K)+2}$, which provide (rough) guidance on how we should balance the exploration and exploitation for actions in $A_1^{(1)}$ and $A_1^{(2)}$. These two series are similar but different from the series $(t_j)_{j=1}^{q(S,K)+1}$ defined in Algorithm 6.1 due to (i) we allow AdaLS to switch to the actions in $A_1^{(2)}$ more frequently and (ii) we need to consider the interplay between the two classes of actions. Then, AdaLS runs in $q(S,K)+1$ epochs. In each epoch $l \in [q(S,K)]$, AdaLS first explores each action in $A_l^{(2)}$ (which consists of all *uneliminated* action in $A_1^{(2)}$) for an equal number of rounds (see Line 2), then explores each action in $A_l^{(1)}$ (which consists of all *uneliminated* action in $A_1^{(1)}$) for a *data-dependent* number of rounds (see Line 3 to Line 10), and finally conducts elimination to determine $A_{l+1}^{(1)}$ and $A_{l+1}^{(2)}$ (see Line 12). At the last epoch $q(S,K)+1$, AdaLS will first explore all actions in $A_{q(S,K)+1}^{(2)}$ (see Line 14), then conducts elimination, and finally commits to a single action (see Line 16). Notably, in every elimination step of AdaLS, the confidence bounds of different actions are at different scales, because they were explored non-uniformly. Compared with LS-SE where each uneliminated action is uniformly explored in epoch $l \in [q(S,K)]$, the exploration scheme of AdaLS requires more delicate design (which is multi-scale in nature) because we need to ensure that the elimination based on *confidence bounds with different scales* are effective.

**Higher adaptivity via more frequent queries to the data.** A significant difference between AdaLS and LS-SE is that AdaLS utilizes data more frequently and is not a batched policy — while AdaLS runs in $q(S,K)+1$ epochs, each epoch does not correspond to a batch because *actions selected during epoch $l$ depends on the latest data collected during epoch $l$* (see Lines 6, 10 and 15, where AdaLS utilizes the latest data to determine whether to switch or not). Moreover, the actual epoch schedule $(T_l^{(1)})_{l=1}^{q(S,K)+1}$ is also data-dependent, i.e., AdaLS decides when to start and end each epoch only *after* gradual access to the data. Such additional adaptivity is critical for AdaLS to achieve better performance; otherwise, one cannot have careful control over the exploration and may over-explore the actions in $A_1^{(1)}$ (as they are visited relatively less frequently).

We provide a rigorous analysis of AdaLS, verify that it is an $S$-switch policy, and show that it attains an improved regret bound; see the statement in Theorem 6.1. The proof of Theorem 6.1 (which closely follows the intuition that we provide above) is considerably more

challenging than Proposition 6.1; see Appendix E.8 for details.

**Theorem 6.1.** *Let $\pi$ be the AdaLS policy, then $\pi \in \Pi_S$. There exists an absolute constant $C \geq 0$ such that for all $K > 1$, $S \geq 0$ and $T \geq K$,*

$$R^\pi(K,T) \leq C(\log T)^2 \cdot \max\left\{ \frac{(K - r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}},\ K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}} \right\},$$

*where $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$ and $r(S,K) = (S-1)\%(K-1)$.*

To illustrate the regret guarantee given by Theorem 6.1, we use it to calculate the exact regret rates (in terms of both $K$ and $T$) of AdaLS under different concrete values of $S$; see Table 6.1 for details. As benchmarks, we also calculate the exact regret rates of LS-SE using Proposition 6.1, and compare the regret of LS-SE and AdaLS under each single value of $S$; see the detailed comparisons in Table 6.1.

We make two observations. First, AdaLS shares the same regret rate as LS-SE when $K - r(S,K) = \Omega(K)$; see the second column in Table 6.1, where the regret rate of both AdaLS and LS-SE is $K^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}}$ (within logarithmic factors). This implies that AdaLS does not attain a fundamentally better rate than LS-SE when $K = \widetilde{\mathcal{O}}(1)$ or when $r(S,K)$ is not close to $K$. Second, AdaLS can attain a significantly better regret rate when $r(S,K) = K - o(K)$; see the last two columns in Table 6.1, where the regret rate of AdaLS is always better than LS-SE. Note that the closer $r(S,K)$ is to $K$, the better AdaLS's regret rate can be. In particular, when $r(S,K) = K - 2$ (i.e, when $S$ is a *multiple* of $K-1$), AdaLS attains a rate of $K^{-1} T^{\frac{1}{2-2^{-q(S,K)}}} \vee K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}}$ — if the first term dominates, then AdaLS improves upon the regret of LS-SE by a multiplicative factor of $\widetilde{\Theta}(K^{2-\frac{1}{2-2^{-q(S,K)}}})$; if the second term dominates (which is not uncommon when $K$ is large), then the regret of AdaLS has a better growth rate in $T$, which enables it to perform arbitrarily better than LS-SE when $T \to \infty$.

*Remark.* We make two remarks for Table 6.1. First, for brevity, we only present the regret rates for $S \in [0, 4K-4]$ in Table 6.1 — regret rates for larger $S$ follows the same pattern. Second, since it is quite easy to show *algorithm-specific* lower bounds for LS-SE and AdaLS which match the upper bounds in Proposition 6.1 and Theorem 6.1 respectively (up to logarithmic factors), we directly use $\widetilde{\Theta}$ (rather than $\widetilde{\mathcal{O}}$) to describe the regret of LS-SE and AdaLS; of course, the *algorithm-specific* lower bounds for LS-SE and LS-SE do not imply

Table 6.1: Regret of LS-SE and AdaLS under different switching budgets. Here $\epsilon \in (0,1)$ is an arbitrary constant independent of $K$ and $T$ (it can be arbitrarily close to 0, as long as it is fixed).

| | | | |
|---|---|---|---|
| $S \in \{0, 1, \dots, K-1\}$ | | | |
| $S$ | $0, 1, \dots, (1-\epsilon)(K-1)$ | $K-1-\widetilde{\Theta}(K^{\delta}),\ \delta \in (0,1)$ | $K-1$ |
| LS-SE | $\widetilde{\Theta}(T)$ | $\widetilde{\Theta}(T)$ | $\widetilde{\Theta}(T)$ |
| AdaLS | $\widetilde{\Theta}(T)$ | $\widetilde{\Theta}\left(K^{\delta-1}T \vee K^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ | $\widetilde{\Theta}\left(K^{-1}T \vee K^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ |
| $S \in \{K, \dots, 2K-2\}$ | | | |
| $S$ | $K, K+1, \dots, (1-\epsilon)(2K-2)$ | $2K-2-\widetilde{\Theta}(K^{\delta}),\ \delta \in (0,1)$ | $2K-2$ |
| LS-SE | $\widetilde{\Theta}\left(K^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ |
| AdaLS | $\widetilde{\Theta}\left(K^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{4}{3}\delta-1}T^{\frac{2}{3}} \vee K^{\frac{3}{7}}T^{\frac{4}{7}}\right)$ | $\widetilde{\Theta}\left(K^{-1}T^{\frac{2}{3}} \vee K^{\frac{3}{7}}T^{\frac{4}{7}}\right)$ |
| $S \in \{2K-1, \dots, 3K-3\}$ | | | |
| $S$ | $2K-1, 2K, \dots, (1-\epsilon)(3K-3)$ | $3K-3-\widetilde{\Theta}(K^{\delta}),\ \delta \in (0,1)$ | $3K-3$ |
| LS-SE | $\widetilde{\Theta}\left(K^{\frac{3}{7}}T^{\frac{4}{7}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{3}{7}}T^{\frac{4}{7}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{3}{7}}T^{\frac{4}{7}}\right)$ |
| AdaLS | $\widetilde{\Theta}\left(K^{\frac{3}{7}}T^{\frac{4}{7}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{10}{7}\delta-1}T^{\frac{4}{7}} \vee K^{\frac{7}{15}}T^{\frac{8}{15}}\right)$ | $\widetilde{\Theta}\left(K^{-1}T^{\frac{4}{7}} \vee K^{\frac{7}{15}}T^{\frac{8}{15}}\right)$ |
| $S \in \{3K-2, \dots, 4K-4\}$ | | | |
| $S$ | $3K-2, 3K-1, \dots, (1-\epsilon)(4K-4)$ | $4K-4-\widetilde{\Theta}(K^{\delta}),\ \delta \in (0,1)$ | $4K-4$ |
| LS-SE | $\widetilde{\Theta}\left(K^{\frac{7}{15}}T^{\frac{8}{15}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{7}{15}}T^{\frac{8}{15}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{7}{15}}T^{\frac{8}{15}}\right)$ |
| AdaLS | $\widetilde{\Theta}\left(K^{\frac{7}{15}}T^{\frac{8}{15}}\right)$ | $\widetilde{\Theta}\left(K^{\frac{22}{15}\delta-1}T^{\frac{8}{15}} \vee K^{\frac{15}{31}}T^{\frac{16}{31}}\right)$ | $\widetilde{\Theta}\left(K^{-1}T^{\frac{8}{15}} \vee K^{\frac{15}{31}}T^{\frac{16}{31}}\right)$ |

fundamental limits for *other* algorithms (which can be arbitrarily more complicated) — proving an universal *algorithm-independent* lower bound is the task of Section 6.3.3.

### 6.3.3 Lower Bound on Regret

The AdaLS algorithm, though being a significantly refined version of the LS-SE algorithm, still seems to leave plenty of room for improvement. For example, while AdaLS has higher adaptivity than LS-SE, it learns from data for at most $Kq(S,K)+1$ times, leaving an open question of whether one can utilize even more adaptivity to achieve lower regret. Moreover, as discussed in Section 6.3.2, AdaLS only improves the regret with the help of $K$, failing to *directly* improve the regret's dependence on the most important parameter $T$ when $K=\widetilde{\mathcal{O}}(1)$. This motivates the following natural questions: Is it possible to directly improve the regret in terms of $T$? Can the dependence on $K$ be further improved? What is the fundamental limit of the U-BwSC problem?

We answer the above questions by establishing a strong (and quite surprising) information-theoretic lower bound on the regret incurred by *any* admissible policy; see Theorem 6.2. The lower bound directly match the upper bound in Theorem 6.1, indicating that AdaLS is optimal up to logarithmic factors. Notably, our lower bound holds for any $K$, any $S$, any $T$, thus is substantially stronger than a specific lower bound demonstrated for special choices of $S,K,T$.

**Theorem 6.2.** *There exists an absolute constant $C>0$ such that for all $K>1, S\geq 0, T\geq 2K$ and for all policy $\pi\in\Pi_S$,*

$$R^{\pi}(K,T)\geq \frac{C}{\log T}\cdot\max\left\{\frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K}T^{\frac{1}{2-2^{-q(S,K)}}},\ K^{1-\frac{1}{2-2^{-q(S,K)-1}}}T^{\frac{1}{2-2^{-q(S,K)-1}}}\right\},$$

*where $q(S,K)=\left\lfloor\frac{S-1}{K-1}\right\rfloor$ and $r(S,K)=(S-1)\%(K-1)$.*

As we introduced in Section 6.1.3, the proof of Theorem 6.2 is non-trivial, and will be elaborated on in a separate section (Appendix E.4). Combining Theorem 6.1 and Theorem 6.2, we completely characterize the optimal regret of U-BwSC as follows.

**Corollary 6.1.** *For all $S \geq 0, K > 1, T \geq 2K$, we have*

$$R_S^\star(K,T) = \widetilde{\Theta}(1)\cdot\max\left\{ \frac{(K - r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}},\ K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}} \right\}.$$

*If $K = \widetilde{\mathcal{O}}(1)$, then we have $R_S^\star(T) = \widetilde{\Theta}\left(T^{\frac{1}{2-2^{-q(S,K)}}}\right)$.*

**On the Necessity of Switching in MAB**

The lower bound in Theorem 6.2 also leads to new results for the classical MAB problem.

**Corollary 6.2.** *The following properties hold for the classical MAB: (i) $\Theta(K \log\log T)$ switches are necessary and sufficient for achieving $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret, (ii) for any fixed $N \in \mathbb{N}_{>0}$, $N(K-1)+1$ switches are necessary and sufficient for achieving $\widetilde{\mathcal{O}}(K^{1-\frac{1}{2-2^{-N}}} T^{\frac{1}{2-2^{-N}}})$ regret, and (iii) $\Omega(K)$ switches are necessary for achieving sublinear regret.*

Note that the number of switches stated in Corollary 6.2 refers to the maximum number of switches that a policy can make. While Cesa-Bianchi et al. (2013) has proposed policies that achieve $\widetilde{\mathcal{O}}(\sqrt{KT})$ (near-optimal) regret with $\mathcal{O}(K \log\log T)$ switches, no prior work has answered the question of how many switches are *necessary* for a near-optimal learning policy in MAB. To the best of our knowledge, this paper is the first to show $\Omega(K \log\log T)$ lower bound on the number of switches.

### 6.3.4 Phase Transitions

Corollary 6.2 provides a non-asymptotic (i.e., finite-time) characterization on the optimal regret of the U-BwSC problem. While such non-asymptotic characterization is very general (e.g., it holds for arbitrary $S, K, T$ and does not rely on any assumption on their orders), if we want to obtain deeper insights on the optimal regret's *growth rate* (which is easier to understand when defined as concrete limiting behavior), then the asymptotic regime may be more appropriate in terms of making our statements rigorous and precise. In this subsection, we use asymptotics to rigorously define the optimal regret rate, and characterize the trade-off between the optimal regret rate and the switching budget. As we shall see, the trade-off reveals surprising *phase transitions*[39] (to be defined shortly). For ease of presentation, all

[39]The terminology of phase transitions originates from physics (see, e.g., Domb 2000), and has been used in various fields in probability theory and statistics. We note that most of rigorous definitions of phase

the "rates" defined in this subsection do not involve logarithmic factors.

### Phase Transitions in the Fixed-$K$ Asymptotic Regime

We first consider the most natural asymptotic regime where we let the time horizon $T \to \infty$ and keep the number of arms $K$ fixed; we refer to this regime as the "fixed-$K$" asymptotic regime. For any fixed switching budget $S \geq 0$, we are interested in the growth rate of the optimal regret $R_S^\star(T)$ as $T \to \infty$. Following the convention of statistics and machine learning (see, e.g., Tsybakov 2009), we define the optimal regret rate (i.e., minimax rate) as the power function that best approximates $R_S^\star(T)$ as $T \to \infty$; see below.

**Definition 6.2.** *For any fixed $K > 1, S \geq 0$, there exists a unique constant $p \in [0,1]$ such that*

$$\lim_{T\to\infty} \frac{R_S^\star(T)}{T^{p+\epsilon}} = 0, \quad \lim_{T\to\infty} \frac{R_S^\star(T)}{T^{p-\epsilon}} = \infty, \qquad \forall \epsilon > 0.$$

*We call $T^p$ the optimal regret rate under switching budget $S$, and $p$ the optimal regret rate exponent.*

Note that an equivalent definition is to directly let $p := \lim_{T\to\infty} \frac{\log R_S^\star(T)}{\log T}$ and let the power function $T^p$ be the optimal regret rate (see Hu et al. 2020).

By Corollary 6.2, $R_S^\star(T) = \widetilde{\Theta}\left(T^{\frac{1}{2-2^{-\lfloor (S-1)/(K-1) \rfloor}}}\right)$ when $K$ is fixed. To the best of our knowledge, this is the first time that a floor function naturally arises in the exponent of $T$ in the optimal regret of an online learning problem. Consequently, we know that the optimal regret rate under switching budget $S$ is $T^{\frac{1}{2-2^{-\lfloor (S-1)/(K-1) \rfloor}}}$, which exhibits surprising *phase transitions* described below.

**Definition 6.3** (Phases & Transition Points)**.** *In the fixed-$K$ regime, we call the interval $[(j-1)(K-1)+1, j(K-1)+1)$ the $j$-th phase, and call $j(K-1)+1$ the $j$-th transition point ($j \in \mathbb{N}_{>0}$).*

**Observation 6.1** (Phase Transitions)**.** *As $S$ increases from 0 to infinity, $S$ will leave the $j$-th phase and enter the $(j+1)$-th phase at the $j$-th transition point ($j \in \mathbb{N}_{>0}$). Each time $S$ arrives at a transition point, the optimal regret rate will change abruptly, and then remain the same until $S$ arrives at the next transition point.*

transitions in probability theory and statistics require asymptotics; see, e.g., Wainwright (2009), Bayati et al. (2015).

Table 6.2: Optimal regret rate under different switching budgets for a fixed $K$.

| $S$ | $[0, K)$ | $[K, 2K-1)$ | $[2K-1, 3K-2)$ | $[3K-2, 4K-3)$ | $[4K-3, 5K-4)$ | $[5K-4, 6K-5)$ |
|---|---|---|---|---|---|---|
| Rate | $T$ | $T^{2/3}$ | $T^{4/7}$ | $T^{8/15}$ | $T^{16/31}$ | $T^{32/63}$ |

Phase transitions are illustrated in Table 6.2. This phenomenon seems counter-intuitive, as it suggests that in the fixed-$K$ regime, increasing switching budget would not help reduce the best achievable regret rate, as long as the budget does not reach the next transition point. Moreover, the abrupt change happens at each transition point is very interesting — at this point, a minimal difference in the switching budget can fundamentally change the statistical nature of the problem. Along with phase transitions, we also observe an interesting property: the length of each phase is always equal to $K-1$. This property is elegant and reveals some favorable features of the U-BwSC problem: as we will show in Section 6.4, this property does not hold under the general switching cost structure.

*Remark.* While phase transitions are intriguing and theoretically interesting, we would like to make some comments on the scope of the above reuslts. First, one should keep in mind that for any $S$, the above analysis concerns the growth rate (i.e. scaling behavior) of $R_S^\star(T)$ as $T$ grows (which is a statistical property), rather than the numerical value of $R_S^\star(T)$ for a specific $T$. In particular, we are less interested in describing how $R_S^\star(T)$ changes with respect to $S$ for a fixed, specifically chosen $T$; instead, we seek to understand how the minimax rate of the problem (which reflects the "learnability" of the problem when the sample size grows) changes with respect to $S$. Second, phase transitions are more relevant to practice when $S$ belongs to the first 4 or 5 phases. Indeed, the regret rate exponent in Table 6.2 decreases dramatically as $S$ goes over more phases — when $S$ is relatively large, the difference between two phases may be too small to make the rate reduction happens in the "ideal" asymptotic world really make a difference in reality. Recall that in the non-asymptotic regime where $S$ can depend on $T$, $\mathcal{O}(\log\log T)$ switches are sufficient for one to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret — this also indicates that the most interesting transitions should happen when $S$ is small.

### Phase Transition in the Growing-$K$ Asymptotic Regime

We now consider a second asymptotic regime which allows $K$ to grow with $T$ in a moderate rate, which corresponds to the "growing-dimension" asymptotic regime in statistics (Portnoy 1984, 1988). Specifically, we consider the following "growing-$K$" asymptotic regime: $K, T \to$

$\infty$ and $K/T^\alpha \to c$ for some $\alpha \in (0,1), c \in (0,\infty)$. By Corollary 6.2, $\Omega(K)$ switches are necessary for achieving sublinear regret in MAB; thus, in the "growing-$K$" regime of U-BwSC, a fixed $S$ cannot avoid $\Omega(T)$ regret, and the values of $S$ that we are most interested in should range from $\Omega(K)$ to $o(K \log\log T)$. This indicates that in the growing-$K$ regime, $S$ should be naturally understood as "a function of $K$": the dependence between $S$ and $K$ is necessary, while $S/K$ should have no or extremely small dependence on $T$. We thus only consider $S$ such that $S/K \to \theta$ for some constant $\theta \in [0,\infty)$, and we call $\theta := \lim_{K\to\infty} S/K$ the "budget-to-arm ratio" (BAR). The optimal regret rate in the "growing-$K$" regime can be then defined similar to Definition 6.2 (with $K$ and $S$ scales proportional to $T$), or equivalently by calculating $p := \lim_{T\to\infty} \frac{\log R^\star_S(K,T)}{\log T}$ and denote $T^p$ as the optimal regret rate.

The first finding in the growing-$K$ regime is that the (original form of) phase transitions described in Section 6.3.4 may not hold any more, and the existence of "abrupt rate changes" depends on the magnitude of $K$ relative to $T$. To see this, let us focus on a small range of $S$ at the end of the first phase and at the start of the second phase: $S = K - 1 - \widetilde{\Theta}(K^\delta)$ with $\delta \in (0,1)$, $S = K-1$ and $S = K$. By Table 6.1 and simple calculation, the corresponding optimal regret rate exponents for them are $\max\{1-\alpha(1-\delta), \frac{\alpha+2}{3}\}$, $\max\{1-\alpha, \frac{\alpha+2}{3}\}$ and $\frac{\alpha+2}{3}$ respectively. By letting $\delta$ move smoothly from 1 to 0, one can find that the optimal regret rate exponent associated with $S = K - 1 - \Theta(K^\delta)$ smoothly decays from 1 to $\max\{1-\alpha, \frac{\alpha+2}{3}\}$: while $S$ is always in the first phase defined in Section 6.3.4, the optimal regret rate does not "remain the same" any more. Moreover, whether there is an abrupt change when $S$ moves from $K-1$ to $K$ depends on the magnitude of $\alpha$. If $\alpha < \frac{1}{4}$, then there is still an abrupt change in the optimal regret rate (e.g., if $\alpha = 0.1$, then the rate jumps from $T^{0.9}$ to $T^{0.7}$); if $\alpha \geq \frac{1}{4}$, then the optimal regret rate under $S = K-1$ remains unchanged when $S$ reaches the next transition point. One can keep conducting such analysis for other phases, and find similar examples on the ending range of each phase where $r(S,K) = K - o(K)$ — interestingly, these are also the ranges that AdaLS significantly improves the regret; see the last two columns in Table 6.1.

The second finding in the growing-$K$ regime is that when we consider how the optimal regret rate changes with respect to the budget-to-arm ratio $\theta$ (rather than $S$), we can still discover phase transitions similar to Section 6.3.4. Indeed, the counter-examples described in the above paragraphs correspond to the ranges of $S$ where the remainder function $r(S,K)$ moves from $K - o(K)$ to $K-2$ to 0, i.e., the ranges where $S$ moves from $N(K-1) - o(K)$

to $N(K-1)$ to $N(K-1)+1$ for some integer $N \in \mathbb{N}_{>0}$. All valid $S$ in these ranges have the property that the BAR $\theta = \lim_{K\to\infty} S/K = N \in \mathbb{N}_{>0}$. This indicates that the complicated behavior of the optimal regret rate described in the above paragraph might only happen in scenarios where the BAR $\theta$ is exactly a positive integer. In fact, for all $S$ such that $\theta$ exists and $\theta$ is not an integer (e.g., consider $S = \lfloor 2.5K \rfloor$, then $\theta = 2.5$ is not an integer), one can find that the (unique) optimal regret rate exponent is $\alpha + \frac{1}{2-2^{-\lfloor\theta\rfloor}}(1-\alpha)$, which contains a floor function. Consequently, we can define the interval $(j-1, j)$ as the $j$-th phase ($j \in \mathbb{N}_{>0}$) for $\theta$, and discover phase transitions of the optimal regret rate as illustrated in Table 6.3. A difference in this new notion of phase transitions is that in the growing-$K$ regime, each "transition point" $j \in \mathbb{N}_{>0}$ can be associated with infinitely many optimal regret rates which interpolate between the rates of the previous and the next phase (see the previous paragraph for the example of $\theta = 1$).

Table 6.3: Optimal regret rate under different BAR $\theta$ when $K$ grows as $T^\alpha$.

| $\theta$ | $[0,1)$ | 1 | $(1,2)$ | 2 | $(2,3)$ | 3 | $(3,4)$ |
|---|---|---|---|---|---|---|---|
| Rate | $T$ | – | $T^{(2+\alpha)/3}$ | – | $T^{(4+3\alpha)/7}$ | – | $T^{(8+7\alpha)/15}$ |

### 6.3.5 Relationship Between Limited Switches and Limited Adaptivity

In this subsection, we discuss the relationship between limited switches and limited adaptivity in bandit problems. As discussed in Section 6.1.3, in the U-BwSC problem, the constraint is on the number of switches and is defined in the "action world," hence the learner has full adpativity. By contrast, in the batched bandit problem, the constraint is on adaptivity and is defined in the "observation world," hence the learner has full switching power. Since the two constraints in the two problems are defined in two different "worlds," the relationship between the two problems is interesting.

We first claim that the U-BwSC problem can be seen as a strict relaxation of the batched bandit problem (no matter with the "static grid" restriction, like Perchet et al. 2016, or without such a restriction, like Gao et al. 2019), in the sense that U-BwSC admits more flexible policies and can enjoy a fundamentally better optimal regret rate. The LS-SE and AdaLS algorithms help establish this claim. First, one can easily show that any $M$-batch policy that achieves certain regret in the $M$-batch $K$-armed bandit problem can be transformed, using the LS-SE ingredients and randomization, to an $S$-switch policy that

achieves exactly the same regret in the $S$-switch $K$-armed U-BwSC problem, as long as $q(S,K) = M-1$, i.e., $S \in [(M-1)(K-1)+1 : M(K-1)]$. This implies that the admissible policy class of the $S$-switch U-BwSC problem essentially contains the admissible policy class of the $(q(S,K)+1)$-batch bandit problem as a subset. Moreover, since an $S$-switch policy can utilize data much more flexibly than an $(q(S,K)+1)$-batch policy (see Section 6.3.2), $(q(S,K)+1)$-batch algorithms *necessarily* suffer from sub-optimal rates for U-BwSC in general. Note that the regret lower bound for the $(q(S,K)+1)$-batch bandit problem is $\widetilde{\Omega}\left(K^{1-\frac{1}{2-2^{-q(S,K)}}}T^{\frac{1}{2-2^{-q(S,K)}}}\right)$[40], which is a fundamental limit for all $(q(S,K)+1)$-batch algorithms; AdaLS's regret bound (6.2) surpasses this limit, indicating that the admissible policy class of U-BwSC indeed contains lower-regret policies, and U-BwSC can have a fundamentally better optimal regret rate.

On the other hand, the performance improvement that one can benefit from the above relaxation also has a limit, as demonstrated by our lower bound (Theorem 6.2). In fact, one can find that the optimal regret rate of the $S$-switch U-BwSC (Corollary 6.2) *interpolates* between the optimal regret rates of the $(q(S,K)+1)$-batch and the $(q(S,K)+2)$-batch bandit problems; moreover, when $K = \widetilde{\mathcal{O}}(1)$, the optimal regret rate of the $S$-switch U-BwSC coincides with the optimal regret rate of the $(q(S,K)+1)$-batch bandit problem. The above findings provide very useful managerial insights:

1. The switching constraint is a more relaxed constraint than the batch constraint, and enables better performance guarantees when $K$ is large.

2. Limiting switches (in the "action" world) implicitly limits adaptivity (in the "observation" world), in the sense that the optimal $S$-switch policy's regret rate lies between the optimal $(q(S,K)+1)$-batch and $(q(S,K)+2)$-batch policies' regret rates; when $K = \widetilde{\mathcal{O}}(1)$, the regret rate of the optimal $S$-switch policy and the optimal $(q(S,K)+1)$-batch policy coincide (up to logarithmic factors).

Finally, we would like to point out that our lower bound result (Theorem 6.2) is theoretically stronger and more general than the lower bounds for batched bandits. Indeed, since any $M$-batch policy can be transformed to an equivalent $((M-1)(K-1)+1)$-switch policy, any lower bound proved for U-BwSC implies a lower bound for batched bandits. In

[40] This is shown in Gao et al. (2019) with slight difference dependence on $K$ as they allow $\sup_{i,j}|\mu_i - \mu_j| = \sqrt{K}$. Since our setting is $\sup_{i,j}|\mu_i - \mu_j| \in [0,1]$, we write the associated "right" dependence on $K$ for ease of comparison.

particular, by plugging $S = (M-1)(K-1)+1$ in Theorem 6.2 (this value of $S$ actually corresponds to the "easier-to-prove" part of Theorem 6.2, as it is the start of a "phase"), one recovers the minimax lower bound result of Gao et al. (2019) as a corollary (the order of $K, T$ will be the same under their conditions).

## 6.4 General Switching Costs

We now proceed to the general setting of BwSC, where $c_{i,j}$ can be any non-negative real number and even $\infty$ $(i \neq j)$ . For this general setting, a new and fundamental resarch question is how the structure of switching costs $(c_{i,j})$ affects the statistical nature of BwSC. Since this question is interesting and challenging even when $K = \widetilde{\mathcal{O}}(1)$, in this section, we only seek to derive algorithms and regret bounds that are effective when $K = \widetilde{\mathcal{O}}(1)$. We believe that the techniques developed in Section 6.3 should be helpful for one to obtain refined algorithms and results (for general switching costs) when $K$ is large, but we leave it for future work. In what follows, we consider two switching cost structures: a general symmetric one in Section 6.4.1 and an asymmetric one in Section 6.4.2.

### 6.4.1 Symmetric Switching Costs

We first consider the *general symmetric switching cost* structure where $c_{i,j} = c_{j,i}$ for all $i, j \in [K]$. The corresponding BwSC problem is referred to as the G-BwSC problem. To start with, we need to enhance the framework of Section 6.1.2 to better represent the switching costs. We do this by representing switching costs via a weighted graph. Let $G = (V, E)$ be a (weighted) complete graph, where $V = [K]$ (i.e., each vertex corresponds to an action), and the edge between $i$ and $j$ is assigned a weight $c_{i,j}$ $(\forall i \neq j)$. We call the weighted graph $G$ the *switching graph*. In this subsection, we assume the switching costs satisfy the triangle inequality: $\forall i, j, l \in [k]$, $c_{i,j} \leq c_{i,l} + c_{l,j}$. We relax this assumption in Appendix E.5.1.

The results in Section 6.3 suggest that a simple and effective learning strategy (when $K = \widetilde{\mathcal{O}}(1)$) is to repeatedly visit all actions for many times and then commit to the best action, in a manner similar to LS-SE. This indicates that in the G-BwSC problem, one should consider how to repeatedly visit all vertices in the switching graph, in a most economical way to stay within budget. This implies a connection between G-BwSC and the celebrated shortest Hamiltonian path problem. Motivated by this connection, we propose the *Hamiltonian-*

*Switching Successive Elimination* (HS-SE) algorithm, and present it in Algorithm 6.3. The algorithm enhances the original LS-SE algorithm by adding an additional ingredient: a pre-specified switching order determined by the shortest Hamiltonian path of the switching graph $G$. Note that while the shortest Hamiltonian path problem is NP-hard, solving this problem is entirely an "offline" step in the HS-SE algorithm, i.e., for a given switching graph, the learner only needs to solve this problem once. We also comment that one may use the techniques presented in Section 6.3.2 to design a refined algorithm (analogous to AdaLS) that achieves better performance; however, we leave this for future work, as the simple HS-SE algorithm is already sufficient for revealing important properties of G-BwSC when $K = \widetilde{\mathcal{O}}(1)$.

**Algorithm 6.3** Hamiltonian-Switching Successive Elimination (HS-SE)

**Input:** Switching budget $S$, switching graph $G$, horizon $T$.

**Initialization:** Let $A_1 = [K]$. Find a shortest Hamiltonian path in $G$: $i_1 \to \cdots \to i_K$. Denote the total weight of the shortest Hamiltonian path as $H$. Compute $q'(S,G) = \left\lfloor \frac{S - \max_{i,j\in[K]} c_{i,j}}{H} \right\rfloor$. Divide the entire time horizon $T$ into $q'(S,G)+1$ epochs: $(t_0 : t_1], (t_1 : t_2], \ldots, (t_{q'(S,G)} : t_{q'(S,G)+1}]$, where the endpoints are defined by $t_0 = 0$ and

$$t_j = \left\lfloor K^{1-\frac{2-2^{-(j-1)}}{2-2^{-q'(S,G)}}} T^{\frac{2-2^{-(j-1)}}{2-2^{-q'(S,G)}}} \right\rfloor, \quad \forall j = 1, \ldots, q'(S,G)+1.$$

**Policy:**

**for** $l = 1, \ldots, q'(S,G)$ **do**
  **if** $l$ is odd **then**
    **for** $i = i_1, \ldots, i_K$ **do** ▷ along the direction of $i_1 \to \cdots \to i_K$
      If $i \in A_l$ (i.e., uneliminated), choose action $i$ for $\frac{t_l - t_{l-1}}{|A_l|}$ *consecutive* rounds.
  **else**
    **for** $i = i_K, \ldots, i_1$ **do** ▷ along the direction of $i_K \to \cdots \to i_1$ (the reverse of the above)
      If $i \in A_l$ (i.e., uneliminated), choose action $i$ for $\frac{t_l - t_{l-1}}{|A_l|}$ *consecutive* rounds.
  Elimination: compute $\mathsf{UCB}_i(t_l)$ and $\mathsf{LCB}_i(t_l)$ for all $i \in A_l$ and let ▷ learn from data

$$A_{l+1} = \left\{ i \in A_l \mid \mathsf{UCB}_i(t_l) \geq \max_{j \in A_l} \mathsf{LCB}_j(t_l) \right\}.$$

For $l = q'(S,G)+1$, find an action $i \in A_l$ that maximizes $\bar{\mu}_i(t_{l-1})$. Keep choosing this action until round $T$.

Let $H$ denote the total weight of the shortest Hamiltonian path of $G$. It is not difficult to verify that HS-SE is an $S$-switching-budget policy and ensures the following upper bound on regret; see Appendix E.9 for a proof.

**Theorem 6.3.** *Let $\pi$ be the HS-SE policy, then $\pi \in \Pi_S$. There exists an absolute constant $C \geq 0$ such that for all $G$ (with $H > 0$), $K = |G|$, $S \geq 0$, $T \geq K$,*

$$R^{\pi}(T) \leq C(\log K \log T) K^{1-\frac{1}{2-2^{-q'(S,G)}}} T^{\frac{1}{2-2^{-q'(S,G)}}},$$

*where $q'(S,G) = \left\lfloor \frac{S - \max_{i,j\in[K]} c_{i,j}}{H} \right\rfloor$.*

In Theorem 6.4, we provide a lower bound that is very close to the above upper bound. The proof of Theorem 6.4 builds on the proof of Theorem 6.2, but has two notable differences: (i) it involves several new techniques to deal with the general switching cost structure, and (ii) it pays less attention to the dependence on $K$; see Appendix E.13 for details.

**Theorem 6.4.** *There exists an absolute constant $C > 0$ such that for all $G$ (with $H > 0$), $K = |G|$, $S \geq 0$, $T \geq 2K$ and for all policy $\pi \in \Pi_S$,*

$$R^{\pi}(K,T) \geq \frac{C}{K \log T} \cdot T^{\frac{1}{2-2^{-q''(S,G)}}},$$

*where $q''(S,G) = \left\lfloor \frac{S - \max_{i\in[K]} \min_{j\neq i} c_{i,j}}{H} \right\rfloor$.*

Let us focus on the case of $K = \widetilde{\mathcal{O}}(1)$ and compare the upper and lower bounds given by Theorem 6.3 and Theorem 6.4. When the switching costs satisfy the condition $\max_{i,j\in[K]} c_{i,j} = \max_{i\in[K]} \min_{j\neq i} c_{i,j}$, we have $q'(S,G) = q''(S,G)$, thus the two bounds directly match (up to polylog$(T)$). This reveals an interesting fact: when $\max_{i,j\in[K]} c_{i,j} = \max_{i\in[K]} \min_{j\neq i} c_{i,j}$, the optimal regret rate of G-BwSC is completely characterized by the floor function $\left\lfloor \frac{S - \max_{i,j\in[K]} c_{i,j}}{H} \right\rfloor$, which further depends on $H$. The fact implies that the length of the shortest Hamiltonian path is indeed a fundamental quantity associated with the G-BwSC problem, and conveys an important message: the structure of switching costs may affect the optimal regret rate of BwSC through some key quantities associated with graph traversal problems. We now provide a concrete switching cost structure satisfying the condition $\max_{i,j\in[K]} c_{i,j} = \max_{i\in[K]} \min_{j\neq i} c_{i,j}$ below.

**Example 6.1** (Isolated Action Model)**.** *Consider a set of $K-1$ "close" or "similar" actions $\{1, \cdots, K-1\}$, and another "isolated" action $K$, such that $c_{K,1} = \cdots = c_{K,K-1} \geq \max_{i,j\in[K-1]} c_{i,j}$ (i.e., action $K$ is isolated from other actions such that its distance to every other action is a large constant). This model always satisfies the condition $\max_{i,j\in[K]} c_{i,j} =$*

$\max_{i\in[K]} \min_{j\neq i} c_{i,j}$, *and subsumes the unit-switching-cost model as a special case. As an example, in promotion planning,* $1, \cdots, K-1$ *can be different variants of a standard promotion strategy, while* $K$ *can be an aggressive clearance strategy.*

When the condition $\max_{i,j\in[K]} c_{i,j} = \max_{i\in[K]} \min_{j\neq i} c_{i,j}$ is not satisfied, for any switching graph $G$, the upper and lower bounds still match for a wide range of $S$:

$$\left[0, H + \max_{i\in[k]} \min_{j\neq i} c_{i,j}\right) \bigcup \left\{ \bigcup_{n=1}^{\infty} \left[ nH + \max_{i,j\in[k]} c_{i,j}, (n+1)H + \max_{i\in[k]} \min_{j\neq i} c_{i,j} \right) \right\}.$$

Even when $S$ is not in this range, we still have $q'(S,G) \leq q''(S,G) \leq q'(S.G) + 1$ for any $G$ and any $S$, which means that the difference between the two indices is at most 1 and the upper and lower bounds are always close. In fact, it can be shown that as $S$ increases, the gap between the upper and lower bounds decreases *doubly exponentially*. Therefore, the HS-SE algorithm is quite effective for the G-BwSC problem when $K = \widetilde{\mathcal{O}}(1)$.

### 6.4.2 Asymmetric Switching Costs: The Departure Cost Structure

We now consider another switching cost structure that allows asymmetry. Since the general asymmetric case is only more complicated than the case studied in Section 6.4.1, we consider a special case of asymmetric switching costs: there exists $\boldsymbol{c} = (c_1, \ldots, c_K) \in \mathbb{R}^K_{\geq 0}$ such that $c_{i,j} = c_i$ for all $i \in [K], j \neq i$. That is, the switching cost between any pair of actions only depends on the action that the learner departs from. We refer to this switching cost structure as the *departure cost* structure (with $c_i$ called the departure cost of action $i$), and the corresponding BwSC problem as the D-BwSC problem. As we shall see, for this fairly general problem, we can fully characterize the optimal regret when $K = \widetilde{\mathcal{O}}(1)$.

We provide an algorithm (AS-SE) for the D-BwSC problem; see Algorithm 6.4. The algorithm follows the same main steps of HS-SE, but has some important differences in the initialization step: it calculates the key actions $i_1, i_K$ and the key index $q(S, \boldsymbol{c})$ differently. We provide some intuition for this new configuration. First, we can still construct a switching graph $G$ associated with the switching costs, but this time being a *directed* graph. Since $i_K$ is the action with the maximum departure cost (denoted by $c^{(1)}$), the path $i_1 \to \cdots \to i_K$ is a shortest Hamiltonian path of the switching graph $G$. The choice of $i_K$ is thus consistent with HS-SE. However, since the switching costs are asymmetric now, one cannot guarantee that the reverse path $i_K \to \cdots \to i_1$ also has a small length. In Algorithm 6.4, based on the

departure cost structure, we optimize the reverse path by letting $i_1$ be the action with the second largest departure cost (denoted by $c^{(2)}$). The determination of the key index $q(S, \boldsymbol{c})$ is a little more complicated, as we need to consider the alternation of two directions; thanks to the departure cost structure, it is easy to compute.

**Algorithm 6.4** Asymmetric-Switching Successive Elimination (AS-SE)

**Input:** Switching budget $S$, switching costs $\boldsymbol{c}$, horizon $T$.

**Initialization:** Let $A_1 = [K]$. Find an action $i_K \in \arg\max_{i\in[K]} c_i$ and an action $i_1 \in \arg\max_{i\in[K]\setminus\{i_1\}} c_i$. Let $(i_2, \ldots, i_{K-1})$ be an arbitrary permutation of $[K] \setminus \{i_1, i_K\}$. Let $\Sigma = \sum_{i=1}^{K} c_i$, $c^{(1)} = c_{i_K}$ and $c^{(2)} = c_{i_1}$. Compute $q(S, \boldsymbol{c}) = \max\left\{1 + 2\left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor, 2\left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor\right\}$. Divide the entire time horizon $T$ into $q(S, \boldsymbol{c}) + 1$ epochs: $(t_0 : t_1], (t_1 : t_2], \ldots, (t_{q(S,\boldsymbol{c})} : t_{q(S,\boldsymbol{c})+1}]$, where $t_0 = 0$ and

$$t_j = \left\lfloor K^{1 - \frac{2-2^{-(j-1)}}{2-2^{-q(S,\boldsymbol{c})}}} T^{\frac{2-2^{-(j-1)}}{2-2^{-q(S,\boldsymbol{c})}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S, \boldsymbol{c}) + 1.$$

**Policy:** The same as Lines 1 to 9 of Algorithm 6.3.

**Theorem 6.5.** *When $K = \widetilde{\mathcal{O}}(1)$, the optimal regret of D-BwSC is $\widetilde{\Theta}\left(T^{\frac{1}{2-2^{-q(S,\boldsymbol{c})}}}\right)$, where $q(S, \boldsymbol{c})$ is given by Algorithm 6.4. Furthermore, the AS-SE algorithm attains this regret rate.*

Theorem 6.5 shows that when $K = \widetilde{\mathcal{O}}(1)$, the optimal regret rate of D-BwSC can be completely characterized by the key index $q(S, \boldsymbol{c})$, and AS-SE is rate-optimal. This reveals surprising phase transitions similar to Section 6.3.4; the difference is that each phase may not have an equal length anymore. See Table 6.4 for an illustration (the quantities $\Sigma, c^{(1)}, c^{(2)}$ are given in Algorithm 6.4).

Table 6.4: Optimal regret rate (of D-BwSC) under different switching budgets for fixed $K$ and $\boldsymbol{c}$.

| $S$ | $[0, \Sigma)$ | $[\Sigma, 2\Sigma - c^{(1)})$ | $[2\Sigma - c^{(1)}, 3\Sigma - c^{(1)} - c^{(2)})$ | $[3\Sigma - c^{(1)} - c^{(2)}, 4\Sigma - 2c^{(1)} - c^{(2)})$ |
|---|---|---|---|---|
| Rate | $T$ | $T^{2/3}$ | $T^{4/7}$ | $T^{8/15}$ |

## 6.5 Numerical Experiments

In this section, we conduct numerical experiments to show the practicality and effectiveness of our algorithms. We focus on a U-BwSC setting where the reward distribution is $\mathcal{N}(\Delta, 1)$ for the optimal action and is $\mathcal{N}(0, 1)$ for all other actions. We consider $K = 8$, $T \in \{10000, 50000\}$,

and $\Delta \in \{0.1, 0.2, 0.3, \ldots, 1\}$. The ten choices of $\Delta$ correspond to ten different types of underlying environments. For each $T$, for each $\Delta$, we generate 1000 i.i.d. synthetic datasets; hence for each $T$, there are 10000 synthetic datasets in total. We consider 15 different switching budgets ranging from 8 to 22.

We test the performance of four algorithms on the synthetic datasets. The four algorithms include the LS-SE and AdaLS ($\lambda = 0.5$) algorithms proposed in Section 6.3, and two natural heuristics which serve as benchmarks: (1) running the celebrated UCB algorithm (e.g., the UCB1 algorithm of Auer et al. 2002a) until the number of switches exceeds $S$ — we denote this heuristic by UCB, and (2) separating exploration and exploitation (see Algorithm 1.1 of Slivkins 2019) and only making $K - 1$ switches during exploration — this heuristic corresponds to LS-SE with $S = K$, and is denoted by E&E. Our implementations of LS-SE and E&E use randomization to decide the order to explore uneliminated actions in each epoch. For practical purposes, the confidence bounds in (6.4) are modified to $\mathsf{UCB}_i(t) := \bar{\mu}_i(t) + \sqrt{\frac{\gamma \log T}{N_i(t)}}$ and $\mathsf{LCB}_i(t) := \bar{\mu}_i(t) - \sqrt{\frac{\gamma \log T}{N_i(t)}}$, with $\gamma$ being a tuning parameter. In our experiments, we set $\gamma = 0.1$ after fine tuning.

We first fix $T = 10000$. For each switching budget $S \in [8 : 22]$, we test the performance of the four algorithms on the 10000 synthetic datasets, and compute the *empirical average-case regret* by taking an average of the empirical regret incurred over 10000 datasets. The performance of the four algorithms under different switching budgets are presented in Parts (a1) and (a2) of Figure 6-1. The quantity $q$ appearing in Figure 6-1 stands for the quotient index $q(S, K)$.

We make the following observations from Parts (a1) and (a2) of Figure 6-1. First, both LS-SE and AdaLS perform better than the two heuristics (UCB and E&E). UCB performs significantly worse than other algorithms, while AdaLS performs the best overall. Second, the performance of LS-SE only depends on the quotient index $q(S, K)$, irrespective of the remainder index $r(S, K)$. As a result, LS-SE only improves upon E&E when $S$ is large enough such that $q(S, K) \geq 2$. By contrast, AdaLS uniformly improves upon both E&E and LS-SE for all $S \in [8 : 22]$. The superiority of AdaLS demonstrates the value of *adaptivity* in achieving better empirical performance. In particular, learning from data more frequently and adpatively determining the epoch sizes seem to help a lot in our experiments. Third, although AdaLS is already quite adaptive and designed to *behave differently* for different $r(S, K)$, it still seems to face some performance barriers when $q(S, K)$ is fixed, even when $r(S, K)$ is large

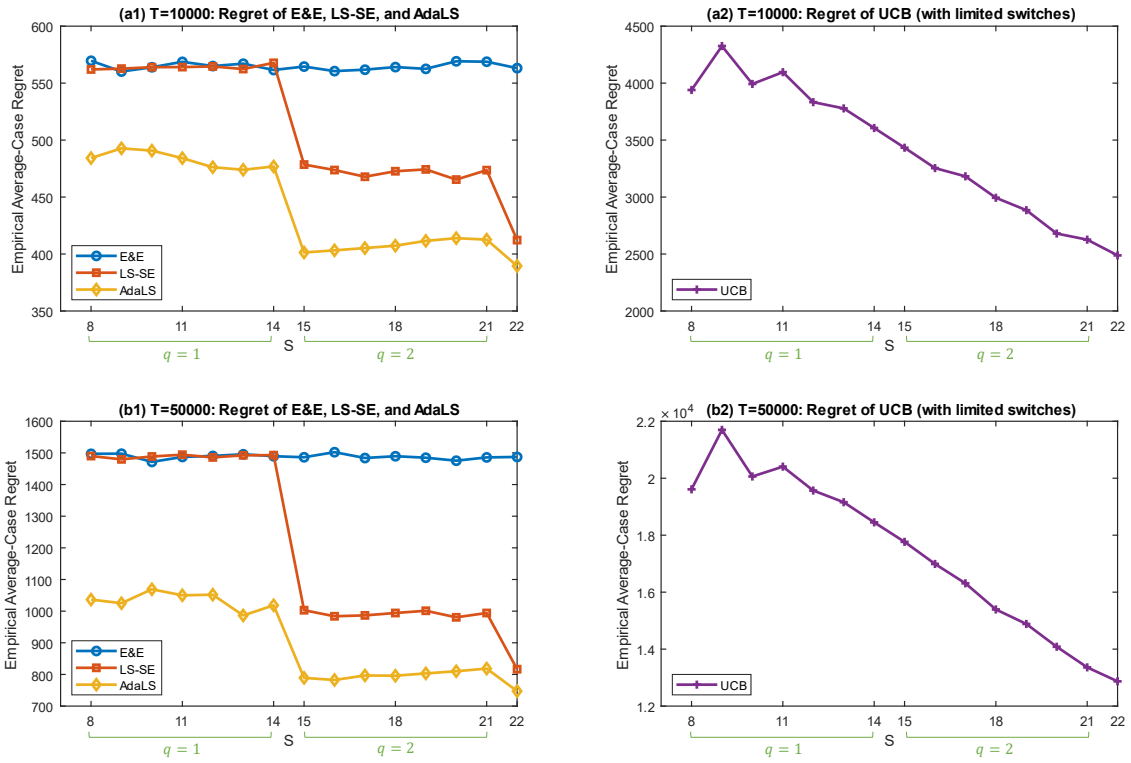


Figure 6-1: Empirical average-case regret v.s. the switching budget $S$, for the four considered algorithms. The regret of UCB has to be plotted separately because it is too large.

(e.g., when $S = 14$ or 21); in our experiments, the most significant performance improvements of AdaLS are observed when there is a new increment in $q(S, K)$ (e.g., when $S = 15$ or 22). Such phenomena are very interesting and can be related to the phase transitions (under a fixed $K$) discussed in Section 6.3.4.

We then repeat the experiments for $T = 50000$ to have robustness checks; see the results plotted in Parts (b1) and (b2) of Figure 6-1. The previous observations still hold. Finally, we repeat all of the above experiments for $K = 4$ (a smaller $K$) and $K = 16$ (a larger $K$) for additional robustness checks. The detailed results are deferred to Appendix E.6.

*Remark.* Although the numerical performance barriers of AdaLS under fixed $q(S, K)$ can be related to phase transitions, they should *not* be seen as direct evidence of phase transitions, because (i) phase transitions concern the optimal worst-case regret optimized over *all* algorithms, while we only compute the empirical average-case regret (averaged over specific instances) of some specific algorithms, and (ii) phase transitions concern the growth rate of the optimal regret as $T$ grows, rather then the numerical value of the regret under a fixed $T$. One might hope to design better numerical experiments to make phase

transitions show up numerically, though it seems some real new idea is needed for designing such experiments (the naive way to numerically compute the optimal regret rate requires enumerations over *all* possible algorithms and *all* instances for a wide range of $T$, which is computationally intractable).

## 6.6 Concluding Remarks

We study the stochastic multi-armed bandit problem with a constraint on the total cost incurred by switching between actions. Under different switching cost structures, we prove matching (or almost matching) upper and lower bounds on regret and provide near-optimal algorithms for the problem. The results enable us to fully characterize the trade-off between the regret rate and the incurred switching cost in the stochastic multi-armed bandit problem, contributing new insights to this fundamental problem. Under the general switching cost structure, the results reveal interesting connections between bandit problems and graph traversal problems, such as the shortest Hamiltonian path problem.

# Chapter 7

# Blind Network Revenue Management and Bandits with Knapsacks Under Limited Switches

## 7.1 Introduction

In this work, we study the classical price-based blind network revenue management (BNRM) problem (Besbes and Zeevi 2012) and its extensions to the bandits with knapsacks (BwK) problem (Badanidiyuru et al. 2018). In the BNRM problem, a firm is endowed with finite inventory of multiple resources to sell over a finite time horizon. The starting inventory is unreplenishable and exogenously given. The firm can control its sales through sequential decisions on the prices offered, which come from a discrete set of candidates[41]. The firm's objective is to maximize its expected cumulative revenue.

We consider the setup in which demand is stochastic, independent and time homogeneous over the time horizon. Yet the distributional information is unknown to the firm, and has to be sequentially learned over the selling horizon. Such a setup is well studied in the literature, see Besbes and Zeevi (2012), Badanidiyuru et al. (2018), Ferreira et al. (2018). In such a setup, there are two sources of trade-offs that the firm needs to consider.

1. The *exploitation-exploration* trade-off. The firm must trade-off between *exploitation* decisions which utilize the learned information to maximize the expected revenue as if it was in the distributionally-known setup, and *exploration* decisions which discover the

[41] In this work, we focus on the discrete price setup of the BNRM problem, instead of the continuous price setup. We explain this distinction in Section 7.2.

demand distributions of the less certain price decisions, regardless of how rewarding they are. Exploitation decisions tend to favor the more rewarding price decisions (with respect to resource constraints), while exploration decisions tend to favor the less discovered price decisions. Intuitively, the optimal policy alternates between exploitation and exploration decisions based on the information learned from the realized demands. Such a trade-off has been extensively studied in the literature of multi-armed bandit (MAB) and dynamic pricing with demand learning; see Bubeck and Cesa-Bianchi (2012), den Boer (2015), Slivkins (2019), Lattimore and Szepesvári (2020) for some overviews and book chapters.

2. The *revenue-inventory* trade-off. Even when the demand distribution has been perfectly learned, the firm faces a trade-off between *revenue-centric* decisions which maximize immediate expected revenue irrespective of resource constraints, and *inventory-centric* decisions which maximize the revenue from the remaining inventory. Revenue-centric decisions tend to be myopic and favor the revenue-maximizing items, while inventory-centric decisions tend to be conservative and favor the highly stocked items. Intuitively, the optimal policy alternates between revenue-centric and inventory-centric decisions based on the remaining inventory and the remaining time periods. Such a trade-off (when separated from the exploration-exploitation trade-off) has been extensively studied in the literature of network revenue management and stochastic control; see Bertsekas (1995), Bitran and Caldentey (2003), Elmaghraby and Keskinocak (2003), Phillips (2005), Talluri and Van Ryzin (2006) for some overviews and book chapters. The revenue-inventory trade-off becomes even more challenging when it is integrated with the exploration-exploitation trade-off; see Besbes and Zeevi (2012), Badanidiyuru et al. (2018), Agrawal (2019) for more discussions.

In the face of the above two trade-offs, any optimal policy must adjust its decisions and instantaneously switch between actions over the time horizon. However, not all firms have the infrastructure to query the realized demand in real-time, to adjust their decisions instantaneously, or to switch between actions as freely as possible. Because changing the posted prices is too costly for many firms (Levy et al. 1998, Zbaracki et al. 2004, Stamatopoulos et al. 2020, Bray and Stamatopoulos 2022), and frequent price changes may confuse the customers (Jørgensen et al. 2003). A common practice for many firms is that they restrict

the number of price changes to be within a budgeted number (Netessine 2006, Chen et al. 2015, Cheung et al. 2017, Perakis and Singhvi 2019, Chen et al. 2020, Simchi-Levi and Xu 2023).

Motivated by this challenge, we analyze the impact of limited switches on the above two trade-offs. In this paper, we primarily consider the classical blind network revenue management (BNRM) problem as described above. We incorporate an additional constraint of limited switching budget into the classical BNRM model and formulate a new problem: blind network revenue management under limited switches (BNRM-LS). For the BNRM-LS problem, we establish tight upper and lower bounds on the optimal regret, and design limited-switch algorithms that achieve the optimal regret rate. Our results reveal a characterization of the optimal regret rate as a function of the switching budget, which further depends on the number of resources.

Moreover, we extend our results to the more general bandits with knapsacks setup (Badanidiyuru et al. 2018, Slivkins and Vaughan 2014) (see Appendices F.1 and F.2). The bandits with knapsacks setup generalizes the blind network revenue management setup in the sense that the reward (revenue) and the costs (consumption of resources) can have an arbitrary relationship, i.e., they are not necessarily connected through demand variables.

### 7.1.1 Contributions

To the best of our knowledge, this paper is one of the first papers to study online learning problems with both resource and switching constraints. In this paper, we formulate and study the BNRM-LS problem, which extends the classical BNRM model by taking into account an additional hard constraint of limited switching budget.

In Appendices F.1 and F.2, we extend our results to the bandits with knapsacks under limited switches (BwK-LS) problem, which generalizes the classical bandits with knapsacks (BwK) problem.

The main contributions of this paper lie in fully characterizing the statistical complexity of the BNRM-LS problem (as well as the generalization to the BwK-LS problem in the appendix), and providing optimal and efficient algorithms that are relevant to practice. We show matching upper and lower bounds on the optimal regret, and design novel limited-switch algorithms to achieve such regret. We use big $O, \Omega, \Theta$ notation to hide constant factors, and use $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$ notation to hide both constant and logarithmic factors. Using the above

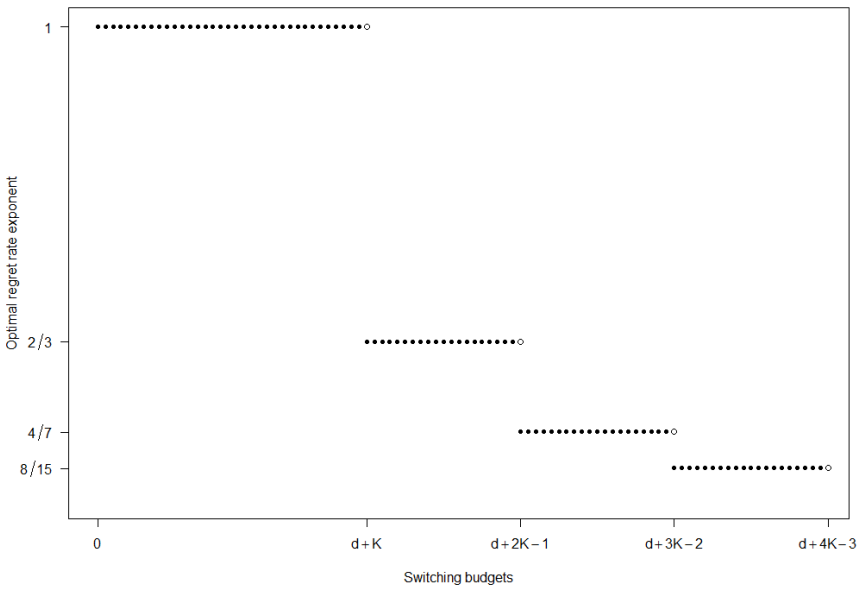


Figure 7-1: Optimal regret rate exponent $\lim_{T\to\infty} \log R^\star(T)/\log T$ as a function of switching budget $s$ in the BNRM-LS problem. Here $R^\star(T)$ stands for the optimal (i.e., minimax) regret.

notation, our main results can be summarized as follows.

1. We provide a computationally efficient limited-switch algorithm and show that the regret is upper bounded by $\widetilde{O}\left(T^{\frac{1}{2-2^{-\nu(s,d)}}}\right)$, where $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$. Here $s$ stands for the switching budget, $d$ for the number of resources, and $K$ for the number of price vectors in the BNRM-LS setup (the number of arms in the BwK-LS setup).

2. We provide matching lower bounds (i.e., impossibility results) on the optimal regret. Specifically, for any algorithm with switching budget $s$, we construct a class of BNRM instances such that the algorithm must suffer $\widetilde{\Omega}\left(T^{\frac{1}{2-2^{-\nu(s,d)}}}\right)$ expected revenue loss on one of these instances.

3. Combining the above upper and lower bounds, we show that the optimal regret is in the order of $\widetilde{\Theta}\left(T^{\frac{1}{2-2^{-\nu(s,d)}}}\right)$. Notably, the optimal regret rate is completely characterized by a piece-wise constant function of the switching budget $s$, which further depends on the number of resources $d$. See Figure 7-1 for an illustration.

Our results lead to the following two implications. First, our results show that a total number of $\Theta(\log\log T)$ switching budget is necessary and sufficient to achieve the optimal $\widetilde{\Theta}(\sqrt{T})$ regret for the classical BNRM problem (and the BwK problem as well). Compared with existing optimal algorithms for BNRM and BwK which require $\Omega(T)$ switching cost in

the worst case, our algorithm achieves a *doubly exponential* (and best possible) improvement on the switching cost.

Second, our results reveal a separation on the optimal regret between the resource-constrained problem (BNRM) and the resource-unconstrained problem (MAB). Under the standard regime where $T$ and $B_{\mathsf{min}}$ are in the same order, prior literature has demonstrated that both MAB and BNRM have the same optimal regret rate $\widetilde{\Theta}(\sqrt{T})$. Our paper shows that when there is a switching budget, the resource-constrained problems can exhibit larger regret rates than the resource-unconstrained problems. We explicitly characterize how the optimal regret rate depends on the number of resource constraints (given any switching budget). If we fix all the other problem primitives unchanged and only add in one more resource constraint, then the optimal regret rate is going to be larger (or the same), illustrating that resource constraints can indeed increase the optimal regret — they make the problem "harder."

In addition, we conduct experiments to examine the performance of our algorithms in a numerical setup that is widely used in the literature[42]. Compared with benchmark algorithms from the literature (Besbes and Zeevi 2012, Badanidiyuru et al. 2018, Ferreira et al. 2018), our proposed algorithms achieve promising performance with clear advantages on the number of incurred switches. Our numerical results also provide practical suggestions to firms on how to design their switching budgets. The detailed numerical results can be found in the full version of our paper (Simchi-Levi et al. 2019).

### 7.1.2 Challenges and Approaches

#### Algorithmic Techniques

We note that it is the co-existence of both resource and switching constraints that make our problems particularly challenging. Indeed, when there is no resource constraint, the topic of switching cost has been well-studied in the online learning literature (for both limited and unlimited switching budgets setups); see Agrawal et al. (1988, 1990), Guha and Munagala (2009), Cesa-Bianchi et al. (2013), Cheung et al. (2017), Perakis and Singhvi (2019), Chen and Chao (2019), Chen et al. (2020), Dong et al. (2020), Simchi-Levi and Xu (2023) for various models and results. In particular, prior literature has established

[42]The setup is a BNRM setup without any switching constraint. We note that when there is a (hard) switching constraint, most existing algorithms would incur $\Omega(T)$ regret, and our algorithms are obviously better.

an "epoch-based elimination" framework (see, e.g., Perchet et al. 2016, Gao et al. 2019) which enables researchers to design optimal algorithms for MAB with switching constraints (Simchi-Levi and Xu 2023). The epoch-based elimination framework divides the time horizon into multiple pre-determined epochs; at each epoch, the algorithm constructs confidence bounds for (the expected reward of) each action, eliminates actions that are obviously sub-optimal, and then uniformly explores each remaining action for an equal number of periods.

However, the above framework cannot be directly applied to BNRM and BwK due to the following two reasons. First, it is hard to eliminate actions in BNRM and BwK, as each action's "value" to the decision maker depends on the remaining inventory and is changing over time (this is because resource constraints are long-term constraints, see Agrawal (2019)). Second, unlike in MAB, uniform exploration over actions is not a good idea in BNRM and BwK, as "choosing actions uniformly" could be arbitrarily worse than "choosing actions non-uniformly" in terms of exploitation, due to the existence of resource constraints (this is because the collected reward is no longer a linear combination of individual actions' rewards, and the decision maker is competing with an optimal dynamic policy rather than an optimal fixed action).

To address the above challenges, we develop a novel "two-stage linear programming" (2SLP) approach in Section 7.4.1, which provides guidance on how to conduct efficient elimination and exploration. This approach has two features. First, the elimination and exploration is conduced over *action combinations* rather than individual actions. Second, the complicated optimization tasks involved in computing the elimination and exploration strategies are reduced to solving $(K+1)$ simple linear programs in two stages. The 2SLP approach builds on and improves upon the algorithmic principle of Balanced Exploration (Badanidiyuru et al. 2018), which is the first optimal algorithm proposed for BNRM and BwK but unfortunately suffers from a severe drawback of computationally inefficiency[43]. Since the 2SLP approach is computationally fast, easy to modify, and provides the first practical generalization of the celebrated *successive elimination* principle from MAB to BNRM and BwK, we believe this approach is of independent interests and has the potential to be a meta-algorithm for efficiently solving more complex resource-constrained online learning

[43]Due to this reason, Balanced Exploration is only considered to be a proof of concept rather than an implementable algorithm (see Remark 4.2 of Badanidiyuru et al. 2018), and has been much less favored in literature compared with subsequent algorithms developed based on primal-dual and UCB principles.

problems.

It is worth mentioning that, without developing 2SLP as a new algorithmic principle for BNRM and BwK, it is not apparent whether one can directly modify existing BNRM and BwK algorithms to obtain efficient limited-switch algorithms. Note that Badanidiyuru et al. (2018) and Immorlica et al. (2019) design computationally efficient algorithms for BwK using adversarial online learning subroutines, which does not seem to work for our purpose as adversarial online learning is shown to require frequent switches (Dekel et al. 2014, Altschuler and Talwar 2021). Note also that by incorporating the delayed update techniques (Auer et al. 2002a) into the UCB-type algorithms (Agrawal and Devanur 2014), one may design a modified UCB-type algorithm that achieves $\widetilde{\Theta}(\sqrt{T})$ regret using $O(\log T)$ switches. This guarantee is exponentially worse than our guarantee, as our algorithm achieves $\widetilde{\Theta}(\sqrt{T})$ regret using only $O(\log \log T)$ switches.

### Lower Bound Techniques

Our lower bound proof builds on the "tracking the cover time" argument of Simchi-Levi and Xu (2023), which establish regret lower bounds for MAB with a single switching constraint by tracking carefully-defined stopping times and constructing hard MAB instances based on (algorithm-dependent) realizations of the stopping times. Extending the argument of Simchi-Levi and Xu (2023) from their resource-unconstrained setting to the resource-constrained setting of BNRM-LS is non-trivial, due to the following two reasons. First, the argument of Simchi-Levi and Xu (2023) critically utilizes the fact that the regret is measured against a single fixed action, but in BNRM-LS the regret is measured against a complex dynamic policy (which itself requires switches). Second, the analysis of Simchi-Levi and Xu (2023) is not sensitive to the number of resource constraints $d$ at all, but in order to match the upper bound of BNRM-LS, we need to establish a strengthened lower bound that gradually increases with $d$. We address the above two challenges by developing an LP-based analysis framework to construct hard BNRM instances with specially-designed resource constraints and demand structures, and measuring several revenue gaps based on clean event analysis of the demand realization process.

It is worth mentioning that no prior work in BNRM and BwK has tried to construct hard instances that involve multiple ($> 1$) resource constraints[44]. Moreover, prior lower

[44]One reason is that using zero or one resource constraint is already sufficient for their purposes.

bound constructions that involve a single resource constraint (Badanidiyuru et al. 2018, Sankararaman and Slivkins 2020) are all in the BwK setup instead of the BNRM setup. Since BwK is more general than BNRM, constructing lower bound examples for BNRM is more challenging than for BwK. All prior constructions break the specific reward-cost structure of BNRM, thus failing to provide lower bounds for BNRM. Compared with prior work, our lower bound instance construction is considerably more complicated, as we have to deal with $d$ resource constraints and we cannot break the BNRM structure.

### 7.1.3 Organization and Notation

We develop our results in the following manner. In Section 7.2 we start with the classical discrete price BNRM model and then introduce its limited switching budget variant BNRM-LS. In Section 7.3 we introduce the deterministic linear program, and present the results in the (simple) distributionally-known case. While the techniques and results in this case are standard, they build intuitions for our main results. In Section 7.4 we introduce our main results in the distributionally-unknown case. We prove matching upper and lower bounds on the optimal regret and provide optimal and efficient algorithms that achieve the optimal regret. We conclude the paper in Section 7.5. All the extensions to the BwK-LS problem are deferred to Appendices F.1 and F.2.

The theoretical results that we present are complemented by an extensive numerical study, which can be found in the full version of our paper (Simchi-Levi et al. 2019).

Let $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_{>0}$ and $\mathbb{R}_{\geq 0}$ be the set of positive integers, real numbers, positive real numbers and non-negative real numbers, respectively. For any $N \in \mathbb{N}$, define $[N] = \{1, 2, ..., N\}$. We use bold font letters for vectors, where we do not explicitly indicate how large the dimension is. For any vector $\boldsymbol{x} \in \mathbb{R}^K$, let $\|\boldsymbol{x}\|_0 = \sum_{k\in[K]} \mathbb{1}\{x_k \neq 0\}$ be the $L_0$ norm of $\boldsymbol{x}$, i.e. the number of non-zero elements in vector $\boldsymbol{x}$. For any vector $\boldsymbol{x} \in \mathbb{R}^K$ and any $k \in [K]$, let $(\boldsymbol{x})_k$ be the $k^{\text{th}}$ element of vector $\boldsymbol{x}$. For any positive real number $x \in \mathbb{R}_{>0}$, let $\lfloor x \rfloor$ be the largest integer that is smaller or equal to $x$; for any non-positive real number $x \in \mathbb{R} \setminus \mathbb{R}_{>0}$, let $\lfloor x \rfloor = 0$. For any set $X$, let $\Delta(X)$ be the set of all probability distributions over $X$. We use big $O, \Omega, \Theta$ notation to hide constant factors, and use $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$ notation to hide both constant and logarithmic factors.

## 7.2 Problem Formulation

From now on in the main paper, we introduce the blind network revenue management (BNRM) model, and defer introducing the bandits with knapsacks (BwK) model to Appendix F.1. The BNRM problem is the online learning version of the classical price-based network revenue management (NRM) problem. The NRM problem is a (full-information) stochastic control problem which originates from the airline industry (Gallego and Van Ryzin 1997, Talluri and Van Ryzin 1998), and has been extensively studied in the revenue management literature (Jasin 2014, Adelman 2007, Topaloglu 2009, Ma et al. 2020) with diverse applications. The BNRM problem extends the classical NRM problem by assuming that the demand distribution is unknown and has to be sequentially learned over time.

The NRM problem and the BNRM problem have two distinct setups: a discrete price setup and a continuous price setup. In this work, we focus on the discrete price setup. We refer to Chen et al. (2019), Chen and Shi (2019), Miao and Wang (2021) for detailed discussions on the continuous price setup.

#### BNRM Setup

Let there be a discrete, finite time horizon with $T$ periods. Time starts from period 1 and ends in period $T$. Let there be $n$ different products generated by $d$ different resources. Each resource is endowed with finite initial inventory $B_i \in \mathbb{R}_{\geq 0}$, $\forall i \in [d]$, and $B_{\mathsf{min}} = \min_{i\in[d]} B_i$. Let $A = (a_{ij})_{i\in[d],j\in[n]}$ be the consumption matrix. Each entry $a_{ij} \in \mathbb{R}_{\geq 0}$ stands for the amount of inventory $i \in [d]$ used, if one unit of product $j \in [n]$ is sold. Let $A_i$ denote the $i$-th row of $A$. Let $a_{\mathsf{max}} = \max_{i,j} a_{ij}$ to be some bounded constant.

In each period $t$, a decision maker can post prices for the $n$ products by selecting a price vector from a finite set of $K$ price vectors $P := \{\boldsymbol{p}_1, ..., \boldsymbol{p}_K\}$, which we denote using $z_t \in [K]$. A price vector is $\boldsymbol{p}_k = (p_{1,k}, ..., p_{n,k})$, and $p_{j,k} \in [0, p_{\mathsf{max}}]$ is the price for product $j$ under $\boldsymbol{p}_k$. This captures situations where a few price points have been pre-determined by market standards, e.g., a common menu of prices that end in \$9.99: \$69.99, \$79.99, \$99.99.

Given price (vector) $\boldsymbol{p}_k$, the demand for each product $j \in [n]$ is an unknown but bounded random variable[45], $Q_{j,k} := Q_j(\boldsymbol{p}_k) \in [0, 1]$, which has to be sequentially learned over time. Let $q_{j,k} := \mathbb{E}[Q_{j,k}]$ denote the unknown mean demand for product $j$ under price $\boldsymbol{p}_k$, and

[45] All our results can be easily extended to the more general sub-Gaussian random variables.

$\boldsymbol{Q} = (Q_{j,k})_{j\in[n],k\in[K]}$, $\boldsymbol{q} = (q_{j,k})_{j\in[n],k\in[K]}$. For each unit of demand generated for product $j \in [n]$ under price vector $\boldsymbol{p}_k$, the decision maker generates $p_{j,k}$ revenue by depleting $a_{ij}$ units of each inventory $i \in [d]$. If no demand is generated, all the remaining inventory is carried over into the next period. The selling process stops immediately when the total cumulative demand of any resource exceeds its initial inventory; see Section F.1.2 for discussions of alternative stopping rules. We use $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$ to stand for a BNRM problem instance.

The objective of the decision maker is to maximize the expected total cumulative revenue (collected before exhausting the resources) over $T$ periods. The performance is measured by the *regret*, which is defined as the worst-case expected revenue loss compared with a clairvoyant decision maker who knows the true demand distributions (but not the realizations). The revenue maximization problem is equivalent to a regret minimization problem.

### Regime for Regret Analysis

We derive non-asymptotic bounds on the regret of policies in terms of the number of time periods $T$. For all of our results (except Theorem 7.1, which we will discuss a different scaling regime), we adopt the following regret analysis regime: there exists an arbitrary constant $\underline{b} > 0$, such that $B_{\mathsf{min}} \geq \underline{b}T$. In other words, we do not assume any specific form of dependence between $T$ and $\boldsymbol{B}$. We only require that inventory is not too scarce compared to the time horizon. This regime generalizes the standard linear scaling regime in the network revenue management literature; see, e.g., Gallego and Van Ryzin (1997), Liu and Van Ryzin (2008), Besbes and Zeevi (2012), Jasin (2014), Ferreira et al. (2018), Chen et al. (2019), Chen and Shi (2019), Bumpensanti and Wang (2020).

Following the literature, we assume $n, p_{\mathsf{max}}, a_{\mathsf{max}}$ are all absolute constants that do not depend on $T$ or $\boldsymbol{B}$. The other parameters $K$ and $d$ do not depend on $T$ or $\boldsymbol{B}$, either. Yet we write out our regret bounds' exact dependence on $K$ and $d$ in our main Theorems and all the proofs, for better managerial insights. Obtaining regret upper and lower bounds that are tight in the orders of $K$ and $d$ is an interesting future direction. For ease of presentation, we also assume that $d < K - 1$. Note that this assumption is only for the purpose of avoiding repeatedly using the notation $\min\{d, K-1\}$; all our results straightforwardly extend to the general case without assuming $d < K - 1$ by replacing $d$ with $\min\{d, K-1\}$ and replacing $\nu(s, d)$ with $\left\lfloor \frac{s-\min\{d,K-1\}-\mathbb{1}\{d<K-1\}}{K-1} \right\rfloor$.

### New Constraint to BNRM

We model the business constraint of limited price changes as a hard constraint, and define the blind network revenue management under limited switches (BNRM-LS) problem as the BNRM problem with an extra constraint of limited switches. Specifically, on top of the initial resource capacities, the decision maker is initially endowed with a fixed number of switching budget $s$, to change the price vector from one to another. When two consecutive price vectors are different, i.e., $z_t \neq z_{t+1}$, one unit of switching budget is consumed. For example, if there is only one product, a sequence of prices $(\$79.99, \$89.99, \$79.99, \$89.99)$ uses two distinct prices, and makes three price changes. When there is no switching budget remaining, the decision maker cannot change the price vector anymore, and has to keep using the last price vector used. There are other ways to model the business constraint of limited switches, but all are beyond the scope of this paper. We can view the BNRM problem as the BNRM-LS problem under an infinite switching budget. Since a limited switching budget restricts the family of admissible policies, any admissible algorithm for the BNRM-LS problem is also an admissible algorithm for the BNRM problem.

### The Impact of Limited Switches

The BNRM problem is extensively studied in the literature, with multiple algorithms developed, e.g., the explore-then-exploit algorithm in Besbes and Zeevi (2012), the Balanced Exploration algorithm in Badanidiyuru et al. (2018), the primal-dual algorithms in Badanidiyuru et al. (2018) and Immorlica et al. (2019), the UCB-type algorithm in Agrawal and Devanur (2014), and the Thompson Sampling algorithm in Ferreira et al. (2018). Under the standard linear scaling regime where $T$ and $B_{\mathsf{min}}$ are in the same order, it has been shown that the optimal regret rate of the BNRM problem is $\widetilde{\Theta}(\sqrt{T})$, which is the same as the optimal regret rate of the classical MAB problem. Existing results thus characterize a relatively complete picture of the statistical complexity and algorithmic principles for stochastic online learning problems with resource constraints.

The new constraint of limited switches, however, has not been explored in the BNRM and BwK literature. Notably, all existing near-optimal algorithms for BNRM and BwK require frequently switching between actions — they all incur $\Omega(T)$ switching cost over $T$ periods. The only exception is the explore-then-exploit algorithm of Besbes and Zeevi (2012), which

controls its number of switches (i.e., price changes) within $K + d$, but unfortunately suffers from $\Omega(T^{2/3})$ regret rate — much worse than $\widetilde{\Theta}(\sqrt{T})$. Prior to our work, general algorithms and regret bounds applicable for an arbitrarily given switching budget remain unknown.

## 7.3 Warm-Up: Network Revenue Management Under Limited Switches

Before we proceed to consider the learning problem, we study the distributionally known case to build better intuitions. In such case, the distributions of $\boldsymbol{Q}$ are known to the decision maker, and the learning problem reduces to a stochastic control problem. In this section, since the distributions are known, the *regret* of any policy (with limited switching budget) refers to the expected revenue loss compared to the optimal policy endowed with unlimited switching budget.[46] Our techniques and results in this section are standard, yet serve as a foundation of Section 7.4.

For any problem instance $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$, we adopt the general notation $\pi : \mathbb{R}^d \times [s] \times [T] \to \Delta([K])$ to denote any policy with full information of $\boldsymbol{Q}$, which suggests a (possibly randomized) price vector to use given the remaining inventory, remaining switching budget, and the remaining periods. For any $s \in \mathbb{N}$, let $\Pi[s]$ be the set of policies that change prices for no more than $s$ times on this problem instance $\mathcal{I}$. For any $s, s' \in \mathbb{N}$ such that $s \leq s'$, we know that $\Pi[s] \subseteq \Pi[s']$. Let $\Pi[\infty]$ be the set of all admissible policies. Let $\mathsf{Rev}(\pi)$ be the expected revenue that policy $\pi$ generates on this problem instance $\mathcal{I}$. Let $\pi^\star[s] \in \arg\max_{\pi \in \Pi[s]} \mathsf{Rev}(\pi)$ be one of the optimal dynamic policies with switching budget $s$, and $\pi^\star[\infty]$ be one of the optimal dynamic policies with an infinite switching budget (i.e., without a switching constraint).

### 7.3.1 The Deterministic Linear Program

For any problem instance $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$, the literature have extensively studied the following deterministic linear program (DLP) in the NRM setup. See Gallego and

[46] Unlike the learning problem where the regret is defined by taking a worst case over $\boldsymbol{Q}$, the regret considered in this section is instance-dependent because $\boldsymbol{Q}$ is already given.

Van Ryzin (1997), Cooper (2002), Maglaras and Meissner (2006), Liu and Van Ryzin (2008).

$$\mathsf{J}^{\mathsf{DLP}} = \max_{(x_1,\ldots,x_K)} \sum_{k\in[K]} \sum_{j\in[n]} p_{j,k}\ q_{j,k}\ x_k \tag{7.1}$$

$$\text{s.t.} \sum_{k\in[K]} \sum_{j\in[n]} a_{ij}\ q_{j,k}\ x_k \le B_i \qquad \forall\ i \in [d] \tag{7.2}$$

$$\sum_{k\in[K]} x_k \le T \tag{7.3}$$

$$x_k \ge 0 \qquad \forall\ k \in [K] \tag{7.4}$$

It is well known that in the NRM setup, the above DLP serves as an upper bound on the expected revenue of any policy, even an optimal policy with an infinite switching budget (i.e., $\pi^\star[\infty]$). When $B_{\mathsf{min}} = \Omega(T)$, it is well known that the gap between the expected revenue obtained by the optimal policy and the DLP upper bound is bounded by $O(\sqrt{T})$ for all instances, i.e., $\mathsf{Rev}(\pi^\star[\infty]) = \mathsf{J}^{\mathsf{DLP}} - O(\sqrt{T})$.

Let the set of optimal solutions to the DLP be

$$X^\star = \arg\max_{\boldsymbol{x}\in\mathbb{R}^K} \{(7.1)\,|(7.2), (7.3), (7.4) \text{ are satisfied}\}.$$

For any vector $\boldsymbol{x} \in \mathbb{R}^K$, let $\|\boldsymbol{x}\|_0 = \sum_{k\in[K]} \mathbb{1}\{x_k \neq 0\}$ be the $L_0$ norm of $\boldsymbol{x}$, i.e. the number of non-zero elements in vector $\boldsymbol{x}$. Let $\Lambda = \min\{\|\boldsymbol{x}\|_0\,|\boldsymbol{x} \in X^\star\}$ be the least number of non-zero variables of any optimal solution. Let $\mathcal{X} = \arg\min\{\|\boldsymbol{x}\|_0\,|\boldsymbol{x} \in X^\star\}$ be the set of such solutions. For any $\boldsymbol{x}^\star \in \mathcal{X}$, let $\mathcal{Z}(\boldsymbol{x}^\star) = \{k \in [K]\,|\boldsymbol{x}^\star \neq 0\} \subseteq [K]$ be the subset of dimensions that are non-zero in $\boldsymbol{x}^\star$. Note that $\Lambda$ is an instance-dependent quantity such that $\Lambda \le d+1$, where $d+1$ is the number of all constraints (resource constraints and time constraint) in the linear program. When DLP is non-degenerate, then equality holds and $\Lambda = d+1$.

In Sections 7.3.2 and 7.3.3 we show that for any problem instance, the instance-dependent quantity $(\Lambda - 1)$ is a critical threshold for the switching budget $s$: if $s \ge \Lambda - 1$, then the optimal regret is $\widetilde{O}(\sqrt{T})$; if $s < \Lambda - 1$, then the optimal regret is $\Theta(T)$.

### 7.3.2 Lower Bound

In this subsection, we show that when the switching budget is below $\Lambda - 1$ (at most $\Lambda - 2$), then a linear regret rate is inevitable. Recall that $\Pi[\Lambda - 2]$ stands for the family of admissible policies that make no more than $\Lambda - 2$ price changes.

**Theorem 7.1.** *Let $\boldsymbol{b} \in \mathbb{R}^d_{>0}$ be any arbitrary vector of positive constants. For any NRM instance $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$ with $\boldsymbol{B} = T \cdot \boldsymbol{b}$, $d \geq 0$, $K > d + 1$, $n \geq 1$, there is an associated $\Lambda$ number (defined in Section 7.3.1), such that any policy $\pi \in \Pi[\Lambda - 2]$ earns an expected revenue:*

$$\mathsf{Rev}(\pi) \leq \mathsf{J}^{\mathsf{DLP}} - c \cdot T,$$

*where $c > 0$ is some distribution-dependent constant that is independent of $T$.*

As a direct implication of Theorem 7.1, we combine the inequality in Theorem 7.1 with the known fact that $\mathsf{Rev}(\pi^\star[\infty]) \geq \mathsf{J}^{\mathsf{DLP}} - O(\sqrt{T})$ and have $\mathsf{Rev}(\pi) \leq \mathsf{Rev}(\pi^\star[\infty]) - \Omega(T)$. That is, the regret scales linearly with $(T, \boldsymbol{B})$ when other parameters are fixed.

The lower bound established in Theorem 7.1 holds for any $\boldsymbol{Q}$. Such a result is much stronger than the worst-case type results which only require finding a single $\boldsymbol{Q}$ that makes the statement hold.

We outline three key steps here. The detailed proof can be found in the full version of our paper (Simchi-Levi et al. 2019). We first identify a clean event, such that the realized demands are close to the expected demands that the LP suggests. This clean event happens with high probability $(1 - \frac{2}{T^3})$. In the second step, conditioning on such event, the maximum amount of revenue we generate is no more than $O(\sqrt{T})$ compared to what the LP suggests; and the minimum amount of inventory demanded is no less than $O(\sqrt{T})$ compared to what the LP suggests, resulting in no more than $O(\sqrt{T})$ of realized revenue. In the third step, we show that the regret from insufficient price changes scales in the order of $\Omega(T)$, which dominates the $O(\sqrt{T})$ amount revenue due to randomness. Such clean event analysis, originating from the online learning literature to prove upper bounds (Badanidiyuru et al. 2018, Slivkins 2019, Lattimore and Szepesvári 2020), was recently used in Arlotto and Gurvich (2019) to prove lower bounds.

### 7.3.3 Upper Bound

In this subsection, we show that when the switching budget is greater or equal to $\Lambda - 1$, then the regret is $\widetilde{O}(\sqrt{T})$. Such a sub-linear guarantee is achieved by tweaking the well-known static control policy in the network revenue management literature (Gallego and Van Ryzin 1997, Cooper 2002, Maglaras and Meissner 2006, Liu and Van Ryzin 2008, Ahn et al. 2021).

We tweak the static control policy, so that with high probability the selling horizon never stops earlier than the last period $T$. See Algorithm 7.1 below[47]. This is achieved by selecting the value of $\gamma$ in the first step of Algorithm 7.1. Similar ideas have been used in Hajiaghayi et al. (2007), Ma et al. (2021), Balseiro et al. (2021) to prove asymptotic results in different setups.

**Algorithm 7.1** Tweaked LP Policy for the NRM Problem

**Input:** $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$.

**Policy:**

1: Define $\gamma = 1 - 2\frac{a_{\max}}{B_{\min}}\sqrt{nT \log T}$.
2: Solve the DLP as defined by (7.1), (7.2), (7.3), and (7.4). Find an optimal solution with the least number of non-zero variables, $\boldsymbol{x}^\star \in \mathcal{X}$.
3: Arbitrarily choose any permutation $\sigma : [\Lambda] \to \mathcal{Z}(\boldsymbol{x}^\star)$ from all $(\Lambda)!$ possibilities.
4: Execute: Set the price vector to be $\boldsymbol{p}_{\sigma(1)}$ for the first $\gamma \cdot x^\star_{\sigma(1)}$ periods, then $\boldsymbol{p}_{\sigma(2)}$ for the next $\gamma \cdot x^\star_{\sigma(2)}$ periods, ..., and finally $\boldsymbol{p}_{\sigma(\Lambda)}$ for the last $T - \gamma \cdot \sum_{l=1}^{\Lambda-1} x^\star_{\sigma(l)}$ periods.

We explain the third step permutation. Suppose $\mathcal{Z}(\boldsymbol{x}^\star) = \{1, 3, 4\}$. In this case, $\Lambda = 3$ and there are 6 permutations. There are 6 possible policies as suggested in Algorithm 7.1. While some of these policies may have better empirical performance than others, they all achieve $\widetilde{O}(\sqrt{T})$ regret.

**Theorem 7.2.** *Let $\underline{b} > 0$ be an arbitrary constant. For any NRM instance $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$ with $T \geq 1, d \geq 0, K > d + 1$ and $B_{\min}/T \geq \underline{b}$, any policy $\pi$ as defined in Algorithm 7.1 satisfies $\pi \in \Pi[\Lambda - 1]$ and earns an expected revenue:*

$$\begin{aligned}\mathsf{Rev}(\pi) &\geq \mathsf{J}^{\mathsf{DLP}} - \max\{c/\underline{b}, c'd\}\sqrt{n^3}\sqrt{T \log T} \\ &\geq \mathsf{Rev}(\pi^\star[\infty]) - \max\{c/\underline{b}, c'd\}\sqrt{n^3}\sqrt{T \log T}\end{aligned}$$

*where $c, c' > 0$ are some absolute constants.*

[47] In Algorithm 7.1, we assume that $x^\star_k, \forall k \in [K]$ are integers, because rounding issues incur a regret of at most $(d \cdot \max_k \boldsymbol{p}_k^\intercal \cdot \boldsymbol{q}_k)$, which is negligible compared with $\sqrt{T}$.

As we will see in Section 7.4, since the loss from an unknown distribution is in the order of $\widetilde{\Omega}(\sqrt{T})$, the $\widetilde{O}(\sqrt{T})$ regret from Algorithm 7.1 suffices to serve as a sub-routine in the last epoch of the main algorithm. Even though there are many advanced techniques that improve the $\widetilde{O}(\sqrt{T})$ result, they are beyond the scope of this paper.

## 7.4 Blind Network Revenue Management Under Limited Switches

In this section, we study the BNRM-LS problem, introduce an efficient algorithm, and provide matching upper and lower bounds of the optimal regret. We start with some definitions.

### Admissible Policies and Clairvoyant Policies

In this section, we distinguish between a BNRM *instance* $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$ and a BNRM *problem* $\mathcal{P} = (T, \boldsymbol{B}, K, d, n, P, A)$ based on whether the underlying demand distributions $\boldsymbol{Q}$ are specified or not. Consider a BNRM problem $\mathcal{P} = (T, \boldsymbol{B}, K, d, n, P, A)$. Let $\phi$ denote any non-anticipating learning policy; specifically, $\phi$ consists of a sequence of (possibly randomized) decision rules $(\phi^t)_{t\in[T]}$, where each $\phi^t$ establishes a probability kernel acting from the space of historical actions and observations in periods $1, \ldots, t-1$ to the space of actions at period $t$. For any $s \in \mathbb{N}$, let $\Phi[s]$ be the set of learning policies that change price vectors for no more than $s$ times almost surely under all possible demand distributions $\boldsymbol{Q}$. For any $s, s' \in \mathbb{N}$ such that $s \leq s'$, $\Phi[s] \subseteq \Phi[s']$. Let $\Phi[\infty]$ be the set of all admissible learning policies. Let $\mathsf{Rev}_{\boldsymbol{Q}}(\phi)$ be the expected revenue that a learning policy $\phi$ generates under demand distributions $\boldsymbol{Q}$.

As we have defined in Section 7.2, $\pi$ refers to a *clairvoyant* policy with full distributional information about the true distributions $\boldsymbol{Q}$. For any $s \in \mathbb{N}$, let $\Pi_{\boldsymbol{Q}}[s]$ be the set of clairvoyant policies that change price vectors for no more than $s$ times under the true distributions $\boldsymbol{Q}$. For any $s, s' \in \mathbb{N}$ such that $s \leq s'$, $\Pi_{\boldsymbol{Q}}[s] \subseteq \Pi_{\boldsymbol{Q}}[s']$. Let $\Pi_{\boldsymbol{Q}}[\infty]$ be the set of all admissible clairvoyant policies. Let $\mathsf{Rev}_{\boldsymbol{Q}}(\pi)$ be the expected revenue that a clairvoyant policy $\pi \in \Pi_{\boldsymbol{Q}}$ generates under distributions $\boldsymbol{Q}$. Let $\pi^\star_{\boldsymbol{Q}}[s] \in \arg\sup_{\pi\in\Pi_{\boldsymbol{Q}}[s]} \mathsf{Rev}(\pi)$ be one optimal clairvoyant policy with switching budget $s$, and $\pi^\star_{\boldsymbol{Q}}$ be one of the optimal dynamic policies with an infinite switching budget (i.e., without a switching constraint).

**Performance Metrics**

The performance of an $s$-switch learning policy $\phi \in \Phi[s]$ is measured against the performance of the optimal $s$-switch clairvoyant policy $\pi^\star_{\boldsymbol{Q}}[s]$. Specifically, for any BNRM problem $\mathcal{P}$ and switching budget $s$, we define the *$s$-switch regret* of a learning policy $\phi \in \Phi[s]$ as the worst-case difference between the expected revenue of the optimal $s$-switch clairvoyant policy $\pi^\star_{\boldsymbol{Q}}[s]$ and the expected revenue of policy $\phi$:

$$R^\phi_s(T) = \sup_{\boldsymbol{Q}} \left\{ \mathsf{Rev}_{\boldsymbol{Q}}(\pi^\star_{\boldsymbol{Q}}[s]) - \mathsf{Rev}_{\boldsymbol{Q}}(\phi) \right\}.$$

We also measure the performance of policy $\phi$ against the performance of the optimal unlimited-switch clairvoyant policy $\pi^\star_{\boldsymbol{Q}}$. Specifically, we define the *overall regret* of a learning policy $\phi \in \Phi[s]$ as the worst-case difference between the expected revenue of the optimal unlimited-switch clairvoyant policy $\pi^\star_{\boldsymbol{Q}}$ and the expected revenue of the policy $\phi$:

$$R^\phi(T) = \sup_{\boldsymbol{Q}} \left\{ \mathsf{Rev}_{\boldsymbol{Q}}(\pi^\star_{\boldsymbol{Q}}) - \mathsf{Rev}_{\boldsymbol{Q}}(\phi) \right\}.$$

Intuitively, the $s$-switch regret $R^\phi_s(T)$ measures the "informational revenue loss" due to not knowing the demand distributions, while the overall regret $R^\phi(T)$ measures the "overall revenue loss" due to not knowing the demand distributions and not being able to switch freely. Clearly, the overall regret $R^\phi(T)$ is always larger than the $s$-switch regret $R^\phi_s(T)$. Interestingly (and quite surprisingly), as we will show later, for all $s$, $R^\phi(T)$ and $R^\phi_s(T)$ are always in the same order in terms of the dependence on $T$.

### 7.4.1 Upper Bound

We propose a computationally efficient algorithm that provides an upper bound on both the $s$-switch regret and the overall regret. Our algorithm, called *Limited-Switch Learning via Two-Stage Linear Programming* (LS-2SLP), is described in Algorithm 7.2.

The design of our algorithm builds on the insights from the LS-SE algorithm proposed in Simchi-Levi and Xu (2023), the Balanced Exploration algorithm proposed in Badanidiyuru et al. (2018), and the Tweaked LP policy defined in Algorithm 7.1. To address the fundamental challenges inherent in our problems (as illustrated in Section 7.1.2), we go beyond the above algorithms and develop novel ingredients for efficient exploration and exploitation under

both resource and switching constraints. We provide more details and insights below.

**Description of the Algorithm**

Our algorithm runs in an epoch schedule which generalizes and improves the epoch schedule of the LS-SE algorithm (Simchi-Levi and Xu 2023). Specifically, our LS-2SLP algorithm first computes a key index $\nu(s,d)$ based on the switching budget $s$, the number of actions $K$, and (importantly) the number of resource constraints $d$, then computes a series of fixed time points $\{t_l\}_{l=1}^{\nu(s,d)+1}$ according to formula (7.5) (see also Perchet et al. 2016, Gao et al. 2019 for the use of such formulas in MAB), which provides important guidance on how to divide the $T$ selling periods into $\nu(s,d)+1$ epochs. Compared with the LS-SE algorithm which directly uses the pre-determined sequence $\{t_l\}_{l=1}^{\nu(s,d)+1}$ as its epoch schedule, our algorithm exhibits two notable differences in determining the epoch schedule. First, the parameter $\nu(s,d)$ takes account of the number of resource constraints $d$. Second, our algorithm uses an *adaptive* epoch schedule $\{T_l\}_{l=1}^{\nu(s,d)+1}$ rather than the pre-determined schedule $\{t_l\}_{l=1}^{\nu(s,d)+1}$ — in particular, our algorithm decides the length of the next epoch only *after* the current epoch ends, and the length would be determined by both $\{t_l\}_{l=1}^{\nu(s,d)+1}$ and the data collected so far. Such an adaptive epoch schedule is crucial for our algorithm to achieve the desired theoretical guarantee[48].

During each epoch except for the last one, our algorithm strikes a balance between exploration and exploitation via a Two-Stage Linear Programming (2SLP) scheme. Specifically, our algorithm first builds high-probability upper and lower confidence bounds on the purchase probability of each price vector, based on the demand data collected so far. Then, the algorithm solves a first-stage pessimistic LP, which is a "pessimistic" variant of the DLP studied in Section 7.3, with the reward of each action being as underestimated as possible and the consumption of each action being as overestimated as possible. Intuitively, the optimal value of this pessimistic LP serves as a conservative estimate on how much accumulated revenue should be generated by a "plausible enough" policy. The algorithm then moves to the second stage, where it solves $K$ linear programs. For any $j \in [K]$, the $j^{\text{th}}$

[48] We provide a more detailed explanation for this point. Almost all existing BNRM and BwK literature assumes a "null arm" with zero reward and zero costs and allows the algorithms to repetitively switch to the null arm. While such an assumption is completely fine (and without loss of generality) in the classical BNRM/ BwK setting, the behavior of repetitively switching to a null arm would cause a waste of switching budget in our limited-switch setting, and this would make the algorithm suboptimal. In order to bypass the need to switch to a null arm, our algorithm has to plan for an "early stopping" of each epoch, which requires the epoch schedule to be adaptive.

**Algorithm 7.2** Limited-Switch Learning via Two-Stage Linear Programming (LS-2SLP)

**Input:** Problem parameters $(T, \boldsymbol{B}, K, d, n, P, A)$; switching budget $s$; discounting factor $\gamma$.

**Initialization:** Calculate $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$. Define $t_0 = 0$ and

$$t_l = \left\lfloor K^{1-\frac{2-2^{-(l-1)}}{2-2^{-\nu(s,d)}}} T^{\frac{2-2^{-(l-1)}}{2-2^{-\nu(s,d)}}} \right\rfloor, \quad \forall l = 1, \ldots, \nu(s,d)+1. \tag{7.5}$$

Set $\gamma = 1 - 3\frac{a_{\max}\sqrt{dn\log[dKT]}\log T}{B_{\min}} t_1$.

**Notation:** Let $T_l$ denote the ending period of epoch $l$ (which will be determined by the algorithm). Let $z_t$ denote the algorithm's selected price vector at period $t$. Let $z_0$ be a random price vector in $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K\}$.

**Policy:**

1: **for** epoch $l = 1, \ldots, \nu(s,d)$ **do**

2: **if** $l = 1$ **then**

3: Set $T_0 = L_k^{\mathsf{rew}}(0) = L_{i,k}^{\mathsf{cost}}(0) = 0$ and $U_k^{\mathsf{rew}}(0) = U_{i,k}^{\mathsf{cost}}(0) = \infty$, $\forall i \in [d], \forall k \in [K]$.

4: **else**

5: Let $n_k(T_{l-1})$ be the total number of periods that price vector $\boldsymbol{p}_k$ is chosen up to period $T_{l-1}$, and $\bar{q}_{j,k}(T_{l-1})$ be the empirical mean demand of product $j$ sold at price vector $\boldsymbol{p}_k$ up to period $T_{l-1}$. Calculate $\nabla_k(T_{l-1}) = \sqrt{\frac{\log[(d+1)KT]}{n_k(T_{l-1})}}$ and

$$\begin{cases} U_k^{\mathsf{rew}}(T_{l-1}) = \min\left\{\sum_{j\in[n]} p_{j,k}\bar{q}_{j,k}(T_{l-1}) + ||\boldsymbol{p}_k||_2 \nabla_k(T_{l-1}), U_k^{\mathsf{rew}}(T_{l-2})\right\}, \\ L_k^{\mathsf{rew}}(T_{l-1}) = \max\left\{\sum_{j\in[n]} p_{j,k}\bar{q}_{j,k}(T_{l-1}) - ||\boldsymbol{p}_k||_2 \nabla_k(T_{l-1}), L_k^{\mathsf{rew}}(T_{l-2})\right\}, \end{cases} \quad \forall k \in [K],$$

$$\begin{cases} U_{i,k}^{\mathsf{cost}}(T_{l-1}) = \min\left\{\sum_{j\in[n]} a_{ij}\bar{q}_{j,k}(T_{l-1}) + ||A_i||_2 \nabla_k(T_{l-1}), U_{i,k}^{\mathsf{cost}}(T_{l-2})\right\}, \\ L_{i,k}^{\mathsf{cost}}(T_{l-1}) = \max\left\{\sum_{j\in[n]} a_{ij}\bar{q}_{j,k}(T_{l-1}) - ||A_i||_2 \nabla_k(T_{l-1}), L_{i,k}^{\mathsf{cost}}(T_{l-2})\right\}, \end{cases} \quad \forall i \in [d], \forall k \in [K].$$

6: Solve the first-stage pessimistic LP:

$$\begin{aligned} \mathsf{J}_l^{\mathsf{PES}} = \max_{(x_1,\ldots,x_K)} & \sum_{k\in[K]} L_k^{\mathsf{rew}}(T_{l-1}) x_k \\ \text{s.t.} & \sum_{k\in[K]} U_{i,k}^{\mathsf{cost}}(T_{l-1}) x_k \le B_i && \forall i \in [d] \\ & \sum_{k\in[K]} x_k \le T \\ & x_k \ge 0 && \forall k \in [K] \end{aligned}$$

7: For each $j \in [K]$, solve the second-stage exploration LP:

$$\begin{aligned}
\boldsymbol{x}^{l,j} = \arg \max_{(x_1,\ldots,x_K)} \; & x_j \\
\text{s.t.} \; \sum_{k\in[K]} & U_k^{\mathsf{rew}}(T_{l-1}) x_k \geq \mathsf{J}_l^{\mathsf{PES}} \\
\sum_{k\in[K]} & L_{i,k}^{\mathsf{cost}}(T_{l-1}) x_k \leq B_i && \forall i \in [d] \\
\sum_{k\in[K]} & x_k \leq T \\
& x_k \geq 0 && \forall\, k \in [K]
\end{aligned}$$

8: For all $k \in [K]$, let $N_k^l = \frac{(t_l - t_{l-1})}{T} \sum_{j=1}^{K} \frac{1}{K} (\boldsymbol{x}^{l,j})_k$. Let $z_{T_{l-1}+1} = z_{T_{l-1}}$. Starting from this action, choose each price vector $\boldsymbol{p}_k$ for $\gamma N_k^l$ *consecutive* periods, $k \in [K]$ (we overlook the rounding issues here, which are easy to fix in regret analysis). Stop the algorithm once time horizon is met or one of the resources is exhausted (in which case we stop the algorithm by keeping selling at the last price vector).

9: End epoch $l$. Mark the last period in epoch $l$ as $T_l$.

10: For epoch $\nu(s,d)+1$ (the last epoch), let $\bar{q}_{j,k}(T_{\nu(s,d)})$ be the empirical mean demand of product $j$ sold at price vector $\boldsymbol{p}_k$ up to period $T_{\nu(s,d)}$. Solve the following deterministic LP

$$\begin{aligned}
\max_{(x_1,\ldots,x_K)} \; & \sum_{k\in[K]} \sum_{j\in[n]} p_{j,k} \; \bar{q}_{j,k}(T_{\nu(s,d)}) \; x_k \\
\text{s.t.} \; & \sum_{k\in[K]} \sum_{j\in[n]} a_{ij} \; \bar{q}_{j,k}(T_{\nu(s,d)}) \; x_k \leq B_i && \forall\; i \in [d] \\
& \sum_{k\in[K]} x_k \leq T \\
& x_k \geq 0 && \forall\; k \in [K],
\end{aligned}$$

and find an optimal solution with the least number of non-zero variables, $\boldsymbol{x}_{\tilde{\boldsymbol{q}}}^{\star}$. Let $N_k^{\nu(s,d)+1} = \frac{(T - t_{\nu(s,d)})}{T} (\boldsymbol{x}_{\tilde{\boldsymbol{q}}}^{\star})_k$ for all $k \in [K]$. First let $z_{T_{\nu(s,d)}+1} = z_{T_{\nu(s,d)}}$. Start from this action, choose each price vector $\boldsymbol{p}_k$ for $\gamma N_k^{\nu(s,d)+1}$ *consecutive* periods, $k \in [K]$ (we overlook the rounding issues here, which are easy to fix in regret analysis). Stop the algorithm once time horizon is met or one of the resources is exhausted (in which case we stop the algorithm by keeping selling at the last price vector). End epoch $\nu(s,d)+1$.

linear program considers how to execute each action, with an objective of exploring action $j$ as many periods as possible, subject to $d$ constraints on the inventory consumption, and an extra constraint on generating at least as much revenue as the pessimistic LP suggests — such a constraint ensures that the exploration of action $j$ cannot be too extensive to hurt the revenue. Contrary to the pessimism in the first stage, all the constraints in the second stage are specified in an "optimistic" manner, with the reward of each action being as overestimated as possible and the consumption of each action being as underestimated as possible — by doing so, the $j^{\text{th}}$ linear program implicitly encourages exploring action $j$ more (while still keeping its solution approximately plausible enough). Finally, our algorithm makes decisions by exploring each arm in a "balanced" fashion, i.e., it computes an average of the $K$ linear programming solutions obtained in the second stage, and then determines the total number of periods to execute each action in this epoch based on the average.

In the very last epoch, our algorithm implements Algorithm 7.1 to conduct pure exploitation, by using the empirical distributions estimated from the data as the stochastic distributions of $\boldsymbol{Q}$. It is worth noting that, in the special case of $\nu(s, d) = 1$, there are only two epochs, and the LS-2SLP algorithm becomes essentially the same as the explore-then-exploit algorithm in Besbes and Zeevi (2012), with some slight difference such as the epoch schedule's dependence on $K$.

### Discussion of the 2SLP Scheme

The 2SLP scheme builds on the insights from the Balanced Exploration algorithm (Badanidiyuru et al. 2018), which extends the celebrated *successive elimination* idea by choosing over *mixtures* of actions (rather than individual actions) and ensuring that the choice is "balanced" across actions. The Balanced Exploration algorithm is however computationally inefficient, as it conducts elimination over mixtures of actions in an explicit and exact manner (which requires one to solve infinite linear programs), and requires an "approximate optimization over a (complicated) infinite-dimensional set" step for which the authors do not provide an implementation.

The 2SLP approach bypasses the computational difficulty inherent in Balanced Exploration by making the elimination procedure *implicit* and *relaxed*, and reducing the hard optimization tasks to simply solving $(K + 1)$ linear programs (which is kind of surprising). Such improvements are brought by our design of the first-stage pessimistic LP and the second-stage

exploration LP's. Interestingly, the 2SLP approach starts with pessimism in the first stage and turns to optimism in the second stage — such a combination of pessimism and optimism seems novel and may be of independent interest (as a comparison, the UCB-type algorithm in Agrawal and Devanur (2014) is fully optimistic).

**Theoretical Guarantees**

The LS-2SLP algorithm is indeed a limited-switch algorithm. During each epoch except for the last one, the LS-2SLP policy chooses at most $K$ actions, thus making at most $K-1$ switches between them. During the last epoch (when the algorithm does purely exploitation), since there are $d+1$ constraints in the deterministic LP, the optimal solution contains at most $d+1$ non-zero solutions. Yet the last action executed during the second last epoch is not necessarily among the non-zero solutions, thus it requires at most $d+1$ switches. There are $\nu(s,d) = \lfloor \frac{s-d-1}{K-1} \rfloor$ epochs before the last exploitation epoch, so there are at most $\nu(s,d) \cdot (K-1) + (d+1) \leq s$ switches, satisfying the definition of the $s$-switch learning policy.

We establish the following upper bound on the regret incurred by the LS-2SLP policy.

**Theorem 7.3.** *Let $\phi$ be the LS-2SLP policy as suggested by Algorithm 7.2. Let $\underline{b} > 0$ be an arbitrary constant. For any BNRM problem $\mathcal{P} = (T, \boldsymbol{B}, K, d, n, P, A)$ with $T \geq 1, d \geq 0, K > d+1$ and $B_{\mathsf{min}}/T \geq \underline{b}$, $\phi$ is guaranteed to be a $s$-switch learning policy, and the regret incurred by $\phi$ satisfies*

$$R_s^{\phi}(T) \leq R^{\phi}(T) \leq \left( \max\{c/\underline{b}, c'\} \cdot n\sqrt{nd \log[dKT]} K^{1-\frac{1}{2-2^{-\nu(s,d)}}} \log T \right) \cdot T^{\frac{1}{2-2^{-\nu(s,d)}}},$$

*where $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$, and $c, c' > 0$ are some absolute constants.*

It is worth noting that the above upper bound holds in a *non-asymptotic* sense: it holds for all finite $T$ and $\boldsymbol{B}$, as long as $B_{\mathsf{min}}/T$ is lower bounded by a positive constant $\underline{b}$. The detailed proof can be found in the full version of our paper (Simchi-Levi et al. 2019).

### 7.4.2 Lower Bound

In this subsection, we prove a matching lower bound, which holds for both the $s$-switch regret and the overall regret. This lower bound is based on the construction of a family of

problem instances in the BNRM-LS setup, combined with non-trivial information-theoretic arguments. Note that our lower bound is established for all $T, \boldsymbol{B}, K, d, n, s$ (under certain weak conditions), which is substantially stronger (and more challenging to prove) than a single lower bounds demonstrated for specific values of $T, \boldsymbol{B}, K, d, n, s$.

**Theorem 7.4.** *Let $\underline{b} > 0$ be an arbitrary constant. For any $T \geq 1, d \geq 0, K \geq 2(d+1), n \geq K(d+1)$ and $\boldsymbol{B}$ such that $B_{\mathsf{min}}/T \geq \underline{b}$, there exist $P, A$, such that for the BNRM problem $\mathcal{P} = (T, \boldsymbol{B}, K, d, n, P, A)$, for any switching budget $s \geq 0$ and any $\phi \in \Phi[s]$,*

$$R^{\phi}(T) \geq R_s^{\phi}(T) \geq \left( \min\{c\underline{b}, c'\} \cdot (d+1)^{-3} K^{-\frac{3}{2} - \frac{1}{2-2^{-\nu(s,d)}}} (\log T)^{-\frac{5}{2}} \right) \cdot T^{\frac{1}{2-2^{-\nu(s,d)}}},$$

*where $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$, and $c, c' > 0$ are some absolute constants.*

The proof of Theorem 7.4 is non-trivial. We outline the proof idea below. The complete proof can be found in the full version of our paper (Simchi-Levi et al. 2019).

Proof Idea. We construct a BNRM problem $\mathcal{P}$ as follows. Let $\boldsymbol{b} = \boldsymbol{B}/T$ (i.e., $b_1 = B_1/T, \ldots, b_d = B_d/T$). Let there be $n \geq K(d+1)$ products. Let the $d \times n$ consumption matrix $A$ be

$$2 \cdot [\underbrace{\mathbf{0}_{d\times 1} \quad \mathrm{diag}(b_1, \ldots, b_d) \quad \mathbf{0}_{d\times 1} \quad \mathrm{diag}(b_1, \ldots, b_d) \quad \cdots \quad \mathbf{0}_{d\times 1} \quad \mathrm{diag}(b_1, \ldots, b_d)}_{K \text{ times}} \quad \mathbf{0}_{d\times(n-K(d+1))}],$$

where $\mathrm{diag}(b_1, \ldots, b_d)$ stands for the $d \times d$ diagonal matrix whose diagonal entries are $b_1, \ldots, b_d$. For any $j \in [n]$, $k \in [K]$, let the price be

$$p_{j,k} = \begin{cases} 1, & \text{if } j = (k-1)(d+1) + 1, \\ 0, & \text{otherwise.} \end{cases}$$

Based on the above BNRM problem $\mathcal{P}$, we will construct different BNRM instances by specifying different demand distributions $\boldsymbol{Q}$.

We prove Theorem 7.4 even when we restrict $\boldsymbol{Q}$ to Bernoulli demand distributions. Recall that we use $q_{j,k} = \mathbb{E}[Q_{j,k}]$ to stand for the mean value of the distribution $Q_{j,k}$. When restricted to Bernoulli distributions, such a $q_{j,k}$ uniquely describes the distribution of $Q_{j,k}$. Thus every $\boldsymbol{q} \in [0,1]^{n\times K}$ uniquely determines a BNRM instance $\mathcal{I}_{\boldsymbol{q}} := (T, \boldsymbol{B}, K, d, n, P, A, s; \boldsymbol{q})$. Specifically, we parameterize $\boldsymbol{q}$ by matrix $\boldsymbol{\mu} = (\mu_{i,k})_{i\in[2],k\in[K]} \in [-\frac{1}{2}, \frac{1}{2}]^{2\times K}$, such that for all

$k \in [K]$ and $j \in [n]$,

$$q_{j,k} = \begin{cases} \frac{1}{2} + \mu_{1,k}, & \text{if } j = (k-1)(d+1)+1, \\ \frac{1}{2} - \mu_{2,k}, & \text{else if } j = (k-1)(d+1) + (k-1)\%(d+1) + 1, \\ \frac{1}{2} + \mu_{2,k}, & \text{else if } j = (k-1)(d+1) + k\%(d+1) + 1, \\ \frac{1}{2}, & \text{else if } j \in [(k-1)(d+1)+1, k(d+1)], \\ 0, & \text{else,} \end{cases}$$

where % stands for the modulo operation (which returns the non-negative remainder of a division). In the lower bound proof, we will assign different values to $\boldsymbol{\mu}$ to construct different BNRM instances. Below we will use $\mathcal{I}_{\boldsymbol{\mu}} := (T, \boldsymbol{B}, K, d, n, P, A, s; \boldsymbol{\mu})$ to stand for a BNRM instance, which highlights the dependence on $\boldsymbol{\mu}$. Let $\mathsf{DLP}_{\boldsymbol{\mu}}$ denote the DLP as defined by (7.1), (7.2), (7.3), (7.4), on instance $\mathcal{I}_{\boldsymbol{\mu}}$:

$$\begin{aligned} \mathsf{J}^{\mathsf{DLP}_{\boldsymbol{\mu}}} = \max_{\boldsymbol{x}} \ & \sum_{k \in [K]} \left( \frac{1}{2} + \mu_{1,k} \right) x_k \\ \text{s.t. } & b_i \left( \frac{1}{2} \sum_{k \in [K]} x_k + \mu_{2,k} \sum_{k' : k'\%(d+1) = i} x_{k'} - \mu_{2,k} \sum_{k'' : k''\%(d+1) = i+1} x_{k''} \right) \le \frac{b_i T}{2}, \quad \forall\, i \in [d], \\ & \sum_{k \in [K]} x_k \le T, \\ & x_k \ge 0, \qquad \forall k \in [K]. \end{aligned}$$

In our analysis, we conduct a thorough analysis of the above type of linear programs, and show that for a family of properly structured $\boldsymbol{\mu}$, $\mathsf{DLP}_{\boldsymbol{\mu}}$ is always non-degenerate — this means that $\Lambda$, the least number of non-zero components of any optimal solution to $\mathsf{DLP}_{\boldsymbol{\mu}}$, is equal to $d+1$. By Theorem 7.1 (which is proved via a clean event analysis of the demand realization process in the distributionally-known setting), for any such BNRM instance $\mathcal{I}_{\boldsymbol{\mu}}$, even for a clairvoyant policy that knows $\boldsymbol{\mu}$ in advance, it needs to make at least $d$ switches to guarantee a sublinear regret (note that $\boldsymbol{\mu}$ is treated as a fixed quantity independent of $T$ in this statement); in our analysis, we strengthen the above statement and show that even when $\boldsymbol{\mu}$ is not fixed and can depend on $T$, any policy still needs to make at least $d$ switch to avoid a large (non-asymptotic) revenue loss. Moreover, since a learning policy $\phi$ does not know the value of $\boldsymbol{\mu}$ in advance, it has to make much more switches than the clairvoyant policy to learn $\boldsymbol{\mu}$. In the rest of the proof, we take into account the effect of unknown demand distributions

and show a lower bound for both $R^{\phi}(T)$ and $R_s^{\phi}(T)$, under any switching budget $s$ — the basic idea is that any limited-switch learning policy faces some fundamental difficulties in distinguishing similar but different $\boldsymbol{\mu}$, which necessarily leads to certain worst-case revenue loss, and we can measure it using certain linear programs.

Specifically, for any $s$-switch learning policy $\phi \in \Phi[s]$, we construct a class of BNRM instances by adversarially choosing a class of $\boldsymbol{\mu}$, such that policy $\phi$ must incur an expected revenue loss of $\widetilde{\Omega}\left(T^{\frac{1}{2-2^{-\nu(s,d)}}}\right)$ under one of the constructed instances. A challenge here is that $\phi$ is an arbitrary and abstract $s$-switch policy — we need more information about $\phi$ to design $\boldsymbol{\mu}$. We address this challenge by developing an enhanced version of the "tracking the cover time" argument. The "tracking the cover time" argument was originally proposed in Simchi-Levi and Xu (2023) to establish regret lower bounds for MAB with limited switches (which can be viewed as a special case of our problem when $d = 0$) by tracking carefully-defined stopping times and constructing hard MAB instances based on (algorithm-dependent) realizations of the stopping times. Since we are proving a larger regret lower bound here (our lower bound gradually increases as $d$ increases), we have to utilize the structure of resource constraints and incorporate their "complexity" into the construction of hard instances which lead to our lower bound. We thus develop novel techniques beyond Simchi-Levi and Xu (2023), i.e., incorporating the complexity of resource constraints into the "tracking the cover time" argument (which requires us to strategically design $\boldsymbol{\mu}$ and utilize several LP benchmarks); see the complete proof in the full version of our paper (Simchi-Levi et al. 2019).

### 7.4.3 Remarkable Implications

#### $O(\log\log T)$ Switches are Sufficient for Optimal BNRM and BwK

Plugging $s = (K-1)\lceil \log_2 \log_2 T \rceil + d + 1$ into Algorithm 7.2 (resp. Algorithm F.1) and Theorem 7.3 (resp. Theorem F.1), we obtain an algorithm that achieves the optimal $\widetilde{\Theta}(\sqrt{T})$ regret for the classical BNRM (resp.BwK) problem, while using only $O(\log\log T)$ switching budget. Note that $\Omega(\log\log T)$ switching budget is necessary for obtaining the $\widetilde{\Theta}(\sqrt{T})$ regret even in the simpler MAB setting, see Simchi-Levi and Xu (2023). Our algorithm and result thus show that a total number of $\Theta(\log\log T)$ switching budget is necessary and sufficient to achieve the optimal $\widetilde{\Theta}(\sqrt{T})$ regret for the classical BNRM (resp. BwK) problem. Compared with existing optimal algorithms that require $\Omega(T)$ switching cost in the worst case, our

algorithm achieves a *doubly exponential* improvement on the switching cost.

### Regret Impact of Resource Constraints

Combining Theorem 7.3 and Theorem 7.4, we can see that for any switching budget $s$, the optimal regret of the BNRM-LS problem is in the order of $\widetilde{\Theta}\left(T^{\overline{2-2^{-\left\lfloor \frac{1}{\frac{s-d-1}{K-1}} \right\rfloor}}}\right)$. This suggests that given a fixed switching budget $s$, an increase in the number of resources $d$ may result in an increase in the order of the optimal regret. To the best our knowledge, this is the first result of such kind that explicitly characterizes how the dimension of the resource constraints makes a stochastic online learning problem "harder". In other words, an increase in the number of resources increases the required number of switches to achieve a given regret rate. Note that both the classical MAB problem and the BNRM problem studied in the literature essentially exhibit the same optimal regret rate in the order of $\widetilde{\Theta}(\sqrt{T})$, where the regret rate is not affected by the dimension of the resource constraints. Our results indicate a separation of the optimal regret rates associated with the dimension of the resource constraints, due to the existence of a switching constraint.

### Managerial Implications

There are two managerial implications regarding our algorithm. First, the length of each epoch is increasing, i.e., more of the price changes occur early in the selling season. This implication suggests managers should be more cautious in the earlier phase of the selling season. Informally, this is because in the earlier epochs, we do not want to incur huge regret by sticking to each price vector for too long. Later on for each epoch, we have relatively better understanding of the underlying demand distributions, so we can be more confident in choosing some effective price vectors for a longer duration of time. Second, our algorithm crucially relies on constructing upper and lower confidence bounds for both revenue and cost of each price vector. We then use such confidence bounds to solve some linear programs, which would suggest which prices to use. Practitioners should keep in mind that for an inventory-constrained problem, it is not necessarily the ratio between revenue and cost, but both revenue and cost that matter.

## 7.5 Concluding Remarks

In this paper we have studied the blind network revenue management problem under the constraint that the decision maker cannot change the price vector for more than a certain number of times. We characterize the best-achievable regret as a function of the switching budget, and provide optimal and efficient algorithms that are relevant to practice. Real-world decision makers can benefit from our study and achieve better performance when they simultaneously face demand uncertainty, resource constraints, and switching constraints.

We conclude this paper by acknowledging two limitations of our paper which could lead to future research questions. First, if we compare Theorem 7.3 and Theorem 7.4, there is a gap in the dependence on $K$ (the number of price vectors). It is an interesting future research question to close this gap. In the "MAB under limited switches" setting, Simchi-Levi and Xu (2023) introduced various techniques to obtain regret bounds that are tight in $K$. However, those techniques do not directly extend to our setting, due to the existence of resource constraints.

Second, we note that in Theorem 7.4 the choices of the parameters are not fully general. We allow for general choices of $\boldsymbol{B}$ (the initial inventory) and $n$ (the number of products), but we cannot allow for general choices of $P$ (the price vectors) and $A$ (the consumption structure). It is an interesting future research question to discuss how general choices of $P$ and $A$ would impact the hardness of the BNRM problem.

# Chapter 8

# Conclusion and Future Plan

This thesis aims to advance the theory and practice of data-driven dynamic decision making, by synergizing ideas from machine learning and operations research. Throughout this thesis, we focus on three aspects: (i) developing new, practical *algorithms* that systematically empower data-driven dynamic decision making, (ii) identifying and utilizing key problem *structures* that lead to statistical and computational efficiency, and (iii) contributing to a general understanding of the statistical and computational *complexity* of data-driven dynamic decision making, which parallels our understanding of supervised machine learning and also accounts for the crucial roles of model structures and constraints for decision making.

Moving forward, there are many interesting future research directions. This chapter highlights three of them, which involve both methodological development and practical implementation.

**Towards more tractable RL by exploiting model structures via operations research** At the end of Section 1.1, we emphasized the importance of problem structures in RL. Operations research has long-established expertise of leveraging fine-grained problem structures to optimize decisions, and has provided us with sophisticated understanding of various complex, stochastic models (e.g., inventory models, queueing networks, transportation systems). Turning our focus to RL (where model parameters are *unknown*), since generic, unstructured RL can be fundamentally harder than supervised learning (c.f. Chapter 4), some natural questions arise: *Can we utilize the expertise of OR to identify broad classes of structured MDPs, whose associated RL tasks can be reduced to supervised learning or similarly tractable tasks? Can we develop a systematic understanding of the roles of different*

*model structures in determining the statistical and computational complexity of RL?*

**Bridging online decision making and offline causal inference** A crucial challenge of (offline) causal inference, beyond (offline) supervised learning, is the requirement of predicting *counterfactual* outcomes for different treatments/interventions. A natural extension of the "online decision making with offline data" framework that we developed in Chapter 5 is "online decision making with observational data," where the offline data are *observational* in the sense that the collection of such data is not well-controlled and there may be unobserved confounding that makes discovering the ground truth more challenging. Given the close relationship between offline causal inference and offline RL, the aspiration to bridge online decision making and offline causal inference can also be expressed as bridging online RL and offline RL. There are many interesting questions: *Can we utilize the pre-existence of (possibly multi-source and confounded) offline observational data to improve online decision making? In a reverse direction, can we use a few online experiments/interactions to help with difficult offline causal inference/offline RL tasks?*

**Broader impact of data-driven decision making in our society** To build impactful data-driven decision-making systems that can operate in the long run, the *societal* consequences of data-driven decision making cannot be overlooked. This motivates the study of the societal aspects of data-driven decision making. Some interesting topics may lie in the interfaces between machine learning and fairness, sustainability, economics, social sciences, etc. A promising direction is to combine data-driven decision making with mechanism design, with a long-term goal of developing better data-driven decision-making systems that can significantly benefit our society.

# Appendix A

# Supplementary Material for Chapter 2

## A.1 Proof of Theorems 2.1 and 2.2

Since Theorem 2.2 is almost a strict generalization of Theorem 2.1 (and Corollary 2.1), we prove them together by conducting a unified analysis. To ensure that the notations are compatible, in some lemmas we will state our results under two different setups, each of which provides consistent definitions of $(\widehat{f}_m)_{m\in\mathbb{N}}$ and $(\gamma_m)_{m\in\mathbb{N}}$ and makes consistent assumptions.

**Setup A.1.** *We consider the learning model in §2.2, where $|\mathcal{F}| < \infty$, $f^\star \in \mathcal{F}$, and all rewards are $[0,1]$-bounded. In this setup, $(\widehat{f}_m)_{m\in\mathbb{N}}$ and $(\gamma_m)_{m\in\mathbb{N}}$ are given by Algorithm 2.1 with $c = 1/30$.*

**Setup A.2.** *We consider the learning model in §2.3 and do not make any assumption on $\mathcal{F}$, $f^\star$, or the reward distribution. We only assume Assumption 2.2 and condition (2.5) for now (as a result, we allow $f^\star \notin \mathcal{F}$, and allow rewards to be unbounded/heavy-tailed). In this setup, $(\widehat{f}_m)_{m\in\mathbb{N}}$ and $(\gamma_m)_{m\in\mathbb{N}}$ are given by Algorithm 2.2 with $c = 1/2$. Moreover, we assume that $\tau_m \geq 2^m$ for $m \in \mathbb{N}$; for such epoch schedules, it is essentially without loss of generality to assume that $(\gamma_m)_{m\in\mathbb{N}}$ are non-decreasing.*

As we can see, Setup A.2 involves a much more general learning model. However, Setup A.1 allows $(\widehat{f}_m)_{m\in\mathbb{N}}$ to be generated with data reuse (i.e., one can feed the data collected in all previous epochs into the offline regression oracle) and does not require the epoch length to grow geometrically. One can understand Setup A.1 as an example of utilizing martingale concentration results (Lemma A.1) obtained under additional model assumptions to show some additional properties.

Before we start our proof of Theorem 2.1 (under Setup A.1) and Theorem 2.2 (under Setup A.2), we make an important remark regarding our presentation. In the main part of this section (Appendix A.1.1 to Appendix A.1.6), we give a detailed proof (of both theorems) which closely follows the proof sketch in Section 2.4, *under the condition that* $|\mathcal{X}| < \infty$. Such a condition enables us to give an insightful proof without the need to worry about measurability issues. However, when $\mathcal{X}$ is infinite or (more generally) uncountable, one has to deal with the measurability issues arising from the presence of uncountable probability spaces. While rigorous discussions of measurability issues are usually omitted in the contextual bandit literature (for brevity or for simplicity), we feel that a discussion of such issues is important here due to our extensive use of the *universal policy space* $\mathcal{A}^{\mathcal{X}}$, which

would easily contain non-measurable policies when $\mathcal{X}$ is uncountable. Therefore, in the last part of this section (Appendix A.1.7), we consider general uncountable $\mathcal{X}$ and present a simple fix to the associated measurability issues, showing that our results are indeed general.

### A.1.1 Definitions

For notational convenience, we make some definitions. Some of the definitions have appeared in the main article. For all $t = 1, \ldots, T$, we let $\Upsilon_t := \sigma((x_1, r_1, a_1), \cdots, (x_t, r_t, a_t))$ denote the sigma-algebra generated by the history up to round $t$ (inclusive), and let $m(t) := \min\{m \in \mathbb{N} : t \leq \tau_m\}$ denote the epoch that round $t$ belongs to. Let $\Psi := \mathcal{A}^{\mathcal{X}}$ be the *universal policy space* (which is defined via taking the *Cartesian product*). For any action selection kernel $p$ and any policy $\pi \in \Psi$, define

$$V(p, \pi) := \mathbb{E}_{x \sim \mathcal{D}} \left[ \frac{1}{p(\pi(x) \mid x)} \right],$$

$$\mathcal{V}_t(\pi) := \max_{1 \leq m \leq m(t)-1} \{V(p_m, \pi)\}.$$

For any $\pi \in \Psi$, define

$$\mathcal{R}(\pi) := \mathbb{E}_{x \sim \mathcal{D}} \left[ f^\star(x, \pi(x)) \right],$$

$$\widehat{\mathcal{R}}_t(\pi) := \mathbb{E}_{x \sim \mathcal{D}} \left[ \widehat{f}_{m(t)}(x, \pi(x)) \right],$$

$$\mathrm{Reg}(\pi) := \mathcal{R}(\pi_{f^\star}) - \mathcal{R}(\pi),$$

$$\widehat{\mathrm{Reg}}_t(\pi) := \widehat{\mathcal{R}}_t(\pi_{\widehat{f}_{m(t)}}) - \widehat{\mathcal{R}}_t(\pi).$$

### A.1.2 High-probability Events

Lemma A.1 and Lemma A.2 present some basic concentration results.

**Lemma A.1.** *Consider Setup A.1. For all $m \geq 2$, with probability at least $1 - \delta/(2m^2)$, we have:*

$$\begin{aligned}
&\sum_{t=1}^{\tau_{m-1}} \mathbb{E}_{x_t, a_t} \left[ (\widehat{f}_m(x_t, a_t) - f^\star(x_t, a_t))^2 \mid \Upsilon_{t-1} \right] \\
&= \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_{x_t, r_t, a_t} \left[ (\widehat{f}_m(x_t, a_t) - r_t(a_t))^2 - (f^\star(x_t, a_t) - r_t(a_t))^2 \mid \Upsilon_{t-1} \right] \\
&\leq 100 \log \left( \frac{2|\mathcal{F}| m^2 \log_2(\tau_{m-1})}{\delta} \right) \leq 225 \log \left( \frac{|\mathcal{F}| m \log(\tau_{m-1})}{\delta} \right) = \frac{K}{4\gamma_m^2}.
\end{aligned}$$

*Therefore (by a union bound), the following event $\Gamma_1$ holds with probability at least $1 - \delta/2$:*

$$\Gamma_1 := \left\{ \forall m \geq 2, \quad \frac{1}{\tau_{m-1}} \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_{x_t, a_t} \left[ (\widehat{f}_m(x_t, a_t) - f^\star(x_t, a_t))^2 \mid \Upsilon_{t-1} \right] \leq \frac{K}{4\gamma_m^2} \right\}.$$

Lemma A.1 follows from Lemma 4.1 (see equation (4.1) therein) and Lemma 4.2 in Agarwal et al. (2012), and we omit the proof here (compared to their proof, we just slightly change the way of taking union bounds, and plug in the definition of $\gamma_m$).

**Lemma A.2.** *Consider Setup A.2. For all $m \geq 2$, with probability at least $1 - \delta/(2m^2)$, we have:*

$$\forall t \in \{\tau_{m-2}+1, \cdots, \tau_{m-1}\},$$
$$\mathbb{E}_{x_t,a_t}\left[(\widehat{f}_m(x_t,a_t) - f^\star(x_t,a_t))^2 \mid \Upsilon_{t-1}\right] \leq \mathcal{E}_{\mathcal{F},\delta/(2m^2)}(\tau_{m-1}-\tau_{m-2}) = \frac{K}{4\gamma_m^2}.$$

*Therefore (by a union bound), the following event $\Gamma_2$ holds with probability at least $1-\delta/2$:*

$$\Gamma_2 := \left\{\forall m \geq 2 \text{ and } t \text{ in epoch } m, \quad \mathbb{E}_{x_t,a_t}\left[(\widehat{f}_m(x_t,a_t) - f^\star(x_t,a_t))^2 \mid \Upsilon_{t-1}\right] \leq \frac{K}{4\gamma_m^2}\right\}.$$

*Proof of Lemma A.2.* Note that Algorithm 2.2 always sends $(x_t, a_t; r_t(a_t))$-type data to $\texttt{OffReg}_{\mathcal{F}}$, where $(x_t, r_t) \sim \mathcal{D}$ and $a_t \sim p_{m(t)-1}(\cdot \mid x_t)$. By Assumption 2.2 and the specification of Algorithm 2.2, we have $\forall t \in \{\tau_{m-2}+1, \cdots, \tau_{m-1}\}$,

$$\begin{aligned}\mathbb{E}_{x_t,a_t}\left[(\widehat{f}_m(x_t,a_t) - f^\star(x_t,a_t))^2 \mid \Upsilon_{t-1}\right] &= \mathbb{E}_{x_t\sim\mathcal{D},a_t\sim p_{m-1}(\cdot|x_t)}\left[(\widehat{f}_m(x_t,a_t) - f^\star(x_t,a_t))^2 \mid p_{m-1}\right] \\ &\leq \mathcal{E}_{\mathcal{F},\delta/(2m^2)}(\tau_{m-1}-\tau_{m-2}) = K/(4\gamma_m^2),\end{aligned}$$

where the inequality simply follows from Assumption 2.2. □

### A.1.3 Per-Epoch Properties of the Algorithm

We start to utilize the specific properties of our algorithm's action selection kernels to prove our regret bound. We start with some per-epoch properties that always hold for our algorithm regardless of its performance in other epochs. All results in this part hold for both Setup A.1 and Setup A.2.

As we mentioned in the main article, a starting point of our proof is to translate the action selection kernel $p_m(\cdot \mid \cdot)$ into an equivalent distribution over policies $Q_m(\cdot)$. Lemma A.3 provides a justification of such translation by showing the existence of an equivalent $Q_m$ for every $p_m(\cdot \mid \cdot)$.

**Lemma A.3.** *Fix any epoch $m \in \mathbb{N}$. The action selection scheme $p_m(\cdot \mid \cdot)$ is a valid probability kernel $\mathcal{B}(\mathcal{A}) \times \mathcal{X} \to [0,1]$. There exists a probability measure $Q_m$ on $\Psi$ such that*

$$\forall a \in \mathcal{A}, \forall x \in \mathcal{X}, \quad p_m(a \mid x) = \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x) = a\} Q_m(\pi).$$

*Proof of Lemma A.3.* This proof is straightforward when $|\mathcal{X}| < \infty$. Since $(\mathcal{A}, \mathcal{B}(\mathcal{A}), p_m(\cdot \mid x))$ is a probability space for each $x \in \mathcal{X}$, by the existence and uniqueness of finite product probability measures, there exists a unique probability measure $Q_m := \prod_{x\in\mathcal{X}} p_m(\cdot \mid x)$ on $(\Psi, (\mathcal{B}(\mathcal{A}))^{\mathcal{X}}) = (\mathcal{A}^{\mathcal{X}}, \mathcal{B}(\mathcal{A})^{\mathcal{X}})$ with the property that

$$Q_m\left(\prod_{x\in\mathcal{X}} E_x\right) = \prod_{x\in\mathcal{X}} p_m(E_x \mid x)$$

whenever one has $E_x \in \mathcal{B}(\mathcal{A})$ for all $x \in \mathcal{X}$. For any $a_0 \in \mathcal{A}, x_0 \in \mathcal{X}$, by letting $E_{x_0} = \{a_0\}$ and $E_x = \mathcal{A}$ for all $x \neq x_0$, we have $p_m(a_0 \mid x_0) = Q_m(\{\pi : \pi(x_0) = a_0\}) = \sum_{\pi\in\Psi} \mathbb{I}\{\pi(x_0) = a_0\} Q_m(\pi)$.

*Remark.* In fact, Lemma A.3 generally holds for an arbitrary (possibly uncountable) $\mathcal{X}$, if the

equation $p_m(a \mid x) = \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x) = a\} Q_m(\pi)$ in the statement is replaced by its more general form $p_m(a \mid x) = \mathbb{E}_{\pi \sim Q_m}[\mathbb{I}\{\pi(x) = a\}]$. Such a result can be easily obtained by applying the Kolmogorov extension theorem (see, e.g., Theorem 2.4.4 in Tao 2011). □

We call the $Q_m$ determined in the proof of Lemma A.3 the "equivalent randomized policy" induced by $p_m(\cdot \mid \cdot)$. Lemma A.4 states a key property of $Q_m$: the expected instantaneous regret incurred by $p_m(\cdot \mid \cdot)$ is equal to the implicit regret of the randomized policy $Q_m$. Thus, to analyze our algorithm's expected regret, we only need to analyze the induced randomized policies' implicit regret.

**Lemma A.4.** *Fix any epoch $m \in \mathbb{N}$, for any round $t$ in epoch $m$, we have*

$$\mathbb{E}_{x_t, r_t, a_t}\left[r_t(\pi_{f^\star}) - r_t(a_t) \mid \Upsilon_{t-1}\right] = \sum_{\pi \in \Psi} Q_m(\pi) \mathrm{Reg}(\pi).$$

*Proof of Lemma A.4.* By Lemma A.3, we have

$$\begin{aligned}
\mathbb{E}_{x_t, r_t, a_t}\left[r_t(\pi_{f^\star}) - r_t(a_t) \mid \Upsilon_{t-1}\right] &= \mathbb{E}_{x_t, a_t}\left[f^\star(x_t, \pi_{f^\star}(x_t)) - f^\star(x_t, a_t) \mid \Upsilon_{t-1}\right] \\
&= \mathbb{E}_{x \sim \mathcal{D}, a \sim p_m(\cdot \mid x)}\left[f^\star(x, \pi_{f^\star}(x)) - f^\star(x, a)\right] \\
&= \mathbb{E}_x\left[\sum_{a \in \mathcal{A}} p_m(a \mid x)\left(f^\star(x, \pi_{f^\star}(x)) - f^\star(x, a)\right)\right] \\
&= \mathbb{E}_x\left[\sum_{a \in \mathcal{A}} \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x) = a\} Q_m(\pi)\left(f^\star(x, \pi_{f^\star}(x)) - f^\star(x, a)\right)\right] \\
&= \mathbb{E}_x\left[\sum_{\pi \in \Psi} Q_m(\pi)\left(f^\star(x, \pi_{f^\star}(x)) - f^\star(x, \pi(x))\right)\right] \\
&= \sum_{\pi \in \Psi} Q_m(\pi)\, \mathbb{E}_x\left[f^\star(x, \pi_{f^\star}(x)) - f^\star(x, \pi(x))\right] \\
&= \sum_{\pi \in \Psi} Q_m(\pi) \mathrm{Reg}(\pi).
\end{aligned}$$

□

Lemma A.5 states another key property of $Q_m$. It says that $Q_m$ controls its predicted implicit regret (relative to the greedy policy based on $\widehat{f}_m$) within $K/\gamma_m$. Note that the controlled error $K/\gamma_m$ is gradually shrinking as the algorithm finishes more epochs.

**Lemma A.5.** *Fix any epoch $m \in \mathbb{N}$, for any round $t$ in epoch $m$, we have*

$$\sum_{\pi \in \Psi} Q_m(\pi) \widehat{\mathrm{Reg}}_t(\pi) \leq \frac{K}{\gamma_m}.$$

*Proof of Lemma A.5.* We have

$$\begin{aligned}\sum_{\pi\in\Psi} Q_m(\pi)\widehat{\mathrm{Reg}}_t(\pi) &= \sum_{\pi\in\Psi} Q_m(\pi)\,\mathbb{E}_{x\sim\mathcal{D}}\left[\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,\pi(x))\right]\\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{\pi\in\Psi} Q_m(\pi)\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,\pi(x))\right)\right]\\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}\sum_{\pi\in\Psi} \mathbb{I}\{\pi(x)=a\}Q_m(\pi)\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,a)\right)\right]\\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}} p_m(a\mid x)\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,a)\right)\right].\end{aligned}$$

Given any context $x\in\mathcal{X}$,

$$\begin{aligned}\sum_{a\in\mathcal{A}} p_m(a\mid x)\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,a)\right) &= \sum_{a\neq\widehat{a}_m(x)} \frac{\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,a)}{K+\gamma_m\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,a)\right)}\\ &\le \frac{K-1}{\gamma_m}.\end{aligned}$$

Lemma A.5 follows immediately. □

Lemma A.6 states another key per-epoch property of our algorithm. For any deterministic policy $\pi\in\Psi$, the quantity $V(p_m,\pi) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_m(\pi(x)|x)}\right]$ is the expected inverse probability that the algorithm's decision generated by $p_m$ (i.e., the decision generated by the randomized policy $Q_m$) is the same as the decision generated by the deterministic policy $\pi$, over the randomization of context $x$. This can be intuitively understood as a measure of the "decisional divergence" between the randomized policy $Q_m$ and the deterministic policy $\pi$. Lemma A.6 states that this divergence can be bounded by the predicted implicit regret of policy $\pi$.

**Lemma A.6.** *Fix any epoch $m\in\mathbb{N}$, for any round $t$ in epoch $m$, for any policy $\pi\in\Psi$,*

$$V(p_m,\pi)\le K+\gamma_m\widehat{\mathrm{Reg}}_t(\pi).$$

*Proof of Lemma A.6.* For any policy $\pi\in\Psi$, given any context $x\in\mathcal{X}$,

$$\frac{1}{p_m(\pi(x)\mid x)}\begin{cases} = K+\gamma_m\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,\pi(x))\right), & \text{if } \pi(x)\neq\widehat{a}_m(x);\\ \le \frac{1}{1/K} = K = K+\gamma_m\left(\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,\pi(x))\right), & \text{if } \pi(x)=\widehat{a}_m(x).\end{cases}$$

Thus

$$\begin{aligned}V(p_m,\pi) &= \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_m(\pi(x)\mid x)}\right]\\ &\le K+\gamma_m\,\mathbb{E}_{x\sim\mathcal{D}}\left[\widehat{f}_m(x,\widehat{a}_m(x)) - \widehat{f}_m(x,\pi(x))\right]\\ &= K+\gamma_m\widehat{\mathrm{Reg}}_t(\pi)\end{aligned}$$

for every round $t$ in epoch $m$. $\square$

### A.1.4 Bounding the Prediction Error of Implicit Rewards

Lemma A.7 relates the prediction error of the implicit reward of any policy $\pi$ at round $t$ to the value of $\mathcal{V}_t(\pi)$. Recall that $\Gamma_1$ and $\Gamma_2$ are defined in Appendix A.1.2.

**Lemma A.7.** *Assume $\Gamma_1$ (resp. $\Gamma_2$) holds under Setup A.1 (resp., Setup A.2). For any round $t > \tau_1$, for any $\pi \in \Psi$,*

$$|\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)| \le \frac{\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{2\gamma_{m(t)}}.$$

*Proof of Lemma A.7.* Fix any policy $\pi \in \Psi$, and any round $t > \tau_1$. By the definitions of $\widehat{\mathcal{R}}_t(\pi)$ and $\mathcal{R}(\pi)$, we have

$$\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\left[\widehat{f}_{m(t)}(x,\pi(x)) - f^\star(x,\pi(x))\right].$$

Given a context $x$, define

$$\Delta_x = \widehat{f}_{m(t)}(x,\pi(x)) - f^\star(x,\pi(x)),$$

then $\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi) = \mathbb{E}_{x\sim\mathcal{D}}[\Delta_x]$. For all $s = 1, 2, \ldots, \tau_{m(t)-1}$, we have

$$\begin{aligned}
\mathbb{E}_{a_s|x_s}\left[\left(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s)\right)^2 \mid \Upsilon_{s-1}\right] &= \sum_{a\in\mathcal{A}} p_{m(s)}(a \mid x_s)\left(\widehat{f}_{m(t)}(x_s,a) - f^\star(x_s,a)\right)^2 \\
&\ge p_{m(s)}(\pi(x_s) \mid x_s)\left(\widehat{f}_{m(t)}(x_s,\pi(x_s)) - f^\star(x_s,\pi(x_s))\right)^2 \\
&= p_{m(s)}(\pi(x_s) \mid x_s)\left(\Delta_{x_s}\right)^2. \qquad \text{(A.1)}
\end{aligned}$$

Thus for both $s_0 = 1$ and $s_0 = \tau_{m(t)-2} + 1$, we have

$$\begin{aligned}
&\mathcal{V}_t(\pi)\sum_{s=s_0}^{\tau_{m(t)-1}} \mathbb{E}_{x_s,a_s}\left[(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s))^2 \mid \Upsilon_{s-1}\right] \\
&\overset{\text{(i)}}{\ge} \sum_{s=s_0}^{\tau_{m(t)-1}} V(p_{m(s)},\pi)\,\mathbb{E}_{x_s,a_s}\left[(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s))^2 \mid \Upsilon_{s-1}\right] \\
&= \sum_{s=s_0}^{\tau_{m(t)-1}} \mathbb{E}_{x_s}\left[\frac{1}{p_{m(s)}(\pi(x_s) \mid x_s)}\right]\mathbb{E}_{x_s}\,\mathbb{E}_{a_s|x_s}\left[\left(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s)\right)^2 \mid \Upsilon_{s-1}\right] \\
&\overset{\text{(ii)}}{\ge} \sum_{s=s_0}^{\tau_{m(t)-1}} \left(\mathbb{E}_{x_s}\left[\sqrt{\frac{1}{p_{m(s)}(\pi(x_s) \mid x_s)}\,\mathbb{E}_{a_s|x_s}\left[\left(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s)\right)^2 \mid \Upsilon_{s-1}\right]}\right]\right)^2 \\
&\overset{\text{(iii)}}{\ge} \sum_{s=s_0}^{\tau_{m(t)-1}} \left(\mathbb{E}_{x_s}\left[\sqrt{\frac{1}{p_{m(s)}(\pi(x_s) \mid x_s)} p_{m(s)}(\pi(x_s) \mid x_s)\left(\Delta_{x_s}\right)^2}\,\right]\right)^2 \\
&= \sum_{s=s_0}^{\tau_{m(t)-1}} \left(\mathbb{E}_{x_s}\left[|\Delta_{x_s}|\right]\right)^2 \\
&\overset{\text{(iv)}}{\ge} \sum_{s=s_0}^{\tau_{m(t)-1}} |\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)|^2 = (\tau_{m(t)-1} - s_0 + 1)|\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)|^2,
\end{aligned}$$

where (i) follows from the definition of $\mathcal{V}_t(\pi)$, (ii) follows from the Cauchy-Schwarz inequality, (iii) follows from (A.1), and (iv) follows from the convexity of the $\ell_1$ norm.

If we are in Setup A.1 and $\Gamma_1$ holds, then by letting $s_0 = 1$, we have

$$\begin{aligned}|\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)| &\le \sqrt{\mathcal{V}_t(\pi)}\sqrt{\frac{\sum_{s=1}^{\tau_{m(t)-1}} \mathbb{E}_{x_s,a_s}\left[(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s))^2 \mid \Upsilon_{s-1}\right]}{\tau_{m(t)-1}}} \\ &\le \frac{\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{2\gamma_{m(t)}},\end{aligned}$$

where the last inequality follows from the definition of $\Gamma_1$. If we are in Setup A.2 and $\Gamma_2$ holds, then by letting $s_0 = \tau_{m(t)-2} + 1$, we have

$$\begin{aligned}|\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)| &\le \sqrt{\mathcal{V}_t(\pi)}\sqrt{\frac{\sum_{s=\tau_{m(t)-2}+1}^{\tau_{m(t)-1}} \mathbb{E}_{x_s,a_s}\left[(\widehat{f}_{m(t)}(x_s,a_s) - f^\star(x_s,a_s))^2 \mid \Upsilon_{s-1}\right]}{\tau_{m(t)-1} - \tau_{m(t)-2}}} \\ &\le \frac{\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{2\gamma_{m(t)}},\end{aligned}$$

where the last inequality follows from the definition of $\Gamma_2$. □

### A.1.5 Bounding the Prediction Error of Implicit Regret

Lemma A.8 establishes an important relationship between the predicted implicit regret and the true implicit regret of any policy $\pi$ at round $t$. This lemma ensures that the predicted implicit regret of "good policies" are becoming more and more accurate, while the predicted implicit regret of "bad policies" do not need to have such property.

Recall that $\Gamma_1$ and $\Gamma_2$ are defined in Appendix A.1.2.

**Lemma A.8.** *Assume that $\Gamma_1$ (resp. $\Gamma_2$) holds under Setup A.1 (resp. Setup A.2). Let $c_0 := 5.15$. For all epochs $m \in \mathbb{N}$, all rounds $t$ in epoch $m$, and all policies $\pi \in \Psi$,*

$$\mathrm{Reg}(\pi) \le 2\widehat{\mathrm{Reg}}_t(\pi) + c_0 K/\gamma_m,$$

$$\widehat{\mathrm{Reg}}_t(\pi) \le 2\mathrm{Reg}(\pi) + c_0 K/\gamma_m.$$

*Proof of Lemma A.8.* We prove Lemma A.8 via induction on $m$. We first consider the base case where $m = 1$ and $1 \le t \le \tau_1$. In this case, since $\gamma_1 = 1$, we know that $\forall \pi \in \Psi$,

$$\mathrm{Reg}(\pi) \le 1 \le c_0 K/\gamma_1, \quad \widehat{\mathrm{Reg}}_t(\pi) \le 1 \le c_0 K/\gamma_1; \qquad \text{(under Setup A.1)}$$

$$\mathrm{Reg}(\pi) \le \sqrt{K} \le c_0 K/\gamma_1, \quad \widehat{\mathrm{Reg}}_t(\pi) = 0 \le c_0 K/\gamma_1. \qquad \text{(under Setup A.2)}$$

Note that we use condition (2.5) for Setup A.2 here. Thus the claim holds in the base case.

For the inductive step, fix some epoch $m > 1$. We assume that for all epochs $m' < m$, all rounds

$t'$ in epoch $m'$, and all $\pi \in \Psi$,

$$\mathrm{Reg}(\pi) \le 2\widehat{\mathrm{Reg}}_{t'}(\pi) + c_0 K/\gamma_{m'}, \tag{A.2}$$

$$\widehat{\mathrm{Reg}}_{t'}(\pi) \le 2\mathrm{Reg}(\pi) + c_0 K/\gamma_{m'}. \tag{A.3}$$

We first show that for all rounds $t$ in epoch $m$ and all $\pi \in \Psi$,

$$\mathrm{Reg}(\pi) \le 2\widehat{\mathrm{Reg}}_{t}(\pi) + c_0 K/\gamma_{m}.$$

We have

$$\begin{aligned}
\mathrm{Reg}(\pi) - \widehat{\mathrm{Reg}}_t(\pi) &= (\mathcal{R}(\pi_{f^\star}) - \mathcal{R}(\pi)) - (\widehat{\mathcal{R}}_t(\pi_{\widehat{f}_m}) - \widehat{\mathcal{R}}_t(\pi)) \\
&\le (\mathcal{R}(\pi_{f^\star}) - \mathcal{R}(\pi)) - (\widehat{\mathcal{R}}_t(\pi_{f^\star}) - \widehat{\mathcal{R}}_t(\pi)) \\
&\le |\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)| + |\widehat{\mathcal{R}}_t(\pi_{f^\star}) - \mathcal{R}(\pi_{f^\star})| \\
&\le \frac{\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{2\gamma_m} + \frac{\sqrt{\mathcal{V}_t(\pi_{f^\star})}\sqrt{K}}{2\gamma_m} \\
&\le \frac{\mathcal{V}_t(\pi)}{5\gamma_m} + \frac{\mathcal{V}_t(\pi_{f^\star})}{5\gamma_m} + \frac{5K}{8\gamma_m}
\end{aligned} \tag{A.4}$$

where the first inequality is by the optimality of $\pi_{\widehat{f}_m}$ for $\widehat{\mathcal{R}}_t(\cdot)$, the second inequality is by the triangle inequality, the third inequality is by Lemma A.7, and the fourth inequality is by the AM-GM inequality. By the definitions of $\mathcal{V}_t(\pi), \mathcal{V}_t(\pi_{f^\star})$ and Lemma A.6, there exist epochs $i, j < m$ such that

$$\mathcal{V}_t(\pi) = V(p_i, \pi) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_i(\pi(x) \mid x)}\right] \le K + \gamma_i \widehat{\mathrm{Reg}}_{\tau_i}(\pi),$$

$$\mathcal{V}_t(\pi_{f^\star}) = V(p_j, \pi_{f^\star}) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_j(\pi_{f^\star}(x) \mid x)}\right] \le K + \gamma_j \widehat{\mathrm{Reg}}_{\tau_j}(\pi_{f^\star}).$$

Combining the above two inequalities with (A.3), we have

$$\frac{\mathcal{V}_t(\pi)}{5\gamma_m} \le \frac{K + \gamma_i \widehat{\mathrm{Reg}}_{\tau_i}(\pi)}{5\gamma_m} \le \frac{K + \gamma_i(2\mathrm{Reg}(\pi) + c_0 K/\gamma_i)}{5\gamma_m} \le \frac{(1+c_0)K}{5\gamma_m} + \frac{2}{5}\mathrm{Reg}(\pi), \tag{A.5}$$

$$\frac{\mathcal{V}_t(\pi_{f^\star})}{5\gamma_m} \le \frac{K + \gamma_j \widehat{\mathrm{Reg}}_{\tau_j}(\pi_{f^\star})}{5\gamma_m} \le \frac{K + \gamma_j(2\mathrm{Reg}(\pi_{f^\star}) + c_0 K/\gamma_j)}{5\gamma_m} = \frac{(1+c_0)K}{5\gamma_m}, \tag{A.6}$$

where the last inequality in (A.5) follows from $\gamma_i \le \gamma_m$ and the last inequality in (A.6) follows from $\mathrm{Reg}(\pi_{f^\star}) = 0$. Combining (A.4), (A.5) and (A.6), we have

$$\mathrm{Reg}(\pi) \le \frac{5}{3}\widehat{\mathrm{Reg}}_t(\pi) + \frac{2c_0 K}{3\gamma_m} + \frac{1.71K}{\gamma_m} \le 2\widehat{\mathrm{Reg}}_t(\pi) + \frac{c_0 K}{\gamma_m}. \tag{A.7}$$

We then show that for all rounds $t$ in epoch $m$ and all $\pi \in \Psi$,

$$\widehat{\mathrm{Reg}}_t(\pi) \le 2\mathrm{Reg}_t(\pi) + c_0 K/\gamma_m.$$

Similar to (A.4), we have

$$\begin{aligned}\widehat{\mathrm{Reg}}_t(\pi) - \mathrm{Reg}(\pi) &= (\widehat{\mathcal{R}}(\pi_{\widehat{f}_m}) - \widehat{\mathcal{R}}_t(\pi)) - (\mathcal{R}(\pi_{f^\star}) - \mathcal{R}(\pi)) \\ &\le (\widehat{\mathcal{R}}_t(\pi_{\widehat{f}_m}) - \widehat{\mathcal{R}}_t(\pi)) - (\mathcal{R}(\pi_{\widehat{f}_m}) - \mathcal{R}(\pi)) \\ &\le |\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}(\pi)| + |\widehat{\mathcal{R}}_t(\pi_{\widehat{f}_m}) - \mathcal{R}(\pi_{\widehat{f}_m})| \\ &\le \frac{\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{\gamma_m} + \frac{\sqrt{\mathcal{V}_t(\pi_{\widehat{f}_m})}\sqrt{K}}{\gamma_m} \\ &\le \frac{\mathcal{V}_t(\pi)}{5\gamma_m} + \frac{\mathcal{V}_t(\pi_{\widehat{f}_m})}{5\gamma_m} + \frac{5K}{8\gamma_m}. \end{aligned} \tag{A.8}$$

By the definition of $\mathcal{V}_t(\pi_{\widehat{f}_m})$ and Lemma A.6, there exist epoch $l < m$ such that

$$\mathcal{V}_t(\pi_{\widehat{f}_m}) = V(p_l, \pi_{\widehat{f}_m}) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{p_l(\pi_{\widehat{f}_m}|x)}\right] \le K + \gamma_l \widehat{\mathrm{Reg}}_{\tau_l}(\pi_{\widehat{f}_m}).$$

Using (A.3), $\gamma_l \le \gamma_m$, (A.7) and $\widehat{\mathrm{Reg}}_t(\pi_{\widehat{f}_m}) = 0$, we have

$$\begin{aligned}\frac{\mathcal{V}_t(\pi_{\widehat{f}_m})}{5\gamma_m} &\le \frac{K + \gamma_l \widehat{\mathrm{Reg}}_{\tau_l}(\pi_{\widehat{f}_m})}{5\gamma_m} \le \frac{K + \gamma_l(2\mathrm{Reg}(\pi_{\widehat{f}_m}) + c_0 K/\gamma_l)}{5\gamma_m} \le \frac{(1+c_0)K}{5\gamma_m} + \frac{2}{5}\mathrm{Reg}(\pi_{\widehat{f}_m}) \\ &\le \frac{(1+c_0)K}{5\gamma_m} + \frac{2}{5}\left(\widehat{\mathrm{Reg}}_t(\pi_{\widehat{f}_m}) + \frac{c_0 K}{\gamma_m}\right) = \frac{(1+3c_0)K}{5\gamma_m}. \end{aligned} \tag{A.9}$$

Combining (A.5), (A.8) and (A.9), we have

$$\widehat{\mathrm{Reg}}_t(\pi) \le \frac{7}{5}\mathrm{Reg}(\pi) + \frac{4c_0 K}{5\gamma_m} + \frac{1.03K}{\gamma_m} \le 2\mathrm{Reg}(\pi) + \frac{c_0 K}{\gamma_m}.$$

Thus we complete the inductive step, and the claim proves to be true for all $m \in \mathbb{N}$. □

### A.1.6 Bounding the True Regret

In this part, we put everything together and finally prove Lemma A.10, which holds for both Setups A.1 and A.2, and simultaneously implies Theorem 2.1, Corollary 2.1, and Theorem 2.2. Moreover, Lemma A.10 implies that bounded rewards are not required if we only want to bound the expected regret.

**Lemma A.9.** *Recall that $\Gamma_1$ and $\Gamma_2$ are defined in Appendix A.1.2. Assume that $\Gamma_1$ (resp. $\Gamma_2$) holds under Setup A.1 (resp. Setup A.2). For every epoch $m \in \mathbb{N}$,*

$$\sum_{\pi\in\Psi} Q_m(\pi)\mathrm{Reg}(\pi) \le 7.15K/\gamma_m.$$

*Proof of Lemma A.9.* Fix any epoch $m \in \mathbb{N}$. Since $\tau_{m-1}+1$ belongs to epoch $m$, we have

$$\begin{aligned}\sum_{\pi\in\Psi} Q_m(\pi)\mathrm{Reg}(\pi) &\le \sum_{\pi\in\Psi} Q_m(\pi)\left(2\widehat{\mathrm{Reg}}_{\tau_{m-1}+1}(\pi) + \frac{c_0 K}{\gamma_m}\right)\\ &= 2\sum_{\pi\in\Psi} Q_m(\pi)\widehat{\mathrm{Reg}}_{\tau_{m-1}+1}(\pi) + \frac{c_0 K}{\gamma_m}\\ &\le \frac{(2+c_0)K}{\gamma_m},\end{aligned}$$

where the first inequality follows from Lemma A.8, and the second inequality follows from Lemma A.5. We then take in $c_0 = 5.15$. □

**Lemma A.10.** *For any $T \in \mathbb{N}$, the expected regret of our algorithm after $T$ rounds is at most $\sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \sqrt{K}\tau_1 + T\delta/2$. Furthermore, if all rewards are $[0,1]$-bounded, then with probability at least $1-\delta$, the regret after $T$ rounds is at most*

$$\sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \tau_1 + \sqrt{8T\log(2/\delta)}.$$

*Proof of Lemma A.10.* Fix $T \in \mathbb{N}$. Since $\Gamma_1$ (resp. $\Gamma_2$) holds under Setup A.1 (resp. Setup A.2) with probability at least $1-\delta/2$, by Lemma A.4 and Lemma A.9, we can bound the expected regret:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\left(r_t(\pi_{f^\star}) - r_t(a_t)\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{x_t,r_t,a_t}\left[r_t(\pi_{f^\star}) - r_t(a_t) \mid \Upsilon_{t-1}\right]\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Psi} Q_{m(t)}(\pi)\mathrm{Reg}(\pi)\right] \le \sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \sqrt{K}\tau_1 + T\delta/2.\end{aligned}$$

We now assume $r_t \in [0,1]^K$ and turn to the high-probability bound. For each round $t \in \{1,\dots,T\}$, define $M_t := r_t(\pi_{f^\star}) - r_t(a_t) - \sum_{\pi\in\Psi} Q_{m(t)}(\pi)\mathrm{Reg}(\pi)$. By Lemma A.4 we have

$$\mathbb{E}_{x_t,r_t,a_t}\left[r_t(\pi_{f^\star}) - r_t(a_t) \mid \Upsilon_{t-1}\right] = \sum_{\pi\in\Psi} Q_{m(t)}(\pi)\mathrm{Reg}(\pi), \quad \mathbb{E}_{x_t,r_t,a_t}[M_t \mid \Upsilon_{t-1}] = 0,$$

Since $|M_t| \le 2$, $M_t$ is a martingale difference sequence. By Azuma's inequality,

$$\sum_{t=1}^{T} M_t \le 2\sqrt{2T\log(2/\delta)} \tag{A.10}$$

with probability at least $1-\delta/2$. By Lemma A.1 (resp. Lemma A.2), with probability at least $1-\delta/2$, the event $\Gamma_1$ (resp. $\Gamma_2$) holds. By a union bound, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T}\left(r_t(\pi_{f^\star}) - r_t(a_t)\right) &\le \sum_{t=1}^{T}\sum_{\pi\in\Psi} Q_{m(t)}(\pi)\mathrm{Reg}(\pi) + \sqrt{8T\log(2/\delta)}\\ &\le \sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \tau_1 + \sqrt{8T\log(2/\delta)}\end{aligned}$$

where the first inequality follows from (A.10), and the second inequality follows from Lemma A.9. □

Finally, we assume that all rewards are $[0,1]$-bounded and use Lemma A.10 to derive Theorem 2.1, Corollary 2.1, and Theorem 2.2.

*Proof of Theorem 2.1.* We are in Setup A.1, and we have $\tau_m \leq 2^m$, $\forall m \in \mathbb{N}$ and $\tau_m \leq 2\tau_{m-1}$, $\forall m > 1$. By Lemma A.10, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}(r_t(\pi_{f^\star}) - r_t(a_t)) &\leq \sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \tau_1 + \sqrt{8T\log(2/\delta)} \\
&= 215\sum_{m=2}^{m(T)}\sqrt{\frac{K\log(|\mathcal{F}|\log(\tau_{m-1})m/\delta)}{\tau_{m-1}}}(\tau_m - \tau_{m-1}) + \tau_1 + \sqrt{8T\log(2/\delta)} \\
&\overset{\text{(i)}}{\leq} 215\sqrt{K\log(|\mathcal{F}|m(T)^2/\delta)}\sum_{m=2}^{m(T)}\frac{\tau_m - \tau_{m-1}}{\sqrt{\tau_m/2}} + \sqrt{8T\log(2/\delta)} + \tau_1 \\
&\overset{\text{(ii)}}{\leq} 215\sqrt{2K\log(|\mathcal{F}|m(T)^2/\delta)}\sum_{m=2}^{m(T)}\int_{\tau_{m-1}}^{\tau_m}\frac{dx}{\sqrt{x}} + \sqrt{8T\log(2/\delta)} + \tau_1 \\
&= 215\sqrt{2K\log(|\mathcal{F}|m(T)^2/\delta)}\int_{\tau_1}^{\tau_{m(T)}}\frac{dx}{\sqrt{x}} + \sqrt{8T\log(2/\delta)} + \tau_1 \\
&\leq 430\sqrt{2\tau_{m(T)}K\log(|\mathcal{F}|m(T)^2/\delta)} + \sqrt{8T\log(2/\delta)} + \tau_1 \\
&\overset{\text{(iii)}}{\leq} 860\sqrt{KT\log(|\mathcal{F}|m(T)^2/\delta)} + \sqrt{8T\log(2/\delta)} + \tau_1,
\end{aligned}$$

where (i) follows from $\log(\tau_{m-1}) \leq m-1 \leq m(T)$ and $\tau_m \leq 2\tau_{m-1}$, (ii) follows from an integral bound, and (iii) follows from $\tau_{m(T)} \leq 2\tau_{m(T)-1} < 2T$. We thus finish our proof of Theorem 2.1. □

*Proof of Corollary 2.1.* We are in Setup A.1, and we have $\tau_m = \left\lfloor 2T^{1-2^{-m}} \right\rfloor$, $\forall m \in \mathbb{N}$. Without loss of generality, assume that $T > 1000$.

By Lemma A.10, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{t=1}^{T}(r_t(\pi_{f^\star}) - r_t(a_t)) &\leq \sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \tau_1 + \sqrt{8T\log(2/\delta)} \\
&= 215\sum_{m=2}^{m(T)}\sqrt{\frac{K\log(|\mathcal{F}|m\log(\tau_{m-1})/\delta)}{\tau_{m-1}}}(\tau_m - \tau_{m-1}) + \tau_1 + \sqrt{8T\log(2/\delta)} \\
&\leq 215\sqrt{K\log(|\mathcal{F}|m(T)\log T/\delta)}\sum_{m=2}^{m(T)}\frac{\tau_m - \tau_{m-1}}{\sqrt{\tau_{m-1}}} + \sqrt{8T\log(2/\delta)} + \tau_1 \\
&\leq 215\sqrt{K\log(|\mathcal{F}|m(T)\log T/\delta)}\sum_{m=2}^{m(T)}\frac{\tau_m}{\sqrt{\tau_{m-1}}} + \sqrt{8T\log(2/\delta)} + \tau_1 \\
&\leq 215\sqrt{K\log(|\mathcal{F}|m(T)\log T/\delta)}\left(2\sqrt{T}\right)(m(T)-1) + \sqrt{8T\log(2/\delta)} + 2\sqrt{T},
\end{aligned}$$

where the last inequality follows from

$$\frac{\tau_m}{\sqrt{\tau_{m-1}}} \leq \frac{\tau_m}{\sqrt{(\tau_{m-1}+1)/2}} \leq \frac{2T^{1-2^{-m}}}{T^{\frac{1}{2}(1-2^{-m+1})}} = 2\sqrt{T}, \quad \forall m > 1$$

and $\tau_1 \leq 2\sqrt{T}$. Corollary 2.1 follows from the fact that $m(T) = O(\log\log T)$. □

*Proof of Theorem 2.2.* We are in Setup A.2, and we have assumed that all rewards are $[0,1]$-bounded.

By Lemma A.10, with probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{t=1}^{T}(r_t(\pi_{f^\star}) - r_t(a_t)) &\leq \sum_{t=\tau_1+1}^{T} 7.15K/\gamma_{m(t)} + \tau_1 + \sqrt{8T\log(2/\delta)} \\ &= 14.3\sum_{m=2}^{m(T)}\sqrt{K\mathcal{E}_{\mathcal{F},\delta/(2m^2)}(\tau_{m-1}-\tau_{m-2})}(\tau_m - \tau_{m-1}) + \tau_1 + \sqrt{8T\log(2/\delta)},\end{aligned}$$

where we directly plug in the definition of $\gamma_m$. □

*Remark.* Note that the entire proof of Theorem 2.2 holds without assuming $f^\star \in \mathcal{F}$. Therefore, the extension to the misspecified setting described in §2.3.2 (where the misspecification error is known) is straightforward.

### A.1.7 Dealing with Uncountable Context Spaces

We have proved Theorem 2.1 and Theorem 2.2 under the condition that $|\mathcal{X}| < \infty$ (note that we allow $|\mathcal{X}|$ to be arbitrarily large in this setting, as the regret bound does not depend on $|\mathcal{X}|$). The main role of the condition $|\mathcal{X}| < \infty$ is that it makes all policies of the form $\pi : \mathcal{X} \to \mathcal{A}$ automatically measurable with respect to $\mathcal{D}$, enabling us to define $\mathcal{R}(\cdot), \widehat{\mathcal{R}}_t(\cdot), \mathrm{Reg}(\cdot), \widehat{\mathrm{Reg}}_t(\cdot)$ for all policies $\pi \in \mathcal{A}^{\mathcal{X}}$ without worrying about any measurability issues. Since everything is well-defined in the universal policy space, we are able to provide an illuminating analysis in this space, which not only explains the value and role of realizability (or relaxed notions of realizability) but also establishes new connections among several research lines of contextual bandits.

Nevertheless, to ensure that Theorem 2.1 and Theorem 2.2 indeed hold for a generic, possibly uncountable $\mathcal{X}$, we need to deal with the potential measurability issues associated with $\mathcal{A}^{\mathcal{X}}$ when $\mathcal{X}$ is arbitrary. In what follows, we discuss such issues and give a simple resolution.

Before we proceed, we make two remarks. First, to ensure that the contextual bandit problem that we study is meaningful, some basic conditions are required, e.g., the true reward function $f^\star$ should be measurable with respect to $\mathcal{D}$ for each fixed $a \in \mathcal{A}$, and the regression oracle should always generate predictors with such a property. Such conditions are necessary for the regret of the algorithm to be well-defined, and are directly assumed here.

Second, in terms of our analysis, the only issue that is worthy of special attention when $\mathcal{X}$ is uncountable is that, not all policies in $\Psi = \mathcal{A}^{\mathcal{X}}$ are guaranteed to be measurable with respect to $\mathcal{D}$ (as a consequence, $\mathcal{R}(\cdot), \widehat{\mathcal{R}}_t(\cdot), \mathrm{Reg}(\cdot), \widehat{\mathrm{Reg}}_t(\cdot)$ may not be "everywhere defined" on $\Psi$). All other issues, such as the existence and properties of $Q_m(\cdot)$, can be easily addressed by standard tools from

measure theory (e.g., Theorem 2.4.4 of Tao 2011).

We now focus on the key issue that $\Psi$ may contain non-measurable policies (when $\mathcal{X}$ is uncountable). It turns out that the affect of this issue on our previous analysis is mostly "notational." Namely, since $\mathcal{R}(\cdot), \widehat{\mathcal{R}}_t(\cdot), \mathrm{Reg}(\cdot), \widehat{\mathrm{Reg}}_t(\cdot)$ are not necessarily well-defined for all $\pi \in \Psi$, Lemma A.7 and Lemma A.8—which involves the universal quantifier "for all $\pi \in \Psi$"—require slight modifications.

There are multiple ways to address such an issue. We provide a a simple resolution below, which works for general uncountable $\mathcal{X}$ and generates additional insights about our proof.

The resolution is based on the following observation: while $\mathcal{R}(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\left[f^\star(x, \pi(x))\right]$ is not guaranteed to be well-defined for an arbitrary deterministic policy $\pi \in \Psi$, the quantity $\mathcal{R}(Q_m) = \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{\pi\sim Q_m}\left[f^\star(x, \pi(x))\right]$ is always well-defined, as the algorithm's adopted randomized policy $Q_m(\cdot)$ in epoch $m$ is "measurable." Thus, we can directly define $\mathcal{R}(Q) = \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{\pi\sim Q}\left[f^\star(x, \pi(x))\right]$ for all "measurable" randomized policies $Q(\cdot)$, and similarly define $\widehat{\mathcal{R}}_t(Q), \mathrm{Reg}(Q), \widehat{\mathrm{Reg}}_t(Q)$. This would enable us to restate Lemma A.7 and Lemma A.8 by replacing "for all deterministic policies $\pi \in \Psi$" with "for all measurable randomized policies $Q$," and there would be no measurability issues any more. Moreover, observing that an action selection kernel $p : \mathcal{B}(\mathcal{A}) \times \mathcal{X} \to [0, 1]$ exactly corresponds to a "measurable" randomized policy (as the definition of a probability kernel clearly requires $p(E \mid x)$ to be measurable with respect to $\mathcal{D}$ for any fixed $E \in \mathcal{B}(\mathcal{A})$), we can actually directly use the kernel $p(\cdot \mid \cdot)$ to denote a "measurable" randomized policy, which would lead to a new version of our proof without explicit use of the notation $Q_m(\cdot)$ or $Q(\cdot)$.

Specifically, the new general proof proceeds as follows. Recall that $\mathcal{P}$ is the space of all action selection kernels, which corresponds to the space of all *measurable randomized policies*. Analogous to Appendix A.1.1, for all (measurable) randomized policies $p, p' \in \mathcal{P}$ and $t = 1, \ldots, T$, define

$$V(p, p') := \mathbb{E}_{x\sim\mathcal{D}, a\sim p'(\cdot|x)}\left[\frac{1}{p(a \mid x)}\right], \qquad \mathcal{V}_t(p) := \max_{1\le m\le m(t)-1}\{V(p_m, p)\},$$

$$\mathcal{R}(p) := \mathbb{E}_{x\sim\mathcal{D}, a\sim p(\cdot|x)}\left[f^\star(x, a)\right], \qquad \widehat{\mathcal{R}}_t(p) := \mathbb{E}_{x\sim\mathcal{D}, a\sim p(\cdot|x)}\left[\widehat{f}_{m(t)}(x, a)\right],$$

$$\mathrm{Reg}(p) := \mathcal{R}(\pi_{f^\star}) - \mathcal{R}(p), \qquad \widehat{\mathrm{Reg}}_t(p) := \widehat{\mathcal{R}}_t(\pi_{\widehat{f}_{m(t)}}) - \widehat{\mathcal{R}}_t(p),$$

where $V(p, p')$ is the "decisional divergence" between two randomized policies $p$ and $p'$. Analogous to Appendix A.1.3, we can show that in each epoch $m \in \mathbb{N}$, the algorithm's adopted randomized policy $p_m \in \mathcal{P}$ is a solution to the following "Implicit Optimization Problem":

$$\widehat{\mathrm{Reg}}_t(p_m) \le K/\gamma_m,$$
$$\forall p \in \mathcal{P}, \quad \mathbb{E}_{x\sim\mathcal{D}, a\sim p(\cdot|x)}\left[\frac{1}{p_m(a \mid x)}\right] \le K + \gamma_m \widehat{\mathrm{Reg}}_t(p).$$

Analogous to Lemma A.7 to Lemma A.8, we have the following guarantees.

**Lemma A.11.** *Assume $\Gamma_1$ (resp. $\Gamma_2$) holds under Setup A.1 (resp., Setup A.2). For all rounds $t > \tau_1$, for all (measurable) randomized policies $p \in \mathcal{P}$,*

$$|\widehat{\mathcal{R}}_t(p) - \mathcal{R}(p)| \le \frac{\sqrt{\mathcal{V}_t(p)}\sqrt{K}}{2\gamma_{m(t)}}.$$

**Lemma A.12.** *Assume that* $\Gamma_1$ *(resp.* $\Gamma_2$*) holds under Setup A.1 (resp. Setup A.2). Let* $c_0 := 5.15$*. For all epochs* $m \in \mathbb{N}$*, all rounds* $t$ *in epoch* $m$*, and all (measurable) randomized policies* $p \in \mathcal{P}$*,*

$$\mathrm{Reg}(p) \le 2\widehat{\mathrm{Reg}}_t(p) + c_0 K/\gamma_m,$$

$$\widehat{\mathrm{Reg}}_t(p) \le 2\mathrm{Reg}(p) + c_0 K/\gamma_m.$$

Lemma A.11 and Lemma A.12 are almost identical to Lemma A.7 and Lemma A.8, except for the slight difference that "for all $\pi \in \Psi$" is replaced with "for all $p \in \mathcal{P}$," which ensures that there are no measurability issues (as $\mathcal{R}(\cdot), \widehat{\mathcal{R}}_t(\cdot), \mathrm{Reg}(\cdot), \widehat{\mathrm{Reg}}_t(\cdot)$ are "everywhere defined" on $\mathcal{P}$). We then easily have $\mathbb{E}_{x_t, r_t, a_t}\left[r_t(\pi_{f^\star}) - r_t(a_t) \mid \Upsilon_{t-1}\right] = \mathrm{Reg}(p_{m(t)}) \le 7.15K/\gamma_{m(t)}$, and Theorem 2.1 and Theorem 2.2 immediately follow. The above proof is essentially the same as the proof that we provide in Appendix A.1.1 to Appendix A.1.6, but under a different set of notations.

# Appendix B

# Supplemental Material for Chapter 3

## B.1 Details for Results of Contextual Bandits

In this section of the appendix, we overview the motivation behind the design of AdaCB, and give regret bounds for the algorithm based on the policy and value function disagreement coefficients, as well as matching lower bounds. We then show how to relate these quantities to other structural parameters for the distribution-free and adversarial settings, and instantiate our bounds for concrete settings of interest.

The proofs of the results mentioned in this section can be found in the full version of our paper (Foster et al. 2020).

### B.1.1 Overview of AdaCB

AdaCB (Algorithm 3.1) operates in a doubling epoch schedule. Letting $\tau_m = 2^m$ with $\tau_0 = 0$, each epoch $m \geq 1$ consists of rounds $\tau_{m-1} + 1, \ldots, \tau_m$, and there are $M = \lceil \log_2 T \rceil$ epochs in total. At the beginning of each epoch $m$, we compute an estimator $\widehat{f}_m$ for the Bayes regression function $f^\star$ by performing least-squares regression on data collected so far (Line 2). We also maintain a version space $\mathcal{F}_m$, which is the set of all plausible predictors that cannot yet be eliminated based on square loss confidence bounds (Line 3). Based on $\mathcal{F}_m$, we select the learning rate $\gamma_m$ for the current epoch adaptively by estimating a parameter called the *instance-dependent scale factor* ($\lambda_m$) which is closely related to the policy disagreement coefficient (Option I) and the value function disagreement coefficient (Option II). Then, when a context $x_t$ in epoch $m$ arrives, AdaCB first computes the *candidate action set* $\mathcal{A}_t := \mathcal{A}(x_t; \mathcal{F}_m)$ (Line 8), which is the set of actions that are optimal for some predictor $f \in \mathcal{F}_m$, and thus could plausibly be equal to $\pi^\star(x_t)$. The algorithm then sets $p_t = \mathsf{IGW}_{\mathcal{A}_t, \gamma_m}(x_t; \widehat{f}_m)$ (Line 9), samples $a_t \sim p_t$, and proceeds to the next round.

The adaptive learning rate $\gamma_m$ balances the algorithm's efforts between exploration and exploitation: a larger learning rate leads to more aggressive exploitation (following the least-squares predictor $\widehat{f}_m$), while a smaller learning rate leads to more conservative exploration over the candidate action set. AdaCB's learning rate $\gamma_m$ (Line 5) has two components: the instance-dependent scale factor $\lambda_m$, which is adaptively determined by the collected data; and a non-adaptive component propositional to $\sqrt{An_{m-1}/\log|\mathcal{F}|}$, where $n_{m-1}$ is the length of the epoch $m-1$. While the non-adaptive component

is the same as the learning rate in FALCON and is sufficient if one only aims to achieve the minimax regret, the adaptive factor $\lambda_m$, combined with the action elimination procedure above, is essential for AdaCB to achieve near-optimal instance-dependent regret. We offer two different schemes to select $\lambda_m$: The first adapts to the policy disagreement coefficient, while the second adapts to the value function disagreement coefficient.

- Option I (policy-based exploration). This option selects $\lambda_m$ as a sample-based approximation to the quantity

$$\mathbb{P}_{\mathcal{D}}(|\mathcal{A}(x, \mathcal{F}_m)| > 1)/\sqrt{\mathbb{P}_{\mathcal{D}}(|\mathcal{A}(x, \mathcal{F}_{m-1})| > 1)},$$

  where $\mathbb{P}_{\mathcal{D}}(|\mathcal{A}(x, \mathcal{F}_m)| > 1)$ and $\mathbb{P}_{\mathcal{D}}(|\mathcal{A}(x, \mathcal{F}_{m-1})| > 1)$ are *disagreement probabilities* (i.e., the probability that we encounter a context on which we cannot yet determine the true optimal action) for epoch $m$ and epoch $m-1$, respectively. Intuitively, this configuration asserts that we should adaptively discount the learning rate if either 1) the current disagreement probability is small, or 2) the disagreement probability is decreasing sufficiently quickly across epochs. This scheme is natural because if we expect that no exploration is required for a large portion future contexts, then we have flexibility to perform more thorough exploration on other contexts where the true optimal action cannot yet be determined. This accelerates AdaCB's exploration of more effective policies.

- Option II (value-based exploration). While the disagreement probability used in Option I is a useful quantity that provides information on the hardness of the problem instance, it does not fully utilize the value function structure. In particular, it is only sensitive to the occurrence of disagreement on each context, but is not sensitive to the *scale* of disagreement (i.e., how much it would cost if we chose a disagreeing action) on each context. This motivates Option II, which is based on a refined *confidence width* $w(x; \mathcal{F}_m)$ that accounts for both the occurrence and the scale of disagreement. Specifically, $w(x; \mathcal{F}_m)$ measures the worst-case cost of exploring a sub-optimal action in the candidate action set for $x$, and Option II selects $\lambda_m$ as a sample-based approximation to the quantity

$$\mathbb{I}\{\mathbb{E}_{\mathcal{D}}[w(x; \mathcal{F}_m)] \geq \sqrt{AT \log |\mathcal{F}|}/n_{m-1}\}.$$

  In other words, we adaptively zero out the learning rate and perform uniform exploration if $\mathbb{E}_{\mathcal{D}}[w(x; \mathcal{F}_m)]$ is smaller than an epoch-varying threshold. This is reasonable because if $\mathbb{E}_{\mathcal{D}}[w(x; \mathcal{F}_m)]$ is small, then the average cost of exploration is small for the underlying instance, so we should take advantage of this and explore as much as possible.

Finally, since $\mathcal{D}$ is unknown, to obtain $\lambda_m$ we compute an empirical approximation to $\mathbb{E}_{\mathcal{D}}[\cdot]$ using sample splitting. That is, we use separate sample to compute $\mathcal{F}_m$ and to approximate $\mathcal{D}$ to ensure independence; this is reflected in the sample splitting schedule $\{t_m\}_{m=1}^M$ in AdaCB. The smoothing parameter $\mu_m$ is designed to correct the approximation error incurred by this procedure.

We make a few additional remarks. First, the learning rate and confidence width parameters in Algorithm 3.1 (and consequently our main theorems) consider a general finite class $\mathcal{F}$. This is only a stylistic choice: AdaCB works as-is for general function classes, with the dependence on $\log |\mathcal{F}|$ in these parameters replaced by standard learning-theoretic complexity measures such as

the pseudodimension; see Appendix B.1.7. Second, Algorithm 3.1 takes $T$ as input. One can straightforwardly extend Algorithm 3.1 to work with unknown $T$ using the standard *doubling trick*. Finally, we emphasize that Option I and Option II are designed based on different techniques and lead to different instance-dependent guarantees. Designing a single option that simultaneously achieving the goals of Option I and Option II is an interesting future direction.

**Oracle efficiency** AdaCB can be implemented efficiently with a weighted least squares regression oracle Oracle (see equation (RO) in the full version of our paper (Foster et al. 2020)) as follows.

- At each epoch $m$, call Oracle to compute the square loss empirical risk minimizer $\widehat{f}_m$.
- For any given context $x$, the candidate action set $\mathcal{A}(x; \mathcal{F}_m)$ can be computed using either $\widetilde{O}(A)$ oracle calls when $\mathcal{F}$ is convex or $\widetilde{O}(AT^2)$ oracle calls for general (in particular, finite) classes.
- For Option II, the function $w(x; \mathcal{F}_m)$ can be computed in a similar fashion to $\mathcal{A}(x; \mathcal{F}_m)$ using $\widetilde{O}(A)$ or $\widetilde{O}(AT^2)$ oracle calls in the convex and general case, respectively.

Altogether, since $\mathcal{A}(x; \mathcal{F}_m)$ and $w(x; \mathcal{F}_m)$ are computed for $O(1)$ different contexts per round amortized, the algorithm requires $O(AT)$ calls to Oracle overall when $\mathcal{F}$ is convex. The reduction is described in detail in the full version of our paper (Foster et al. 2020).

### B.1.2 Disagreement-Based Guarantees

We are now ready to state our first main regret guarantee for AdaCB, which is based on the policy disagreement coefficient (3.5). The theorem also includes a more general result in terms of an intermediate quantity we call the cost-sensitive policy disagreement coefficient, which we define by

$$\boldsymbol{\theta}^{\mathsf{csc}}(\Pi, \varepsilon_0) = \sup_{\varepsilon \geq \varepsilon_0} \frac{\mathbb{P}_{\mathcal{D}}(x : \exists \pi \in \Pi^{\mathrm{csc}}_{\varepsilon} : \pi(x) \neq \pi^\star(x))}{\varepsilon}, \tag{B.1}$$

where $\Pi^{\mathrm{csc}}_{\varepsilon} = \{\pi \in \Pi : R(\pi^\star) - R(\pi) \leq \varepsilon\}$ for $R(\pi) := \mathbb{E}[r(\pi(x))]$.[49] The cost-sensitive policy disagreement coefficient grants finer control over the cost-sensitive structure of the problem and—beyond leading to our main gap-based result—leads to instance-dependent guarantees even when the instance does not have uniform gap.

**Theorem B.1** (Full Version)**.** *For any instance with uniform gap $\Delta$, Algorithm 3.1 with* Option I *ensures that*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \min_{\varepsilon > 0} \max \left\{ \varepsilon \Delta T, \frac{\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) A \log |\mathcal{F}|}{\Delta} \right\} + \widetilde{O}(1). \tag{B.2}$$

*More generally, Algorithm 3.1 with* Option I *ensures that for every instance, without any gap assumption,*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \min_{\varepsilon > 0} \max \left\{ \varepsilon T, \boldsymbol{\theta}^{\mathsf{csc}}(\Pi, \varepsilon) A \log |\mathcal{F}| \right\} + \widetilde{O}(1). \tag{B.3}$$

Let us describe some key features of Theorem B.1.

[49] The acronym CSC refers to *cost-sensitive classification*.

- Whenever $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \leq \mathrm{polylog}(1/\varepsilon)$, we may choose $\varepsilon \propto 1/T$ in (B.2) so that

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \frac{A \log|\mathcal{F}|}{\Delta}.$$

  For example, for the classical multi-armed bandit setup where $\mathcal{X}$ is a singleton, we have $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) = 1$, recovering the usual instance-dependent rate (up to logarithmic factors). We give some more examples where logarithmic regret can be attained in a moment.

- More generally, since the function $\varepsilon \mapsto \varepsilon \Delta T$ is increasing in $\varepsilon$ and $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon)$ is decreasing, the best choice for the bound (B.2) (up to constant factors) is the critical radius $\varepsilon_T$ that satisfies the balance

$$\varepsilon_T \Delta T \propto \frac{\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon_T) A \log|\mathcal{F}|}{\Delta}. \tag{B.4}$$

  For example, if $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \propto \varepsilon^{-\rho}$ for some $\rho \in (0, 1)$, then choosing

$$\varepsilon_T \propto (A \log|\mathcal{F}|(\Delta^2 T)^{-1})^{\frac{1}{1+\rho}}$$

  leads to

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \frac{(A \log|\mathcal{F}|)^{\frac{1}{1+\rho}} \cdot T^{\frac{\rho}{1+\rho}}}{\Delta^{\frac{1-\rho}{1+\rho}}}.$$

  The critical radius also plays an important role in the *proof* of Theorem B.1.

- With no assumption on the gap or $\boldsymbol{\theta}^{\mathsf{pol}}$, we may always take $\boldsymbol{\theta}^{\mathsf{csc}}(\Pi, \varepsilon) \leq 1/\varepsilon$, so that (B.3) implies the minimax rate $\sqrt{AT \log|\mathcal{F}|}$.

The general bound (B.3) can be seen to imply (B.2), since $\Pi^{\mathrm{csc}}_{\varepsilon} \subseteq \Pi_{\varepsilon/\Delta}$ whenever the gap is $\Delta$. More generally, the cost-sensitive policy disagreement coefficient can also lead to instance-dependent regret bounds under other standard assumptions which go beyond the uniform gap; these are discussed at the end of the section.

**Optimality** We now show that the regret bound attained by AdaCB in Theorem B.1 is near-optimal, in the sense that it cannot be improved beyond log factors without making additional assumptions on the class $\mathcal{F}$ or the contextual bandit instance.

Formally, we model a contextual bandit algorithm $\mathsf{A}$ as a sequence of mappings $\mathsf{A}_t : (\mathcal{X} \times \mathcal{A} \times [0,1])^{t-1} \times \mathcal{X} \to \Delta(\mathcal{A})$, so that

$$\mathsf{A}_t(x_t; (x_1, a_1, \ell_1(a_1)), \dots, (x_{t-1}, a_{t-1}, \ell_{t-1}(a_{t-1}))) \tag{B.5}$$

is the algorithm's action distribution after observing context $x_t$ at round $t$.

For a given function class $\mathcal{F}$, we define

$$\mathfrak{M}^{\mathsf{pol}}(\mathcal{F}, \varepsilon, \theta) = \inf_{\mathsf{A}} \sup_{(\mathcal{D}, \mathbb{P}_r)} \left\{ \mathbb{E}[\mathrm{Reg}] \mid f^\star \in \mathcal{F},\ \boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \leq \theta \right\} \tag{B.6}$$

to be the *constrained minimax complexity*, which measures the worst-case performance of any algorithm (B.5) across all instances realizable by $\mathcal{F}$ for which the policy disagreement coefficient at

scale $\varepsilon$ is at most $\theta$.[50]

Our main lower bound shows that there exists a function class $\mathcal{F}$ for which the constrained minimax complexity matches the upper bound (B.2).

**Theorem B.2** (Full Version)**.** *Let parameters $A, F \in \mathbb{N}$ and $\Delta \in (0, 1/4)$ be given. For any $\varepsilon \in (0,1)$ and $1 \leq \theta \leq \min\{1/\varepsilon, e^{-2}A/F\}$, there exists a function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathcal{A})$ with $A$ actions and $|\mathcal{F}| \leq F$ such that:*

- *All $f \in \mathcal{F}$ have uniform gap $\Delta$.*
- *The constrained minimax complexity is lower bounded by*

$$\mathfrak{M}^{\mathsf{pol}}(\mathcal{F}, \varepsilon, \theta) = \widetilde{\Omega}(1) \cdot \min\left\{\varepsilon\Delta T, \frac{\theta A \log F}{\Delta}\right\},$$

*where $\widetilde{\Omega}(\cdot)$ hides factors logarithmic in $A$ and $\varepsilon^{-1}$.*

This lower bound has a simple interpretation: The term $\varepsilon\Delta T$ is the regret incurred if we commit to playing a particular policy $\pi \in \Pi_\varepsilon$ for any "simple" instance in which the gap is no larger than $O(\Delta)$ for all actions, while the term $\frac{\boldsymbol{\theta}^{\mathsf{pol}}(\Pi,\varepsilon) A \log|\mathcal{F}|}{\Delta}$ is the cost of exploration to find such a policy.

The first implication of this lower bound is that without an assumption such as the disagreement coefficient, logarithmic regret is impossible even when the gap is constant; this alone is not surprising since Foster and Rakhlin (2020) already showed a similar impossibility for non-stochastic contexts, but Theorem B.2 strengthens this result since it holds for stochastic contexts. More importantly, the lower bound shows that the tradeoff in Theorem B.1 is tight as a function of $\Delta, \varepsilon, A, \log|\mathcal{F}|$, and $\boldsymbol{\theta}^{\mathsf{pol}}$, so additional assumptions are required to attain stronger instance-dependent regret bounds for specific classes. We explore such assumptions in the sequel.

We mention one important caveat: Compared to instance-dependent lower bounds for multi-armed bandits (e.g., Garivier et al. (2019)), the quantification for Theorem B.2 is slightly weaker. Rather than lower bounding the regret for *any* particular instance (assuming uniformly good performance in a neighborhood), we only show existence of a *particular* realizable instance with gap for which the regret lower bound holds. We suspect that strengthening the lower bound in this regard will be difficult unless one is willing to sacrifice dependence on $\log F$.

**Examples** The (policy) disagreement coefficient has been studied extensively in active learning, and many bounds are known for different function classes and distributions of interest. We refer to Hanneke (2014) for a comprehensive survey and summarize some notable examples here (restricting to the binary/two-action case, which has been the main focus of active learning literature).

- When $\mathcal{F}$ is a $d$-dimensional linear function class, $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) \leq \widetilde{O}(d^{1/2}\log(1/\varepsilon))$ whenever $\mathcal{D}$ is isotropic log-concave (Balcan and Long 2013). More generally, $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) = o(1/\varepsilon)$ as long as $\mathcal{D}$ admits a density (Hanneke 2014).

[50] We leave implicit that rewards are restricted to the range $[0, 1]$. In fact, for our lower bound it suffices to only consider reward distributions $\mathbb{P}_r$ for which $r(a)$ is Bernoulli with mean $f^\star(x, a)$ given $x$ in (B.6).

- $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) = \mathrm{polylog}(1/\varepsilon)$ whenever $\mathcal{F}$ is smoothly parameterized by a subset of euclidean space, subject to certain regularity conditions (Friedman 2009). This includes, for example, axis-aligned rectangles.

- When $\Pi$ is a class of depth-limited decision trees, we have $\boldsymbol{\theta}^{\mathsf{pol}}(\Pi, \varepsilon) = \mathrm{polylog}(1/\varepsilon)$ (Balcan et al. 2010).

For an example which leverages the more general parameter $\boldsymbol{\theta}^{\mathsf{csc}}$, Langford and Zhang (2008) give logarithmic regret bounds for finite-class contextual bandits based on a different notion of gap called the *policy gap* defined by $\Delta_{\mathrm{pol}} = R(\pi^\star) - \max_{\pi \neq \pi^\star} R(\pi)$, where $R(\pi) = \mathbb{E}_{x,r}[r(\pi(x))]$. It is simple to see that $\boldsymbol{\theta}^{\mathsf{csc}}(\Pi, \varepsilon) \leq \Delta_{\mathrm{pol}}^{-1}$, so that Theorem B.1 gives $\mathbb{E}[\mathrm{Reg}] \leq \widetilde{O}\Big(\frac{A \log|\mathcal{F}|}{\Delta_{\mathrm{pol}}}\Big)$, which improves upon the gap dependence of their result.

### B.1.3 Scale-Sensitive Guarantees

We now give instance-dependent regret guarantees based on the value function disagreement coefficient, which is defined via

$$\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta_0, \varepsilon_0) = \sup_{\Delta > \Delta_0, \varepsilon > \varepsilon_0} \sup_{p:\mathcal{X}\to\Delta(\mathcal{D})} \frac{\Delta^2}{\varepsilon^2} \mathbb{P}_{\mathcal{D},p}\Big(\exists f \in \mathcal{F} : |f(x,a) - f^\star(x,a)| > \Delta,\ \|f - f^\star\|_{\mathcal{D},p} \leq \varepsilon\Big). \tag{B.7}$$

Compared to the policy disagreement coefficient, the value function disagreement coefficient is somewhat easier to bound directly when the value function class $\mathcal{F}$ has simple structure. For example, when $\mathcal{F}$ is linear, we can bound $\boldsymbol{\theta}^{\mathsf{val}}$ in terms of the dimension for *any* distribution with a simple linear algebraic calculation.

**Proposition B.1.** *Let $\phi(x,a) \in \mathbb{R}^d$ be a fixed feature map, and let $\mathcal{F} = \{(x,a) \mapsto \langle w, \phi(x,a)\rangle \mid w \in \mathcal{W}\}$, where $\mathcal{W} \subseteq \mathbb{R}^d$ is any fixed set. Then for all $\mathcal{D}$, $\Delta$, $\varepsilon$,*

$$\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta, \varepsilon) \leq d.$$

*Furthermore, if $\mathcal{F} = \{(x,a) \mapsto \sigma(\langle w, \phi(x,a)\rangle) \mid w \in \mathcal{W}\}$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is any fixed link function with $0 < c_l \leq \sigma' \leq c_u$ almost surely, we have*

$$\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta, \varepsilon) \leq \left(\frac{c_u}{c_l}\right)^2 \cdot d.$$

More generally—as we show in the next section—the value function disagreement coefficient is always bounded by the so-called eluder dimension for $\mathcal{F}$, allowing us to leverage existing results for this parameter (Russo and Van Roy 2013). However, the value function disagreement coefficient can be significantly tighter because—among other reasons—it can leverage benign distributional structure.

We now show that AdaCB can simultaneously attain the minimax regret bound and adapt to the value function disagreement coefficient.

**Theorem B.3** (Full Version)**.** *For any instance, Algorithm 3.1 with* Option II *ensures that*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1) \cdot \min\left\{\sqrt{AT\log|\mathcal{F}|}, \frac{\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon_T) A \log|\mathcal{F}|}{\Delta}\right\} + O(1), \tag{B.8}$$

*where* $\varepsilon_T \propto \sqrt{\log(|\mathcal{F}|T)/T}$.

This rate improves over the minimax rate asymptotically whenever $\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon) = o(1/\varepsilon)$, and is logarithmic whenever $\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon) = \mathrm{polylog}(1/\varepsilon)$.

As with our policy disagreement-based result, we complement Theorem B.3 with a lower bound. To state the result, we define

$$\mathfrak{M}^{\mathsf{val}}(\mathcal{F}, \Delta, \varepsilon, \theta) = \inf_{\mathsf{A}} \sup_{(\mathcal{D}, \mathbb{P}_r)} \left\{\mathbb{E}[\mathrm{Reg}] \mid f^\star \in \mathcal{F},\ \boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta, \varepsilon) \le \theta\right\}, \tag{B.9}$$

which is the value-based analogue of the constrained minimax complexity (B.6). Our main lower bound is as follows.

**Theorem B.4.** *Let parameters* $A, F \in \mathbb{N}$ *and* $\Delta \in (0, 1/4)$ *be given. For any* $\varepsilon \in (\Delta, 1)$ *and* $0 \le \theta \le \min\{\Delta^2/\varepsilon^2, e^{-2}F/A\}$*, there exists a function class* $\mathcal{F} : \mathcal{X} \to \mathcal{A}$ *with* $A$ *actions and* $|\mathcal{F}| \le F$ *such that:*

- *All* $f \in \mathcal{F}$ *have uniform gap* $\Delta$*.*
- *The constrained minimax complexity is lower bounded by*

$$\mathfrak{M}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon, \theta) = \widetilde{\Omega}(1) \cdot \min\left\{\frac{\varepsilon^2}{\Delta}T, \frac{\theta A \log F}{\Delta}\right\}, \tag{B.10}$$

  *where* $\widetilde{\Omega}(\cdot)$ *hides factors logarithmic in* $A$ *and* $\Delta/\varepsilon$*.*

As with Theorem B.2, the lower bound (B.10) has a simple interpretation: The term $\frac{\varepsilon^2}{\Delta}T$ is an upper bound on the regret of any policy $\pi_f$ for which the predictor $f$ is within $L_2$-radius $\varepsilon$ of $f^\star$ (under gap $\Delta$), and the term $\frac{\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon) A \log|\mathcal{F}|}{\Delta}$ is the exploration cost to find such a predictor.

The most important implication of Theorem B.4 is as follows: Suppose that $\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon) = \mathrm{polylog}(1/\varepsilon)$. Then by taking $\varepsilon_T \propto (A\log|\mathcal{F}|/T)^{\frac{1}{2}+\rho}$ for any $\rho > 0$, we conclude that for sufficiently large $T$, any algorithm on the lower bound instance must have

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{\Omega}(1) \cdot \frac{\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta/2, \varepsilon_T) A \log|\mathcal{F}|}{\Delta}.$$

This implies that the instance-dependent term in (B.8) is nearly optimal in this regime, in that the parameter $\varepsilon_T = \sqrt{\log|\mathcal{F}|/T}$ used by the algorithm can at most be increased by a sub-polynomial factor. In general, however, (B.8) does not exactly match the tradeoff in (B.10), but we suspect that AdaCB can be improved to close the gap.[51]

[51] With a-priori knowledge of $\boldsymbol{\theta}^{\mathsf{val}}$, this is fairly straightforward.

### B.1.4 Distribution-Free Guarantees

The disagreement coefficients introduced in the previous section depend strongly on the context distribution $\mathcal{D}$. On one hand, this is a desirable feature, since it means we may pay very little to adapt to the gap $\Delta$ for benign distributions. On the other hand, in practical applications, we may not have prior knowledge of how favorable $\mathcal{D}$ is, or whether we should expect to do any better than the minimax rate. A natural question then is for what function classes we can guarantee logarithmic regret for *any* distribution $\mathcal{D}$. An important result of Hanneke and Yang (2015) shows that in the binary setting, the policy disagreement coefficient is always bounded by a combinatorial parameter called the (policy) *star number*. We give distribution-free results based on two multiclass generalizations of this parameter

**Definition B.1** (Policy star number (weak))**.** *For any policy $\pi^\star$ and policy class $\Pi$, let the weak policy star number $\underline{\mathfrak{s}}^{\mathsf{pol}}_{\pi^\star}(\Pi)$ denote the largest number $m$ such that there exist contexts $x^{(1)},\dots,x^{(m)}$ and policies $\pi^{(1)},\dots,\pi^{(m)}$ such that for all $i$,*

$$\pi^{(i)}(x^{(i)}) \neq \pi^\star(x^{(i)}), \quad \textit{and} \quad \pi^{(i)}(x^{(j)}) = \pi^\star(x^{(j)}) \quad \forall j \neq i.$$

**Definition B.2** (Policy star number (strong))**.** *For any policy $\pi^\star$ and policy class $\Pi$, let the strong policy star number $\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi)$ denote the largest number $m$ such that there exist context-action pairs $(x^{(1)},a^{(1)}),\dots,(x^{(m)},a^{(m)})$ and policies $\pi^{(1)},\dots,\pi^{(m)}$ such that for all $i$,*

$$\pi^{(i)}(x^{(i)}) = a^{(i)} \neq \pi^\star(x^{(i)}), \quad \textit{and} \quad \pi^{(i)}(x^{(j)}) = \pi^\star(x^{(j)}) \quad \forall j \neq i : x^{(j)} \neq x^{(i)}.$$

These definitions are closely related: It is simple to see that

$$\underline{\mathfrak{s}}^{\mathsf{pol}}_{\pi^\star}(\Pi) \leq \mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi) \leq (A-1)\cdot \underline{\mathfrak{s}}^{\mathsf{pol}}_{\pi^\star}(\Pi), \tag{B.11}$$

and that both of these inequalities can be tight in the worst case. To obtain distribution-free bounds based on the star number, we recall the following key result of Hanneke and Yang (2015).[52]

**Theorem B.5** (Star number bounds disagreement coefficient (Hanneke and Yang 2015))**.** *For all policies $\pi^\star$,*

$$\sup_{\mathcal{D}} \sup_{\varepsilon>0} \boldsymbol{\theta}^{\mathsf{pol}}_{\mathcal{D},\pi^\star}(\Pi,\varepsilon) \leq \underline{\mathfrak{s}}^{\mathsf{pol}}_{\pi^\star}(\Pi). \tag{B.12}$$

This result immediately implies that AdaCB enjoys logarithmic regret for any function class with bounded policy star number.

**Corollary B.1** (Distribution-free bound for AdaCB)**.** *For any function class $\mathcal{F}$, AdaCB with* Option I *has*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}\left(\frac{\underline{\mathfrak{s}}^{\mathsf{pol}}_{\pi^\star}(\Pi)\cdot A\log|\mathcal{F}|}{\Delta}\right). \tag{B.13}$$

[52] Technically, the original theorem in Hanneke and Yang (2015) only holds the binary case, but the multiclass case here follows immediately by applying the theorem with the collection of binary classifiers $\mathcal{H} = \{x \mapsto \mathbb{I}\{\pi(x) \neq \pi^\star(x)\} \mid \pi \in \Pi\}$.

One slightly unsatisfying feature of our lower bounds based on the disagreement coefficient (Theorem B.2/Theorem B.4) is that they are worst-case in nature, and rely on an adversarially constructed policy class. Our next theorem shows that (B.13) is near-optimal for *any* policy class $\Pi$ (albeit, in the worst case over all value function classes $\mathcal{F}$ inducing $\Pi$). This means that if we take the policy class $\Pi$ as a given rather than the value function class $\mathcal{F}$, bounded policy star number is both necessary and sufficient for logarithmic regret.

**Theorem B.6.** *Let a policy class $\Pi$, $\pi^\star \in \Pi$, and gap $\Delta \in (0, 1/8)$ be given. Then there exists a value function class $\mathcal{F}$ such that*

1. *$\Pi = \{\pi_f \mid f \in \mathcal{F}\}$, and in particular some $f^\star \in \mathcal{F}$ has $\pi^\star = \pi_{f^\star}$.*
2. *Each $f \in \mathcal{F}$ has uniform gap $\Delta$.*
3. *For any algorithm with $\mathbb{E}[\mathrm{Reg}] \leq \frac{\Delta T}{16\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi)}$ for all instances realizable by $\mathcal{F}$, there exists an instance with $f^\star$ as the Bayes reward function such that*

$$\mathbb{E}[\mathrm{Reg}] = \Omega\left(\frac{\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi)}{\Delta}\right). \tag{B.14}$$

This bound scales with the strong variant of the policy disagreement coefficient rather than the (smaller) weak variant, but does not directly scale with the number of actions. Hence, the dependence matches the upper bound of AdaCB in (B.13) whenever the second inequality in (B.11) saturates (since $\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi)$ can itself scale with the number of actions). We suspect that the lower bound is tight and that the upper bound can be improved to scale with $\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi)$, with no explicit dependence on the number of actions.

Unlike the upper bound (B.13), the lower bound (B.14) does not scale with $\log|\mathcal{F}|$. This does not appear to be possible to resolve without additional assumptions, as there are classes for which (B.14) is tight (consider $d$ independent multi-armed bandit problems), as well as classes for which (B.13) is tight (cf. Theorem B.2). Similar issues arise in lower bounds for active learning (Hanneke and Yang 2015). However in the full version of Theorem B.6 (Theorem D.1 of Foster et al. (2020)), we are able to strengthen the lower bound to roughly $\Omega\left(\frac{\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi)+\log|\mathcal{F}|}{\Delta}\right)$ for Natarajan classes.

### Scale-Sensitive Guarantees for the Distribution-Free Setting

We now extend our development based on the star number to give distribution-free upper bounds on the value function disagreement coefficient. Compared to the policy-based setting, where we were able to simply appeal to upper bounds from Hanneke and Yang (2015), scale-sensitive analogues of the star number have not been studied in the literature to our knowledge. This leads us to introduce the following definition.

**Definition B.3** (Value function star number)**.** *Let $\check{\mathfrak{s}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta)$ be the length of the longest sequence of context-action pairs $(x^{(1)}, a^{(1)}), \ldots, (x^{(m)}, a^{(m)})$ such that for all $i$, there exists $f^{(i)} \in \mathcal{F}$ such that*

$$|f^{(i)}(x^{(i)}, a^{(i)}) - f^\star(x^{(i)}, a^{(i)})| > \Delta, \quad \text{and} \quad \sum_{j \neq i}(f^{(i)}(x^{(j)}, a^{(j)}) - f^\star(x^{(j)}, a^{(j)}))^2 \leq \Delta^2.$$

*The value function star number is defined as* $\mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta_0) := \sup_{\Delta > \Delta_0} \check{\mathfrak{s}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta)$.

When the function class $\mathcal{F}$ is $\{0,1\}$-valued, the value function star number coincides with the policy star number, i.e. $\mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, 1) = \mathfrak{s}^{\mathsf{pol}}_{f^\star}(\mathcal{F})$. In general though, for a given class $\mathcal{F}$, the policy star number for the induced class can be arbitrarily large compared to the value function star number.[53]

**Proposition B.2.** *For every* $d \in \mathbb{N}$*, there exists a class* $\mathcal{F}$ *and* $f^\star \in \mathcal{F}$ *such that* $\sup_\Delta \mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta) \leq 5$ *and* $\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi) \geq d$.

Generalizing the result of Hanneke and Yang (2015), we show that the value function star number bounds the value function disagreement coefficient for all distributions and all scale levels.

**Theorem B.7** (Value function star number bounds disagreement coefficient)**.** *For any uniform Glivenko-Cantelli class* $\mathcal{F}$ *and* $f^\star : \mathcal{X} \times \mathcal{A} \to [0,1]$,

$$\sup_{\mathcal{D}} \sup_{\varepsilon > 0} \boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D}; f^\star}(\mathcal{F}, \Delta, \varepsilon) \leq 4(\mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta))^2, \quad \forall \Delta > 0. \tag{B.15}$$

Compared to the bound for the policy star number (Theorem B.5), Theorem B.7 is worse by a quadratic factor when specialized to discrete function classes. Improving (B.15) to be linear in the star number is an interesting technical question. The assumption that $\mathcal{F}$ is uniform Glivenko-Cantelli is quite weak and arises for technical reasons: compared to the policy star number, which always bounds the VC/Natarajan dimension, boundedness of the value function star number is not sufficient to ensure that $\mathcal{F}$ enjoys uniform convergence.

The main takeaway from Theorem B.7 is that AdaCB with Option II guarantees

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}\left( \frac{(\mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2))^2 \cdot A \log|\mathcal{F}|}{\Delta} \right), \tag{B.16}$$

for any distribution. Following our development for the policy star number, we now turn our attention to establishing the necessity of the value function star number for gap-dependent regret bounds. Our lower bound depends on the following "weak" variant of the parameter.

**Definition B.4** (Value function star number (weak variant))**.** *For any* $\Delta \in (0,1)$ *and* $\varepsilon \in (0, \Delta/2)$*, define* $\underline{\mathfrak{s}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta, \varepsilon)$ *be the length of the largest sequence of points* $x^{(1)}, \ldots, x^{(m)}$ *such that for all* $i$*, there exists* $f^{(i)} \in \mathcal{F}$*, such that*

1. $f^{(i)}(x^{(i)}, \pi_{f^{(i)}}(x^{(i)})) \geq \max_{a \neq \pi_{f^{(i)}}(x^{(i)})} f^{(i)}(x^{(i)}, a) + \Delta$ *and* $\pi_{f^{(i)}}(x^{(i)}) \neq \pi^\star(x^{(i)})$.
2. $\max_a |f^{(i)}(x^{(i)}, a) - f^\star(x^{(i)}, a)| \leq 2\Delta$
3. $\sum_{j \neq i} \max_a |f^{(i)}(x^{(j)}, a) - f^\star(x^{(j)}, a)|^2 < \varepsilon^2$.

Relative to the basic value function star number, the key difference above is that we allow a separate scale parameter to control the sum constraint in Item 3 above. This is important to prevent

[53] Interestingly, this construction also shows that in general, the value function star number for $\mathcal{F}$ can be arbitrarily small compared to the fat-shattering dimension. This is somewhat counterintuitive because the star number for a policy class always upper bounds its VC dimension.

passive information leakage in our lower bound construction, but we suspect this condition can be relaxed to more closely match Definition B.3. Our main lower bound is as follows.[54]

**Theorem B.8.** *Let a function class $\mathcal{F}$ and $f^\star \in \mathcal{F}$ with uniform gap $\Delta$ be given. Let $\varepsilon_T \in (0, \Delta/4)$ be the largest solution to the equation*[55]

$$\varepsilon_T^2 T \leq \underline{\mathfrak{s}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2, \varepsilon_T). \tag{B.17}$$

*Then there exists a distribution $\mathcal{D}$ such that for any algorithm with $\mathbb{E}[\mathrm{Reg}] \leq 2^{-6}\frac{\Delta T}{\underline{\mathfrak{s}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2, \varepsilon_T)}$ on all instances realizable by $\mathcal{F}$, there exists an instance with $f^\star$ as the Bayes reward function such that*

$$\mathbb{E}[\mathrm{Reg}] = \Omega\left(\frac{\underline{\mathfrak{s}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2, \varepsilon_T)}{\Delta}\right). \tag{B.18}$$

As mentioned before, we suspect that the linear scaling in (B.18) is correct and that (B.16) can be improved to match. The dependence on the additional scale parameter $\varepsilon_T$ is more subtle, and requires further investigation.

### B.1.5 Adversarial Contexts and the Eluder Dimension

The *eluder dimension* (Russo and Van Roy 2013) is another combinatorial parameter which was introduced to analyze the regret for general function class variants of the UCB algorithm and Thompson sampling for contextual bandits with adversarial contexts.[56] We recall the definition here.[57]

**Definition B.5** (Value function eluder dimension)**.** *Let $\check{\mathfrak{e}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta)$ be the length of the longest sequence of context-action pairs $(x^{(1)}, a^{(1)}), \dots, (x^{(m)}, a^{(m)})$ such that for all $i$, there exists $f^{(i)} \in \mathcal{F}$ such that*

$$|f^{(i)}(x^{(i)}, a^{(i)}) - f^\star(x^{(i)}, a^{(i)})| > \Delta, \quad \text{and} \quad \sum_{j<i}(f^{(i)}(x^{(j)}, a^{(j)}) - f^\star(x^{(j)}, a^{(j)}))^2 \leq \Delta^2. \tag{B.19}$$

*The value function eluder dimension is defined as $\mathfrak{e}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta_0) = \sup_{\Delta > \Delta_0} \check{\mathfrak{e}}^{\mathsf{val}}(\mathcal{F}, \Delta)$.*

The only difference between the value function star number and the value function eluder dimension is whether the sum in (B.19) takes the form "$\sum_{j \neq i}$" or "$\sum_{j<i}$"; the latter reflects the stronger sequential structure present when contexts are adversarial. It is immediate that

$$\mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta) \leq \mathfrak{e}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta). \tag{B.20}$$

However, the separation between the two parameters can be arbitrarily large in general.

[54] To avoid technical conditions involving the boundary of the interval $[0, 1]$, we allow for unit Gaussian rewards with means in $[0, 1]$ for this lower bound.

[55] There is always at least one solution to (B.17), since we can take $\varepsilon_T = 0$.

[56] In the adversarial context setting we allow the contexts $x_1, \dots, x_T$ to be chosen by an adaptive adversary, but we still assume that $r_t \sim \mathbb{P}_r(\cdot \mid x_t)$ at each round.

[57] This definition differs slightly from that of Russo and Van Roy (2013), and in fact is always smaller (yet still sufficient to analyze UCB and Thompson sampling). The original definition allows $\Delta$ in (B.19) to vary as a function of the index $i$.

**Proposition B.3.** *For every $d \in \mathbb{N}$ and $\Delta \in (0,1)$ there exists $\mathcal{F}$ and $f^\star \in \mathcal{F}$ such that $\sup_{\Delta'} \mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta') \leq 2$ and $\mathfrak{e}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2) \geq d$.*

While (B.20) shows that boundedness of the eluder dimension is sufficient for AdaCB achieve logarithmic regret for stochastic contexts (via (B.16)), Proposition B.3, shows that it may lead to rather pessimistic upper bounds. This is not surprising, since the eluder dimension was designed to accomodate adversarially chosen contexts. The next result, which is a small refinement of the analysis of Russo and Van Roy (2013), shows that bounded eluder dimension indeed suffices to guarantee logarithmic regret for the adversarial setting; we defer a precise description of the algorithm to the proof.

**Proposition B.4.** *For the adversarial context setting, the general function class UCB algorithm—when configured appropriately—guarantees that*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}\left(\frac{\mathfrak{e}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2) \cdot \log|\mathcal{F}|}{\Delta}\right) \tag{B.21}$$

*for any instance with uniform gap $\Delta$.*

Paralleling our results for the value function star number, we show that boundedness of a weak variant of the value function eluder dimension is required for logarithmic regret with adversarial contexts.

**Definition B.6** (Value function eluder dimension (weak variant))**.** *For any $\Delta \in (0,1)$ and $\varepsilon \in (0, \Delta/4)$, define $\underline{\mathfrak{e}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta, \varepsilon)$ be the length of the largest sequence of contexts $x^{(1)}, \ldots, x^{(m)}$ such that for all $i$, there exists $f^{(i)} \in \mathcal{F}$, such that*

1. $f^{(i)}(x^{(i)}, \pi_{f^{(i)}}(x^{(i)})) \geq \max_{a \neq \pi_{f^{(i)}}(x^{(i)})} f^{(i)}(x^{(i)}, a) + \Delta$ *and* $\pi_{f^{(i)}}(x^{(i)}) \neq \pi^\star(x^{(i)})$.
2. $\max_a |f^{(i)}(x^{(i)}, a) - f^\star(x^{(i)}, a)| \leq 2\Delta$
3. $\sum_{j<i} \max_a |f^{(i)}(x^{(j)}, a) - f^\star(x^{(j)}, a)|^2 < \varepsilon^2$.

Our main lower bound here shows that—with the same caveats as Theorem B.8—the scaling in (B.21) is near-optimal.[58]

**Theorem B.9.** *Let a function class $\mathcal{F}$ and $f^\star \in \mathcal{F}$ with uniform gap $\Delta$ be given. Let $\varepsilon_T \in (0, \Delta/4)$ be the largest solution to the equation*

$$\varepsilon_T^2 T \leq \underline{\mathfrak{e}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2, \varepsilon_T). \tag{B.22}$$

*Then there exists a distribution $\mathcal{D}$ such that for any algorithm with $\mathbb{E}[\mathrm{Reg}] \leq 2^{-6} \frac{\Delta T}{\underline{\mathfrak{e}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2, \varepsilon_T)}$ on all instances realizable by $\mathcal{F}$, there exists an a sequence $\{x_t\}_{t=1}^T$ and instance with $f^\star$ as the Bayes reward function such that*

$$\mathbb{E}[\mathrm{Reg}] = \Omega\left(\frac{\underline{\mathfrak{e}}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta/2, \varepsilon_T)}{\Delta}\right). \tag{B.23}$$

[58] As with Theorem B.8, we allow for unit Gaussian rewards with means in $[0,1]$ for this lower bound.

**Relating the eluder dimension to the disagreement coefficient** An immediate consequence of Theorem B.7 and (B.20) is that we always have $\boldsymbol{\theta}^{\mathsf{val}}(\mathcal{F}, \Delta, \varepsilon) \leq O(\mathfrak{e}^{\mathsf{val}}(\mathcal{F}, \Delta)^2)$. While this bound scales quadratically, we can show through a more direct argument that the value function disagreement coefficient grows at most linearly with the eluder dimension.

**Theorem B.10** (Value function eluder dimension bounds disagreement coefficient)**.** *For any uniform Glivenko-Cantelli class $\mathcal{F}$ and $f^\star : \mathcal{X} \times \mathcal{A} \to [0,1]$,*

$$\sup_{\mathcal{D}} \sup_{\varepsilon > 0} \boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D}; f^\star}(\mathcal{F}, \Delta, \varepsilon) \leq 4\mathfrak{e}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta), \quad \forall \Delta > 0. \tag{B.24}$$

This result strongly suggests that the quadratic dependence on the value function star number in Theorem B.7 can be improved.

### The Policy Eluder Dimension

Previous work which uses the eluder dimension to analyze algorithms for contextual bandits and reinforcement learning (Russo and Van Roy 2013, Osband and Van Roy 2014, Ayoub et al. 2020, Wang et al. 2020b) only works with the value function-based formulation in Definition B.5. In light of our results for the disagreement coefficient and star number, we propose the following policy-based variant of the eluder dimension.

**Definition B.7** (Policy eluder dimension)**.** *For any policy $\pi^\star$ and policy class $\Pi$, let the policy eluder dimension $\mathfrak{e}^{\mathsf{pol}}_{\pi^\star}(\Pi)$ denote the largest number $m$ such that there exist context-action pairs $(x^{(1)}, a^{(1)}), \ldots, (x^{(m)}, a^{(m)})$ and policies $\pi^{(1)}, \ldots, \pi^{(m)}$ such that for all $i$,*

$$\pi^{(i)}(x^{(i)}) = a^{(i)} \neq \pi^\star(x^{(i)}), \quad \text{and} \quad \pi^{(i)}(x^{(j)}) = \pi^\star(x^{(j)}) \quad \forall j < i : x^{(j)} \neq x^{(i)}$$

Of course, we immediately have that

$$\mathfrak{s}^{\mathsf{pol}}_{\pi^\star}(\Pi) \leq \mathfrak{e}^{\mathsf{pol}}_{\pi^\star}(\Pi). \tag{B.25}$$

We are not yet aware of any upper bounds based on the policy eluder dimension, but we can show that boundedness of this parameter is indeed necessary for logarithmic regret in the adversarial context setting (in a worst-case sense).

**Theorem B.11.** *Consider the adversarial context setting. Let a policy class $\Pi$, $\pi^\star \in \Pi$, and gap $\Delta \in (0, 1/8)$ be given. Then there exists a value function class $\mathcal{F}$ such that:*

1. *$\Pi = \{\pi_f \mid f \in \mathcal{F}\}$, and in particular some $f^\star \in \mathcal{F}$ has $\pi^\star = \pi_{f^\star}$.*
2. *Each $f \in \mathcal{F}$ has uniform gap $\Delta$.*
3. *For any algorithm with $\mathbb{E}[\mathrm{Reg}] \leq \frac{\Delta T}{32\mathfrak{e}^{\mathsf{pol}}_{\pi^\star}(\Pi)}$ for all instances realizable by $\mathcal{F}$, there exists a sequence $\{x_t\}_{t=1}^T$ and instance with $f^\star$ as the Bayes reward function such that*

$$\mathbb{E}[\mathrm{Reg}] = \Omega\left(\frac{\mathfrak{e}^{\mathsf{pol}}_{\pi^\star}(\Pi)}{\Delta}\right). \tag{B.26}$$

### B.1.6 Discussion

#### Proof Techniques

**Policy disagreement-based upper bound (Theorem B.1)** Our proof of Theorem B.1 builds on the regret analysis framework established in Simchi-Levi and Xu (2022), which interprets IGW as maintaining a distribution over policies in the *universal policy space* $\mathcal{A}^{\mathcal{X}}$, and shows that the induced distribution of policies is a solution to an *implicit optimization problem* which (when configured appropriately) provides a sufficient condition for minimax contextual bandit learning. Following this framework, we also view AdaCB's sequential IGW procedure as implicitly maintaining a sequence of distributions over policies, but with an additional key property: the support of the implicit distribution over policies is *adaptively shrinking*. This is enabled by AdaCB's elimination procedure and is essential to our instance-dependent analysis. We show that the implicit distribution over policies given by AdaCB is a solution to a novel *data-driven* implicit optimization problem, which, when configured appropriately by adaptively selecting the learning rate with Option I, provides a sufficient condition for optimal policy disagreement-based instance-dependent contextual bandit learning. Our proof introduces several new techniques to instance-dependent analysis of contextual bandits, including using disagreement-based indicators and disagreement probability to obtain faster policy convergence rates. We also remark that the selection of the adaptive learning rate is non-trivial, and we derive the schedule Option I by carefully balancing key quantities appearing in our analysis. See Foster et al. (2020) for the detailed proof.

**Value function disagreement-based upper bound (Theorem B.3)** The proof of Theorem B.3 consists of two steps. In the first step, we build on the minimax analysis framework of Simchi-Levi and Xu (2022) and show that AdaCB with Option II always guarantees the minimax rate. A new trick that we use here is to carefully track the adaptive value of $\lambda_m$ and use it to infer the exploration cost under the current instance. In the second step, we establish the $\frac{\boldsymbol{\theta}^{\mathsf{val}} \cdot A \log |\mathcal{F}|}{\Delta}$-type instance-dependent upper bound for regret. The analysis is driven by a key inequality (Lemma C.21 of Foster et al. (2020)), which provides a sharp upper bound on $\mathbb{E}_{\mathcal{D}}[w(x; \mathcal{F}_m)]$ in terms of the ratio $\frac{\boldsymbol{\theta}^{\mathsf{val}}}{\Delta}$. Beyond giving a means to bound the (expected) instantaneous regret in terms of $\boldsymbol{\theta}^{\mathsf{val}}$ and $\Delta$, this allows us to adaptively maintain an estimated lower bound for $\frac{\boldsymbol{\theta}^{\mathsf{val}}}{\Delta}$ based on empirical data. We then use an induction argument to show that the specification of $\gamma_m$ in Option II enables AdaCB to enjoy a near-optimal instance-dependent guarantee.

**Lower bounds** Our lower bounds build on the work of Raginsky and Rakhlin (2011), which provides information-theoretic lower bounds for passive and active learning in terms of the disagreement coefficient. As in this work, we rely on a specialized application of the Fano method using the *reverse* KL-divergence, but with some refinements to make the technique more suited for *regret* lower bounds. For Theorem B.2, we also incorporate improvements to the method suggested by Hanneke (2014) to obtain the correct dependence on $\log|\mathcal{F}|$.

**Value function star number bounds disagreement coefficient (Theorem B.7)** The proof of Theorem B.7 is somewhat different from the proof of the analogous policy-based result

by Hanneke and Yang (2015). The key step toward proving Theorem B.7 is to prove an empirical analogue of the result that holds whenever $\mathcal{D}$ is uniform over a finite sequence of examples. This result is given in Lemma E.1 of the full version of our paper (Foster et al. 2020), and is motivated by a property of the eluder dimension established in Proposition 3 of Russo and Van Roy (2013), with their "$\sum_{j<i}$"-based definition changed to our "$\sum_{j\neq i}$"-based definition. The proof of our result is trickier, however, as our "$\sum_{j\neq i}$"-based definition breaks several combinatorial properties utilized in the proof of Russo and Van Roy (2013). We address this challenge by proving a new combinatorial lemma (Lemma E.3 of Foster et al. (2020)), which is fairly general and may be interesting on its own right. Nevertheless, our upper bound is quadratic in $\mathfrak{s}^{\mathsf{val}}_{f^\star}(\mathcal{F}, \Delta)$ rather than linear, and we hope that this dependence can be improved in future work.

## Related Work

Gap-dependent regret bounds for contextual bandits have not been systematically studied at the level of generality we consider here, and we are not aware of any prior lower bounds beyond the linear setting. Most prior work has focused on structured function classes such as linear (Dani et al. 2008, Abbasi-Yadkori et al. 2011, Hao et al. 2019) and nonparametric Lipschitz/Hölder classes (Rigollet and Zeevi 2010, Perchet and Rigollet 2013, Hu et al. 2020).

Our work draws inspiration from Krishnamurthy et al. (2017), who defined variants of the disagreement coefficient which depend on scale-sensitive properties of the class $\mathcal{F}$ in the context of cost-sensitive multiclass active learning. Compared to these results, the key difference is that our value function disagreement coefficient is defined in terms of the $L_2$ ball for the class $\mathcal{F}$ rather than the excess risk ball for the induced policy class. This change is critical to ensure that the value function disagreement coefficient is bounded by the value function star number, and in particular that it is always bounded for linear classes.

Our work also builds on Foster et al. (2018), who give instance-dependent guarantees for the generalized UCB algorithm and an action elimination variant for general function classes based on the cost-sensitive multiclass disagreement coefficients introduced in Krishnamurthy et al. (2017). We improve upon this result on several fronts: 1) As mentioned above, our notion of value function disagreement coefficient is tighter, and is always bounded by the value function star number and value function eluder dimension 2) we attain optimal dependence on the gap, 3) our algorithms are guaranteed to attain the minimax rate in the worst case, and 4) we complement these results with lower bounds.

Lastly, we mention that while there are no prior lower bounds for contextual bandits based on the eluder dimension, Wen and Van Roy (2017) give an eluder-based lower bound for reinforcement learning with deterministic transitions and known rewards. This result is closer in spirit to our disagreement-based lower bounds (Theorems B.2 and B.4), is it applies to a carefully constructed function class rather than holding for all function classes, and mainly serves to demonstrate the worst-case tightness of a particular upper bound.

### B.1.7 Extensions

We conclude this section by presenting some basic extensions of our contextual bandit results, including extensions of our regret bounds to handle infinite classes and weaker noise conditions.

**Infinite function classes** As we have mentioned, Algorithm 3.1, Theorem B.1 and Theorem B.3 trivially extend to infinite $\mathcal{F}$, with the dependence on $\log|\mathcal{F}|$ in the algorithm's parameters and the regret bounds replaced by standard learning-theoretic complexity measures such as the pseudodimension, (localized) Rademacher complexity, or metric entropy. This is because the analysis of AdaCB (see Appendix C of Foster et al. (2020)) does not rely on any complexity assumptions for $\mathcal{F}$, except for Lemma C.1 of Foster et al. (2020), which uses a standard uniform martingale concentration bound for the square loss to show that the empirical risk minimizer $\widehat{f}_m$ has low excess risk at each epoch. Therefore, to extend our results to infinite $\mathcal{F}$, one only needs to replace Lemma C.1 of Foster et al. (2020) with an analogous uniform martingale concentration inequality for infinite classes. Such results have already been established in the literature, see, e.g., Krishnamurthy et al. (2017) and Foster et al. (2018).

**Alternative noise conditions** Beyond uniform gap, AdaCB can also adapt to the Tsybakov noise condition (Mammen and Tsybakov 1999, Tsybakov 2004, Audibert and Tsybakov 2007, Rigollet and Zeevi 2010, Hu et al. 2020), as the following proposition shows.

**Proposition B.5** (Regret under the Tsybakov noise condition)**.** *Suppose there exist constants* $\alpha, \beta \geq 0$ *such that*

$$\mathbb{P}_{\mathcal{D}}\left(f^\star(x,\pi^\star(x)) - \max_{a\neq\pi^\star(x)} f^\star(x,a) \leq \gamma\right) \leq \beta\gamma^\alpha, \quad \forall \gamma \geq 0.$$

*Then Algorithm 3.1 with* Option I *ensures that*

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1)\cdot \min_{\varepsilon>0} \max\left\{\varepsilon T, \left(\boldsymbol{\theta}^{\mathsf{pol}}(\Pi,\varepsilon)A\log|\mathcal{F}|\right)^{\frac{1+\alpha}{2+\alpha}} T^{\frac{1}{2+\alpha}}\right\} + \widetilde{O}(1).$$

**Tightening the value function disagreement coefficient** The supremum over the action distribution $p$ in the definition (B.7) of the value function disagreement coefficient is more pessimistic than what is actually required to analyze AdaCB. Consider the following action distribution-dependent definition:

$$\boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D},p;f^\star}(\mathcal{F},\Delta_0,\varepsilon_0) = \sup_{\Delta>\Delta_0,\varepsilon>\varepsilon_0} \frac{\Delta^2}{\varepsilon^2}\mathbb{P}_{\mathcal{D},p}\Big(\exists f\in\mathcal{F} : |f(x,a)-f^\star(x,a)|>\Delta,\ \|f-f^\star\|_{\mathcal{D},p}\leq\varepsilon\Big). \tag{B.27}$$

The regret bound in Theorem B.3 can be tightened to depend on $\sup_{p\in\mathcal{P}} \boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D},p;f^\star}(\mathcal{F},\Delta/2,\varepsilon_T)$, where $\mathcal{P}$ is a set of action distributions with favorable properties that can lead to tighter bounds. In particular, the proof of Theorem B.3 implies that for any instance with uniform gap $\Delta$, if $\boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D},p;f^\star}(\mathcal{F},\Delta/2,\varepsilon_T)\leq\theta$ for all $p$ such that $p(\pi^\star(x)|x)\geq A^{-1}$ for all $x$, then AdaCB with Option II ensures that

$$\mathbb{E}[\mathrm{Reg}] = \widetilde{O}(1)\cdot\min\left\{\sqrt{AT\log|\mathcal{F}|}, \frac{\theta A\log|\mathcal{F}|}{\Delta}\right\} + O(1).$$

The following result shows that this property leads to dimension-independent bounds for sparse linear function classes.

**Proposition B.6.** *Consider the function class* $\mathcal{F} = \big\{(x,a) \mapsto \langle w, \phi(x,a)\rangle \mid w \in \mathbb{R}^d, \|w\|_0 \leq s\big\}$, *where* $\|\phi(x,a)\|_\infty \leq 1$. *Define* $\Sigma^\star = \mathbb{E}_{\mathcal{D}}\big[\phi(x,\pi^\star(x))\phi(x,\pi^\star(x))^\top\big]$, *and let* $\lambda_{\mathrm{re}} = \inf_{w\neq 0, \|w\|_0 \leq 2s} \langle w, \Sigma^\star w\rangle / \|w\|_2^2$ *be the restricted eigenvalue. Then* $\forall \Delta, \varepsilon > 0$,

$$\boldsymbol{\theta}^{\mathsf{val}}_{\mathcal{D},p;f^\star}(\mathcal{F},\Delta,\varepsilon) \leq 2\alpha^{-1}\lambda_{\mathrm{re}}^{-1} s$$

*for all* $p$ *such that* $p(\pi^\star(x)|x) \geq \alpha$ *for all* $x$.

As a concrete example, if $\phi(x,\pi^\star(x)) \sim \mathrm{Unif}(\{\pm 1\}^d)$ we have $\lambda_{\mathrm{re}} = 1$, so that the bound is indeed dimension-independent.

**Handling multiple optimal actions** For simplicity, we assume that $\arg\max_{a\in\mathcal{A}} f^\star(x,a)$ is unique for all $x$ in the main body of the paper. When such assumption does not hold, we keep the original definition of $\pi^\star(x)$ (which makes $\pi^\star(x)$ unique for each $x \in \mathcal{X}$), while defining

$$\pi^\star_{\mathrm{set}}(x) := \{a \in \mathcal{A} \mid f^\star(x,a) = \max_{a'\in\mathcal{A}} f^\star(x,a')\}, \quad \forall x \in \mathcal{X}.$$

We then make the following modifications to our framework. First, we modify the uniform gap condition (3.2) to require that for all $x \in \mathcal{X}$,

$$f^\star(x,\pi^\star(x)) - f^\star(x,a) \geq \Delta \quad \forall a \notin \pi^\star_{\mathrm{set}}(x).$$

Second, we modify the definition of policy disagreement coefficient to

$$\boldsymbol{\theta}^{\mathsf{pol}}_{\mathcal{D},\pi^\star}(\Pi,\varepsilon_0) = \sup_{\varepsilon\geq\varepsilon_0} \frac{\mathbb{P}_{\mathcal{D}}(x : \exists \pi \in \Pi_\varepsilon : \pi(x) \neq \pi^\star(x))}{\varepsilon},$$

where $\Pi_\varepsilon := \{\pi \in \Pi : \mathbb{P}_{\mathcal{D}}(\pi(x) \notin \pi^\star_{\mathrm{set}}(x)) \leq \varepsilon\}$. By doing so, all our guarantees for AdaCB extend to the general setting where $\arg\max_{a\in\mathcal{A}} f^\star(x,a)$ may not be unique for some $x \in \mathcal{X}$.

## B.2 Details for Results of Reinforcement Learning

In this section, we give disagreement-based guarantees for reinforcement learning with function approximation in the block MDP setting (cf. Section 3.3.1). The proofs of the results mentioned in this section can be found in the full version of our paper (Foster et al. 2020).

Before proceeding, let us introduce some additional notation.

**Additional notation** For any Markov policy $\pi(x)$, let $\mathbf{Q}^\pi_h(x,a) = \mathbb{E}\Big[\sum_{h'\geq h}^H r_{h'} \mid x_h = x, a_h = a\Big]$ be the corresponding Q-function. We likewise define $\mathbf{V}^\pi_h(x) = \max_{a\in\mathcal{A}} \mathbf{Q}^\pi(x,a)$, as well as $\mathbf{V}^\pi = \mathbb{E}_{x_1}[\mathbf{V}^\pi_1(x_1)]$ and $V^\star = \mathbb{E}_{x_1}[V^\star_1(x_1)]$. Next, for any function $V : \mathcal{X} \to \mathbb{R}$ we define the transition operator by

$$[P^\star_h V](x,a) = \mathbb{E}[V(x_{h+1}) \mid x_h = x, a_h = a].$$

We also define the Bayes reward function as

$$f^\star(x,a) = \mathbb{E}[r_h \mid x_h = x, a_h = a]$$

for each $x \in \mathcal{X}_h$. Finally, let $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_H$ be the full regression function class, and define $F_{\max} = \max_h |\mathcal{F}_h|$.

### B.2.1 The Algorithm

Our main reinforcement learning algorithm, RegRL, is presented in Algorithm B.1. The algorithm follows the optimistic least-squares value iteration framework (Jin et al. 2020, Wang et al. 2019, 2020b), with a few key changes that allow us to prove guarantees based on a suitable notion of value function disagreement coefficient rather than stronger complexity measures such as the eluder dimension. The most interesting aspect of the algorithm is a feature we call the *star hull upper confidence bound*: Compared to the classical UCB approach, which computes an optimistic $Q$-function by taking largest predicted reward amongst all value function in an $L_2$ ball around an empirical risk minimizer, we add an additional step which first "lightly convexifies" this set. This step is based on techniques from the literature on aggregation in least squares (Audibert 2008, Liang et al. 2015), and leads to more stable predictions.

In more detail, the algorithm proceeds in $K$ iterations.[59] In each iteration $k$, we compute an optimistic Q-function $\overline{\mathbf{Q}}^{(k)}$ such that

$$\overline{\mathbf{Q}}_h^{(k)}(x,a) \geq Q_h^\star(x,a) \quad \text{for all } x,\, a,\, h. \tag{B.28}$$

We then take the greedy argmax policy defined by $\pi^{(k)}(x) = \arg\max_{a\in\mathcal{A}} \overline{\mathbf{Q}}_h^{(k)}(x,a)$ for $x \in \mathcal{X}_h$, and gather $H$ trajectories as follows: For each $h$, we roll in to layer $h$ with $\pi^{(k)}$, then choose actions uniformly at random for the rest of the episode. These trajectories are used to refine our value function estimates for subsequent iterations, with the $h$th trajectory used for estimation at layer $h$. Choosing actions uniformly ensures that the data gathered from these trajectories is useful regardless of the action distribution in subsequent iterations.

Let us now elaborate on the upper confidence bound computation. Let iteration $k$ and layer $h$ be fixed, and suppose we have already computed $\overline{\mathbf{Q}}_{h+1}^{(k)}$ and $\overline{\mathbf{V}}_{h+1}^{(k)}(x) := \max_{a\in\mathcal{A}} \overline{\mathbf{Q}}_{h+1}^{(k)}(x,a)$. The first step, following the usual optimistic LSVI schema, is to estimate a value function for layer $h$ by regressing onto the empirical Bellman backups from the next layer (Line 5):

$$\widehat{f}_h^{(k)} = \arg\min_{f\in\mathcal{F}_h} \sum_{j<k} \left(f(x_h^{(j,h)}, a_h^{(j,h)}) - \left(r_h^{(j,h)} + \overline{\mathbf{V}}_{h+1}^{(k)}(x_{h+1}^{(j,h)})\right)\right)^2; \tag{B.29}$$

here the $(j,h)$ superscript on $(x_h^{(j,h)}, a_h^{(j,h)}, r_h^{(j,h)}, x_{h+1}^{(j,h)})$ indicates that the example was collected in the $h$th trajectory at iteration $j$. Assumption 3.3 ensures that this regression problem is well-specified.

[59] We use the term "iteration" distinctly from the term "episode", as each iteration consists of multiple episodes.

Let $\mathcal{Z}_h^{(k)} = \{(x_h^{(j,h)}, a_h^{(j,h)})\}_{j<k}$, and define

$$\|f - f'\|_{\mathcal{Z}}^2 = \sum_{(x,a)\in\mathcal{Z}} (f(x,a) - f'(x,a))^2. \tag{B.30}$$

At this point, the usual optimistic value function for layer $h$ (cf. Russo and Van Roy (2013), Foster et al. (2018) for contextual bandits and Jin et al. (2020), Wang et al. (2019, 2020b) for RL) is defined as

$$\overline{\mathbf{Q}}_h(x,a) = \sup\Big\{f(x,a) \mid f \in \mathcal{F}_h, \big\|f - \widehat{f}_h^{(k)}\big\|_{\mathcal{Z}_h^{(k)}} \le \beta_h\Big\},$$

where $\beta_h$ is a confidence parameter. As observed in Jin et al. (2020), Wang et al. (2020b), however, this UCB function can be unstable, leading to issues with generalization when we use it as a target for least squares at layer $h-1$. Our approach to address this problem is to expand the supremum above to include the *star hull* of $\mathcal{F}_h$ centered at $\widehat{f}_h^{(k)}$. Define the star hull of $\mathcal{F}_h$ centered at $f \in \mathcal{F}_h$ by

$$\mathrm{star}(\mathcal{F}, f) = \bigcup_{f'\in\mathcal{F}} \mathrm{conv}(\{f', f\}) = \{t(f' - f) + f \mid f' \in \mathcal{F}, t \in [0,1]\}. \tag{B.31}$$

We define the *star hull upper confidence bound* (Line 6) by

$$\overline{\mathbf{Q}}_h^{(k)}(x,a) = \sup\Big\{f(x,a) \mid f \in \mathrm{star}(\mathcal{F}_h, \widehat{f}_h^{(k)}), \big\|f - \widehat{f}_h^{(k)}\big\|_{\mathcal{Z}_h^{(k)}} \le \beta_h\Big\}. \tag{B.32}$$

When $\mathcal{F}_h$ is convex this coincides with the usual upper confidence bound, but in the star hull operation convexifies $\mathcal{F}_h$ along rays emanating from $\widehat{f}_h^{(k)}$. This small amount of convexification (note that $\mathrm{star}(\mathcal{F}_h, f)$ is still non-convex if, e.g., $\mathcal{F}_h$ is a finite class), ensures that $\overline{\mathbf{Q}}_h^{(k)}(x,a)$ is Lipschitz as a function of the confidence radius $\beta_h$, which stabilizes the predictions and facilitates a tight generalization analysis.

**Oracle efficiency** RegRL is oracle-efficient, and can be implemented using an offline regression oracle as follows.

- At each iteration, the empirical risk minimizer in Line 5 can be computed with a single oracle call.
- For any $(x,a)$ pair, the star hull UCB function in Line 6 can be computed by reduction to a regression oracle. In particular, to compute an $\varepsilon$-approximate UCB:
  - For convex function classes, $O(\log(1/\varepsilon))$ calls are required.
  - For general (in particular, finite) classes, $\widetilde{O}(\varepsilon^{-3})$ oracle calls are required. The key idea here is that we can reduce ERM over the star hull to ERM over the original class.

See Section 4 of the full version of our paper (Foster et al. 2020) for more details.

**Algorithm B.1** RegRL

**input:** Value function classes $\mathcal{F}_1, \dots, \mathcal{F}_H$. Number of iterations $K$.
**initialization:**

- Let $\delta = 1/KH$. `// Failure probability.`

- Let $\beta_H^2 = 400H^2 \log(F_{\max} HK\delta^{-1})$ `// Confidence radius.`

  and $\beta_h^2 = \frac{1}{2}\beta_{h+1}^2 + 60^4 H^2 A^2 \boldsymbol{\theta}^{\mathsf{val}}_{h+1}(\mathcal{F}_{h+1}, \beta_{h+1}K^{-1/2})^2 \log^2(HKe) \log(2F_{\max}HK\delta^{-1}) + 700H^2 S \log(2eK)$ for all $1 \le h \le H-1$.

**algorithm:**

**for** iteration $k = 1, \dots, K$ **do**
  Set $\overline{\mathbf{V}}^{(k)}_{H+1}(x) = 0$.
  Define $\mathcal{Z}_h^{(k)} = \{(x_h^{(j,h)}, a_h^{(j,h)})\}_{j<k}$.
  **for** $h = H, \dots, 1$ **do**
    Set $\widehat{f}_h^{(k)} = \arg\min_{f \in \mathcal{F}_h} \sum_{j<k} \big(f(x_h^{(j,h)}, a_h^{(j,h)}) - \big(r_h^{(j,h)} + \overline{\mathbf{V}}^{(k)}_{h+1}(x_{h+1}^{(j,h)})\big)\big)^2$.
    `// Compute optimistic value function via star-hull upper confidence bound.`
    Define

$$\overline{\mathbf{Q}}_h^{(k)}(x,a) = \sup\Big\{ f(x,a) \mid f \in \mathrm{star}(\mathcal{F}_h, \widehat{f}_h^{(k)}), \big\|f - \widehat{f}_h^{(k)}\big\|_{\mathcal{Z}_h^{(k)}} \le \beta_h \Big\}.$$

    $\pi^{(k)}(x) := \arg\max_{a \in \mathcal{A}} \overline{\mathbf{Q}}_h^{(k)}(x,a)$ for all $x \in \mathcal{X}_h$.
    $\overline{\mathbf{V}}_h^{(k)}(x) := \max_{a \in \mathcal{A}} \overline{\mathbf{Q}}_h^{(k)}(x,a)$.
  **for** $h = 1, \dots, H$ **do**
    Gather trajectory $(x_1^{(k,h)}, a_1^{(k,h)}, r_1^{(k,h)}), \dots, (x_H^{(k,h)}, a_H^{(k,h)}, r_H^{(k,h)})$ by rolling in with $\pi^{(k)}$ for layers $1, \dots, h-1$ and selecting actions uniformly for layers $h, \dots, H$.
**return** $\pi^{(k)}$ for $k \sim \mathrm{Unif}([K])$.

### B.2.2 Main Result

We now state the main guarantee for RegRL. Our guarantee depends on the following "per-state" gap and worst-case gap:

$$\Delta(s) = \min_a \inf_{x \in \mathrm{supp}(\psi(s))} \{\Delta(x,a) \mid \Delta(x,a) > 0\},$$
$$\Delta_{\min} = \min_s \Delta(s),$$

where we recall that $\Delta(x,a) := V_h^\star(x) - Q_h^\star(x,a)$. We adapt the value function disagreement coefficient to the block MDP setting as follows. Let $\pi_{\mathrm{unif}}$ be the policy that selects actions uniformly

from $\mathcal{A}$. For each latent state $s \in \mathcal{S}_h$, we define

$$\boldsymbol{\theta}^{\mathsf{val}}_s(\mathcal{F}_h, \varepsilon_0) = \sup_{f^\star \in \mathcal{F}_h} \sup_{\varepsilon \geq \varepsilon_0} \frac{1}{\varepsilon^2} \mathbb{E}_{x \sim \psi(s), a \sim \pi_{\mathrm{unif}}} \sup\Big\{ |f(x,a) - f^\star(x,a)|^2 \mid f \in \mathcal{F}_h, \|f - f^\star\|_s \leq \varepsilon \Big\}, \tag{B.33}$$

where $\|f\|_s^2 := \mathbb{E}_{x \sim \psi(s), a \sim \pi_{\mathrm{unif}}}[f^2(x,a)]$. This notion is closely related to the value function disagreement coefficient (B.7) for the contextual bandit setting via Markov's inequality, with the context distribution $\mathcal{D}$ replaced by the latent state's emmission distribution $\psi(s)$. We additionally define the total disagreement for layer $h$ by

$$\boldsymbol{\theta}^{\mathsf{val}}_h(\mathcal{F}_h, \varepsilon) = \sum_{s \in \mathcal{S}_h} \boldsymbol{\theta}^{\mathsf{val}}_s(\mathcal{F}_h, \varepsilon),$$

and define $\boldsymbol{\theta}^{\mathsf{val}}_{\max}(\mathcal{F}, \varepsilon) = \max_h \max_{s \in \mathcal{S}_h} \boldsymbol{\theta}^{\mathsf{val}}_s(\mathcal{F}_h, \varepsilon)$.

Our main theorem bounding the error of RegRL is as follows. As with our contextual bandit results, we focus on finite classes $\mathcal{F}$ for simplicity, but the result trivially extends to general function classes.

**Theorem B.12.** *Algorithm B.1 guarantees that*

$$V^\star - \mathbb{E}[\mathbf{V}^\pi] = \widetilde{O}\Big( \frac{\boldsymbol{\theta}^{\mathsf{val}}_{\max}(\mathcal{F}, \beta_H K^{-1/2})^3 \cdot H^5 A^3 S^2 \log|\mathcal{F}|}{\Delta_{\min} K} \Big),$$

*and does so using at most $HK$ trajectories. More generally, the algorithm guarantees that*

$$V^\star - \mathbb{E}[\mathbf{V}^\pi] = \widetilde{O}\Big( C_{\mathcal{M}} \cdot \frac{H^2 A^3 \max_h \boldsymbol{\theta}^{\mathsf{val}}_h(\mathcal{F}, \beta_h K^{-1/2})^2 \log|\mathcal{F}| + H^3 SA}{K} \Big),$$

*where* $C_{\mathcal{M}} := \sum_{h=1}^H \sum_{s \in \mathcal{S}_h} \frac{\boldsymbol{\theta}^{\mathsf{val}}_s(\mathcal{F}_h, \beta_h K^{-1/2})}{\Delta(s)}$.

Let us describe a few key features of this theorem and interpret the result.

- First, if $\boldsymbol{\theta}^{\mathsf{val}}_{\max}(\mathcal{F}, \varepsilon) \propto \mathrm{polylog}(1/\varepsilon)$ (e.g., for a linear function class), then—ignoring other parameters—we can attain an $\varepsilon$-optimal policy using $\frac{1}{\Delta_{\min}\varepsilon}$ trajectories. This *fast rate* improves over the minimax optimal $\varepsilon^{-2}$ rate, and is optimal even for bandits. This is the first fast rate result we are aware of for reinforcement learning in block MDPs.

- In light of the results in Appendix B.1 this implies that one can attain the fast $\frac{1}{\Delta_{\min}\varepsilon}$ rate whenever the value function star number for $\mathcal{F}$ is bounded.

- More generally, if $\boldsymbol{\theta}^{\mathsf{val}}_{\max}(\mathcal{F}, \varepsilon) \propto \varepsilon^{-\rho}$ for $\rho < 2/3$, then $(\Delta\varepsilon)^{-\frac{2}{2-3\rho}}$ trajectories suffice for an $\varepsilon$-optimal policy. However, the guarantee becomes vacuous once $\rho \geq 2/3$. In other words, bounded disagreement coefficient is essentially for the algorithm to have low error, and it does not necessarily attain the minimax rate if this fails to hold. Achieving a best-of-both-worlds guarantee similar to our results for contextual bandits is an interesting direction for future work.

- To the best of our knowledge this is the first oracle-efficient algorithm that attains near-optimal statistical performance in terms of $\varepsilon$ for block MDPs. Of course, this is only achieved in the low-noise regime where $\Delta_{\min} > 0$, and when $\boldsymbol{\theta}^{\mathsf{val}}_{\max}(\mathcal{F}, \varepsilon) \propto \mathrm{polylog}(1/\varepsilon)$.

We emphasize that while the dependence on all of the parameters in Theorem B.12 can almost certainly be improved, we hope this result will open the door for further disagreement-based algorithms and analysis techniques in reinforcement learning.

### B.2.3 Discussion

#### Proof Techniques

The proof of Theorem B.12 has two main components. The first part of the proof shows that with high probability, for all iterations $k$ and layers $h$, the set $\mathcal{F}_h^{(k)} := \{f \in \mathcal{F}_h \mid \|f - \widehat{f}_h^{(k)}\|_{\mathcal{Z}_h^{(k)}} \leq \beta_h\}$ contains the Bellman backup $\big[P_h^\star \overline{\mathbf{V}}_{h+1}^{(k)}\big](x,a) + f^\star(x,a)$ of the value function from the next layer, which ensures that $\overline{\mathbf{Q}}_h^{(k)}$ is optimistic in the sense of (B.28) and leads to exploration. Then, in the second part, we prove a regret decomposition which shows that whenever the optimistic property holds, the suboptimality of $\pi^{(k)}$ is controlled by the gap $\Delta$ and the value function disagreement coefficient $\boldsymbol{\theta}^{\mathsf{val}}$.

The first part of the proof (Appendix G of Foster et al. (2020)) boils down to showing that the empirical risk minimizer in $\widehat{f}_h^{(k)}$ in (B.29) has favorable concentration properties. This is highly non-trivial because the targets $\overline{\mathbf{V}}_{h+1}^{(k)}$ in (B.29) depend on the entire dataset, which breaks the independence assumptions required to apply standard generalization bounds for least squares. Instead, following Jin et al. (2020), Wang et al. (2020b), we opt for a *uniform* generalization bound which holds uniformly over all possible choices of $\overline{\mathbf{V}}_{h+1}^{(k)}$. To do so, we must show that $\overline{\mathbf{V}}_{h+1}^{(k)}$ is approximated by a relatively low complexity function class, which we accomplish as follows. First, we show that—thanks to a certain Lipschitz property granted by the star hull—$\overline{\mathbf{Q}}_{h+1}^{(k)}$ is well approximated by a function

$$\widetilde{\mathbf{Q}}_{h+1}^{(k)}(x,a) := \sup\Big\{f(x,a) \mid f \in \mathrm{star}(\mathcal{F}_{h+1}, \widehat{f}_{h+1}^{(k)}), \big\|f - \widehat{f}_{h+1}^{(k)}\big\|_{\mathcal{L}_{h+1}^{(k)}} \leq \widetilde{\beta}_{h+1}\Big\},$$

where $\widetilde{\beta}_{h+1} \approx \beta_{h+1}$, and where

$$\|f\|^2_{\mathcal{L}_{h+1}^{(k)}} := \sum_{j<k} \mathbb{E}_{x_{h+1} \sim \psi(s_{h+1}^{(j,h+1)}), a \sim \pi_{\mathrm{unif}}}[f^2(x,a)]$$

is the *latent state norm*, which measures the expected squared error conditioned on the sequence of latent states $\mathcal{L}_{h+1}^{(k)} := (s^{(1,h+1)}, \ldots, s^{(k-1,h+1)})$ encountered in the trajectories gathered for layer $h+1$. This approximation argument is rather non-trivial, and involves a recursion across all layers that we manage using the disagreement coefficient. With this taken care of, the next step is to use the block MDP structure to argue that $\widetilde{\mathbf{Q}}_{h+1}^{(k)}$ has low complexity. To see this, observe that $\widetilde{\mathbf{Q}}_{h+1}^{(k)}$ is completely determined by the center $\widehat{f}_{h+1}^{(k)}$ and the latent state sequence above. Since the latent state constraint does not depend on the ordering of the latent states, we can use a counting argument to show that there are at most $|\mathcal{F}|K^{O(S)}$ possible choices for $\widetilde{\mathbf{Q}}_{h+1}^{(k)}$ overall. This suffices to prove the desired concentration guarantee.

The second part of the proof (Appendix F of Foster et al. (2020)) proceeds as follows. Define the Bellman surplus as

$$\overline{\mathbf{E}}_h^{(k)}(x,a) = \overline{\mathbf{Q}}_h^{(k)}(x,a) - \big(f^\star(x,a) + [P_h^\star \overline{\mathbf{V}}_{h+1}^{(k)}](x,a)\big),$$

which measures the width for our upper confidence bound. We use a "clipped" regret decomposition from Simchowitz and Jamieson (2019) to show that whenever the concentration event from the first part of the proof holds, the suboptimality of $\pi^{(k)}$ is controlled by the confidence widths:

$$V^\star - \mathbf{V}^{\pi^{(k)}} \lesssim \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h} \mathbb{P}_{\pi^{(k)}}(s_h = s) \cdot \frac{\mathbb{E}_{x_h \sim \psi(s_h), a_h \sim \pi_{\mathrm{unif}}}\big[\overline{\mathbf{E}}_h^{(k)}(x_h,a_h)^2\big]}{\Delta(s)}.$$

In particular, let $n^{(k,h)}(s)$ denote the number of times the latent state $s$ was encountered in the layer $h$ trajectories prior to iteration $k$. Our key observation is that bounded disagreement coefficient implies that for each state $s$,

$$\mathbb{E}_{x_h \sim \psi(s), a_h \sim \pi_{\mathrm{unif}}}\big[\overline{\mathbf{E}}_h^{(k)}(x_h,a_h)^2\big] \lesssim \boldsymbol{\theta}_s^{\mathsf{val}} \cdot \frac{\beta_h^2}{n^{(k,h)}(s)}.$$

In other words, the disagreement coefficient controls the rate at which the confidence width shrinks. Moreover, since the width for latent state $s$ is proportional to the number of times we have visited the state (even though the algorithm cannot observe this quantity), we can bound the overall suboptimality across all iterations using similar arguments to those employed in the tabular setting (Azar et al. 2017, Simchowitz and Jamieson 2019).

## Related Work

Our result is closely related to that of Wang et al. (2020b), who gave regret bounds for a variant of optimistic LSVI based on the eluder dimension of $\mathcal{F}$. Compared to this result, we require the additional block MDP assumption and finite actions, but our bounds scale with the value function disagreement coefficient, which can be arbitrarily small compared to the eluder dimension (Proposition B.3). On the technical side, their algorithm stabilizes the upper confidence bounds using a sensitivity sampling procedure, whereas we address this issue using the star hull. Ayoub et al. (2020) give similar eluder dimension-based guarantees for a model-based algorithm, though the notion of eluder dimension is somewhat stronger, and it is not clear whether this algorithm can be made oracle-efficient.

Reinforcement learning with function approximation in block MDPs has been the subject of extensive recent investigation (Krishnamurthy et al. 2016, Jiang et al. 2017, Dann et al. 2018, Du et al. 2019a,b, Misra et al. 2019, Feng et al. 2020, Agarwal et al. 2020a). In terms of assumptions, we require the rather strong optimistic completeness condition, but do not require any reachability conditions or any clusterability-type assumptions that facilitate the use of unsupervised learning. The main advantages of our results are 1) we require only a basic regression oracle for the value function class, and 2) we attain the optimal $\varepsilon^{-1}$ fast rate in the presence of the gap and bounded disagreement coefficient.

We should also mention that the gap for $Q^\star$ has been used in a number of recent results on

reinforcement learning with function approximation (Du et al. 2019a, 2020a,b), albeit for a somewhat different purpose. These results use the gap to prove that certain "non-optimistic" algorithms succeed, whereas we use it to beat the minimax rate.

Lastly, we note that the value function disagreement coefficient is similar to the "low variance" parameter used in Du et al. (2019a) to give guarantees for reinforcement learning with linear function approximation, but can be considerably smaller when applied to block MDPs. For example, in the trivial case in which each emission distribution $\psi(s)$ is a singleton, the value function disagreement coefficient is automatically bounded by 1, while the low variance assumption may not be satisfied unless the latent MDP is near-deterministic.

# Appendix C

# Suppmentary Material for Chapter 4

## C.1 General Scheme to Construct Hard Families of Instances

Recall that Section 4.2.1 gives specific numerical values for the parameters that define the model class $\mathcal{M}$ used in our lower bound construction. The precise values are not critical for our proof, and in this section we give general conditions on the parameters under which one can derive a similar lower bound. In doing so, we also provide some intuition behind the specific choice of parameters used for Theorem 4.1.

In more detail, for any tuple of parameters $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; w)$, we consider the family of MDPs $\mathcal{M}$ given by

$$\mathcal{M}_1 := \bigcup_{I \in \mathcal{I}_{\theta_1}} M_{\alpha_1, \beta_1, w, I}, \quad \mathcal{M}_2 := \bigcup_{I \in \mathcal{I}_{\theta_2}} M_{\alpha_2, \beta_2, w, I}, \quad \mathcal{M} := \mathcal{M}_1 \cup \mathcal{M}_2.$$

There are 7 independent scalars in the tuple, all of which lie in $[0, 1]$: $\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2, w$; note that the parameter $w$ is shared between $\mathcal{M}_1$ and $\mathcal{M}_2$. The family $\mathcal{M}$ above can be used to derive a hardness result similar to Theorem 4.1 as long as the following three general equality and inequality constraints are satisfied.

- All $M \in \mathcal{M}$ have the same marginal distribution for $s'$ under the process $s \sim \mathrm{Unif}(\mathcal{S}^1)$, $s' \sim P(s, \mathfrak{a})$:

  $$\theta_1 \alpha_1 = \theta_2 \alpha_2, \quad \text{and} \quad (1 - \theta_1)\beta_1 = (1 - \theta_2)\beta_2. \tag{C.1}$$

  This ensures that the learner cannot trivially test whether $M \in \mathcal{M}_1$ or $\mathcal{M}_2$ using marginals, which is tacitly used in the proof of Lemma 4.2.

- The parameters $\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2$ are bounded away from 0 and 1:

  $$\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2 \in (0, 1). \tag{C.2}$$

  In particular, the distance from the boundary should be a constant independent of $\frac{1}{|\mathcal{S}|}$ and $\gamma$.

- In state $\mathfrak{s}$ (i.e., the only state where the two actions have distinct effects), action 1 is strictly better (resp. worse) than action 2 if $M \in \mathcal{M}_1$ (resp. $M \in \mathcal{M}_2$), which means $w/(1 - \gamma) =$

$Q^\star_M(\mathfrak{s},1) > Q^\star_M(\mathfrak{s},2) = \gamma\alpha_1(1-\gamma)$ (resp. $w/(1-\gamma) = Q^\star_M(\mathfrak{s},1) < Q^\star_M(\mathfrak{s},2) = \gamma\alpha_2(1-\gamma)$). This means

$$\gamma\alpha_1 < w < \gamma\alpha_2. \tag{C.3}$$

The final lower bound depends on this separation quantitatively.

Any tuple simultaneously satisfying Eqs. (C.1) to (C.3) is sufficient for our proof (modulo numerical differences). Naturally, the numerical values for the function class $\mathcal{F}$ defined in (4.2) must be changed accordingly so that the class contains $Q^\star$ for both $\mathcal{M}_1$ and $\mathcal{M}_2$.

## C.2 Computation of Value Functions (Proposition 4.1)

In this section, we verify Proposition 4.1, which asserts that for all $\pi$, $Q^\pi_M = f_1$ for all $M \in \mathcal{M}_1$ and $Q^\pi_M = f_2$ for all $M \in \mathcal{M}_2$, where $f_1$ and $f_2$ are defined in (4.2). Note that the calculation we present here is based on the precise values for the parameters $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; w)$ given in Section 4.2.1, not the general scheme given in Appendix C.1.

*Proof of Proposition 4.1.* Suppose $M \in \mathcal{M}_1$. Let $I_M$ denote the planted subset associated with $M$. First, for any self-looping terminal state $s \in \{W, X, Y, Z\}$, since all actions in $\mathcal{A}$ have identical effects, we have

$$V^\pi_M(s) = Q^\pi_M(s,\mathfrak{a}) = \sum_{h=0}^{\infty} \gamma^h R_{\alpha_1,\beta_1,w}(s,\mathfrak{a}) = \frac{1}{1-\gamma} \cdot \begin{cases} \frac{3}{8}\gamma, & s = W \\ 1, & s = X \\ 0, & s = Y \\ \frac{1}{3}, & s = Z \end{cases}$$

for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where we utilize the fact that $R_{\alpha_1,\beta_1,w}(W,\mathfrak{a}) = w = \gamma(\alpha_1+\alpha_2)/2 = 3\gamma/8$ and $R_{\alpha_1,\beta_1,w}(Z,\mathfrak{a}) = \alpha_1/\beta_1 = 1/3$.

Next, for any intermediate state $s \in \mathcal{S}^1$, since all actions in $\mathcal{A}$ have identical effects, we have

$$\begin{aligned} V^\pi_M(s) = Q^\pi_M(s,\mathfrak{a}) &= R_{\alpha_1,\beta_1,w}(s,\mathfrak{a}) + \gamma \mathbb{E}_{s' \sim P_{\alpha_1,\beta_1,w,I_M}(s,\mathfrak{a})}[V^\pi_M(s')] \\ &= \begin{cases} 0 + \gamma[\alpha_1 V^\pi_M(X) + (1-\alpha_1)V^\pi_M(Y)], & s \in I_M \\ 0 + \gamma[\beta_1 V^\pi_M(Z) + (1-\beta_1)V^\pi_M(Y)], & s \in \mathcal{S}^1 \setminus I_M \end{cases} \\ &= \begin{cases} \frac{\gamma}{1-\gamma}(\frac{1}{4} \times 1 + \frac{3}{4} \times 0), & s \in I_M \\ \frac{\gamma}{1-\gamma}(\frac{3}{4} \times \frac{1}{3} + \frac{1}{4} \times 0), & s \in \mathcal{S}^1 \setminus I_M \end{cases} \\ &= \frac{\gamma}{1-\gamma}\frac{1}{4} \end{aligned}$$

for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.

Thus, for the initial state $\mathfrak{s}$, we have

$$Q^\pi_M(\mathfrak{s},1) = R_{\alpha_1,\beta_1,w}(\mathfrak{s},1) + \gamma \mathbb{E}_{s' \sim P_{\alpha_1,\beta_1,w,I_M}(\mathfrak{s},1)}[V^\pi_M(s')] = 0 + \gamma V^\pi_M(W) = \frac{\gamma^2}{1-\gamma}\frac{3}{8},$$

$$Q^\pi_M(\mathfrak{s},2) = R_{\alpha_1,\beta_1,w}(\mathfrak{s},2) + \gamma \mathbb{E}_{s' \sim P_{\alpha_1,\beta_1,w,I_M}(\mathfrak{s},2)}[V^\pi_M(s')] = 0 + \gamma \mathbb{E}_{s' \sim \mathrm{Unif}(I_M)} V^\pi_M(s') = \frac{\gamma^2}{1-\gamma}\frac{1}{4}$$

for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.

Therefore, $Q_M^\pi(s,a) = f_1(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.

Now suppose $M \in \mathcal{M}_2$. Let $I_M$ denote the planted subset associated with $M$. For any self-looping terminal state $s \in \{W, X, Y, Z\}$, since all actions in $\mathcal{A}$ have identical effects, we have

$$V_M^\pi(s) = Q_M^\pi(s, \mathfrak{a}) = \sum_{h=0}^{\infty} \gamma^h R_{\alpha_2,\beta_2,w}(s, \mathfrak{a}) = \frac{1}{1-\gamma} \cdot \begin{cases} \frac{3}{8}\gamma, & s = W \\ 1, & s = X \\ 0, & s = Y \\ 1, & s = Z \end{cases}$$

for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where we utilize the fact that $R_{\alpha_2,\beta_2,w}(W, \mathfrak{a}) = w = \gamma(\alpha_1 + \alpha_2)/2 = 3\gamma/8$ and $R_{\alpha_2,\beta_2,w}(Z, \mathfrak{a}) = \alpha_2/\beta_2 = 1$. For any intermediate state $s \in \mathcal{S}^1$, since all actions in $\mathcal{A}$ have identical effects, we have

$$\begin{aligned} V_M^\pi(s) = Q_M^\pi(s, \mathfrak{a}) &= R_{\alpha_2,\beta_2,w}(s, \mathfrak{a}) + \gamma \mathbb{E}_{s' \sim P_{\alpha_2,\beta_2,w,I_M}(s,\mathfrak{a})}[V_M^\pi(s')] \\ &= \begin{cases} 0 + \gamma[\alpha_2 V_M^\pi(X) + (1-\alpha_2) V_M^\pi(Y)], & s \in I_M \\ 0 + \gamma[\beta_2 V_M^\pi(Z) + (1-\beta_2) V_M^\pi(Y)], & s \in \mathcal{S}^1 \setminus I_M \end{cases} \\ &= \begin{cases} \frac{\gamma}{1-\gamma}(\frac{1}{2} \times 1 + \frac{1}{2} \times 0), & s \in I_M \\ \frac{\gamma}{1-\gamma}(\frac{1}{2} \times 1 + \frac{1}{2} \times 0), & s \in \mathcal{S}^1 \setminus I_M \end{cases} \\ &= \frac{\gamma}{1-\gamma}\frac{1}{2} \end{aligned}$$

for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Thus, for the initial state $\mathfrak{s}$, we have

$$Q_M^\pi(\mathfrak{s}, 1) = R_{\alpha_2,\beta_2,w}(\mathfrak{s}, 1) + \gamma \mathbb{E}_{s' \sim P_{\alpha_2,\beta_2,w,I_M}(\mathfrak{s},1)}[V_M^\pi(s')] = 0 + \gamma V_M^\pi(W) = \frac{\gamma^2}{1-\gamma}\frac{3}{8},$$

$$Q_M^\pi(\mathfrak{s}, 2) = R_{\alpha_2,\beta_2,w}(\mathfrak{s}, 2) + \gamma \mathbb{E}_{s' \sim P_{\alpha_2,\beta_2,w,I_M}(\mathfrak{s},2)}[V_M^\pi(s')] = 0 + \gamma \mathbb{E}_{s' \sim \mathrm{Unif}(I_M)} V_M^\pi(s') = \frac{\gamma^2}{1-\gamma}\frac{1}{2}$$

for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. It follows that $Q_M^\star(s,a) = f_2(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. □

## C.3 Proof of Lemma 4.1

We now prove Lemma 4.1. Before proceeding, let us note that this lemma is proven only for the precise values for the parameters $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; w)$ given in Section 4.2.1. One could establish a more general lemma using the generic parameters introduced in Appendix C.1, but this would require changing the numerical constants appearing in the statement.

We begin the proof by lower bounding the regret for any MDP in the family $\mathcal{M}$. For any

$i \in \{1,2\}$, any MDP $M \in \mathcal{M}_i$, and any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have

$$\begin{aligned} J_M(\pi_M^\star) - J_M(\pi) &= Q_M^\star(\mathfrak{s}, \pi_M^\star(\mathfrak{s})) - Q_M^\pi(\mathfrak{s}, \pi(\mathfrak{s})) \\ &= Q_M^\star(\mathfrak{s}, \pi_M^\star(\mathfrak{s})) - Q_M^\star(\mathfrak{s}, \pi(\mathfrak{s})) \\ &= Q_M^\star(\mathfrak{s}, i) - Q_M^\star(\mathfrak{s}, \pi(\mathfrak{s})) \\ &\geq \frac{\gamma^2}{8(1-\gamma)} \mathbb{P}(\pi(\mathfrak{s}) \neq i), \end{aligned} \tag{C.4}$$

where the second equality follows because $\mathfrak{s}$ is the only state where different actions have distinct effects, and the last inequality follows from (4.3).

Now, consider any fixed offline reinforcement learning algorithm which takes the offline dataset $D_n$ as an input and returns a stochastic policy $\widehat{\pi}_{D_n} : \mathcal{S} \to \Delta(\mathcal{A})$. For each $i \in \{1,2\}$, we apply (C.4) to all MDPs in $\mathcal{M}_i$ and average to obtain

$$\frac{1}{|\mathcal{M}_i|} \sum_{M \in \mathcal{M}_i} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \geq \frac{\gamma^2}{8(1-\gamma)} \frac{1}{|\mathcal{M}_i|} \sum_{M \in \mathcal{M}_i} \mathbb{P}_n^M(\widehat{\pi}_{D_n}(\mathfrak{s}) \neq i).$$

Applying the inequality above for $i = 1$ and $i = 2$ and combining the results, we have

$$\begin{aligned} &\max_{M \in \mathcal{M}} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \\ &\geq \frac{1}{2|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] + \frac{1}{2|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \\ &\geq \frac{\gamma^2}{16(1-\gamma)} \left\{ \frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M(\widehat{\pi}_{D_n}(\mathfrak{s}) \neq 1) + \frac{1}{|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} \mathbb{P}_n^M(\widehat{\pi}_{D_n}(\mathfrak{s}) \neq 2) \right\} \\ &\geq \frac{\gamma^2}{16(1-\gamma)} \left( 1 - D_{\mathrm{TV}} \left( \frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M, \frac{1}{|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} \mathbb{P}_n^M \right) \right), \end{aligned}$$

where the last inequality follows because $\mathbb{P}(E) + \mathbb{Q}(E^c) \geq 1 - D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ for any event $E$. □

## C.4 Proof of Lemma 4.2

This proof is organized as follows. In Appendix C.4.1, we introduce a reference measure and move from the total variation distance to the $\chi^2$-divergence. This allows us to reduce the task of upper bounding $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$ to the task of upper bounding two manageable density ratios (Eqs. (C.6) and (C.7) in the sequel). We develop several intermediate technical lemmas related to the density ratios in Appendix C.4.2, and in Appendix C.4.3 we put everything together to bound the density ratios, thus completing the proof of Lemma 4.2.

For the statement of Lemma 4.2 and the main subsections of this section (Appendices C.4.1 and C.4.3), we only consider the specific values for the parameters $(\theta_1, \alpha_1, \beta_1; \theta_2, \alpha_2, \beta_2; w)$ given in Section 4.2.1. However, in Appendix C.4.2, which contains intermediate technical lemmas, results are presented under a slightly more general setup, as explained at the beginning of the subsection.

### C.4.1 Introducing a Reference Measure and Moving to $\chi^2$-Divergence

Directly calculating the total variation distance $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$ is challenging, so we design an auxillary *reference measure* $\mathbb{P}_n^0$ which serves as an intermediate quantity to help with the upper bound. The reference measure $\mathbb{P}_n^0$ lies in the same measurable space as $\mathbb{P}_n^1$ and $\mathbb{P}_n^2$, and is defined as follows:

$$\mathbb{P}_n^0(\{(s_i, a_i, r_i, s_i')\}_{i=1}^n) := \prod_{i=1}^n \mu(s_i, a_i) \mathbb{1}_{\{r_i = R_0(s_i, a_i)\}} P_0(s_i' \mid s_i, a_i), \quad \forall\, \{(s_i, a_i, r_i, s_i')\}_{i=1}^n,$$

where

$$R_0(s, \mathfrak{a}) := \begin{cases} 0, & s \in \{\mathfrak{s}\} \cup \mathcal{S}^1, \\ w = 3\gamma/8, & s = W, \\ 1, & s = X, \\ 0, & s = Y, \\ 0, & s = Z, \end{cases}$$

and

$$\begin{aligned} & P_0(\cdot \mid \mathfrak{s}, 1) := W, \text{ w.p. } 1, \\ & P_0(\cdot \mid \mathfrak{s}, 2) := \mathrm{Unif}(\mathcal{S}^1), \\ \forall s \in \mathcal{S}^1: \quad & P_0(\cdot \mid s, \mathfrak{a}) := \begin{cases} X, & \text{w.p. } \theta_1 \alpha_1, \\ Y, & \text{w.p. } 1 - \theta_1 \alpha_1 - (1-\theta_1)\beta_1, \\ Z, & \text{w.p. } (1-\theta_1)\beta_1, \end{cases} \\ \forall s \in \{W, X, Y, Z\}: \quad & P_0(\cdot \mid s, \mathfrak{a}) := s, \text{ w.p. } 1. \end{aligned}$$

The reference measure $\mathbb{P}_n^0$ can be understood as the law of $D_n$ when the data collection distribution is $\mu$ and the underlying MDP is $M_0 := (\mathcal{S}, \mathcal{A}, P_0, R_0, \gamma, d_0)$. Note that although we define the transition operator $P_0$ above based on the tuple $(\theta_1, \alpha_1, \beta_1)$, substituting in $(\theta_2, \alpha_2, \beta_2)$ leads to the same operator — this is guaranteed by an important feature of our construction: the families $\mathcal{M}_1$ and $\mathcal{M}_2$ in our construction satisfy the constraint (C.1), so that $\theta_1 \alpha_1 = \theta_2 \alpha_2$ and $(1-\theta_1)\beta_1 = (1-\theta_2)\beta_2$.

In what follows, we provide more explanations on the design of the transition operator $P_0$ and the reward function $R_0$.

**Properties and Intuition of $P_0$** There are two ways to understand $P_0$. Operationally, $P_0$ is simply the pointwise *average* transition operator of the MDPs in $\mathcal{M}_1$ or $\mathcal{M}_2$, in the sense that

$$\forall s \in \mathcal{S}, a \in \mathcal{A}: \quad P_0(\cdot \mid s, a) = \frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} P_M(\cdot \mid s, a) = \frac{1}{|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} P_M(\cdot \mid s, a),$$

where $P_M$ is the transition operator associated with each MDP $M$. More conceptually, $P_0$ is the transition operator obtained by performing state aggregation using the value function class $\mathcal{F} = \{f_1, f_2\}$, where states with the same values for both $f_1$ and $f_2$ are viewed as identical and constrained to share dynamics (which is induced by averaging over the data collection distribution).

**Properties and Intuition of $R_0$** Outside of state $Z$, the reward function $R_0$ is the same as the reward function of any MDP in $\mathcal{M}$, i..e.,

$$\forall s \neq Z, a \in \mathcal{A}: \quad R_0(s,a) = R_M(s,a), \ \forall M \in \mathcal{M},$$

where $R_M$ is the transition operator associated with each MDP $M$. The value of $R_0(Z, \mathfrak{a})$ is immaterial, as the data collection distribution $\mu$ is not supported on $(Z, \mathfrak{a})$ (in other words, different values of $R_0(Z, \mathfrak{a})$ lead to essentially the same reference measure $\mathbb{P}_n^0$); we choose $R_0(Z, a) = 0$ for concreteness.

**Moving to $\chi^2$-Divergence** Equipped with the definition of the reference measure $\mathbb{P}_n^0$, we proceed to bound $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$. By the triangle inequality for the total variation distance, we have

$$\begin{aligned} D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) &\leq D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^0\big) + D_{\mathrm{TV}}\big(\mathbb{P}_n^2, \mathbb{P}_n^0\big) \\ &\leq \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0)} + \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0)}, \end{aligned} \tag{C.5}$$

where the last inequality follows from the fact that $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P} \,\|\, \mathbb{Q})}$ for any $\mathbb{P}, \mathbb{Q}$ (see Proposition 7.2 or Section 7.6 of Polyanskiy (2020)).

In what follows, we derive simplified expressions for $D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0)$ and $D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0)$. We first expand and simplify $D_{\chi^2}(\mathbb{P}_n^1, \mathbb{P}_n^0)$, then obtain a similar expression for $D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0)$.

For each MDP $M \in \mathcal{M}$, let $P_M$ and $R_M$ denote the associated transition and reward functions. Observe that our construction for $P_M$, $R_M$, and $\mu$ (see Section 4.2.1) ensures that for any $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times [0,1] \times \mathcal{S}$ with $\mu(s,a)\mathbb{1}_{\{r = R_0(s,a)\}} P_0(s' \mid s, a) = 0$, we have $\mu(s,a)\mathbb{1}_{\{r = R_M(s,a)\}} P_M(s' \mid s, a) = 0$. As a result, we have $\mathbb{P}_n^M \ll \mathbb{P}_n^0$ for any $M \in \mathcal{M}$, which implies that $\mathbb{P}_n^1, \mathbb{P}_n^2 \ll \mathbb{P}_n^0$. Hence, we can expand the $\chi^2$-divergence as

$$\begin{aligned} &D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0) \\ &= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[ \left( \frac{\frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)}{\mathbb{P}_n^0(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)} \right)^2 \right] - 1 \\ &= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[ \left( \frac{\frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \prod_{i=1}^n \mu(s_i,a_i)\mathbb{1}_{\{r_i = R_M(s_i,a_i)\}} P_M(s_i' \mid s_i, a_i)}{\prod_{i=1}^n \mu(s_i,a_i)\mathbb{1}_{\{r_i = R_0(s_i,a_i)\}} P_0(s_i' \mid s_i, a_i)} \right)^2 \right] - 1 \\ &= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[ \left( \frac{\frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \prod_{i=1}^n P_M(s_i' \mid s_i, a_i)}{\prod_{i=1}^n P_0(s_i' \mid s_i, a_i)} \right)^2 \right] - 1 \\ &= \frac{1}{|\mathcal{M}_1|^2} \sum_{M, M' \in \mathcal{M}_1} \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{P}_n^0} \left[ \frac{\prod_{i=1}^n P_M(s_i' \mid s_i, a_i) P_{M'}(s_i' \mid s_i, a_i)}{\prod_{i=1}^n P_0^2(s_i' \mid s_i, a_i)} \right] - 1 \\ &= \frac{1}{|\mathcal{M}_1|^2} \sum_{M, M' \in \mathcal{M}_1} \left( \mathbb{E}_{(s,a) \sim \mu, s' \sim P_0(\cdot \mid s,a)} \left[ \frac{P_M(s' \mid s, a) P_{M'}(s' \mid s, a)}{P_0^2(s' \mid s, a)} \right] \right)^n - 1, \end{aligned} \tag{C.6}$$

where the third equality follows because (i) $R_M(s,a) = R_0(s,a), \forall M \in \mathcal{M}, \forall a \in \mathcal{A}, \forall s \neq Z$, and (ii) state $Z$ is not covered by $\mu$. Indeed, since the reward function for every MDP in $\mathcal{M}$ is the same

as $R_0$ *for all* $(s,a)$ *covered by* $\mu$, the rewards $r_1,\dots,r_n$ in $D_n$ are completely uninformative in our construction—they have the same distribution regardless of the underlying MDP. This is why the final expression for $D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0)$ in (C.6) is completely independent of the reward distribution for both measures.

Using an identical calculation, we also have

$$D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0) = \frac{1}{|\mathcal{M}_2|^2} \sum_{M,M' \in \mathcal{M}_2} \left( \mathbb{E}_{(s,a)\sim\mu, s'\sim P_0(\cdot|s,a)} \left[ \frac{P_M(s' \mid s,a) P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)} \right] \right)^n - 1. \tag{C.7}$$

Equipped with these expressions for the $\chi^2$-divergence, the next step in the proof of Lemma 4.2 is to upper bound the right-hand side for Eqs. (C.6) and (C.7). This is done in Appendix C.4.3, but before proceeding we require several intermediate technical lemmas.

### C.4.2 Technical Lemmas for Density Ratios

In this subsection, we state a number of technical lemmas which can be used to bound the density ratio appearing inside the square in Eqs. (C.6) and (C.7) for generic MDPs $M_{\alpha,\beta,w,I}$ with $I \in \mathcal{I}_\theta$. The lemmas hold for any choice of $(\theta,\alpha,\beta)$, and are independent of the reward parameter $w$. For this general setup, we work with a variant of the reference operator $P_0$ defined based on the values $(\theta,\alpha,\beta)$ via

$$\begin{aligned}
&P_0(\cdot \mid \mathfrak{s}, 1) := W, \text{ w.p. } 1, \\
&P_0(\cdot \mid \mathfrak{s}, 2) := \mathrm{Unif}(\mathcal{S}^1), \\
\forall s \in \mathcal{S}^1: \quad &P_0(\cdot \mid s, \mathfrak{a}) := \begin{cases} X, & \text{w.p. } \theta\alpha, \\ Y, & \text{w.p. } 1-\theta\alpha-(1-\theta)\beta, \\ Z, & \text{w.p. } (1-\theta)\beta, \end{cases} \\
\forall s \in \{W,X,Y,Z\}: \quad &P_0(\cdot \mid s, \mathfrak{a}) := s, \text{ w.p. } 1.
\end{aligned}$$

In Appendix C.4.3, we instantiate the results from this subsection with $(\theta_i,\alpha_i,\beta_i)$ for $i \in \{1,2\}$. Recall that per the discussion in Appendix C.4.1, our specific parameter choices for the families $\mathcal{M}_1$ and $\mathcal{M}_2$ induce the same reference operator $P_0$.

**Lemma C.1.** *For all* $I, I' \in \mathcal{I}_\theta$, $(2\theta-1)_+ S_1 \le |I \cap I'| \le \theta S_1$.

*Proof of Lemma C.1.* Since $|I| = |I'| = \theta S_1$, we have $|I \cap I'| \le |I| = \theta S_1$ and

$$|I \cap I'| = |I| + |I'| - |I \cup I'| \ge |I| + |I'| - S_1 = (2\theta-1)S_1.$$

Since $|I \cap I'| \ge 0$ trivially, the result follows. □

The next lemma controls the density ratio for states in $\mathcal{S}^1$. To state the result compactly, we define

$$\phi_{\theta,\alpha,\beta} := \theta^2 \left( \frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} \right). \tag{C.8}$$

Since Lemma C.2 is stated for any given $\theta, \alpha, \beta$, we use $P_I$ to denote $P_{\alpha,\beta,I}$ to keep notation compact.

**Lemma C.2.** *For all $I, I' \in \mathcal{I}_\theta$, we have*

$$\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}^1),s'\sim P_0(\cdot|s,\mathfrak{a})}\left[\frac{P_I(s' \mid s,\mathfrak{a})P_{I'}(s' \mid s,\mathfrak{a})}{P_0^2(s' \mid s,\mathfrak{a})}\right] = 1 + \phi_{\theta,\alpha,\beta}\cdot\left(\frac{|I\cap I'|}{\theta^2 S_1} - 1\right).$$

*Proof of Lemma C.2.* For any $I, I' \in \mathcal{I}_\theta$, we observe that

$$\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}^1),s'\sim P_0(\cdot|s,\mathfrak{a})}\left[\frac{P_I(s' \mid s,\mathfrak{a})P_{I'}(s' \mid s,\mathfrak{a})}{P_0^2(s' \mid s,\mathfrak{a})}\right] = \mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}^1)}\left[\sum_{s'\in\{X,Y,Z\}}\frac{P_I(s' \mid s,\mathfrak{a})P_{I'}(s' \mid s,\mathfrak{a})}{P_0(s' \mid s,\mathfrak{a})}\right].$$

To proceed, we calculate the value of the ratio $\frac{P_I(s'|s,\mathfrak{a})P_{I'}(s'|s,\mathfrak{a})}{P_0(s'|s,\mathfrak{a})}$ for each possible choice for $s \in \mathcal{S}^1$ and $s' \in \{X, Y, Z\}$ in Table C.1 below.

| | $s' = X$ | $s' = Y$ | $s' = Z$ |
|---|---|---|---|
| $s \in I \cap I'$ | $\alpha/\theta$ | $(1-\alpha)^2/(\theta(1-\alpha)+(1-\theta)(1-\beta))$ | 0 |
| $s \in (I \cup I') \setminus (I \cap I')$ | 0 | $(1-\alpha)(1-\beta)/(\theta(1-\alpha)+(1-\theta)(1-\beta))$ | 0 |
| $s \notin (I \cup I')$ | 0 | $(1-\beta)^2/(\theta(1-\alpha)+(1-\theta)(1-\beta))$ | $\beta/(1-\theta)$ |

Table C.1: Value of $\frac{P_I(s'|s,2)P_{I'}(s'|s,2)}{P_0(s'|s,2)}$ for all possible pairs $(s, s')$.

Define $t := |I \cap I'|$. From Lemma C.1, we must have $t \in [(2\theta-1)_+S_1, \theta S_1]$. We also have $|I \cup I'| = |I| + |I'| - |I \cap I'| = 2\theta S_1 - t$. Hence, the event in the first row of Table C.1 occurs with probability $|I \cap I'|/S_1 = t/S_1$, the event in the second row occurs with probability $|(I \cup I') \setminus (I \cap I')|/S_1 = (2\theta S_1 - 2t)/S_1$ and the event in the third row occurs with probability $|S_1 \setminus (I \cup I')|/S_1 = ((1-2\theta)S_1 + t)/S_1$. Using these values, we obtain

$$\begin{aligned}
&\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}^1)}\left[\sum_{s'\in\{X,Y,Z\}}\frac{P_I(s' \mid s,\mathfrak{a})P_{I'}(s' \mid s,\mathfrak{a})}{P_0(s' \mid s,\mathfrak{a})}\right] \\
&= \frac{t}{S_1}\cdot\left(\frac{\alpha}{\theta} + \frac{(1-\alpha)^2}{\theta(1-\alpha)+(1-\theta)(1-\beta)}\right) + \left(2\theta - \frac{2t}{S_1}\right)\cdot\frac{(1-\alpha)(1-\beta)}{\theta(1-\alpha)+(1-\theta)(1-\beta)} \\
&\quad + \left(1 - 2\theta + \frac{t}{S_1}\right)\cdot\left(\frac{(1-\beta)^2}{\theta(1-\alpha)+(1-\theta)(1-\beta)} + \frac{\beta}{1-\theta}\right) \\
&= \frac{t}{S_1}\left(\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\alpha}{\theta} + \frac{\beta}{1-\theta}\right) + \frac{2\theta(\beta-\alpha)(1-\beta)+(1-\beta)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{(1-2\theta)\beta}{1-\theta} \\
&= \frac{t}{S_1}\left(\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\alpha}{\theta} + \frac{\beta}{1-\theta}\right) + \frac{2\theta(\beta-\alpha)(1-\beta)+(1-\beta)^2}{\theta(\beta-\alpha)+1-\beta} + \beta - \frac{\theta\beta}{1-\theta} \\
&= \left(\frac{t}{S_1} - \theta^2\right)\left(\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\alpha}{\theta} + \frac{\beta}{1-\theta}\right) \\
&\quad + \underbrace{\theta^2\cdot\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta}}_{\text{(i)}} + \underbrace{\theta^2\cdot\frac{\alpha}{\theta}}_{\text{(ii)}} + \underbrace{\theta^2\cdot\frac{\beta}{1-\theta}}_{\text{(iii)}} + \underbrace{\frac{2\theta(\beta-\alpha)(1-\beta)+(1-\beta)^2}{\theta(\beta-\alpha)+1-\beta}}_{\text{(i)}} + \underbrace{\beta}_{\text{(ii)}} - \underbrace{\frac{\theta\beta}{1-\theta}}_{\text{(iii)}}.
\end{aligned}$$

Grouping the terms in the second line together, we find that (i) $= \theta(\beta-\alpha)+1-\beta$, (ii) $= \theta\alpha+\beta$,

and (iii) $= -\theta\beta$, and by summing,

$$\text{(i)} + \text{(ii)} + \text{(iii)} = 1.$$

Hence, the above expression is equal to

$$\begin{aligned}&\left(\frac{t}{S_1} - \theta^2\right)\left(\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\alpha}{\theta} + \frac{\beta}{1-\theta}\right) + 1\\ &= \left(\frac{t}{S_1} - \theta^2\right)\left(\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)}\right) + 1.\end{aligned}$$

Recalling the definition of $\phi_{\theta,\alpha,\beta}$, this completes the proof. □

The next lemma bounds the magnitude of $\phi_{\theta,\alpha,\beta}$ in terms of the parameter $\theta$.

**Lemma C.3.** *For any $\alpha, \beta, \theta \in (0,1)$, we have*

$$\theta^2|\alpha-\beta| \le \phi_{\theta,\alpha,\beta} \le \frac{\theta}{1-\theta}\max\{\alpha,\beta\} \le \frac{\theta}{1-\theta}.$$

*Proof of Lemma C.3.* Recall that $\phi_{\theta,\alpha,\beta} = \theta^2\Big(\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)}\Big)$. We consider two cases.

**Case 1: $\alpha \le \beta$.** Assume $\alpha < \beta$, as the result is immediate if $\alpha = \beta$. We have

$$\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} \ge 0 + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} \ge \frac{\theta(\beta-\alpha)}{\theta(1-\theta)} = \frac{\beta-\alpha}{1-\theta} > |\alpha-\beta|$$

and

$$\begin{aligned}\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} &= \frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\beta-\alpha}{1-\theta} + \frac{\alpha}{\theta(1-\theta)}\\ &\le \frac{\beta-\alpha}{\theta} + \frac{\beta-\alpha}{1-\theta} + \frac{\alpha}{\theta(1-\theta)}\\ &= \frac{\beta}{\theta(1-\theta)},\end{aligned}$$

where the inequality above follows since $\theta(\beta-\alpha)+1-\beta > \theta(\beta-\alpha) > 0$.

**Case 2: $\alpha > \beta$.** We have

$$\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} \ge 0 + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} \ge \frac{\theta(\beta-\alpha)+\alpha-\beta}{\theta(1-\theta)} = \frac{\alpha-\beta}{\theta} > |\alpha-\beta|$$

and

$$\begin{aligned}\frac{(\beta-\alpha)^2}{\theta(\beta-\alpha)+1-\beta} + \frac{\theta(\beta-\alpha)+\alpha}{\theta(1-\theta)} &= \frac{(\alpha-\beta)^2}{1-\beta-\theta(\alpha-\beta)} + \frac{\beta-\alpha}{1-\theta} + \frac{\alpha}{\theta(1-\theta)}\\ &\le \frac{(\alpha-\beta)^2}{(\alpha-\beta)-\theta(\alpha-\beta)} + \frac{\beta-\alpha}{1-\theta} + \frac{\alpha}{\theta(1-\theta)}\\ &= \frac{\alpha}{\theta(1-\theta)},\end{aligned}$$

where the inequality uses that $1-\beta > \alpha-\beta > \theta(\alpha-\beta)$.

The lemma immediately follows. □

The final lemma in this subsection controls the density ratio for the initial state $\mathfrak{s}$ when action 2 is chosen. Again, we use $P_I$ to denote $P_{\alpha,\beta,I}$ to keep notation compact.

**Lemma C.4.** *For any $I, I' \in \mathcal{I}_\theta$, we have*

$$\mathbb{E}_{s' \sim P_0(\cdot \mid \mathfrak{s}, 2)}\left[\frac{P_I(s' \mid \mathfrak{s}, 2) P_{I'}(s' \mid \mathfrak{s}, 2)}{P_0^2(s' \mid \mathfrak{s}, 2)}\right] = \frac{|I \cap I'|}{\theta^2 S_1}.$$

*Proof of Lemma C.4.* Let $I, I' \in \mathcal{I}_\theta$ be given and observe that

$$\mathbb{E}_{s' \sim P_0(\cdot \mid \mathfrak{s}, 2)}\left[\frac{P_I(s' \mid \mathfrak{s}, 2) \times P_{I'}(s' \mid \mathfrak{s}, 2)}{P_0^2(s' \mid \mathfrak{s}, 2)}\right] = \mathbb{E}_{s' \sim \mathrm{Unif}(\mathcal{S}^1)}\left[\frac{\mathbb{1}_{\{s' \in I \cap I'\}}}{\theta^2}\right] = \frac{|I \cap I'|}{\theta^2 S_1}.$$

□

### C.4.3 Completing the Proof

To keep notation compact, define

$$g_{\theta,\alpha,\beta}(t; n) := \left(\left(\frac{t}{\theta^2 S_1} - 1\right)\frac{8\phi_{\theta,\alpha,\beta} + 1}{16} + 1\right)^n.$$

For all $M \in \mathcal{M}$, $P_M(\cdot \mid s, a)$ and $P_0(\cdot \mid s, a)$ differ only when $(s, a) = (\mathfrak{s}, 2)$ or $(s, a) \in \mathcal{S}^1 \times \mathcal{A}$, so—recalling the value of $\mu$—we have

$$\begin{aligned}
&\frac{1}{|\mathcal{M}_1|^2} \sum_{M, M' \in \mathcal{M}_1} \left(\mathbb{E}_{\substack{(s,a) \sim \mu, \\ s' \sim P_0(\cdot \mid s, a)}}\left[\frac{P_M(s' \mid s, a) P_{M'}(s' \mid s, a)}{P_0^2(s' \mid s, a)}\right]\right)^n \\
&= \frac{1}{|\mathcal{M}_1|^2} \sum_{M, M' \in \mathcal{M}_1} \left(\frac{1}{2}\mathbb{E}_{\substack{s \sim \mathrm{Unif}(\mathcal{S}^1), \\ s' \sim P_0(\cdot \mid s, \mathfrak{a})}}\left[\frac{P_M(s' \mid s, \mathfrak{a}) P_{M'}(s' \mid s, \mathfrak{a})}{P_0^2(s' \mid s, \mathfrak{a})}\right] + \frac{1}{16}\mathbb{E}_{s' \sim P_0(\cdot \mid \mathfrak{s}, 2)}\left[\frac{P_M(s' \mid \mathfrak{s}, 2) P_{M'}(s' \mid \mathfrak{s}, 2)}{P_0^2(s' \mid \mathfrak{s}, 2)}\right] + \frac{7}{16}\right)^n \\
&= \frac{1}{\binom{S_1}{\theta_1 S_1}^2} \sum_t \sum_{I, I' \in \mathcal{I}_{\theta_1} : |I \cap I'| = t} \left(\frac{1}{2}\left(\left(\frac{t}{\theta_1^2 S_1} - 1\right)\phi_{\theta_1,\alpha_1,\beta_1} + 1\right) + \frac{1}{16}\frac{t}{\theta_1^2 S_1} + \frac{7}{16}\right)^n,
\end{aligned}$$

where we have used the expressions for the density ratio from Lemmas C.2 and C.4. We further simplify to

$$\begin{aligned}
&= \frac{1}{\binom{S_1}{\theta_1 S_1}^2} \sum_t \sum_{I, I' \in \mathcal{I}_{\theta_1} : |I \cap I'| = t} \left(\left(\frac{t}{\theta_1^2 S_1} - 1\right)\frac{8\phi_{\theta_1,\alpha_1,\beta_1} + 1}{16} + 1\right)^n \\
&= \sum_{t=(2\theta_1 - 1)_+ S_1}^{\theta_1 S_1} \frac{\binom{\theta_1 S_1}{t}\binom{S_1 - \theta_1 S_1}{\theta_1 S_1 - t}}{\binom{S_1}{\theta S_1}} \left(\left(\frac{t}{\theta_1^2 S_1} - 1\right)\frac{8\phi_{\theta_1,\alpha_1,\beta_1} + 1}{16} + 1\right)^n \\
&= \sum_{t=(2\theta_1 - 1)_+ S_1}^{\theta_1 S_1} \frac{\binom{\theta_1 S_1}{t}\binom{S_1 - \theta_1 S_1}{\theta_1 S_1 - t}}{\binom{S_1}{\theta_1 S_1}} g_{\theta_1,\alpha_1,\beta_1}(t; n),
\end{aligned}$$

where the second equality uses Lemma C.1. Applying the same calculation for $\mathcal{M}_2$, we also have that

$$\frac{1}{|\mathcal{M}_2|^2}\sum_{M,M'\in\mathcal{M}_2}\left(\mathbb{E}_{(s,a)\sim\mu,s'\sim P_0(\cdot|s,a)}\left[\frac{P_M(s'\mid s,a)P_{M'}(s'\mid s,a)}{P_0^2(s'\mid s,a)}\right]\right)^n$$
$$=\sum_{t=(2\theta_2-1)_+S_1}^{\theta_2 S_1}\frac{\binom{\theta_2 S_1}{t}\binom{S_1-\theta_2 S_1}{\theta_2 S_1-t}}{\binom{S_1}{\theta_2 S_1}}g_{\theta_2,\alpha_2,\beta_2}(t;n).$$

Therefore, to upper bound the right-hand sides of Eqs. (C.6) and (C.7), we only need to upper bound the quantity

$$\sum_{t=(2\theta-1)_+S_1}^{\theta S_1}\frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1-t}}{\binom{S_1}{\theta S_1}}g_{\theta,\alpha,\beta}(t;n), \tag{C.9}$$

for both $(\theta,\alpha,\beta)=(\theta_1,\alpha_1,\beta_1)$ and $(\theta,\alpha,\beta)=(\theta_2,\alpha_2,\beta_2)$. To upper bound this quantity, we use the following two lemmas.

**Lemma C.5** (Monotonicity of $g_{\theta,\alpha,\beta}$)**.** *For any $\theta,\alpha,\beta\in(0,1)$ and any $n\in\mathbb{N}$, the function $t\mapsto g_{\theta,\alpha,\beta}(t;n)$ is non-decreasing for $t\in[(2\theta-1)_+S_1,\theta S_1]$.*

*Proof of Lemma C.5.* By Lemma C.2, we have $\left(\frac{t}{\theta^2 S_1}-1\right)\phi_{\theta,\alpha,\beta}+1\geq 0$ for all $t\in[(2\theta-1)_+S_1,\theta S_1]$, and hence

$$\left(\frac{t}{\theta^2 S_1}-1\right)\frac{8\phi_{\theta,\alpha,\beta}+1}{16}+1=\frac{1}{2}\left(\left(\frac{t}{\theta^2 S_1}-1\right)\phi_{\theta,\alpha,\beta}+1\right)+\frac{1}{16}\frac{t}{\theta^2 S_1}+\frac{7}{16}\geq 0$$

for all $t\in[(2\theta-1)_+S_1,\theta S_1]$. This ensures that we are in the domain where $x\mapsto x^n$ is non-decreasing. Next, by Lemma C.3, we know that $\phi_{\theta,\alpha,\beta}\geq 0$, so the coefficient on $t$ is non-negative. It follows that $g_{\theta,\alpha,\beta}(t;n)$ is non-decreasing in $t\in[(2\theta-1)_+S_1,\theta S_1]$. □

**Lemma C.6** (Hypergeometric tail bound)**.** *For any $\theta\in\{\theta_1,\theta_2\}$ and $\epsilon\in(0,\theta^2 S_1)$, we have*

$$\sum_{t\geq(\theta+\epsilon)\cdot\theta S_1}\frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1-t}}{\binom{S_1}{\theta S_1}}\leq\exp(-2\epsilon^2\theta S_1). \tag{C.10}$$

*Proof of Lemma C.6.* Let $\mathrm{Hyper}(t;K,N,N'):=\binom{K}{t}\binom{N-K}{N'-t}/\binom{N}{N'}$ denote the hypergeometric probability mass function, which corresponds to the probability that exactly $t$ balls are blue when $N'$ balls are sampled without replacement from a jar containing $N$ total balls, $K$ of which are blue (see, e.g., Chapter 2.1.4 of Rice (2006) for background). We observe that the term $\frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1-t}}{\binom{S_1}{\theta S_1}}$ arising in Eqs. (C.9) and (C.10) is precisely $\mathrm{Hyper}(t;\theta S_1,S_1,\theta S_1)$, which corresponds to the process in which we sample $\theta S_1$ balls without replacement from a jar with $S_1$ balls, $\theta S_1$ of which are blue.

We now apply a classical tail bound for hypergeometric random variables.

**Lemma C.7** (Hoeffding (1963))**.** *Let $X\sim\mathrm{Hyper}(K,N,N')$ and define $p=K/N$. Then for any $0<\epsilon<pN'$, we have*

$$\Pr[X\geq(p+\epsilon)N']\leq\exp\bigl(-2\epsilon^2 N'\bigr).$$

Instantiating this bound with $\mathrm{Hyper}(\theta S_1, S_1, \theta S_1)$ (since $\theta S_1$ is an integer), we have $p = \theta$ and

$$\sum_{t \geq (\theta+\epsilon)\cdot\theta S_1} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} = \Pr[X \geq (\theta+\epsilon)\cdot\theta S_1] \leq \exp(-2\epsilon^2\theta S_1).$$

□

Returning to the quantity in (C.9), for any $(\theta, \alpha, \beta) \in \{(\theta_1, \alpha_1, \beta), (\theta_2, \alpha_2, \beta_2)\}$ and any $\epsilon \in (0, \theta^2 S_1)$ we can split the sum and upper bound as follows:

$$\begin{aligned}
&\sum_{t=(2\theta-1)_+ S_1}^{\theta S_1} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} g_{\theta,\alpha,\beta}(t; n) \\
&\leq \sum_{t=0}^{\lfloor(\theta+\epsilon)\theta S_1\rfloor} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}} g_{\theta,\alpha,\beta}(t; n) + \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n) \\
&\leq \left(\sum_{t=0}^{\lfloor(\theta+\epsilon)\theta S_1\rfloor} \frac{\binom{\theta S_1}{t}\binom{S_1-\theta S_1}{\theta S_1 - t}}{\binom{S_1}{\theta S_1}}\right) g_{\theta,\alpha,\beta}((\theta+\epsilon)\theta S_1; n) + \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n) \\
&\leq g_{\theta,\alpha,\beta}((\theta+\epsilon)\theta S_1; n) + \exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n), \qquad \text{(C.11)}
\end{aligned}$$

where the first two inequalities follow from Lemmas C.5 and C.6 and the last uses that the sum in the penultimate line is at most 1. We further calculate

$$\begin{aligned}
g_{\theta,\alpha,\beta}((\theta+\epsilon)\theta S_1; n) &= \left(\left(\frac{(\theta+\epsilon)\theta S_1}{\theta^2 S_1} - 1\right)\frac{8\phi_{\theta,\alpha,\beta}+1}{16} + 1\right)^n \\
&= \left(\frac{\epsilon}{\theta}\frac{8\phi_{\theta,\alpha,\beta}+1}{16} + 1\right)^n \\
&\leq \left(\frac{\epsilon}{2(1-\theta)\theta} + 1\right)^n, \qquad \text{(C.12)}
\end{aligned}$$

where the inequality follows from Lemma C.3. Similarly, we have

$$\begin{aligned}
\exp(-2\epsilon^2\theta S_1)\cdot g_{\theta,\alpha,\beta}(\theta S_1; n) &= \exp(-2\epsilon^2\theta S_1)\cdot\left(\left(\frac{\theta S_1}{\theta^2 S_1} - 1\right)\frac{8\phi_{\theta,\alpha,\beta}+1}{16} + 1\right)^n \\
&\leq \exp\bigl(-2\epsilon^2\theta S_1\bigr)\cdot\left(\left(\frac{1}{\theta} - 1\right)\frac{8\theta/(1-\theta)+1}{16} + 1\right)^n \\
&\leq \exp\bigl(-2\epsilon^2\theta S_1\bigr)\cdot\left(1 + \frac{1}{2\theta}\right)^n \\
&= \exp\bigl(n\ln(1+1/(2\theta)) - 2\epsilon^2\theta S_1\bigr) \\
&\leq \exp\bigl(n/(2\theta) - 2\epsilon^2\theta S_1\bigr), \qquad \text{(C.13)}
\end{aligned}$$

where the first inequality follows from Lemma C.3 and the last inequality uses that $\log(1+x) \leq x$.

Combining Eqs. (C.6), (C.7), (C.9) and (C.11) to (C.13) and instantiating the bounds for

$(\theta_1, \alpha_1, \beta_1)$ and $(\theta_2, \alpha_2, \beta_2)$, we have

$$D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{P}_n^0) \le \inf_{\epsilon \in (0, \theta_1^2 S_1)} \left\{ \left( \frac{\epsilon}{2(1-\theta_1)\theta_1} + 1 \right)^n + \exp\bigl(n/(2\theta_1) - 2\epsilon^2 \theta_1 S_1\bigr) \right\} - 1.$$

$$D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{P}_n^0) \le \inf_{\epsilon \in (0, \theta_2^2 S_1)} \left\{ \left( \frac{\epsilon}{2(1-\theta_2)\theta_2} + 1 \right)^n + \exp\bigl(n/(2\theta_2) - 2\epsilon^2 \theta_2 S_1\bigr) \right\} - 1.$$

Let $c \in (0, 1/2)$ be an arbitrary constant. For each $i \in \{1, 2\}$, we set $\epsilon = 2c \cdot \frac{(1-\theta_i)\theta_i}{n}$ (which belongs to $(0, \theta_i^2 S_1)$ because $\epsilon < \theta_i$ since $n \ge 1$ and $\theta_i S_1 \ge 1$ by assumption). Then we have

$$\left( \frac{\epsilon}{2(1-\theta_i)\theta_i} + 1 \right)^n \le \left(1 + \frac{c}{n}\right)^n \le e^c \le 1 + 2c, \quad \forall i \in \{1, 2\},$$

and

$$D_{\chi^2}(\mathbb{P}_n^i \,\|\, \mathbb{P}_n^0) \le 2c + \exp\left( \frac{n}{2\theta_i} - 8c^2 \theta_i \frac{(1-\theta_i)^2 \theta_i^2}{n^2} S_1 \right), \quad \forall i \in \{1, 2\}.$$

In particular, whenever $S_1 \ge \max_{i \in \{1,2\}} \frac{n^3}{8c^2 \theta_i^4 (1-\theta_i)^2}$, we have

$$D_{\chi^2}(\mathbb{P}_n^i \,\|\, \mathbb{P}_n^0) \le 2c + \exp(-n/(2\theta_i)), \quad \forall i \in \{1, 2\}.$$

Plugging in the values $\theta_1 = 1/2$, $\theta_2 = 1/4$ and setting $c = 1/10$, we have that whenever $n \ge 5$ and $S_1 > 6400 n^3$,

$$D_{\chi^2}(\mathbb{P}_n^i \,\|\, \mathbb{P}_n^0) \le \frac{1}{5} + \exp(-n) \le \frac{1}{4}, \quad \forall i \in \{1, 2\}.$$

Combining this with (C.5), we have that $D_{\mathrm{TV}}\bigl(\mathbb{P}_n^1, \mathbb{P}_n^2\bigr) \le \sqrt{1/4} = 1/2$, which proves the lemma. □

## C.5 Theorem 4.2: Lower Bound Construction and Proof

We restate Theorem 4.2 below for convenience.

**Theorem** (Lower bound for admissible data)**.** *For any $S \ge 9$, $\gamma \in (1/2, 1)$, and $C \ge 64$, there exists a family of MDPs $\mathcal{M}$ with $|\mathcal{S}| = S$ and $|\mathcal{A}| = 2$, a value function class $\mathcal{F}$ with $|\mathcal{F}| = 2$, and a data distribution $\mu$ which is a mixture of admissible distributions, such that:*

1. *We have $Q^\pi \in \mathcal{F}$ for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (all-policy realizability) and $C_{\mathrm{conc}} \le C$ (concentrability) for all models in $\mathcal{M}$.*
2. *Any algorithm using less than $c \cdot \min\bigl\{S^{1/3}/(\log S)^2, 2^{C/32}, 2^{1/(1-\gamma)}\bigr\}$ samples must have $J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \ge c'$ for some instance in $\mathcal{M}$, where $c$ and $c'$ are absolute numerical constants.*

### C.5.1 Lower Bound Construction

We begin by specifying the structure of the MDPs in the family $\mathcal{M}$ used to prove Theorem 4.2. Let $\gamma \in (0, 1)$ be fixed, and let $S \in \mathbb{N}$ be given. Let $L \in \mathbb{N}$ be an integer parameter whose value will be chosen at the end of the proof (Appendix C.5.4). Define $L_{\mathrm{div}} := \sum_{l=1}^{L} (2L + 1 - l)(L + 2 - l) \le 4L^3$,

and assume without loss of generality that $S > 5$ and that $(S-5)/L_{\text{div}}$ is an integer.[60] We consider a parameterized class of MDPs illustrated in Figure 4-2. Each MDP takes the form $M_{L,\alpha,w,\boldsymbol{I}} = \{\mathcal{S}, \mathcal{A}, P_{L,\alpha,\boldsymbol{I}}, R_{L,\alpha,w}, \gamma, d_0\}$, and is parametrized by the integer $L \in \mathbb{N}$, a vector of subsets $\boldsymbol{I} = (I^1, \ldots, I^L)$ where $I^l \subseteq \mathcal{S}$, and scalars $\alpha \in (0, 1/L)$ and $w \in [0,1]$. All MDPs in the family $\{M_{L,\alpha,w,\boldsymbol{I}}\}$ share the same state space $\mathcal{S}$, action space $\mathcal{A}$, discount factor $\gamma$, and initial state distribution $d_0$, and differ only in terms of the transition function $P_{L,\alpha,\boldsymbol{I}}$ and the reward function $R_{L,\alpha,w}$.

**State space** We consider a layered[61] state space $\mathcal{S} = \{\mathfrak{s}\} \cup \mathcal{S}^1 \cup \cdots \cup \mathcal{S}^L \cup \{W, X, Y, Z\}$, where $\mathfrak{s}$ is the initial state, $\mathcal{S}^1, \ldots, \mathcal{S}^L$ are $L$ layers of *intermediate* (i.e., neither initial nor terminal) states, and $\{W, X, Y, Z\}$ are self-looping terminal states. The number of intermediate states in layer $l \in [L]$ is $S_l := \frac{S-5}{L_{\text{div}}}(2L+1-l)(L+2-l)$, which ensures that $|\mathcal{S}| = \sum_{l=1}^{L} S_l + 5 = S$.[62]

**Action space** Our action space is given by $\mathcal{A} = \{1, 2\}$. For the initial state $\mathfrak{s}$, the two actions have distinct effects, while for all other states in $\mathcal{S} \setminus \{\mathfrak{s}\}$ both actions have identical effects. As a result, the value of a given policy only depends on the action it selects in $\mathfrak{s}$. As in the proof of Theorem 4.1, we use the symbol $\mathfrak{a}$ as a placeholder to denote either action when taken in $s \in \mathcal{S} \setminus \{\mathfrak{s}\}$, since the choice is immaterial.

**Transition operator** For each MDP $M_{L,\alpha,w,\boldsymbol{I}}$, recalling $\boldsymbol{I} = (I^1, \ldots, I^L)$, we let $I^l \subseteq \mathcal{S}^l$ parameterize a subset of the $l^{\text{th}}$-layer intermediate states. We call each $s \in I^l$ an $l^{\text{th}}$*-layer planted state* and call $s \in \bar{I}^l := \mathcal{S}^l \setminus I^l$ an $l^{\text{th}}$*-layer unplanted state*. The dynamics $P_{L,\alpha,\boldsymbol{I}}$ for $M_{L,\alpha,w,\boldsymbol{I}}$ are determined by $L$, $\alpha \in (0, 1/L)$, and $\boldsymbol{I}$ as follows (cf. Figure 4-2):

- *Initial state* $\mathfrak{s}$. For the dynamics from the initial state $\mathfrak{s}$, we define

$$P_{L,\alpha,\boldsymbol{I}}(\mathfrak{s}, 1) = \text{Unif}(\{W\}),$$

and

$$P_{L,\alpha,\boldsymbol{I}}(\mathfrak{s}, 2) = \frac{1}{2} \cdot \left( \sum_{l=1}^{L} \left( \frac{1}{2^l} \text{Unif}(\mathcal{S}^l) \right) + \frac{1}{2^L} \text{Unif}(\{Z\}) \right) + \frac{1}{2} \cdot \text{Unif}(\{X, Y\}).$$

That is, from the initial state $\mathfrak{s}$, choosing action 1 always leads to state $W$ in the next time step (see the red arrow in Figure 4-2), while choosing 2 leads to all states in $\mathcal{S}^1 \cup \cdots \cup \mathcal{S}^L \cup \{X, Y, Z\}$ (i.e., $\mathcal{S} \setminus \{\mathfrak{s}, W\}$) with certain probability (see the blue arrow in Figure 4-2, but note that transitions from $\mathfrak{s}$ to $\{X, Y\}$ are not displayed).

[60] If $(S-5)/L_{\text{div}}$ is not an integer, then we can simply construct the MDPs using $\underline{S} := \lfloor (S-5)/L_{\text{div}} \rfloor L_{\text{div}} + 5$ states and then add $S - \underline{S}$ arbitrary states that are not reachable by any policy. Since we are considering the case where $\mu$ is admissible, those non-reachable states do not affect the sample complexity of any algorithm (as they do not affect $D_n$ at all). It is easy to show that the conclusion of Theorem 4.2 still holds.

[61] Importantly, one should distinguish the concept of "layer" (which we use to simply refer to a group of states) and the concept of "time step" (which indexes the sequential evolution of the MDP). A state in layer $l \in [L]$ may be reached in any time step. For example, in Figure 4-2, states in $I^3$ (which belongs to layer 3) can be reached in both time step 1 (through the blue arrow) and time step 2 (from $\bar{I}^2$), but cannot be reached in time step 3.

[62] The precise value of $S_l$ given here is not essential to our proof. Its primarily serves to avoid a rounding issue that arises in Appendix C.5.2, which can also be addressed through other methods.

- *Intermediate states.* Transitions from states in $\mathcal{S}^1, \ldots, \mathcal{S}^L$ are defined as follows.
  - For each $l^{\text{th}}$-layer planted state $s \in I^l \subseteq \mathcal{S}^l$, define
$$P_{L,\alpha,\boldsymbol{I}}(s, \mathfrak{a}) = \frac{\gamma^{L-l}\alpha}{1-(l-1)\alpha}\text{Unif}(\{X\}) + \left(1 - \frac{\gamma^{L-l}\alpha}{1-(l-1)\alpha}\right)\text{Unif}(\{Y\}).$$
  - For each $l^{\text{th}}$-layer unplanted states $s \in \bar{I}^l \subseteq \mathcal{S}^l$, define
$$P_{L,\alpha,\boldsymbol{I}}(s, \mathfrak{a}) = \frac{1-l\cdot\alpha}{1-(l-1)\alpha}\text{Unif}(I^{l+1}) + \frac{\alpha}{1-(l-1)\alpha}\text{Unif}(\{Y\}),$$
    with the convention that $I^{L+1} := \{Z\}$.

  Since we restrict to $\alpha \le 1/L$, one can verify that these are valid probability distributions.
- *Terminal states.* All states in $\{W, X, Y, Z\}$ self-loop indefinitely. That is $P_{L,\alpha,\boldsymbol{I}}(s, \mathfrak{a}) = \text{Unif}(\{s\})$ for all $s \in \{W, X, Y, Z\}$.

**Reward function** The initial and intermediate states have no reward, i.e., $R_{L,\alpha,w}(s, a) = 0, \forall s \in \{\mathfrak{s}\} \cup \mathcal{S}^1 \cdots \cup \mathcal{S}^L, \forall a \in \mathcal{A}$. Each of the self-looping terminal states in $\{W, X, Y, Z\}$ has a fixed reward determined by the parameters $L$, $\alpha$ and $w$. In particular, we define $R_{L,\alpha,w}(W, \mathfrak{a}) = w$, $R_{L,\alpha,w}(X, \mathfrak{a}) = 1$, $R_{L,\alpha,w}(Y, \mathfrak{a}) = 0$, and $R_{L,\alpha,w}(Z, \mathfrak{a}) = \alpha/(1-L\alpha)$.

**Initial state distribution** All MDPs in $\{M_{L,\alpha,w,\boldsymbol{I}}\}$ start at $\mathfrak{s}$ deterministically (that is, the initial state distribution $d_0$ places all its probability mass on $\mathfrak{s}$). Since $d_0$ does not vary between instances, it should be thought of as *known* to the learning algorithm.

### C.5.2 Specifying the MDP Family $\mathcal{M}$

We leave $L \in \mathbb{N}$ (we interpret $\mathbb{N}$ to not include 0) as a free parameter until the end of Appendix C.5, where we will give a concrete $L$ that leads to Theorem 4.2. Given $L \in \mathbb{N}$, let $\alpha_1 := \frac{1}{2L}$ and $\alpha_2 := \frac{1}{L+1}$. For $\alpha \in (0, 1)$, define

$$V_\alpha := \sum_{l=1}^{L} \frac{1}{2^{l+1}} \frac{\gamma^{L-(l-1)}\alpha}{1-(l-1)\alpha} + \frac{1}{2^{L+1}} \frac{\alpha}{1-L\alpha} + \frac{1}{2}, \tag{C.14}$$

which has $0 < V_{\alpha_1} < V_{\alpha_2} < 1$, and let $w := \frac{V_{\alpha_1}+V_{\alpha_2}}{2}$. Define $\mathcal{I}_{\boldsymbol{\theta}} := \{\boldsymbol{I} : |I^l| = \theta_l S_l\}$ for any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L) \in (0, 1)^L$ such that $\theta_l S_l$ is an integer for all $l \in [L]$. We define two sub-families of MDPs via

$$\mathcal{M}_1 := \bigcup_{\boldsymbol{I} \in \mathcal{I}_{\boldsymbol{\theta}^{(1)}}} \{M_{L,\alpha_1,w,\boldsymbol{I}}\}, \quad \text{and} \quad \mathcal{M}_2 := \bigcup_{\boldsymbol{I} \in \mathcal{I}_{\boldsymbol{\theta}^{(2)}}} \{M_{L,\alpha_2,w,\boldsymbol{I}}\},$$

where $\mathcal{M}_1$ is specified by $\alpha_1$ and $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \ldots, \theta_L^{(1)})$ with

$$\theta_l^{(1)} := \frac{\alpha_2}{1-(l-1)\alpha_2}, \quad \forall l \in [L],$$

and $\mathcal{M}_2$ is specified by $\alpha_2$ and $\boldsymbol{\theta}^{(2)} = (\theta_1^{(2)}, \ldots, \theta_L^{(2)})$ with

$$\theta_l^{(2)} := \frac{\alpha_1}{1-(l-1)\alpha_1}, \quad \forall l \in [L].$$

Finally, we define the hard family $\mathcal{M}$ via

$$\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2.$$

Note for this construction, that $\boldsymbol{\theta}^{(1)}$ is defined in terms of $\alpha_2$ and vice-versa, which is a crucial to the proof. In addition, recall that we assume without loss of generality that $\frac{S-5}{L_{\text{div}}}$ is an integer, which implies that $\theta_l^{(i)} S_l = \frac{S-5}{L_{\text{div}}}(2L-(l-1))(L+1-(l-1))\theta_l^{(i)}$ is always an integer for any $i \in \{1,2\}$ and $l \in [L]$.

### C.5.3 Finishing the Construction: Value Functions and Data Distribution

**Value function class** Define functions $f_1, f_2 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows, recalling that $w := \frac{V_{\alpha_1}+V_{\alpha_2}}{2}$ (differences are highlighted in blue):

$$f_1(s,a) := \frac{1}{1-\gamma} \cdot \begin{cases} \gamma w, & s = \mathfrak{s},\ a = 1 \\ \gamma V_{\alpha_1}, & s = \mathfrak{s},\ a = 2 \\ \frac{\gamma^{L-(l-1)}\alpha_1}{1-(l-1)\alpha_1}, & s \in \mathcal{S}^l,\ l \in [L] \\ w, & s = W \\ 1, & s = X \\ 0, & s = Y \\ \frac{\alpha_1}{1-L\alpha_1}, & s = Z \end{cases}, \tag{C.15}$$

$$f_2(s,a) := \frac{1}{1-\gamma} \cdot \begin{cases} \gamma w, & s = \mathfrak{s},\ a = 1 \\ \gamma V_{\alpha_2}, & s = \mathfrak{s},\ a = 2 \\ \frac{\gamma^{L-(l-1)}\alpha_2}{1-(l-1)\alpha_2}, & s \in \mathcal{S}^l,\ l \in [L] \\ w, & s = W \\ 1, & s = X \\ 0, & s = Y \\ \frac{\alpha_2}{1-L\alpha_2}, & s = Z \end{cases}. \tag{C.16}$$

The following result is an elementary calculation. See Appendix C.8 for a detailed calculation.

**Proposition C.1.** *For all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have $Q_M^\pi = f_1$ for all $M \in \mathcal{M}_1$ and $Q_M^\pi = f_2$ for all $M \in \mathcal{M}_2$.*

It follows that by choosing $\mathcal{F} = \{f_1, f_2\}$, all-policy realizability holds for all $M \in \mathcal{M}$.

**Data distribution** Recall that in the offline RL setting, the learner is provided with an i.i.d. dataset $D_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ where $(s_i, a_i) \sim \mu$, $s_i' \sim P(\cdot \mid s_i, a_i)$, and $r_i = R(s_i, a_i)$. To ensure

admissibility of $\mu$, we consider the *exploratory policy* $\pi_0$ given by

$$\pi_0(s) = \mathrm{Unif}(\mathcal{A}),\ \forall s \in \mathcal{S}.$$

We define the data collection distribution $\mu$ via:

$$\mu(s,a) := \frac{1}{2}d_0^{\pi_0}(s,a) + \frac{1}{2}d_1^{\pi_0}(s,a),$$

which, by construction, is a mixture of admissible distributions as desired. As a reminder, we use the notation $d_h^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$ to denote the occupancy measure of $\pi$ at time step $h$, that is $d_h^\pi(s,a) := \mathbb{P}^\pi(s_h = s, a_h = a)$, where the dependence on the MDP $M$ is suppressed.

In general, this choice of $\mu$ will depend on the underlying MDP $M \in \mathcal{M}$ through $d_1^{\pi_0}$. However, for our specific construction, we calculate that

$$\begin{aligned}
\mu(\cdot,\mathfrak{a}) &= \frac{1}{2}d_0 + \frac{1}{2}\left(\frac{1}{2}P_{L,\alpha,\boldsymbol{I}}(\mathfrak{s},1) + \frac{1}{2}P_{L,\alpha,\boldsymbol{I}}(\mathfrak{s},2)\right)\\
&= \frac{1}{2}d_0 + \frac{1}{4}\mathrm{Unif}(\{W\}) + \frac{1}{4}\left(\frac{1}{2}\cdot\left(\sum_{l=1}^{L}\left(\frac{1}{2^l}\mathrm{Unif}(\mathcal{S}^l)\right) + \frac{1}{2^L}\mathrm{Unif}(\{Z\})\right) + \frac{1}{2}\cdot\mathrm{Unif}(\{X,Y\})\right)\\
&= \frac{1}{8}\left(\sum_{l=1}^{L}\left(\frac{1}{2^l}\mathrm{Unif}(\mathcal{S}^l)\right) + \frac{1}{2^L}\mathrm{Unif}(\{Z\})\right) + \frac{1}{2}\mathrm{Unif}(\{\mathfrak{s}\}) + \frac{1}{4}\mathrm{Unif}(\{W\}) + \frac{1}{8}\mathrm{Unif}(\{X,Y\}),
\end{aligned}$$

which is in fact independent of the choice of $M \in \mathcal{M}$.

In addition, by a straightforward calculation, we see that this choice of $\mu$ leads to the following bound on the concentrability coefficient. See Appendix C.8 for a detailed calculation.

**Proposition C.2.** *We have $C_{\mathrm{conc}} \le 32L$ for all models in $\mathcal{M}$.*

### C.5.4 Proof of Theorem 4.2

Recall that for each $M \in \mathcal{M}$, we let $\mathbb{P}_n^M$ denote the law of the offline dataset $D_n$ when the underlying MDP is $M$, and we let $\mathbb{E}_n^M$ be the associated expectation operator. Lemma C.8, stated below, reduces the task of proving a policy learning lower bound to the task of upper bounding the total variation distance between the mixture distributions $\mathbb{P}_n^1 := \frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1}\mathbb{P}_n^M$ and $\mathbb{P}_n^2 := \frac{1}{|\mathcal{M}_2|}\sum_{M\in\mathcal{M}_2}\mathbb{P}_n^M$.

**Lemma C.8.** *Consider any fixed $\gamma \in (0,1)$ and $L \in \mathbb{N}_{>1}$. For any offline RL algorithm which takes $D_n = \{(s_i,a_i,r_i,s_i')\}_{i=1}^n$ as input and returns a stochastic policy $\widehat{\pi}_{D_n} : \mathcal{S}\to\Delta(\mathcal{A})$, we have*

$$\sup_{M\in\mathcal{M}}\left\{J_M(\pi_M^\star) - \mathbb{E}_n^M[J_M(\widehat{\pi}_{D_n})]\right\} \ge \frac{\gamma^L}{16L}\frac{\gamma}{(1-\gamma)}\big(1 - D_{\mathrm{TV}}\big(\mathbb{P}_n^1,\mathbb{P}_n^2\big)\big).$$

See Appendix C.6 for the proof of Lemma C.8. We conclude the proof of Theorem 4.2 by bounding the total variation distance $D_{\mathrm{TV}}\big(\mathbb{P}_n^1,\mathbb{P}_n^2\big)$. Because directly calculating the total variation distance is difficult, we proceed in two steps. We first design two auxiliary reference measures $\mathbb{Q}_n^1$ and $\mathbb{Q}_n^2$, and then bound $D_{\mathrm{TV}}\big(\mathbb{P}_n^1,\mathbb{Q}_n^1\big)$, $D_{\mathrm{TV}}\big(\mathbb{P}_n^2,\mathbb{Q}_n^2\big)$ and $D_{\mathrm{TV}}\big(\mathbb{Q}_n^1,\mathbb{Q}_n^2\big)$ separately. For the latter step, as in the proof of Theorem 4.1, we move from total variation distance to $\chi^2$-divergence, which

we bound using similar arguments. Our final bound on the total variation distance, which is proven in Appendix C.7, is as follows.

**Lemma C.9.** *Consider any fixed $\gamma \in (0,1)$ and $L \in \mathbb{N}$. For all $n \leq \sqrt[3]{(S-5)}/(20L^2)$, we have*

$$D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) \leq 1/2 + n/(8 \cdot 2^L).$$

Theorem 4.2 immediately follows by choosing

$$L := \left\lfloor \min\left\{ \frac{C}{32}, \frac{1}{1-\gamma}, \log_2(S) \right\} \right\rfloor$$

and combining Lemma C.8 and Lemma C.9. With this choice of $L$, Lemma C.9 implies that $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) \leq 5/8$ whenever $n \leq c \cdot \min\big\{S^{1/3}/(\log S)^2, 2^{C/32}, 2^{1/(1-\gamma)}\big\}$ for a sufficiently small numerical constant $c$, and we have $C_{\mathrm{conc}} \leq C$ as desired by Proposition C.2. Finally, whenever $\gamma \geq 1/2$, using our choice for $L$ within Lemma C.8 gives

$$\sup_{M \in \mathcal{M}} \big\{ J_M(\pi_M^\star) - \mathbb{E}_n^M[J_M(\widehat{\pi}_{D_n})] \big\} \geq \Omega(1) \cdot \frac{\gamma^L}{L} \frac{\gamma}{(1-\gamma)} = \Omega(1)$$

where we use the fact that

$$\frac{\gamma^L}{L(1-\gamma)} \geq \frac{\gamma^{1/(1-\gamma)}}{(1/(1-\gamma))(1-\gamma)} = \gamma^{1/(1-\gamma)} \geq (1/2)^2$$

when $\gamma \in [1/2, 1)$.

## C.6 Proof of Lemma C.8

We begin the proof by lower bounding the regret for any MDP in the family $\mathcal{M}$. For any $i \in \{1,2\}$, any MDP $M \in \mathcal{M}_i$, and any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have

$$\begin{aligned}
J_M(\pi_M^\star) - J_M(\pi) &= Q_M^\star(\mathfrak{s}, \pi_M^\star(\mathfrak{s})) - Q_M^\pi(\mathfrak{s}, \pi(\mathfrak{s})) \\
&= Q_M^\star(\mathfrak{s}, \pi_M^\star(\mathfrak{s})) - Q_M^\star(\mathfrak{s}, \pi(\mathfrak{s})) \\
&= Q_M^\star(\mathfrak{s}, i) - Q_M^\star(\mathfrak{s}, \pi(\mathfrak{s})) \\
&= \frac{\gamma}{1-\gamma} \frac{|V_{\alpha_1} - V_{\alpha_2}|}{2} \mathbb{P}(\pi(\mathfrak{s}) \neq i) \\
&\geq \frac{\gamma^L}{24L} \frac{\gamma}{(1-\gamma)} \mathbb{P}(\pi(\mathfrak{s}) \neq i),
\end{aligned} \tag{C.17}$$

where the inequality follows because

$$\begin{aligned}
|V_{\alpha_1} - V_{\alpha_2}| &= \sum_{l=1}^{L} \frac{1}{2^{l-1}} \left( \frac{\gamma^{L-(l-1)}\alpha_1}{1-(l-1)\alpha_1} - \frac{\gamma^{L-(l-1)}\alpha_2}{1-(l-1)\alpha_2} \right) + \frac{1}{2^{L+1}} \left( \frac{\alpha_1}{1-L\alpha_1} - \frac{\alpha_2}{1-L\alpha_2} \right) \\
&\geq \frac{1}{2}\gamma^L |\alpha_1 - \alpha_2| = \frac{\gamma^L}{2L} \left( \frac{1}{L+1} - \frac{1}{2L} \right) \geq \frac{\gamma^L}{12L}
\end{aligned}$$

when $L \geq 2$.

Now, consider any fixed offline reinforcement learning algorithm which takes the offline dataset $D_n$ as an input and returns a stochastic policy $\widehat{\pi}_{D_n} : \mathcal{S} \to \Delta(\mathcal{A})$. For each $i \in \{1, 2\}$, we apply (C.17) to all MDPs in $\mathcal{M}_i$ and average to obtain

$$\frac{1}{|\mathcal{M}_i|} \sum_{M \in \mathcal{M}_i} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \geq \frac{\gamma^L}{24L} \frac{\gamma}{(1-\gamma)} \frac{1}{|\mathcal{M}_i|} \sum_{M \in \mathcal{M}_i} \mathbb{P}_n^M(\widehat{\pi}_{D_n}(\mathfrak{s}) \neq i).$$

Applying the inequality above for $i = 1$ and $i = 2$ and combining the results, we have

$$\begin{aligned}
&\max_{M \in \mathcal{M}} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \\
&\geq \frac{1}{2|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] + \frac{1}{2|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} \mathbb{E}_n^M [J_M(\pi_M^\star) - J_M(\widehat{\pi}_{D_n})] \\
&\geq \frac{\gamma^L}{48L} \frac{\gamma}{(1-\gamma)} \left\{ \frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M(\widehat{\pi}_{D_n}(\mathfrak{s}) \neq 1) + \frac{1}{|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} \mathbb{P}_n^M(\widehat{\pi}_{D_n}(\mathfrak{s}) \neq 2) \right\} \\
&\geq \frac{\gamma^L}{48L} \frac{\gamma}{(1-\gamma)} \left( 1 - D_{\mathrm{TV}} \left( \frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} \mathbb{P}_n^M, \frac{1}{|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} \mathbb{P}_n^M \right) \right),
\end{aligned}$$

where the last inequality follows because $\mathbb{P}(E) + \mathbb{Q}(E^c) \geq 1 - D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ for any event $E$. □

## C.7 Proof of Lemma C.9

This proof is organized as follows. In Appendix C.7.1, we introduce two reference measures and move from the total variation distance to the $\chi^2$-divergence. This allows us to reduce the task of upper bounding $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$ to the task of upper bounding two manageable density ratios (Eqs. (C.20) and (C.21) in the sequel). We develop several intermediate technical lemmas related to the density ratios in Appendix C.7.2, and in Appendix C.7.3 we put everything together to bound the density ratios, thus completing the proof of Lemma C.9.

### C.7.1 Introducing Reference Measures and Moving to $\chi^2$-Divergence

Directly calculating the total variation distance $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big)$ is challenging, so we design two auxillary *reference measures* $\mathbb{Q}_n^1$ and $\mathbb{Q}_n^2$ which serves as intermediate quantities to help with the upper bound. The reference measures $\mathbb{Q}_n^1, \mathbb{Q}_n^2$ lies in the same measurable space as $\mathbb{P}_n^1$ and $\mathbb{P}_n^2$, and are defined as follows:

$$\begin{aligned}
\mathbb{Q}_n^1(\{(s_i, a_i, r_i, s_i')\}_{i=1}^n) &:= \prod_{i=1}^n \mu(s_i, a_i) \mathbb{1}_{\{r_i = R_1(s_i, a_i)\}} P_0(s_i' \mid s_i, a_i), \quad \forall \{(s_i, a_i, r_i, s_i')\}_{i=1}^n, \\
\mathbb{Q}_n^2(\{(s_i, a_i, r_i, s_i')\}_{i=1}^n) &:= \prod_{i=1}^n \mu(s_i, a_i) \mathbb{1}_{\{r_i = R_2(s_i, a_i)\}} P_0(s_i' \mid s_i, a_i), \quad \forall \{(s_i, a_i, r_i, s_i')\}_{i=1}^n,
\end{aligned}$$

where

$$R_1(s,\mathfrak{a}) := \begin{cases} 0, & s \in \{\mathfrak{s}\} \cup \mathcal{S}^0 \cup \cdots \cup \mathcal{S}^L, \\ w, & s = W, \\ 1, & s = X, \\ 0, & s = Y, \\ \frac{\alpha_1}{1-L\alpha_1}, & s = Z, \end{cases}, \quad R_2(s,\mathfrak{a}) := \begin{cases} 0, & s \in \{\mathfrak{s}\} \cup \mathcal{S}^0 \cup \cdots \cup \mathcal{S}^L, \\ w, & s = W, \\ 1, & s = X, \\ 0, & s = Y, \\ \frac{\alpha_2}{1-L\alpha_2}, & s = Z, \end{cases}$$

and

$$\begin{aligned}
&P_0(\mathfrak{s},1) = \mathrm{Unif}(\{W\}), \\
&P_0(\mathfrak{s},2) = \frac{1}{2}\cdot\left(\sum_{l=1}^{L}\left(\frac{1}{2^l}\mathrm{Unif}(\mathcal{S}^l)\right) + \frac{1}{2^L}\mathrm{Unif}(\{Z\})\right) + \frac{1}{2}\cdot\mathrm{Unif}(\{X,Y\}), \\
&\forall s \in \mathcal{S}^l,\ \forall l \in [L] : P_0(s,\mathfrak{a}) = \frac{(1-l\alpha_1)(1-l\alpha_2)}{(1-(l-1)\alpha_1)(1-(l-1)\alpha_2)}\mathrm{Unif}(S^{l+1}) \\
&\qquad + \frac{\gamma^{L-l}\alpha_1\alpha_2}{(1-(l-1)\alpha_1)(1-(l-1)\alpha_2)}\mathrm{Unif}(\{X\}) \\
&\qquad + \left(1 - \frac{(1-l\alpha_1)(1-l\alpha_2)}{(1-(l-1)\alpha_1)(1-(l-1)\alpha_2)} - \frac{\gamma^{L-l}\alpha_1\alpha_2}{(1-(l-1)\alpha_1)(1-(l-1)\alpha_2)}\right)\mathrm{Unif}(\{Y\}), \\
&\forall s \in \{W,X,Y,Z\} : P_0(s,\mathfrak{a}) = \mathrm{Unif}(\{s\}).
\end{aligned}$$

The reference measure $\mathbb{Q}_n^1$ is the law of $D_n$ when the data collection distribution is $\mu$ and the underlying MDP is $\overline{M}_1 := (\mathcal{S}, \mathcal{A}, P_0, R_1, \gamma, d_0)$. Notably, $\overline{M}_1$ shares the same reward function with all MDPs in $\mathcal{M}_1$, and differs from the MDPs in $\mathcal{M}_1$ only in terms of the transition operator $P_0$.

There are two ways to understand $P_0$. Operationally, $P_0$ is simply the pointwise *average* transition operator of the MDPs in $\mathcal{M}_1$, in the sense that

$$\forall s \in \mathcal{S}, a \in \mathcal{A} : P_0(\cdot \mid s,a) = \frac{1}{|\mathcal{M}_1|}\sum_{M \in \mathcal{M}_1} P_M(\cdot \mid s,a),$$

where $P_M$ is the transition operator associated with each MDP $M$. For this reason, we call $\overline{M}_1$ the *average MDP* associated with $\mathcal{M}_1$. More conceptually, $P_0$ is the transition operator obtained by performing state aggregation using the value function class $\mathcal{F} = \{f_1, f_2\}$, where states with the same values for both $f_1$ and $f_2$ are viewed as identical and constrained to share dynamics (which is induced by averaging over the data collection distribution).

Similarly, the reference measure $\mathbb{Q}_n^2$ can be understood as the law of $D_n$ when the data collection distribution is $\mu$ and the underlying MDP is $\overline{M}_2 := (\mathcal{S}, \mathcal{A}, P_0, R_2, \gamma, d_0)$, where $\overline{M}_2$ is the average MDP associated with $\mathcal{M}_2$. An important property is that $\overline{M}_1$ and $\overline{M}_2$ share the same transition operator $P_0$ and differs only in terms of the reward on state $Z$. This is a consequence of our

construction, as when we construct $\mathcal{M}_1$ and $\mathcal{M}_2$ we strive to ensure that

$$\forall s \in \mathcal{S}, a \in \mathcal{A} : \frac{1}{|\mathcal{M}_1|} \sum_{M \in \mathcal{M}_1} P_M(\cdot \mid s, a) = P_0(\cdot \mid s, a) = \frac{1}{|\mathcal{M}_2|} \sum_{M \in \mathcal{M}_2} P_M(\cdot \mid s, a),$$

and there is no uncertainty in the reward function outside of state $Z$.

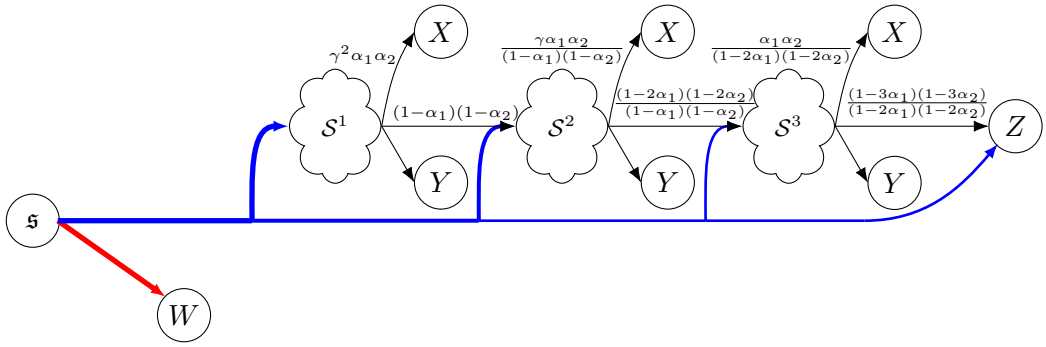


Figure C-1: Illustration of the average MDP with $L = 3$.

Figure C-1 illustrates the average MDPs $\overline{M}_1$ and $\overline{M}_2$ (the only difference between $\overline{M}_1$ and $\overline{M}_2$ is the reward on state $Z$, which is not displayed). Note that for each $l \in [L]$, all intermediate states in $\mathcal{S}^l$ have the same dynamics, so the planted subset structure is erased by averaging/aggregating.

Starting with the triangle inequality for the total variation distance, we have

$$\begin{aligned} D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) &\le D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{Q}_n^1\big) + D_{\mathrm{TV}}\big(\mathbb{P}_n^2, \mathbb{Q}_n^2\big) + D_{\mathrm{TV}}\big(\mathbb{Q}_n^1, \mathbb{Q}_n^2\big) \\ &\le \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1)} + \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{Q}_n^2)} + D_{\mathrm{TV}}\big(\mathbb{Q}_n^1, \mathbb{Q}_n^2\big), \end{aligned} \tag{C.18}$$

where the second inequality follows from the fact that $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \le \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P} \,\|\, \mathbb{Q})}$ for any $\mathbb{P}, \mathbb{Q}$ (see Proposition 7.2 or Section 7.6 of Polyanskiy (2020)).

The next lemma shows that the total variation distance between $\mathbb{Q}_n^1$ and $\mathbb{Q}_n^2$ is small. Intuitively, this is because the average MDPs $\overline{M}_1$ and $\overline{M}_2$ only differ in the reward on state $Z$, but the data distribution $\mu$'s coverage of on $Z$ is very small.

**Lemma C.10.** *For all $n < \infty$, we have $D_{\mathrm{TV}}\big(\mathbb{Q}_n^1, \mathbb{Q}_n^2\big) \le n\mu(Z, \mathfrak{a}) = n/(8 \times 2^L)$.*

*Proof of Lemma C.10.* Let $\mathcal{R} := \{1, 0, \alpha_1/(1 - L\alpha_1), \alpha_2/(1 - L\alpha_2), R(W, \mathfrak{a})\}$, then $R(s, a) \in \mathcal{R}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Since $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$, the realization of the offline dataset $D_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$

only has finitely many possible outcomes, and we have

$$
\begin{aligned}
&D_{\mathrm{TV}}\big(\mathbb{Q}_n^1, \mathbb{Q}_n^2\big) \\
&= \frac{1}{2} \sum_{(s_i,a_i,r_i,s_i') \in \mathcal{S}\times\mathcal{A}\times\mathcal{R}\times\mathcal{S},\, \forall i\in[n]} \left|\mathbb{Q}_n^1(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n) - \mathbb{Q}_n^2(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)\right| \\
&= \frac{1}{2} \sum_{(s_i,a_i,r_i,s_i') \in \mathcal{S}\times\mathcal{A}\times\mathcal{R}\times\mathcal{S},\, \forall i\in[n]} \prod_{i=1}^n \mu(s_i,a_i) P_0(s_i' \mid s_i,a_i) \left|\prod_{i=1}^n \mathbb{1}_{\{r_i = R_1(s_i,a_i)\}} - \prod_{i=1}^n \mathbb{1}_{\{r_i = R_2(s_i,a_i)\}}\right| \\
&= \frac{1}{2} \sum_{(s_i,a_i,r_i) \in \mathcal{S}\times\mathcal{A}\times\mathcal{R},\, \forall i\in[n]} \prod_{i=1}^n \mu(s_i,a_i) \left|\prod_{i=1}^n \mathbb{1}_{\{r_i = R_1(s_i,a_i)\}} - \prod_{i=1}^n \mathbb{1}_{\{r_i = R_2(s_i,a_i)\}}\right| \\
&= \sum_{(s_i,a_i) \in \mathcal{S}\times\mathcal{A},\, \forall i\in[n]} \mathbb{1}_{\{\exists i\in[n] \text{ s.t. } s_i = Z\}} \prod_{i=1}^n \mu(s_i,a_i) \\
&= \mathbb{P}_{s_1,\ldots,s_n\sim\mu}(\{\exists i \in [n] \text{ s.t. } s_i = Z\}) = \mathbb{P}_{s_1,\ldots,s_n\sim\mu}(\{s_1 = Z\} \cup \cdots \cup \{s_n = Z\}) \le n\mu(\{Z\}),
\end{aligned}
$$

where the first equality follows from the well-known identity between the total variation distance and the $L_1$ norm and the last inequality follows from a union bound. □

Using Lemma C.10, we have

$$
D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) \le \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1)} + \frac{1}{2}\sqrt{D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{Q}_n^2)} + n\mu(Z,\mathfrak{a}). \tag{C.19}
$$

Note that $\mu(Z,\mathfrak{a}) = 1/8 \cdot 1/2^L$ which produces the final term in the bound in Lemma C.9.

We now turn our focus to the $\chi^2$-divergence, which we expand as

$$
\begin{aligned}
&D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1) \\
&= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{Q}_n^1} \left[\left(\frac{\frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1} \mathbb{P}_n^M(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)}{\mathbb{Q}_n^1(\{(s_i,a_i,r_i,s_i')\}_{i=1}^n)}\right)^2\right] - 1 \\
&= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{Q}_n^1} \left[\left(\frac{\frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1} \prod_{i=1}^n \mu(s_i,a_i)\mathbb{1}_{\{r_i = R_M(s_i,a_i)\}} P_M(s_i' \mid s_i,a_i)}{\prod_{i=1}^n \mu(s_i,a_i)\mathbb{1}_{\{r_i = R_1(s_i,a_i)\}} P_0(s_i' \mid s_i,a_i)}\right)^2\right] - 1 \\
&= \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{Q}_n^1} \left[\left(\frac{\frac{1}{|\mathcal{M}_1|}\sum_{M\in\mathcal{M}_1} \prod_{i=1}^n P_M(s_i' \mid s_i,a_i)}{\prod_{i=1}^n P_0(s_i' \mid s_i,a_i)}\right)^2\right] - 1 \\
&= \frac{1}{|\mathcal{M}_1|^2} \sum_{M,M'\in\mathcal{M}_1} \mathbb{E}_{\{(s_i,a_i,r_i,s_i')\}_{i=1}^n \sim \mathbb{Q}_n^1} \left[\frac{\prod_{i=1}^n P_M(s_i' \mid s_i,a_i) P_{M'}(s_i' \mid s_i,a_i)}{\prod_{i=1}^n P_0^2(s_i' \mid s_i,a_i)}\right] - 1 \\
&= \frac{1}{|\mathcal{M}_1|^2} \sum_{M,M'\in\mathcal{M}_1} \left(\mathbb{E}_{\substack{(s,a)\sim\mu, \\ s'\sim P_0(\cdot\mid s,a)}} \left[\frac{P_M(s' \mid s,a) P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)}\right]\right)^n - 1,
\end{aligned} \tag{C.20}
$$

where the third equality follows from $R_M(s,a) = R_1(s,a), \forall M \in \mathcal{M}, \forall a \in \mathcal{A}, \forall s \in \mathcal{S}$.

Using an identical calculation, we also have

$$D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{Q}_n^2) = \frac{1}{|\mathcal{M}_2|^2} \sum_{M,M' \in \mathcal{M}_2} \left( \mathbb{E}_{\substack{(s,a)\sim\mu, \\ s'\sim P_0(\cdot|s,a)}} \left[ \frac{P_M(s' \mid s,a) P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)} \right] \right)^n - 1. \tag{C.21}$$

Equipped with these expressions for the $\chi^2$-divergence, the next step in the proof of Lemma C.9 is to upper bound the right-hand side for Eqs. (C.20) and (C.21). This is done in Appendix C.7.3, but before proceeding we require several intermediate technical lemmas.

### C.7.2 Technical Lemmas for Density Ratios

For this subsection only, we focus on MDPs in $\mathcal{M}_1$ and suppress the subscript indexing the subfamily, i.e., we use $\boldsymbol{\theta}$ for $\boldsymbol{\theta}^{(1)}$ and $\alpha$ for $\alpha_1$. Exactly the same calculations apply for $\mathcal{M}_2$, which we will use in the next section. To simplify the presentation and re-use lemmas from Appendix C.4.2, it will be helpful to define the following notation:

$$\begin{aligned} \boldsymbol{\alpha} &= (\alpha_1, \ldots, \alpha_L), \qquad & \alpha_l &= \frac{\gamma^{L-l}\alpha}{1-(l-1)\alpha} \\ \boldsymbol{\beta} &= (\beta_1, \ldots, \beta_L), \qquad & \beta_l &= 1 - \alpha_l \end{aligned}$$

Additionally recall that

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L), \qquad \theta_l = \frac{\alpha}{1-(l-1)\alpha}$$

These vectors parametrize the MDP transitions in the following sense: Let $\boldsymbol{I} \in \mathcal{I}_{\boldsymbol{\theta}}$ denote the choice of planted states for each layer. Then for $l \in [L]$ we have:

$$\begin{aligned} s \in I^l: &\;\; P_{L,\alpha,\boldsymbol{I}}(s, \mathfrak{a}) = \alpha_l \mathrm{Unif}(\{X\}) + \beta_l \mathrm{Unif}(\{Y\}) \\ s \in \bar{I}^l: &\;\; P_{L,\alpha,\boldsymbol{I}}(s, \mathfrak{a}) = (1-\theta_l)\mathrm{Unif}(I^{l+1}) + \theta_l \mathrm{Unif}(\{Y\}) \end{aligned}$$

where $I^{L+1} = \{Z\}$.

To state the results compactly, we define

$$\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} := \theta_l^2 \left( \frac{(\beta_l - \alpha_l)^2}{\theta_l(\beta_l - \alpha_l) + 1 - \beta_l} + \frac{\theta_l(\beta_l - \alpha_l) + \alpha_l}{\theta_l(1-\theta_l)} \right). \tag{C.22}$$

We also use $P_{\boldsymbol{I}}$ to denote $P_{L,\alpha,\boldsymbol{I}}$.

We will bound the density ratio terms for each layer separately. First we control the $L^{\text{th}}$ layer.

**Lemma C.11.** *For any $\boldsymbol{I}, \boldsymbol{J} \in \mathcal{I}_{\boldsymbol{\theta}}$, we have*

$$\mathbb{E}_{\substack{s\sim\mathrm{Unif}(\mathcal{S}^L), \\ s'\sim P_0(\cdot|s,\mathfrak{a})}} \left[ \frac{P_{\boldsymbol{I}}(s' \mid s, \mathfrak{a}) P_{\boldsymbol{J}}(s' \mid s, \mathfrak{a})}{P_0^2(s' \mid s, \mathfrak{a})} \right] = 1 + \phi^L_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} \cdot \left( \frac{|I^L \cap J^L|}{\theta_L^2 S_L} - 1 \right).$$

We omit the proof, which is identical to that of Lemma C.2. Next we turn to intermediate layers.

**Lemma C.12.** *For any $\boldsymbol{I}, \boldsymbol{J} \in \mathcal{I}_{\boldsymbol{\theta}}$, for any $l \in [L-1]$, we have*

$$\mathbb{E}_{\substack{s\sim\mathrm{Unif}(\mathcal{S}^l),\\ s'\sim P_0(\cdot|s,\mathfrak{a})}}\left[\frac{P_{\boldsymbol{I}}(s' \mid s,\mathfrak{a})P_{\boldsymbol{J}}(s' \mid s,\mathfrak{a})}{P_0^2(s' \mid s,\mathfrak{a})}\right] \le 1 + \phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} \cdot \left(\frac{|I^l \cap J^l|}{\theta_l^2 S_l} - 1\right) + \left(\frac{|I^{l+1} \cap J^{l+1}|}{\theta_{l+1}^2 S_{l+1}} - 1\right)_+.$$

*Proof of Lemma C.12.* For any $\boldsymbol{I}, \boldsymbol{J} \in \mathcal{I}_{\boldsymbol{\theta}}$, for any $l \in [L-1]$, we observe that

$$\mathbb{E}_{\substack{s\sim\mathrm{Unif}(\mathcal{S}^l),\\ s'\sim P_0(\cdot|s,\mathfrak{a})}}\left[\frac{P_{\boldsymbol{I}}(s' \mid s,\mathfrak{a})P_{\boldsymbol{J}}(s' \mid s,\mathfrak{a})}{P_0^2(s' \mid s,\mathfrak{a})}\right] = \mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}^l)}\left[\sum_{s'\in\{X,Y\}\cup(I^{l+1}\cap J^{l+1})} \frac{P_{\boldsymbol{I}}(s' \mid s,\mathfrak{a})P_{\boldsymbol{J}}(s' \mid s,\mathfrak{a})}{P_0(s' \mid s,\mathfrak{a})}\right].$$

To proceed, we calculate the value of the ratio $\frac{P_I(s'|s,\mathfrak{a})P_J(s'|s,\mathfrak{a})}{P_0(s'|s,\mathfrak{a})}$ for each possible choice for $s \in \mathcal{S}^1$ and $s' \in \{X, Y\} \cup (I^{l+1} \cap J^{l+1})$ in Table C.2 below.

| | $s' = X$ | $s' = Y$ | $s' \in I^{l+1} \cap J^{l+1}$ |
|---|---|---|---|
| $s \in I^l \cap J^l$ | $\alpha_l/\theta_l$ | $\beta_l^2/(\theta_l\beta_l + (1-\theta_l)\alpha_l)$ | 0 |
| $s \in (I^l \cup J^l) \setminus (I^l \cap J^l)$ | 0 | $\beta_l\alpha_l/(\theta_l\beta_l + (1-\theta_l)\alpha_l)$ | 0 |
| $s \notin (I^l \cup J^l)$ | 0 | $\alpha_l^2/(\theta_l\beta_l + (1-\theta_l)\alpha_l)$ | $\frac{\beta_l}{(1-\theta_l)} \cdot \frac{\lvert I^{l+1}\cap J^{l+1}\rvert}{\theta_{l+1}^2 S_{l+1}}$ |

Table C.2: Value of $\frac{P_{\boldsymbol{I}}(s'|s,\mathfrak{a})P_{\boldsymbol{I}'}(s'|s,\mathfrak{a})}{P_0(s'|s,\mathfrak{a})}$ for all possible pairs $(s, s')$.

Define $t_l := |I^l \cap J^l|$. From Lemma C.1, we must have $t_l \in [(2\theta_l - 1)_+ S_l, \theta_l S_l]$. We also have $|I^l \cup J^l| = |I^l| + |J^l| - |I^l \cap J^l| = 2\theta_l S_l - t_l$. Hence, the event in the first row of Table C.2 occurs with probability $|I^l \cap J^l|/S_l = t_l/S_l$, the event in the second row occurs with probability $|(I^l \cup J^l) \setminus (I^l \cap J^l)|/S_l = (2\theta_l S_l - 2t_l)/S_l$ and the event in the third row occurs with probability $|S_l \setminus (I^l \cup I^l)|/S = ((1-2\theta_l)S_l + t_l)/S_l$. Using these values and performing a similar calculation to the one in the proof of Lemma C.2, we obtain

$$\begin{aligned}
&\mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S}^l)}\left[\sum_{s'\in\{X,Y\}\cup(I^{l+1}\cap J^{l+1})} \frac{P_{\boldsymbol{I}}(s' \mid s,\mathfrak{a})P_{\boldsymbol{J}}(s' \mid s,\mathfrak{a})}{P_0(s' \mid s,\mathfrak{a})}\right] \\
&= 1 + \phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right) + \left(1 - 2\theta_l + \frac{t_l}{S_l}\right)\frac{\beta_l}{1-\theta_l}\left(\frac{t_{l+1}}{\theta_{l+1}^2 S_{l+1}} - 1\right) \\
&\le 1 + \phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right) + \left(\frac{t_{l+1}}{\theta_{l+1}^2 S_{l+1}} - 1\right)_+,
\end{aligned}$$

where the last inequality follows from $(2\theta_l - 1)S_l \le t_l \le \theta_l S_l$ (which implies $0 \le 1 - 2\theta_l + t_l/S_l \le 1 - \theta_l$) and $0 \le \beta_l \le 1$. □

### C.7.3 Completing the Proof

For now, let us also focus on a single MDP subfamily $\mathcal{M}_1$ and suppress the family indices associated with $\alpha$ and $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. As above the same calculations apply to $\mathcal{M}_2$. To keep notation compact, for

any $d \in \Delta(\mathcal{S} \times \mathcal{A})$, define

$$\mathsf{DR}_{M,M'}(d) := \mathbb{E}_{\substack{(s,a)\sim d,\\ s'\sim P_0(\cdot|s,a)}}\left[\frac{P_M(s' \mid s,a)P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)}\right].$$

Consider any $M, M' \in \mathcal{M}_1$. For any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, by Lemmas C.11 and C.12, we have

$$\begin{aligned}
&\sum_{l=1}^{L} \frac{1}{2^l}\mathsf{DR}_{M,M'}(\mathrm{Unif}(\mathcal{S}^l) \times \pi) \le \sum_{l=1}^{L} \frac{1}{2^l} + \sum_{l=1}^{L} \frac{1}{2^l}\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right) + \sum_{l=2}^{L} \frac{1}{2^{l-1}}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+ \\
&\le 1 - \frac{1}{2^L} + \sum_{l=1}^{L} \frac{1}{2^l}\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right) + \sum_{l=2}^{L} \frac{1}{2^{l-1}}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+ \\
&\le 1 - \frac{1}{2^L} + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} + 2}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+. \qquad \text{(C.23)}
\end{aligned}$$

Note that $P_M(\cdot \mid s,a)$ and $P_{M'}(\cdot \mid s,a)$ differ from $P_0(\cdot \mid s,a)$ only when $(s,a) = (\mathfrak{s},2)$ or $s \in (\{\mathfrak{s}\} \cup \mathcal{S}^1 \cup \cdots \cup \mathcal{S}^L)$, so, recalling the value of $\mu$, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mu, s'\sim P_0(\cdot|s,a)}\left[\frac{P_M(s' \mid s,a)P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)}\right] \\
&= \frac{1}{8}\sum_{l=1}^{L} \frac{1}{2^l}\mathsf{DR}_{M,M'}(\mathrm{Unif}(\mathcal{S}^l) \times \pi_0) + \frac{1}{8}\frac{1}{2^L}\mathsf{DR}_{M,M'}(\mathrm{Unif}(\{Z\}) \times \pi_0) \\
&\quad + \frac{1}{2}\mathsf{DR}_{M,M'}(\mathrm{Unif}(\{\mathfrak{s}\}) \times \pi_0) + \frac{1}{4}\mathsf{DR}_{M,M'}(\mathrm{Unif}(\{W\}) \times \pi_0) + \frac{1}{8}\mathsf{DR}_{M,M'}(\mathrm{Unif}(\{X,Y\}) \times \pi_0) \\
&\le \frac{1}{8}\left(1 - \frac{1}{2^L} + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} + 2}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+\right) + \frac{1}{8}\frac{1}{2^L} + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} \\
&= 1 + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+,
\end{aligned}$$

where the first inequality follows from (C.23). As a result, we have

$$\begin{aligned}
&\frac{1}{|\mathcal{M}_1|^2}\sum_{M,M'\in\mathcal{M}_1}\left(\mathbb{E}_{\substack{(s,a)\sim\mu,\\ s'\sim P_0(\cdot|s,a)}}\left[\frac{P_M(s' \mid s,a)P_{M'}(s' \mid s,a)}{P_0^2(s' \mid s,a)}\right]\right)^n \\
&\le \left(\prod_{l=1}^{L}\frac{1}{\binom{S_l}{\theta_l S_l}^2}\right)\sum_{t_1,\ldots,t_L}\sum_{\boldsymbol{I},\boldsymbol{I}'\in\mathcal{I}_{\boldsymbol{\theta}}:|I_l\cap I'_l|=t_l}\left(1 + \sum_{l=1}^{L}\frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 4}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+\right)^n \\
&= \mathbb{E}_{t_l\sim\mathrm{Hyper}(\theta_l S_l, S_l, \theta S_l),\,\forall l\in[L]}\left[\left(1 + \sum_{l=1}^{L}\frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+\right)^n\right], \qquad \text{(C.24)}
\end{aligned}$$

where $\mathrm{Hyper}(\cdot,\cdot,\cdot)$ denotes the hypergeometric distribution (cf. Lemma C.6 for background).

By Lemma C.6, for any $l \in [L]$, the event

$$E_l := \{t_l \ge (\theta_l + \epsilon_l)\theta_l S_l\}.$$

happens with probability at most $\exp\bigl(-2\epsilon_l^2\theta_l S_l\bigr)$. Hence, the event

$$E_{\mathrm{bad}} := \{\exists l \in [L], t_l \geq (\theta_l + \epsilon_l)\theta_l S_l\} = \bigcup_{l=1}^{L} E_l$$

happens with probability at most $\sum_{l=1}^{L} \exp\bigl(-2\epsilon_l^2\theta_l S_l\bigr)$. Conditional on $E_{\mathrm{clean}} := E_{\mathrm{bad}}^c$, i.e., the complement of $E_{\mathrm{bad}}$, we have

$$\begin{aligned}
\left(1 + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+\right)^n &\leq \left(1 + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{\epsilon_l}{\theta_l}\right)\right)^n \\
&\leq \left(1 + \sum_{l=1}^{L} \frac{1/(8(1-\theta_l)) + 1/4}{2^l}\left(\frac{\epsilon_l}{\theta_l}\right)\right)^n \\
&\leq \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l(1-\theta_l)}\epsilon_l\right)^n.
\end{aligned}$$

Here we are using the bound $\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} \leq \frac{1}{1-\theta_l}$, which follows from Lemma C.3. On the other hand, under $E_{\mathrm{bad}}$, we have

$$\begin{aligned}
\left(1 + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+\right)^n &\leq \left(1 + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{1}{\theta_l} - 1\right)\right)^n \\
&\leq \left(1 + \sum_{l=1}^{L} \frac{1/(8(1-\theta_l)) + 1/4}{2^l}\left(\frac{1-\theta_l}{\theta_l}\right)\right)^n \\
&\leq \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l}\right)^n,
\end{aligned}$$

where the first inequality follows from $t_l \leq \theta_l S_l$. Hence we have

$$\begin{aligned}
&\mathbb{E}_{t_l \sim \mathrm{Hyper}(\theta_l S_l, S_l, \theta S_l),\, \forall l \in [L]}\left[\left(1 + \sum_{l=1}^{L} \frac{\phi^l_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}/8 + 1/4}{2^l}\left(\frac{t_l}{\theta_l^2 S_l} - 1\right)_+\right)^n\right] \\
&\leq \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l(1-\theta_l)}\epsilon\right)^n + \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l}\right)^n \cdot \mathbb{P}_{t_l \sim \mathrm{Hyper}(\theta_l S_l, S_l, \theta S_l),\, \forall l \in [L]}(E_{\mathrm{bad}}) \\
&\leq \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l(1-\theta_l)}\epsilon_l\right)^n + \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l}\right)^n \sum_{l=1}^{L} \exp\bigl(-2\epsilon_l^2\theta_l S_l\bigr) \\
&= \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l(1-\theta_l)}\epsilon_l\right)^n + \sum_{l=1}^{L} \exp\left(n\log\left(1 + \sum_{j=1}^{L} \frac{1}{2^{j+1}\theta_j}\right) - 2\epsilon_l^2\theta_l S_l\right) \\
&= \left(1 + \sum_{l=1}^{L} \frac{1}{2^{l+1}\theta_l(1-\theta_l)}\epsilon_l\right)^n + \sum_{l=1}^{L} \exp\left(n\sum_{j=1}^{L} \frac{1}{2^{j+1}\theta_j} - 2\epsilon_l^2\theta_l S_l\right) \qquad \text{(C.25)}
\end{aligned}$$

Combining Eqs. (C.20), (C.24) and (C.25) (note that we are focusing on $\mathcal{M}_1$), we have

$$D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1) \le \inf_{\substack{\epsilon_l \in (0, \theta_l^2 S_l), \\ \forall l \in [L]}} \left\{ \left( 1 + \sum_{l=1}^{L} \frac{\epsilon_l}{2^{l+1}\theta_l(1-\theta_l)} \right)^n + \sum_{l=1}^{L} \exp\left( n \sum_{j=1}^{L} \frac{1}{2^{j+1}\theta_j} - 2\epsilon_l^2 \theta_l S_l \right) \right\} - 1,$$

Let $c \in (0, 1/2)$ be an arbitrary constant. We set $\epsilon_l = 2c \cdot \frac{(1-\theta_l)\theta_l}{n}$ (which belongs to $(0, \theta_l^2 S_l)$ because $\epsilon_l < \theta_l$ since $n \ge 1$ and $\theta_l S_l \ge 1$ by assumption) for all $l \in [L]$. Then we have

$$\left( 1 + \sum_{l=1}^{L} \frac{\epsilon_l}{2^{l+1}(1-\theta_l)\theta_l} \right)^n \le \left( 1 + \frac{c}{n} \right)^n \le e^c \le 1 + 2c,$$

and

$$D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1) \le 2c + \sum_{l=1}^{L} \exp\left( n \sum_{j=1}^{L} \frac{1}{2^{j+1}\theta_j} - 8c^2 \frac{(1-\theta_l)^2 \theta_l^3}{n^2} S_l \right).$$

In particular, whenever $S_l \ge \frac{n^3}{4c^2\theta_l^3(1-\theta_l)^2} \frac{1}{\min_{j\in[L]} \theta_j}$, we have

$$D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1) \le 2c + \exp\left( n \sum_{j=1}^{L} \frac{1}{2^{j+1}\theta_j} - 2n \frac{1}{\min_{j\in[L]} \theta_j} \right) \le 2c + \exp(-n).$$

Since $\theta_l = \frac{\alpha}{1-(l-1)\alpha}$ and the parameter $\alpha \in [\frac{1}{2L}, \frac{1}{L+1}]$ for the MDP family $\mathcal{M}_1$, we have $\theta_l \in [\frac{1}{2L}, \frac{1}{2}]$ for all $l \in [L]$. Setting $c = 1/10$. Whenever $n \ge 5$ and $S - 5 > 3200n^3L^6$, we have $S_l = \frac{S-5}{L_{\mathrm{div}}}(2L + 1 - l)(L + 2 - l) > 1600n^3L^4$ for all $l \in [L]$ (recall that $L_{\mathrm{div}} \le 4L^3$), and hence

$$D_{\chi^2}(\mathbb{P}_n^1 \,\|\, \mathbb{Q}_n^1) \le \frac{1}{5} + \exp(-n) \le \frac{1}{4}.$$

Using the same calculation, whenever $n \ge 5$ and $S - 5 > 800n^3L^6$, it holds that

$$D_{\chi^2}(\mathbb{P}_n^2 \,\|\, \mathbb{Q}_n^2) \le \frac{1}{5} + \exp(-n) \le \frac{1}{4}.$$

Combining th above two inequalities with (C.19), we have $D_{\mathrm{TV}}\big(\mathbb{P}_n^1, \mathbb{P}_n^2\big) \le 1/2 + n/(8 \cdot 2^L)$, which proves the lemma.

□

## C.8 Proofs of Propositions C.1 and C.2

*Proof of Proposition C.1.* Since for all states in $\mathcal{S} \setminus \{\mathfrak{s}\}$ the two actions in $\mathcal{A}$ have identical effects, we have $Q^\pi(s,a) = Q^\star(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and for all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Hence we only need to show $Q_M^\star = f_1$ for all $M \in \mathcal{M}_1$ and $Q_M^\star = f_2$ for all $M \in \mathcal{M}_2$.

Consider an arbitrary $M = M_{\boldsymbol{I},\alpha,w}^L \in \mathcal{M}$. First, for any self-looping terminal state $s \in$

$\{W, X, Y, Z\}$, we have

$$V_M^\star(s) = Q_M^\star(s, \mathfrak{a}) = \sum_{h=0}^{\infty} \gamma^h R_{L,\alpha,w}(s, \mathfrak{a}) = \frac{1}{1-\gamma} \cdot \begin{cases} w, & s = W, \\ 1, & s = X, \\ 0, & s = Y, \\ \frac{\alpha}{1-L\alpha}, & s = Z. \end{cases}$$

Next, for $l = L, \ldots, 1$, for any $l^{\text{th}}$-layer intermediate state $s \in \mathcal{S}^l$, by the Bellman optimality equation, we have

$$\begin{aligned} V_M^\star(s) = Q_M^\star(s, \mathfrak{a}) &= R_{L,\alpha,w}(s, \mathfrak{a}) + \gamma \mathbb{E}_{s' \sim P^L_{\boldsymbol{I},\alpha,w}(s,\mathfrak{a})}[V_M^\star(s')] \\ &= \begin{cases} 0 + \gamma\big[\gamma^{L-l}\alpha V_M^\star(X) + 0\big], & s \in I^l \\ 0 + \gamma\Big[\frac{(1-l\alpha)}{1-(l-1)\alpha}\mathbb{E}_{s' \sim \mathrm{Unif}(I^{l+1})} V_M^\star(s') + 0\Big], & s \in \bar{I}^l \end{cases} \\ &= \begin{cases} \frac{\gamma}{1-\gamma}\frac{\gamma^{L-l}\alpha}{1-(l-1)\alpha}, & s \in I^l \\ \frac{\gamma}{1-\gamma}\frac{\gamma^{L-l}\alpha}{1-(l-1)\alpha}, & s \in \bar{I}^l \end{cases} \\ &= \frac{\gamma}{1-\gamma}\frac{\gamma^{L-l}\alpha}{1-(l-1)\alpha}. \end{aligned}$$

For the initial state $\mathfrak{s}$, we have

$$Q_M^\star(\mathfrak{s}, 1) = R_{L,\alpha,w}(\mathfrak{s}, 1) + \gamma[V_M^\star(W)] = \frac{\gamma w}{1-\gamma},$$

$$Q_M^\star(\mathfrak{s}, 2) = R_{L,\alpha,w}(\mathfrak{s}, 2) + \gamma \mathbb{E}_{s' \sim P^L_{\boldsymbol{I},\alpha,w}(s,2)}[V_M^\star(s')] = \frac{\gamma V_\alpha}{1-\gamma}.$$

Therefore, $Q_M^\star = f_1$ if $M \in \mathcal{M}_1$, and $Q_M^\star = f_2$ if $M \in \mathcal{M}_2$. □

*Proof of Proposition C.2.* We now verify the concentrability condition (4.1).

Consider any $M \in \mathcal{M}_1$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\sup_{\nu \text{ is admissible}} \nu(s,a) \leq \begin{cases} 1, & \text{if } s \in \{\mathfrak{s}, W, X, Y\}, a \in \mathcal{A}, \\ \frac{1}{2} \cdot \frac{1}{2^L}, & \text{if } s = Z, a \in \mathcal{A}, \\ & \text{(maximized when } h = 1) \\ \frac{1}{2} \cdot \frac{1}{2^l}\frac{1}{S_l}, & \text{if } s \in \bar{I}^l, a \in \mathcal{A}, l \in [L], \\ & \text{(maximized when } h = 1) \\ \frac{1}{2} \cdot \frac{1}{2}\frac{1}{S_1}, & \text{if } s \in I^1, a \in \mathcal{A}, \\ & \text{(maximized when } h = 1) \\ \max\left\{\frac{1}{2} \cdot \frac{1}{2^l}\frac{1}{S_l}, \frac{1}{2} \cdot \frac{1}{2^{l-1}}\frac{1-(l-1)\alpha_2}{1-(l-2)\alpha_2}\frac{1-(l)\alpha_1}{1-(l-1)\alpha_1}\frac{1}{\theta_l^{(1)} S_l}\right\}, & \text{if } s \in I^l, a \in \mathcal{A}, 2 \leq l \leq L, \\ & \text{(maximized over } h = 1, 2) \end{cases}$$

Recall the definition of $\mu$ in Appendix C.5.3. We have

$$\mu(s,a) \geq \begin{cases} \frac{1}{16} \cdot \frac{1}{2}, & \text{if } s \in \{\mathfrak{s}, W, X, Y\}, a \in \mathcal{A}, \\ \frac{1}{8} \cdot \frac{1}{2^L} \cdot \frac{1}{2}, & \text{if } s = Z, a \in \mathcal{A}, \\ \frac{1}{8} \cdot \frac{1}{2^l} \frac{1}{S_l} \cdot \frac{1}{2}, & \text{if } s \in \bar{I}^l, a \in \mathcal{A}, l \in [L], \\ \frac{1}{8} \cdot \frac{1}{2} \frac{1}{S_1} \cdot \frac{1}{2}, & \text{if } s \in I^1, a \in \mathcal{A}, \\ \frac{1}{8} \cdot \frac{1}{2^l} \frac{1}{S_l} \cdot \frac{1}{2}, & \text{if } s \in I^l \cdot \frac{1}{2}, a \in \mathcal{A}, 2 \leq l \leq L \end{cases}$$

Combining the above two inequalities, we have

$$\begin{aligned} \sup_{\nu \text{ is admissible}} \left\| \frac{\nu}{\mu} \right\|_\infty &\leq \min_{2 \leq l \leq L} \ \left(8 \cdot 2^l S_l \cdot 2\right) \frac{1}{2} \cdot \frac{1}{2^{l-1}} \frac{1-(l-1)\alpha_2}{1-(l-2)\alpha_2} \frac{1-(l)\alpha_1}{1-(l-1)\alpha_1} \frac{1}{\theta_l^{(1)} S_l} \\ &= \min_{2 \leq l \leq L} \ \frac{16}{\theta_l^{(1)}} = \frac{16}{\theta_2^{(1)}} = \frac{16(1-\alpha_2)}{\alpha_2} \leq 32L, \end{aligned}$$

where the last inequality follows from $\alpha_2 \geq 1/(2L)$.

Similarly, consider any $M \in \mathcal{M}_2$, we have

$$\sup_{\nu \text{ is admissible}} \left\| \frac{\nu}{\mu} \right\|_\infty \leq 32L.$$

We conclude that the construction satisfies concentrability with $C_{\text{conc}} \leq 32L$. □

# Appendix D

# Supplementary Material for Chapter 5

## D.1 Proofs of Statements in Section 5.3

### D.1.1 Proof of Theorem 5.1

As preparations, we introduce two results from Abbasi-Yadkori et al. 2011, which will be used in the analysis.

**Lemma D.1** (Lemma 11 in Abbasi-Yadkori et al. 2011). *Let $\{X_t : t \geq 1\}$ be a sequence in $\mathbb{R}^d$, $V$ be a $d \times d$ positive definite matrix and define $V_t = V + \sum_{s=1}^{t} X_s X_s^\intercal$. If $||X_t||_2 \leq L$ for all $t$ and $\lambda_{\min}(V) \geq \max\{1, L^2\}$, then*

$$\sum_{t=1}^{T} ||X_t||^2_{V_{t-1}^{-1}} \leq 2\Big(d \log \frac{\mathrm{Tr}(V) + TL^2}{d} - \log\det V\Big).$$

**Lemma D.2** (Theorem 2 in Abbasi-Yadkori et al. 2011). *For any $0 < \epsilon < 1$, any $t \geq 1$,*

$$\mathbb{P}\left(||\theta^\star - \hat{\theta}_s||_{V_{s,n}} \leq R\sqrt{2\log\Big(\frac{1 + (1+u^2)(s+n)/\lambda}{\epsilon}\Big)} + \sqrt{\lambda(\alpha^2_{\max} + \beta^2_{\min})}, \forall 1 \leq s \leq t\right) \geq 1 - \epsilon.$$

We now divide the proof for Theorem 5.1 into two steps by proving the instance-independent upper bound $\mathcal{O}(\sqrt{T}\log T)$ and the instance-dependent upper bound $\mathcal{O}(\frac{T(\log T)^2}{(n \wedge T)\delta^2})$.

**Step 1.** In this step, we prove that the regret of O3FU algorithm is $\mathcal{O}(\sqrt{T}\log T)$. Let $x_t = [1\ p_t]^T$ for each $t \geq 1$. For any $t \geq 2$, suppose $\theta^\star \in \mathcal{C}_{t-1}$ (note that in this case, $\mathcal{C}_{t-1} \cap \Theta^\dagger \neq \varnothing$, and thus, $\tilde{\theta}_t$ is well-defined), then we have

$$\begin{aligned}
\psi(\theta^\star)(\alpha^\star + \beta^\star\psi(\theta^\star)) - p_t(\alpha^\star + \beta^\star p_t) &\leq p_t(\tilde{\alpha}_t + \tilde{\beta}_t p_t) - p_t(\alpha^\star + \beta^\star p_t) \\
&\leq u||x_t||_{V_{t-1,n}^{-1}} \cdot ||\tilde{\theta}_t - \theta^\star||_{V_{t-1,n}} \\
&\leq 2u||x_t||_{V_{t-1,n}^{-1}} \cdot w_{t-1} \qquad \text{(D.1)}
\end{aligned}$$

where the first inequality follows from the definition of $(p_t, \tilde{\theta}_t)$ in O3FU algorithm, the second inequality follows from Cauchy-Schwarz inequality, and the last inequality follows from $\theta^\star, \tilde{\theta}_t \in \mathcal{C}_{t-1}$.

Therefore,

$$\sum_{t=2}^{T}\big(\psi(\theta^\star)(\alpha^\star+\beta^\star\psi(\theta^\star))-p_t(\alpha^\star+\beta^\star p_t)\big)\le\sqrt{(T-1)\sum_{t=2}^{T}\big(\psi(\theta^\star)(\alpha^\star+\beta^\star\psi(\theta^\star))-p_t(\alpha^\star+\beta^\star p_t)\big)^2}$$
$$\le 2u\sqrt{(T-1)w_{T-1}^2\sum_{t=2}^{T}||x_t||^2_{V_{t-1,n}^{-1}}}, \tag{D.2}$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from inequality (D.1) and the fact that $w_t$ increases in $t$.

Then we use Lemma D.1 to bound the term $\sum_{t=1}^{T}||x_t||^2_{V_{t-1,n}^{-1}}$. To apply Lemma D.1, let $d=2$, $L=\sqrt{1+u^2}$, $\lambda=1+u^2$,

$$X_t=\begin{bmatrix}1\\ p_t\end{bmatrix},\quad V=\lambda I+n\begin{bmatrix}1 & \hat{p}\\ \hat{p} & \hat{p}^2\end{bmatrix},\quad V_t=V+\sum_{s=1}^{t}\begin{bmatrix}1 & p_s\\ p_s & p_s^2\end{bmatrix}.$$

Then we have

$$\sum_{t=1}^{T}||x_t||^2_{V_{t-1,n}^{-1}}\le 2\Big(2\log\frac{(2\lambda+n(1+\hat{p}^2))+T(1+u^2)}{2}-\log\big(\lambda(\lambda+n(1+\hat{p}^2))\big)\Big)$$
$$\le 2\log\Big(\frac{(1+u^2)(2+n+T)^2}{4(1+l^2)(1+n)}\Big),$$

which, combined with inequality (D.2), the definition of $w_T^2=\mathcal{O}(\log T)$, implies that when $\theta^\star\in\mathcal{C}_{t-1}$ for any $t\ge 2$,

$$\sum_{t=2}^{T}\big(\psi(\theta^\star)(\alpha^\star+\beta^\star\psi(\theta^\star))-p_t(\alpha^\star+\beta^\star p_t)\big)=\mathcal{O}\big(\sqrt{T}\log T\big). \tag{D.3}$$

Then the regret of O3FU algorithm is upper bounded as follows:

$$\begin{aligned}
&\sum_{t=2}^{T}\mathbb{E}\big[r^\star(\theta^\star)-r(p_t;\theta^\star)\big]\\
&=\sum_{t=2}^{T}\mathbb{E}\Big[\big(r^\star(\theta^\star)-r(p_t;\theta^\star)\big)\cdot 1_{\{\forall 2\le s\le t,\theta^\star\in\mathcal{C}_s\}}\Big]+\sum_{t=2}^{T}\mathbb{E}\Big[(-\beta^\star)(\psi(\theta^\star)-p_t)^2\cdot 1_{\{\exists 2\le s\le t,\theta^\star\notin\mathcal{C}_s\}}\Big]\\
&=\mathcal{O}(\sqrt{T}\log T)+|\beta_{\min}|(u-l)^2\sum_{t=2}^{T}\frac{1}{T^2}\\
&=\mathcal{O}(\sqrt{T}\log T),
\end{aligned}$$

where the second identity follows from inequality (D.3) and Lemma D.2 with $\epsilon=\frac{1}{T^2}$ for any $t\ge 2$.

**Step 2.** In this step, we prove that the regret of O3FU algorithm is also $\mathcal{O}(\frac{T(\log T)^2}{(n\wedge T)\delta^2})$. It suffices to show the case when $\delta\ge\frac{2\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{\sqrt{2}\beta_{\max}^2}\cdot\frac{w_T}{n^{1/4}}$, since otherwise, $\frac{T(\log T)^2}{(n\wedge T)\delta^2}\gtrsim\frac{T\sqrt{n}(\log T)^2}{(n\wedge T)\log T}\gtrsim\sqrt{T}\log T$, and the upper bound in Theorem 5.1 becomes $\mathcal{O}(\sqrt{T}\log T)$, which is already proven in Step 1.

Note that it suffices to bound the term $\sum_{t=2}^{T}\mathbb{E}[||\theta^\star-\tilde{\theta}_t||^2]$. Since $T_0$ defined in Lemma 5.1 is an

absolute constant, the result is trivial when $T \le T_0$. We then consider $T \ge T_0$.

$$\begin{aligned}\sum_{t=2}^{T}\mathbb{E}[||\theta^\star-\tilde{\theta}_t||^2] &= \sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{\forall 2\le s\le t,\theta^\star\in\mathcal{C}_s\}}\Big]+\sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{\exists 2\le s\le t,\theta^\star\notin\mathcal{C}_s\}}\Big]\\ &\le \sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{U_{t,2}\}}\Big]+\sum_{t=2}^{T}\big((\alpha_{\max}-\alpha_{\min})^2+(\beta_{\max}-\beta_{\min})^2\big)\frac{1}{T^2}\\ &\le C_2\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2}+((\alpha_{\max}-\alpha_{\min})^2+(\beta_{\max}-\beta_{\min})^2)\frac{1}{T},\end{aligned}$$

where the first inequality follows from the proof of Lemma 5.1 and the concentration inequality in Lemma D.2 with $\epsilon=\frac{1}{T^2}$ for any $t\ge 2$. It is easy to verify that when $n<T$,

$$\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2}=\sum_{t=1}^{n}\frac{w_t^2}{t\delta^2}+\sum_{t=n+1}^{T-1}\frac{w_t^2}{n\delta^2}=\mathcal{O}\Big(\frac{(\log T)^2}{\delta^2}\Big)+\mathcal{O}\Big(\frac{T\log T}{n\delta^2}\Big)=\mathcal{O}\Big(\frac{T(\log T)^2}{(n\wedge T)\delta^2}\Big),$$

and when $n\ge T$,

$$\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2}=\sum_{t=1}^{T-1}\frac{w_t^2}{t\delta^2}=\mathcal{O}\Big(\frac{\log T\log T}{\delta^2}\Big)=\mathcal{O}\Big(\frac{T(\log T)^2}{(n\wedge T)\delta^2}\Big).$$

Combining both cases of $n<T$ and $n\ge T$, we have $\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2}=\mathcal{O}(\frac{T(\log T)^2}{(n\wedge T)\delta^2})$, which completes the proof. □

### D.1.2 Proof of Lemma 5.1

When $t=1$, since $p_1=l\cdot\mathbb{I}\{\hat{p}>\frac{u+l}{2}\}+u\cdot\mathbb{I}\{\hat{p}\le\frac{u+l}{2}\}$, then $|p_1-\hat{p}|\ge\frac{u-l}{2}\ge\frac{1}{2}\delta$. Thus, when $t=1$, $U_{t,1}$ holds.

We next prove the following result: under the assumptions of Lemma 5.1, suppose for each $1\le s\le t-1$ (for a fixed $2\le t\le T$), the event $U_{s,1}$ holds, then $U_{t,1}$ and $U_{t,2}$ also hold. To this end, let $\Delta\alpha_t=\tilde{\alpha}_t-\alpha^\star$, $\Delta\beta_t=\tilde{\beta}_t-\beta^\star$, and $\gamma_t=\frac{\Delta\alpha_t}{\Delta\beta_t}$ (when $\Delta\beta_t\ne 0$). Since $\theta^\star\in\mathcal{C}_{t-1}$ and $\tilde{\theta}_t\in\mathcal{C}_{t-1}$, we have $||\tilde{\theta}_t-\theta^\star||_{V_{t-1,n}}^2\le 2\big(||\tilde{\theta}_t-\hat{\theta}_{t-1}||_{V_{t-1,n}}^2+||\theta^\star-\hat{\theta}_{t-1}||_{V_{t-1,n}}^2\big)\le 2w_{t-1}^2$, which is equivalent to

$$\lambda\big((\Delta\alpha_t)^2+(\Delta\beta_t)^2\big)+n\big(\Delta\alpha_t+\Delta\beta_t\hat{p}\big)^2+\sum_{s=1}^{t-1}\big(\Delta\alpha_t+\Delta\beta_t p_s\big)^2\le 2w_{t-1}^2. \tag{D.4}$$

We next divide the proof into three cases.

**Case 1**: $\Delta\beta_t=0$. In this case, (D.4) becomes $(\Delta\alpha_t)^2(\lambda+n+t-1)\le 2w_{t-1}^2$, and

$$||\theta^\star-\tilde{\theta}_t||^2=(\Delta\alpha_t)^2+(\Delta\beta_t)^2=(\Delta\alpha_t)^2\le\frac{2w_{t-1}^2}{n+t-1}. \tag{D.5}$$

Therefore, (D.5) implies that

$$||\theta^\star-\tilde{\theta}_t||^2\le\frac{2w_{t-1}^2}{n\wedge(t-1)}\le\frac{2(u-l)^2w_{t-1}^2}{(n\wedge(t-1))\delta^2},$$

and

$$\begin{aligned}|\hat{p}-p_t| &\geq |\hat{p}-\psi(\theta^\star)| - |p_t-\psi(\theta^\star)| \\ &\geq |\hat{p}-\psi(\theta^\star)| - \frac{\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{\sqrt{2}\beta_{\max}^2}\cdot\frac{w_{t-1}}{\sqrt{n+t-1}} \\ &\geq |\hat{p}-\psi(\theta^\star)| - \frac{|\psi(\theta^\star)-\hat{p}|}{2n^{\frac{1}{4}}} \\ &\geq \frac{1}{2}\delta,\end{aligned}$$

where the second inequality follows from (D.5) and Lipschitz continuity of the function $\psi(\cdot)$: $|\psi(\theta_1)-\psi(\theta_2)| \leq \frac{1}{2\beta_{\max}^2}\sqrt{\alpha_{\max}^2+\beta_{\max}^2}\cdot||\theta_1-\theta_2||$, and the third inequality holds since the assumption $\delta \geq \frac{2\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{\sqrt{2}\beta_{\max}^2}\cdot\frac{w_T}{n^{1/4}}$ implies

$$\frac{w_{t-1}}{\sqrt{n+t-1}} \leq \frac{w_T}{\sqrt{n}} \leq \frac{\sqrt{2}\beta_{\max}^2 n^{\frac{1}{4}}}{2\sqrt{\alpha_{\max}^2+\beta_{\max}^2}\cdot\sqrt{n}}\delta. \tag{D.6}$$

**Case 2**: $\Delta\beta_t \neq 0$, $|\gamma_t| \geq 4u+1$. In this case, we have

$$||\theta^\star-\tilde{\theta}_t||^2 \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{\lambda(1+\gamma_t^2)+n(\gamma_t+\hat{p})^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2} \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{n(\gamma_t+\hat{p})^2} \leq \frac{4w_{t-1}^2}{n}, \tag{D.7}$$

where the first inequality holds since $||\theta^\star-\tilde{\theta}_t||^2 = (\Delta\beta_t)^2(1+\gamma_t^2)$, and from (D.4), we have

$$(\Delta\beta_t)^2 \leq \frac{2w_{t-1}^2}{\lambda(1+\gamma_t^2)+n(\gamma_t+\hat{p})^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2},$$

and the last inequality follows from $1+\gamma_t^2 \leq 2(\gamma_t+\hat{p})^2$, which is easily verified by noting $(\gamma_t+2\hat{p})^2 \geq (|\gamma_t|-2\hat{p})^2 \geq (2\hat{p}+1)^2 \geq 2\hat{p}^2+1$. Then, (D.7) implies that

$$||\theta^\star-\tilde{\theta}_t||^2 \leq \frac{4(u-l)^2w_{t-1}^2}{(n\wedge(t-1))\delta^2},$$

and

$$|\hat{p}-p_t| \geq |\hat{p}-\psi(\theta^\star)| - |p_t-\psi(\theta^\star)| \geq |\hat{p}-\psi(\theta^\star)| - \frac{\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{2\beta_{\max}^2}\frac{2w_{t-1}}{\sqrt{n}} \geq (1-\frac{\sqrt{2}}{2})\delta,$$

where the second inequality follows from Lipschitz continuity of $\psi(\cdot)$ and (D.7), and the third inequality follows from (D.6).

**Case 3**: $\Delta\beta_t \neq 0$, $|\gamma_t| < 4u+1$. Recall the following definitions of $C_0$, $C_1$ and $T_0$ in Lemma 5.1:

$$C_0 = \frac{l|\beta_{\max}|}{u|\beta_{\min}|}, \quad C_1 = \frac{4(C_0+1)^2}{C_0^2}\big(1+(4u+1)^2\big), \quad T_0 = \min\Big\{t\in\mathbb{N} : w_t \geq \frac{\sqrt{C}_1\beta_{\max}^2}{\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}}\Big\}.$$

**Subcase 3.1**: $1+\gamma_t^2 \le C_1 \frac{(\gamma_t+\hat{p})^2}{\delta^2}$. In this subcase, since

$$||\theta^\star - \tilde{\theta}_t||^2 \le \frac{2w_{t-1}^2(1+\gamma_t^2)}{n(\gamma_t+\hat{p})^2} \le \frac{2C_1 w_{t-1}^2}{n\delta^2},$$

then we have

$$||\theta^\star - \tilde{\theta}_t||^2 \le \frac{2C_1 w_{t-1}^2}{(n \wedge (t-1))\delta^2}.$$

In addition, since $T \ge T_0$, it follows that

$$\begin{aligned}
|p_t - \hat{p}| &\ge |\psi(\theta^\star) - \hat{p}| - |p_t - \psi(\theta^\star)| \\
&\ge |\psi(\theta^\star) - \hat{p}| - \frac{\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{2\beta_{\max}^2}\frac{\sqrt{2C_1}w_{t-1}}{\sqrt{n}\delta} \\
&\ge |\psi(\theta^\star) - \hat{p}| - \frac{\sqrt{C}_1\beta_{\max}^2}{2\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}w_T}\delta \\
&\ge \frac{1}{2}\delta,
\end{aligned}$$

where in the third inequality, we utilize the fact that $\delta \ge \frac{2\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{\sqrt{2}\beta_{\max}^2} \cdot \frac{w_T}{n^{1/4}}$, and the last inequality follows from $T \ge T_0$ and the definition of $T_0$.

**Subcase 3.2**: $1+\gamma_t^2 > C_1 \frac{(\gamma_t+\hat{p})^2}{\delta^2}$. In this subcase, we have

$$\begin{aligned}
||\theta^\star - \tilde{\theta}_t||^2 &\le \frac{2w_{t-1}^2(\gamma_t^2+1)}{n(\gamma_t+\hat{p})^2 + \sum_{s=1}^{t-1}(\gamma_t+p_s)^2} \\
&\le \frac{4w_{t-1}^2(\gamma_t^2+1)}{\sum_{s=1}^{(t-1)\wedge n}(p_s-\hat{p})^2} \\
&\le \frac{4w_{t-1}^2((4u+1)^2+1)}{(n\wedge(t-1))\cdot\min\{(1-\frac{\sqrt{2}}{2})^2, \frac{C_0^2}{4}\}\cdot\delta^2},
\end{aligned}$$

where the second inequality holds since

$$n(\gamma_t+\hat{p})^2 + \sum_{s=1}^{t-1}(\gamma_t+p_s)^2 \ge \sum_{s=1}^{n\wedge(t-1)}\left((\gamma_t+p_s)^2 + (\gamma_t+\hat{p})^2\right) \ge \frac{1}{2}\sum_{s=1}^{n\wedge(t-1)}(p_s-\hat{p})^2,$$

and the last inequality follows from $|\gamma_t| \le 4u+1$ and the inductive assumption:

$$\forall 1 \le s \le t-1, \quad |p_s - \hat{p}| \ge \min\{1-\frac{\sqrt{2}}{2}, \frac{C_0}{2}\}\cdot\delta.$$

Now, it suffices to bound the term $|p_t - \hat{p}|$. If we can prove the following inequality:

$$|\gamma_t + p_t| \ge C_0|\gamma_t + \psi(\theta^\star)|, \tag{D.8}$$

then $|p_t - \hat{p}|$ can be bounded as follows:

$$\begin{aligned}
|p_t - \hat{p}| &\geq |p_t + \gamma_t| - |\gamma_t + \hat{p}| \\
&\geq C_0|\gamma_t + \psi(\theta^\star)| - |\gamma_t + \hat{p}| \\
&\geq C_0(|\psi(\theta^\star) - \hat{p}| - |\gamma_t + \hat{p}|) - |\gamma_t + \hat{p}| \\
&= C_0|\psi(\theta^\star) - \hat{p}| - (C_0 + 1)|\gamma_t + \hat{p}| \\
&\geq \Big(C_0 - (C_0 + 1)\frac{\sqrt{1 + (4u+1)^2}}{\sqrt{C_1}}\Big)|\psi(\theta^\star) - \hat{p}| \\
&\geq \frac{C_0}{2}\delta,
\end{aligned}$$

where the second inequality follows from (D.8), the fourth inequality follows from the assumption of Subcase 3.2, i.e., $1 + \gamma_t^2 > C_1\frac{(\gamma_t + \hat{p})^2}{\delta^2}$ and $|\gamma_t| \leq 4u + 1$, and the last inequality follows from the definition of $C_1$.

Finally, we prove inequality (D.8). We define

$$A_1 = p_t(\tilde{\alpha}_t + \tilde{\beta}_t p_t),\ A_2 = p_t(\alpha^\star + \beta^\star p_t),\ A_3 = \psi(\theta^\star)(\tilde{\alpha}_t + \tilde{\beta}_t\psi(\theta^\star)),\ A_4 = \psi(\theta^\star)(\alpha^\star + \beta^\star\psi(\theta^\star)).$$

Recall that $p_t$ and $\psi(\theta^\star)$ are the maximizers of the following maximization problem:

$$p_t = \arg\max_{p\in[l,u]} p(\tilde{\alpha}_t + \tilde{\beta}p), \quad \psi(\theta^\star) = \arg\max_{p\in[l,u]} p(\alpha^\star + \beta^\star p),$$

then we have the following relationships for $A_i$, $1 \leq i \leq 4$:

$$A_1 \geq A_3, \tag{D.9}$$

$$A_1 \geq A_4 \geq A_2. \tag{D.10}$$

To show inequality (D.8), we consider the following two cases when $A_3 \geq A_2$ and $A_3 < A_2$. If $A_3 \geq A_2$, then we have

$$|\Delta\alpha_t + \Delta\beta_t p_t| = \frac{A_1 - A_2}{p_t} \geq \frac{|A_4 - A_3|}{p_t} = \frac{\psi(\theta^\star)}{p_t}|\Delta\alpha_t + \Delta\beta_t\psi(\theta^\star)| \geq \frac{l}{u}|\Delta\alpha_t + \Delta\beta_t\psi(\theta^\star)|, \tag{D.11}$$

where the first inequality follows from $A_3, A_4 \in [A_2, A_1]$. Without loss of generality, we assume that $\Delta\beta_t > 0$, since otherwise, we can redefine $\Delta\alpha_t$ and $\Delta\beta_t$ as $\alpha^\star - \tilde{\alpha}_t$ and $\beta^\star - \tilde{\beta}_t$ respectively, and the proof will be similar. Therefore, by dividing $\Delta\beta_t$ on both sides of (D.11), we get inequality (D.8). If $A_3 < A_2$,

$$\begin{aligned}
|\Delta\alpha_t + \Delta\beta_t p_t| &= \frac{A_1 - A_2}{p_t} \geq \frac{A_4 - A_2}{p_t} = \frac{-\beta^\star(\psi(\theta^\star) - p_t)^2}{p_t} = \frac{-\beta^\star\psi(\theta^\star)}{-\tilde{\beta}_t p_t} \cdot \frac{-\tilde{\beta}_t(\psi(\theta^\star) - p_t)^2}{\psi(\theta^\star)} \\
&\geq \frac{l|\beta_{\max}|}{u|\beta_{\min}|} \cdot \frac{A_1 - A_3}{\psi(\theta^\star)} \geq \frac{l|\beta_{\max}|}{u|\beta_{\min}|} \cdot \frac{A_4 - A_3}{\psi(\theta^\star)} = \frac{l|\beta_{\max}|}{u|\beta_{\min}|} \cdot |\Delta\alpha_t + \Delta\beta_t\psi(\theta^\star)|,
\end{aligned} \tag{D.12}$$

where the second identity and the second inequality follow from the property of quadratic functions. By dividing $\Delta\beta_t$ ($> 0$ by assumption) on both sides of (D.12), inequality (D.8) holds. It is also

worth noting that from the above arguments, inequality (D.8) holds universally due to the specific property of OFU principle and quadratic structure of the objective function, and does not depend on any inductive assumption.

Combining Cases 1–3, we conclude that

$$|p_t - \hat{p}| \geq \min\Big\{1 - \frac{\sqrt{2}}{2}, \frac{C_0}{2}\Big\} \cdot \delta,$$
$$||\theta^\star - \tilde{\theta}_t||^2 \leq \max\Big\{4(u-l)^2, 2C_1, \frac{4((4u+1)^2+1)}{\min\{\frac{C_0^2}{4}, (1-\frac{\sqrt{2}}{2})^2\}}\Big\} \cdot \frac{w_{t-1}^2}{(n \wedge (t-1))\delta^2},$$

i.e., $U_{t,1}$ and $U_{t,2}$ hold, which completes the inductive arguments. □

### D.1.3 Proof of Theorem 5.2

As preparation, we first present the multivariate van Trees inequality, which will be used in Step 1 of the proof for Theorem 5.2. For simplicity, we focus on the estimation problem for a real-valued function when stating the multivariate van Trees inequality, which is sufficient for our use, and we refer the interested readers to Gill and Levit (2001) for the more general version on estimating a vector-valued function.

**Lemma D.3** (Multivariate van Trees Inequality, Theorem 1, Gill and Levit (2001))**.** *Consider estimating a real-valued function $\psi(\theta)$ with parameter $\theta$ being an $s$-dimensional vector. Suppose we are given $n$ i.i.d. observations $X_1, X_2, \ldots, X_n$ drawn from a common distribution with probability density function $f(x,\theta)$. Suppose $\theta$ is in the compact set $\Theta \subseteq \mathbb{R}^s$, the prior probability density function of $\theta$ is denoted by $\lambda(\theta)$, and $C(\theta)$ is an $s$-dimensional row vector. If $f(x,\theta)$, $\lambda(\theta)$, $C(\theta)$ satisfy certain regularity conditions (see Assumptions in Section 4 of Gill and Levit (2001)), and in particular, $\lambda(\theta)$ is positive in the interior of $\Theta$ and zero on its boundary, then for any estimator $\psi_n$ based on $X_1, X_2, \ldots, X_n$,*

$$\mathbb{E}_\lambda\left[\mathbb{E}_\theta\left[(\psi_n - \psi(\theta))^2\right]\right] \geq \frac{(\mathbb{E}_\lambda[\mathrm{Tr}(C(\theta)(\frac{\partial \psi}{\partial \theta})^\top)])^2}{\widetilde{\mathcal{I}}(\lambda) + n \cdot \mathbb{E}_\lambda[\mathrm{Tr}(C(\theta)\mathcal{I}(\theta)(C(\theta))^\top)]},$$

*where* $\mathrm{Tr}(A)$ *denotes the trace for a square matrix* $A$*, and* $\widetilde{\mathcal{I}}(\lambda) = \int_\Theta \Big(\sum_{k=1}^s \frac{\partial}{\partial \theta_k}\big(C_k(\theta)\lambda(\theta)\big)\Big)\frac{1}{\lambda(\theta)}d\theta$.

It suffices to consider the case when $\varepsilon$ follows a normal distribution with standard deviation $R$. Without loss of generality, we assume $\xi = \frac{1}{2}$, and the analysis can be easily extended to general $\xi \in (0,1)$.

**Step 1.** As the first step, we will prove the following result: for any pricing policy $\pi \in \Pi$,

$$\sup_{\theta \in \Theta_0(\delta)} R_\theta^\pi(T) = \Omega\Big(\big(\sqrt{T} \wedge \big(\frac{T}{\delta^{-2} + (n \wedge T)\delta^2}\big)\big) \vee \log(1 + T\delta^2)\Big), \tag{D.13}$$

where $\Theta_0(\delta) = \big\{\theta \in \Theta^\dagger : \psi(\theta) - \hat{p} \in [\frac{\delta}{2}, \delta]\big\}$. When $\delta \geq \frac{lR}{16|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{2K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$, the above (D.13) implies the desired lower bound in Theorem 5.2. In what follows, we will prove three lower bounds for $\sup_{\theta \in \Theta_0(\delta)} R_\theta^\pi(T)$: $\Omega(\log(1+T\delta^2))$, $\Omega(\sqrt{T} \wedge \frac{T}{\delta^{-2}+T\delta^2})$, and $\Omega(\sqrt{T} \wedge \frac{T}{\delta^{-2}+n\delta^2})$, which, when combined together, imply the lower bound in (D.13).

Before invoking the multivariate van Trees inequality in Lemma D.3, we first note that since $\Theta_0(\delta) = \{\theta \in \Theta^\dagger : -\frac{\alpha}{2\hat{p}+\delta} \le \beta \le -\frac{\alpha}{2\hat{p}+2\delta}\}$, there exist some positive constants $x_0$, $y_0$, $\epsilon$ such that $\Theta_1(\delta) := [x_0 - \frac{1}{2}\epsilon\delta, x_0 + \frac{3}{2}\epsilon\delta] \times [-y_0 - \frac{3}{2}\epsilon\delta, -y_0 + \frac{1}{2}\epsilon\delta] \subseteq \Theta_0(\delta)$. Then we define a prior distribution for $\theta$ on $\Theta_1(\delta)$ as follows:

$$q(x,y) = \frac{1}{(\epsilon\delta)^2}\cos^2\Big(\frac{\pi(x - \frac{2x_0+\epsilon\delta}{2})}{2\epsilon\delta}\Big)\cdot\cos^2\Big(\frac{\pi(y + \frac{2y_0+\epsilon\delta}{2})}{2\epsilon\delta}\Big),\ \forall (x,y)\in\Theta_1(\delta). \tag{D.14}$$

In addition, we have the following inequality:

$$\begin{aligned}\sup_{\theta\in\Theta_0(\delta)} R_\theta^\pi(T) \ge \sup_{\theta\in\Theta_1(\delta)} R_\theta^\pi(T) &= \sup_{\theta\in\Theta_1(\delta)} \sum_{t=1}^T(-\beta)\cdot\sum_{t=1}^T\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\\ &\ge |\beta_{\max}|\cdot\sum_{t=1}^T\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big],\end{aligned} \tag{D.15}$$

where the first inequality holds since $\Theta_1(\delta)\subseteq\Theta_0(\delta)$, the identity follows from the property of quadratic function and optimality of $\psi(\theta)$, and the second inequality holds since $q(\theta)$ is a probability density distribution defined on $\Theta_1(\delta)$. Note that the reason for which we consider a subset of $\Theta_0(\delta)$ is that the Fisher information defined on the rectangle, i.e., $\Theta_1(\theta)$, will be easier to calculate later.

Then for each $t\ge 2$, by letting $n=1$, $X_1 = (\hat{\varepsilon}_1,\dots,\hat{\varepsilon}_n,\varepsilon_1,\dots,\varepsilon_{t-1})$, $\psi_n = p_t$, $\lambda(\cdot) = q(\cdot)$ in Lemma D.3, we have

$$\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big] \ge \frac{(\mathbb{E}_q[C(\theta)^\intercal\frac{\partial\psi}{\partial\theta}])^2}{\mathcal{I}(q)+\mathbb{E}_q[\mathbb{E}_\theta^\pi[C(\theta)^\intercal\mathcal{I}_{t-1}^\pi(\theta)C(\theta)]]}, \tag{D.16}$$

where $C(\theta)$ is any two-dimensional vector to be specified, and

$$\mathcal{I}(q) = \int_{(\theta_1,\theta_2)\in\Theta_1(\delta)}\sum_{i=1}^2\sum_{j=1}^2\frac{\partial}{\partial\theta_i}\big(C_i(\theta_1,\theta_2)\cdot q(\theta_1,\theta_2)\big)\cdot\frac{\partial}{\partial\theta_j}\big(C_j(\theta_1,\theta_2)\cdot q(\theta_1,\theta_2)\big)\cdot\frac{1}{q(\theta_1,\theta_2)}d\theta_1 d\theta_2,$$

and $\mathcal{I}_{t-1}^\pi(\theta)$ is the Fisher information matrix defined as

$$\mathcal{I}_{t-1}^\pi(\theta) = \frac{1}{R^2}\mathbb{E}_\theta^\pi\begin{bmatrix} n+t-1 & n\hat{p}+\sum_{s=1}^{t-1}p_s\\ n\hat{p}+\sum_{s=1}^{t-1}p_s & n\hat{p}^2+\sum_{s=1}^{t-1}p_s^2\end{bmatrix}.$$

We next start from (D.16), and prove the three lower bounds by specifying different $C(\theta)$ and bounding the resulting $\mathbb{E}_q[C(\theta)^\intercal\frac{\partial\psi}{\partial\theta}]$, $\mathcal{I}(q)$, and $\mathbb{E}_q[\mathbb{E}_\theta^\pi[C(\theta)^\intercal\mathcal{I}_{t-1}^\pi(\theta)C(\theta)]]$ in the RHS of (D.16).

To prove the first lower bound $\Omega(\log(1+T\delta^2))$, let $C(\theta) = (-\hat{p},1)$ in (D.16), then we have

$$\sum_{t=1}^T\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big] \ge \sum_{t=2}^T\frac{R^2c_1}{R^2\mathcal{I}(q)+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_s-\hat{p})^2]]} \ge \sum_{t=2}^T\frac{R^2c_1}{R^2\mathcal{I}(q)+(t-1)(u-l)^2},$$

where $c_1 = \big(\min_{\theta\in\Theta^\dagger}\frac{\alpha+\beta\hat{p}}{2\beta^2}\big)^2$. Since $C(\theta) = (-\hat{p},1)$ is independent of $\theta$, by changing variables in the

integrals, we have

$$\mathcal{I}(q) = \frac{\pi}{2\epsilon^2\delta^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial}{\partial\theta_i}\big(C_i(\theta_1,\theta_2)\cdot\tilde{q}(\theta_1,\theta_2)\big)\cdot\frac{\partial}{\partial\theta_j}\big(C_j(\theta_1,\theta_2)\cdot\tilde{q}(\theta_1,\theta_2)\big)\cdot\frac{1}{\tilde{q}(\theta_1,\theta_2)}d\theta_1 d\theta_2,$$

where $\tilde{q}(\theta_1,\theta_2) = \cos^2(\theta_1)\cdot\cos^2(\theta_2)$. Since the integral in the RHS of the above equation is a constant independent of $\delta$, we have $\mathcal{I}(q) = \Theta(\delta^{-2})$, it then follows from (D.15) that

$$\sup_{\theta\in\Theta_0(\delta)}\sum_{t=1}^{T}R_\theta^\pi(T) \geq |\beta_{\max}|\cdot\sup_{\theta\in\Theta_1(\delta)}\sum_{t=1}^{T}\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2] = \Omega(\log(1+T\delta^2)).$$

To prove the second lower bound $\Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+T\delta^2})$, let $C(\theta) = (-\hat{p},1)$ in (D.16) again, then we obtain

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big] &\geq \sum_{t=2}^{T}\frac{R^2c_1}{R^2\mathcal{I}(q)+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_s-\hat{p})^2]]}\\ &\geq \sum_{t=2}^{T}\frac{R^2c_1}{R^2\mathcal{I}(q)+2(t-1)\delta^2+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_s-\psi(\theta))^2]]}\\ &\geq \frac{R^2c_1(T-1)}{R^2\mathcal{I}(q)+2T\delta^2+\sum_{t=1}^{T}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]]},\end{aligned} \tag{D.17}$$

where the second inequality holds since $(p_s-\hat{p})^2 \leq 2(p_s-\psi(\theta)))^2+2(\hat{p}-\psi(\theta))^2 \leq 2(p_s-\psi(\theta))^2+2\delta^2$. It is easily verified that the inequality $x^2+bx+c\geq 0$ for $b>0, c<0, x\geq 0$ implies

$$x \geq \frac{1}{\sqrt{2}+1}\min\big\{\sqrt{|c|},\frac{2|c|}{b}\big\}. \tag{D.18}$$

Applying (D.18) to the inequality (D.17), we obtain from (D.15) that

$$\sup_{\theta\in\Theta_0(\delta)}\sum_{t=1}^{T}R_\theta^\pi(T) \geq |\beta_{\max}|\cdot\sum_{t=1}^{T}\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big] \geq \Omega(\sqrt{T}\wedge\frac{T}{\mathcal{I}(q)+T\delta^2}) = \Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+T\delta^2}),$$

where in the identity, we utilize the fact that $\mathcal{I}(q) = \Theta(\delta^{-2})$.

To prove the third lower bound $\Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+n\delta^2})$, we choose another vector $C(\theta) = (-\psi(\theta),1)$, and the inequality (D.16) becomes

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big] &\geq \sum_{t=2}^{T}\frac{R^2\alpha_{\min}^2/(4\beta_{\min}^2)^2}{R^2\mathcal{I}(q)+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_s-\psi(\theta))^2]]+n\delta^2}\\ &\geq \frac{R^2\alpha_{\min}^2/(4\beta_{\min}^2)^2(T-1)}{R^2\mathcal{I}(q)+\sum_{t=1}^{T}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]]+n\delta^2},\end{aligned}$$

which, combined with (D.15) and (D.18), implies that

$$\sup_{\theta\in\Theta_0(\delta)}R_\theta^\pi(T) \geq |\beta_{\max}|\sum_{t=1}^{T}\mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t-\psi(\theta))^2]\big] \geq \Omega(\sqrt{T}\wedge\frac{T}{\mathcal{I}(q)+n\delta^2}). \tag{D.19}$$

By definition,

$$\mathcal{I}(q) = \frac{\pi}{2\epsilon^2\delta^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial}{\partial\theta_i}\big(\tilde{C}_i(\theta_1,\theta_2)\cdot\tilde{q}(\theta_1,\theta_2)\big)\cdot\frac{\partial}{\partial\theta_j}\big(\tilde{C}_j(\theta_1,\theta_2)\cdot\tilde{q}(\theta_1,\theta_2)\big)\cdot\frac{1}{\tilde{q}(\theta_1,\theta_2)}d\theta_1 d\theta_2,$$

where $\tilde{C}(\theta_1,\theta_2) = \big(\frac{\frac{2\epsilon\delta\theta_1}{\pi}+x_0+\frac{\epsilon\delta}{2}}{2(\frac{2\epsilon\delta\theta_2}{\pi}-y_0-\frac{\epsilon\delta}{2})},1\big)$. Since both $\tilde{C}(\cdot)$ and $C(\theta)$ are bounded by constants independent of $\delta$, it is easily verified that the integral in the RHS of the above identity can be bounded by constant independent of $\delta$. Therefore, $\mathcal{I}(q) = \Theta(\delta^{-2})$, and from inequality (D.19), we have

$$\sup_{\theta\in\Theta_0(\delta)} R_\theta^\pi(T) = \Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+n\delta^2}).$$

**Step 2.** In this step, we complete the proof by showing that when $\delta \le \frac{lR}{16|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{2K_0e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$, for any admissible policy $\pi\in\Pi^\circ$, there exists $\theta\in\Theta^\dagger$ satisfying $|\psi(\theta)-\hat{p}|\in[\frac{1}{2}\delta,\frac{3}{2}\delta]$ such that

$$R_\theta^\pi(T) = \Omega\Big(\frac{\sqrt{T}}{(\log T)^{\lambda_0}}\Big). \tag{D.20}$$

Our proof of (D.20) is based on the concept of KL divergence, which is a quantitative measure of distance between two distributions. The definition is given as follows. For any two distributions $P_1$ and $P_2$, the KL divergence is

$$KL(P_1,P_2) = \mathbb{E}_{X\sim P_1}\Big[\log\frac{P_1(X)}{P_2(X)}\Big].$$

We now consider two vectors of demand parameters $\theta_1$ and $\theta_2$ satisfying the following conditions:

$$-\frac{\alpha_1}{2\beta_1} = \hat{p}+\delta,\quad -\frac{\alpha_2}{2\beta_2} = \hat{p}+\delta+\Delta,\quad (\alpha_1-\alpha_2)+(\beta_1-\beta_2)\hat{p} = 0.$$

where $\Delta>0$ is to be determined. For any policy $\pi\in\Pi^\circ$, let $P_1^\pi$ and $P_2^\pi$ be the following two probability measures induced by the common policy $\pi$ and two parameters $\theta_1$ and $\theta_2$ respectively:

$$P_i^\pi(\hat{D}_1,\ldots,\hat{D}_n,D_1,\ldots,D_T) = \prod_{t=1}^{n}\Big(\frac{1}{R}\phi\big(\frac{\hat{D}_t-(\alpha_i+\beta_i\hat{p})}{R}\big)\Big)\cdot\prod_{t=1}^{T}\Big(\frac{1}{R}\phi\big(\frac{D_t-(\alpha_i+\beta_i p_t)}{R}\big)\Big),\ i=1,2,$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$ is the probability density function of the standard normal distribution. From the definition of KL divergence, we have

$$\begin{aligned}
KL(P_1^\pi,P_2^\pi) &= \frac{(\beta_1-\beta_2)^2}{2R^2}\Big(n\big(\frac{\alpha_1-\alpha_2}{\beta_1-\beta_2}+\hat{p}\big)^2+\sum_{t=1}^{T}\mathbb{E}_{\theta_1}^\pi[(\frac{\alpha_1-\alpha_2}{\beta_1-\beta_2}+p_t)^2]\Big)\\
&= \frac{2\beta_2^2\Delta^2}{(\hat{p}+2\delta)^2R^2}\sum_{t=1}^{T}\mathbb{E}_{\theta_1}^\pi[(p_t-\hat{p})^2]\\
&\le \frac{4\beta_{\min}^2\Delta^2}{l^2R^2}\big(\sum_{t=1}^{T}\mathbb{E}_{\theta_1}^\pi[(p_t-\psi(\theta_1))^2]+T\delta^2\big).
\end{aligned}\tag{D.21}$$

Since $R^{\pi}_{\theta_1}(T) = (-\beta_1)\sum_{t=1}^{T}\mathbb{E}^{\pi}_{\theta_1}[(p_t - \psi(\theta_1))^2]$, it follows that

$$R^{\pi}_{\theta_1}(T) \geq |\beta_{\max}|\Big(\frac{l^2R^2}{4\beta^2_{\min}\Delta^2}KL(P^{\pi}_1, P^{\pi}_2) - T\delta^2\Big). \tag{D.22}$$

We next establish a lower bound on $KL(P^{\pi}_1, P^{\pi}_2)$ and choose a suitable $\Delta$ such that the RHS of (D.22) can be further lower bounded by $\Omega(\frac{\sqrt{T}}{(\log T)^{\lambda_0}})$. Before proceeding, we define two disjoint intervals $I_1 = [\hat{p} + \delta - \frac{1}{4}\Delta, \hat{p} + \delta + \frac{1}{4}\Delta]$, and $I_2 = [\hat{p} + \delta + \frac{3}{4}\Delta, \hat{p} + \delta + \frac{5}{4}\Delta]$. For each $t \geq 1$, let $X_t$ be the following Bernoulli random variable: $X_t = 1$ if $p_t \in I_1$ and $X_t = 0$ otherwise. Then we have

$$\begin{aligned}
R^{\pi}_{\theta_1}(T) + R^{\pi}_{\theta_2}(T) &\geq |\beta_{\max}|\sum_{t=1}^{T}\mathbb{E}^{\pi}_{\theta_1}[(p_t - \psi(\theta_1))^2] + |\beta_{\max}|\sum_{t=1}^{T}\mathbb{E}^{\pi}_{\theta_2}[(p_t - \psi(\theta_2))^2] \\
&\geq \frac{1}{16}|\beta_{\max}|\Delta^2\sum_{t=1}^{T}\big(P^{\pi}_1(p_t \notin I_1) + P^{\pi}_2(p_t \notin I_2)\big) \\
&\geq \frac{1}{16}|\beta_{\max}|\Delta^2\sum_{t=1}^{T}\big(P^{\pi}_1(X_t = 0) + P^{\pi}_2(X_t = 1)\big) \\
&\geq \frac{1}{32}|\beta_{\max}| \cdot e^{-KL(P^{\pi}_1, P^{\pi}_2)} \cdot T\Delta^2,
\end{aligned} \tag{D.23}$$

where the third inequality holds since $I_1$ and $I_2$ are disjoint, and the last inequality follows from the Bretagnolle-Huber inequality (Theorem 2.2 in Tsybakov (2009)). Since $\pi \in \Pi^{\circ}$,

$$R^{\pi}_{\theta_1}(T) + R^{\pi}_{\theta_2}(T) \leq 2K_0\sqrt{T}(\log T)^{\lambda_0},$$

which together with inequality (D.23) implies

$$KL(P^{\pi}_1, P^{\pi}_2) \geq \log(\sqrt{T}\Delta^2) + \log\big(\frac{|\beta_{\max}|}{64K_0}\big) - \lambda_0 \log\log T.$$

Thus, combining the above inequality with (D.22) and letting $\Delta^2 = \frac{64K_0e(\log T)^{\lambda_0}}{|\beta_{\max}|\sqrt{T}}$, we have

$$\begin{aligned}
R^{\pi}_{\theta_1}(T) &\geq |\beta_{\max}|\left(\frac{l^2R^2}{4\beta^2_{\min}\Delta^2}\Big(\log(\sqrt{T}\Delta^2) + \log\big(\frac{|\beta_{\max}|}{64K_0}\big) - \lambda_0\log\log T\Big) - T\delta^2\right) \\
&= |\beta_{\max}|\Big(\frac{l^2R^2|\beta_{\max}|}{256\beta^2_{\min}K_0e} \cdot \frac{\sqrt{T}}{(\log T)^{\lambda_0}} - T\delta^2\Big) \\
&\geq \frac{l^2R^2\beta^2_{\max}}{512\beta^2_{\min}K_0e} \cdot \frac{\sqrt{T}}{(\log T)^{\lambda_0}},
\end{aligned}$$

where the second inequality follows from the choice of $\Delta$ and $\delta \leq \frac{lR}{16|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{2K_0e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$. Thus, $R^{\pi}_{\theta_1}(T) = \Omega(\frac{\sqrt{T}}{(\log T)^{\lambda_0}})$.

Combining Step 1 and Step 2, we conclude that for any admissible policy $\pi \in \Pi^{\circ}$, there exists

$\theta \in \Theta^\dagger$ satisfying $|\psi(\theta) - \hat{p}| \in [\frac{1}{2}\delta, \frac{3}{2}\delta]$, such that

$$R_\theta^\pi(T) = \begin{cases} \Omega\big((\sqrt{T} \wedge \frac{T}{(n\wedge T)\delta^2}) \vee \log T\big), & \text{if } \delta > \frac{lR}{16|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{2K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}; \\ \Omega\big((T\delta^2) \vee \frac{\sqrt{T}}{(\log T)^{\lambda_0}}\big), & \text{if } \delta \le \frac{lR}{16|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{2K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}, \end{cases}$$

which completes the proof of Theorem 5.2. □

## D.2 Proofs of Statements in Section 5.4

### D.2.1 Proof of Theorem 5.3

To prove Theorem 5.3, we first show that O3FU algorithm (after a natural modification to the multiple-historical-price setting) achieves the regret upper bound $\mathcal{O}(\sqrt{T}\log T)$ and $\mathcal{O}\big(\frac{T(\log T)^2}{n\sigma^2+(n\wedge T)\delta^2}+1\big)$ in the following Step 1 and Step 2 respectively. Then in Step 3, we use the results in Steps 1-2 to show that M-O3FU algorithm achieves the desired upper bound.

**Step 1.** In this step, we prove the regret upper bound $\mathcal{O}(\sqrt{T}\log T)$ for O3FU algorithm. Lemma D.2 and inequalities (D.1) and (D.2) continue to hold by replacing each $V_{t-1,n}$ with $\lambda I + \sum_{i=1}^n [1\ \hat{p}_i]^\top [1\ \hat{p}_i] + \sum_{s=1}^{t-1} [1\ p_s]^\top [1\ p_s]$. To apply Lemma D.1 to the RHS of the inequality (D.2), we just let $d = 2$, $L = \sqrt{1+u^2}$, $\lambda = 1 + u^2$,

$$X_t = \begin{bmatrix} 1 \\ p_t \end{bmatrix}, \quad V = \lambda I + \sum_{i=1}^n \begin{bmatrix} 1 & \hat{p}_i \\ \hat{p}_i & \hat{p}_i^2 \end{bmatrix}, \quad V_t = V + \sum_{s=1}^t \begin{bmatrix} 1 & p_s \\ p_s & p_s^2 \end{bmatrix}.$$

Then we get

$$\begin{aligned} \sum_{t=1}^T ||x_t||^2_{V_{t-1,n}^{-1}} &\le 2\Big(2\log\frac{(2\lambda + \sum_{i=1}^n(1+\hat{p}_i^2)) + T(1+u^2)}{2} - \log\big(\lambda(\lambda + \sum_{i=1}^n(1+\hat{p}_i^2))\big)\Big) \\ &\le 2\log\Big(\frac{(1+u^2)(2+n+T)^2}{4(1+l^2)(1+n)}\Big). \end{aligned} \tag{D.24}$$

The remaining proof remains the same as Theorem 5.1, and is therefore omitted.

**Step 2.** In this step, we prove the regret upper bound $\mathcal{O}\big(\frac{T(\log T)^2}{n\sigma^2+(n\wedge T)\delta^2}+1\big)$ for O3FU algorithm. Note that it suffices to consider the case when $n\sigma^2 + (n\wedge T)\delta^2 = \Omega(\sqrt{T}\log T)$, since otherwise, $\mathcal{O}((\sqrt{T}\log T) \wedge \frac{T(\log T)^2}{n\sigma^2+(n\wedge T)\delta^2}) = \mathcal{O}(\sqrt{T}\log T)$, which is already proven in Step 1. Under the assumption $n\sigma^2 + (n\wedge T)\delta^2 = \Omega(\sqrt{T}\log T)$, we consider the following two cases: (1) $n\sigma^2 \lesssim (n\wedge T)\delta^2$; (2) $n\sigma^2 \gtrsim (n\wedge T)\delta^2$.

**Case 1.** $n\sigma^2 \lesssim (n\wedge T)\delta^2$. In this case, the following three inequalities hold: (i) $n\delta^2 \gtrsim \sqrt{T}\log T$; (ii) $\sigma \lesssim \delta$; and (iii) $\delta \gtrsim T^{-1/4}(\log T)^{\frac{1}{2}}$. The reason is as follows. Suppose (i) does not hold, we have $n\sigma^2 + (n\wedge T)\delta^2 \lesssim \sqrt{T}\log T$, leading to contradiction with $n\sigma^2 + (n\wedge T)\delta^2 = \Omega(\sqrt{T}\log T)$. Suppose (ii) does not hold, then we have $n\sigma^2 \gtrsim n\delta^2 \gtrsim (n\wedge T)\delta^2$, leading to contradiction with the assumption of Case 1. Finally, suppose (iii) does not hold, then we have $(n\wedge T)\delta^2 \lesssim \sqrt{T}\log T$, leading to contradiction with $n\sigma^2 + (n\wedge T)\delta^2 = \Omega(\sqrt{T}\log T)$. Thus, when Case 1 happens, the conditions of Lemma 5.2, i.e., $\sigma \lesssim \delta$ and $\delta \gtrsim \max\{T^{\frac{1}{4}}w_T n^{-\frac{1}{2}}, T^{-\frac{1}{4}}\}$, are satisfied. By applying

Lemma 5.2, we have

$$\begin{aligned}\sum_{t=2}^{T}\mathbb{E}[||\theta^\star-\tilde{\theta}_t||^2] &= \sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{\forall 2\le s\le t,\theta^\star\in\mathcal{C}_s\}}\Big]+\sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{\exists 2\le s\le t,\theta^\star\notin\mathcal{C}_s\}}\Big]\\ &\le \sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{U_{t,4}\}}\Big]+\sum_{t=2}^{T}\big((\alpha_{\max}-\alpha_{\min})^2+(\beta_{\max}-\beta_{\min})^2\big)\frac{1}{T^2}\\ &\le C_3\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2}+((\alpha_{\max}-\alpha_{\min})^2+(\beta_{\max}-\beta_{\min})^2)\frac{1}{T},\end{aligned}$$

where the first inequality follows from the proof of Lemma 5.2 and the concentration ineqality in Lemma D.2 with $\epsilon=\frac{1}{T^2}\wedge\frac{1}{n\sigma^2}\le\frac{1}{T^2}$. When $n\ge T$, we have

$$\begin{aligned}\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2} &\le w_T^2\sum_{t=1}^{T-1}\frac{1}{t\delta^2+n\sigma^2}\\ &= \mathcal{O}\big(\frac{\log T\cdot\log(T\delta^2+n\sigma^2)}{\delta^2}\big)\\ &= \mathcal{O}\Big(\frac{T(\log T)^2}{(n\wedge T)\delta^2+n\sigma^2}\Big),\end{aligned}$$

where the second identity follows from $n\sigma^2\lesssim T\delta^2$. When $n<T$, we have

$$\begin{aligned}\sum_{t=2}^{T}\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2} &= \sum_{t=1}^{n}\frac{w_t^2}{t\delta^2+n\sigma^2}+\sum_{t=n+1}^{T-1}\frac{w_t^2}{n\delta^2+n\sigma^2}\\ &= \mathcal{O}\Big(\frac{\log T\cdot\log(n\delta^2+n\sigma^2)}{\delta^2}\Big)+\mathcal{O}\Big(\frac{T\log T}{n\delta^2+n\sigma^2}\Big)\\ &= \mathcal{O}\Big(\frac{T(\log T)^2}{(n\wedge T)\delta^2+n\sigma^2}\Big).\end{aligned}$$

**Case 2.** $n\sigma^2\gtrsim(n\wedge T)\delta^2$. In this case, to prove the upper bound $\mathcal{O}\big(\frac{T(\log T)^2}{(n\wedge T)\delta^2+n\sigma^2}\big)$, we first establish the following lemma, whose proof is deferred to Appendix D.2.3.

**Lemma D.4.** *Suppose $\theta^\star\in\mathcal{C}_t$ for each $t\in[T-1]$, then for each $2\le t\le T$,*

$$||\theta^\star-\tilde{\theta}_t||^2\le 2\big((4u+1)^2+1\big)\frac{w_{t-1}^2}{n\sigma^2}.$$

Based on the above Lemma D.4, we have

$$\begin{aligned}\sum_{t=2}^{T}\mathbb{E}[||\theta^\star-\tilde{\theta}_t||^2] &= \sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{\forall s\in[t-1],\theta^\star\in\mathcal{C}_s\}}\Big]+\sum_{t=2}^{T}\mathbb{E}\Big[||\theta^\star-\tilde{\theta}_t||^2\cdot 1_{\{\exists s\in[t-1],\theta^\star\notin\mathcal{C}_s\}}\Big]\\ &\le 2((4u+1)^2+1)\sum_{t=2}^{T}\frac{w_{t-1}^2}{n\sigma^2}+\big((\alpha_{\max}-\alpha_{\min})^2+(\beta_{\max}-\beta_{\min})^2\big)\sum_{t=2}^{T}\frac{1}{T^2}\wedge\frac{1}{n\sigma^2}\\ &= \mathcal{O}\Big(\frac{T\log T}{n\sigma^2}\Big)\\ &= \mathcal{O}\Big(\frac{T\log T}{(n\wedge T)\delta^2+n\sigma^2}\Big),\end{aligned}$$

where the inequality follows from Lemma D.2 with $\epsilon = \frac{1}{T^2} \wedge \frac{1}{n\sigma^2}$ and Lemma D.4, and in the last identity, we utilize $n\sigma^2 \gtrsim (n \wedge T)\delta^2$.

**Step 3.** In this step, we use the results in Step 1 and Step 2 to show that M-O3FU algorithm achieves the regret upper bound $\mathcal{O}\big(T\delta^2 + 1\big)$ in the corner case, i.e., when $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$ holds, and $\mathcal{O}\big((\sqrt{T}\log T) \wedge \frac{T(\log T)^2}{(n\wedge T)\delta^2 + n\sigma^2} + 1\big)$ in the regular case, i.e., when $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$ does not hold.

Recall that $\hat{\theta}_0$ is the least-square estimator from offline regression, and it follows from Lemma D.2 that with probability $1-\epsilon$, $||\theta^\star - \hat{\theta}_0||^2_{V_{0,n}} \le w_0^2$ holds, where $w_0 = R\sqrt{2\log\frac{n+1}{\epsilon}} + \sqrt{(1+u^2)(\alpha^2_{\max} + \beta^2_{\min})}$. Since $\lambda_{\min}(V_{0,n}) \ge \frac{2}{(1+2u-l)^2} n\sigma^2$ from Lemma 2 in Keskin and Zeevi (2014), it can be verified that when $\theta^\star \in \mathcal{C}_0$, there exists some constant $L_0 > 0$, such that the length of interval $\{\psi(\theta) : \theta \in \mathcal{C}_0\}$ is $\frac{L_0}{2\sqrt{n\sigma^2}}$. In other words, $\mathbb{P}(\max_{\theta_1,\theta_2\in\mathcal{C}_0} |\psi(\theta_1) - \psi(\theta_2)| \le \frac{L_0}{2\sqrt{n\sigma^2}}) \ge \mathbb{P}(\theta^\star \in \mathcal{C}_0) \ge 1-\epsilon$. Let $\mathcal{P}_0 = \{\psi(\theta) : \theta \in \mathcal{C}_0\}$, and $A$ be the event $\big\{\min_{\theta\in\mathcal{C}_0} |\psi(\theta) - \bar{p}_{1:n}| \le \frac{KL_0}{2\sqrt{n\sigma^2}}\big\}$ for some pre-determined constant $K > 1$.

**Corner case: $\delta^2 \le \frac{K^2L_0^2}{4n\sigma^2}$ and $n\sigma^2 \ge \sqrt{T}$.** In this case, if $\theta^\star \in \mathcal{C}_0$, we have

$$\min_{\theta\in\mathcal{C}_0} |\psi(\theta) - \bar{p}_{1:n}| \le |\psi(\theta^\star) - \bar{p}_{1:n}| \le \frac{KL_0}{2\sqrt{n\sigma^2}},$$

and therefore, $\mathbb{P}(A) \ge \mathbb{P}(\theta^\star \in \mathcal{C}_0) \ge 1-\epsilon$, and when $A$ holds, M-O3FU algorithm will use the price $\bar{p}_{1:n}$ for any $1 \le t \le T$ due to $n\sigma^2 \ge \sqrt{T}$. Thus,

$$\begin{aligned} R^\pi_{\theta^\star}(T) &= \mathbb{P}(A)\cdot\sum_{t=1}^T \mathbb{E}\Big[r^\star(\theta^\star) - r(p_t;\theta^\star)\Big|A\Big] + \mathbb{P}(A^\complement)\cdot\sum_{t=1}^T \mathbb{E}\Big[r^\star(\theta^\star) - r(p_t;\theta^\star)\Big|A^\complement\Big] \\ &\lesssim T\delta^2 + \epsilon\sqrt{T}\log T \\ &\lesssim T\delta^2 + 1, \end{aligned}$$

where the first inequality holds since when $A$ does not hold, M-O3FU algorithm directly applies O3FU algorithm, and incurs the regret $\mathcal{O}(\sqrt{T}\log T)$ from the result in Step 1, the second inequality holds since $\epsilon\sqrt{T}\log T = (\frac{1}{T^2} \wedge \frac{1}{n\sigma^2})\sqrt{T}\log T \lesssim 1$.

**Regular case 1: $\delta^2 \le \frac{K^2L_0^2}{4n\sigma^2}$ and $n\sigma^2 < \sqrt{T}$.** In this case, since $n\sigma^2 < \sqrt{T}$, M-O3FU algorithm runs O3FU algorithm from the beginning, and the regret is bounded by $\mathcal{O}\big((\sqrt{T}\log T) \wedge (\frac{T(\log T)^2}{(n\wedge T)\delta^2 + n\sigma^2} + 1)\big)$ from the results in Steps 1 and 2.

**Regular case 2: $\frac{K^2L_0^2}{4n\sigma^2} \le \delta^2 \le \frac{K^2L_0^2}{n\sigma^2}$.** In this case, the condition $\min_{\theta\in\mathcal{C}_0} |\psi(\theta) - \bar{p}_{1:n}| \le \frac{KL_0}{2\sqrt{n\sigma^2}}$ can either hold or not. If the condition holds and $n\sigma^2 \ge \sqrt{T}$, the regret is $\mathcal{O}(T\delta^2)$. Since in this case, $T\delta^2 \lesssim \frac{T}{n\sigma^2} \lesssim \sqrt{T}$ and $n\sigma^2 \gtrsim \sqrt{T} \gtrsim T\delta^2 \gtrsim (n\wedge T)\delta^2$, we have $\mathcal{O}(T\delta^2) = \mathcal{O}((\sqrt{T}\log T) \wedge \frac{T(\log T)^2}{(n\wedge T)\delta^2 + n\sigma^2} + 1)$. If the condition does not hold, the regret is still bounded by $\mathcal{O}((\sqrt{T}\log T) \wedge \frac{T(\log T)^2}{(n\wedge T)\delta^2 + n\sigma^2} + 1)$.

**Regular case 3: $\delta^2 > \frac{K^2L_0^2}{n\sigma^2}$.** In this case, when $\theta \in \mathcal{C}_0$, we have

$$\min_{\theta\in\mathcal{C}_0} |\psi(\theta) - \bar{p}_{1:n}| \ge |\bar{p}_{1:n} - \psi(\theta^\star)| - |\mathrm{Proj}_{\mathcal{P}_0}(\bar{p}_{1:n}) - \psi(\theta^\star)| \ge \frac{KL_0}{\sqrt{n\sigma^2}} - \frac{L_0}{2\sqrt{n\sigma^2}} > \frac{KL_0}{2\sqrt{n\sigma^2}},$$

where the first inequality follows from the triangle inequality ($\mathrm{Proj}_{\mathcal{P}_0}(\bar{p}_{1:n})$ denotes the projection of $\bar{p}_{1:n}$ to set $\mathcal{P}_0$), the second inequality holds since the length of $\mathcal{P}_0$ is $\frac{L_0}{2\sqrt{n\sigma^2}}$ and $\theta^\star \in \mathcal{C}_0$, and the last inequality follows from $K > 1$. In this case, $\theta^\star \in \mathcal{C}_0$ implies $A^\complement$. Therefore, with probability $1-\epsilon$,

$\mathbb{P}(A^{\complement}) \geq 1-\epsilon$. Thus, if $n\sigma^2 \geq \sqrt{T}$, the regret is upper bounded as follows:

$$\begin{aligned} R^{\pi}_{\theta^\star}(T) &= \mathbb{P}(A)\cdot\sum_{t=1}^{T}\mathbb{E}\Big[r^\star(\theta^\star)-r(p_t;\theta^\star)\Big|A\Big] + \mathbb{P}(A^{\complement})\cdot\sum_{t=1}^{T}\mathbb{E}\Big[r^\star(\theta^\star)-r(p_t;\theta^\star)\Big|A^{\complement}\Big] \\ &\lesssim \frac{1}{T^2}\cdot T\delta^2 + (\sqrt{T}\log T)\wedge\frac{T(\log T)^2}{(n\wedge T)\delta^2+n\sigma^2} \\ &\lesssim 1 + (\sqrt{T}\log T)\wedge\frac{T(\log T)^2}{(n\wedge T)\delta^2+n\sigma^2}, \end{aligned}$$

where the first inequality holds since $\epsilon = \frac{1}{T^2}\wedge\frac{1}{n\sigma^2}\leq\frac{1}{T^2}$. If $n\sigma^2<\sqrt{T}$, M-O3FU algorithm runs O3FU algorithm from the beginning, and the regret is bounded by $\mathcal{O}((\sqrt{T}\log T)\wedge\frac{T(\log T)^2}{(n\wedge T)\delta^2+n\sigma^2}+1)$. □

### D.2.2 Proof of Lemma 5.2

When $t=1$, since $p_1 = l\cdot\mathbb{I}\{\bar{p}_{1:n}>\frac{u+l}{2}\}+u\cdot\mathbb{I}\{\bar{p}_{1:n}\leq\frac{u+l}{2}\}$, then $|p_1-\bar{p}_{1:n}|\geq\frac{u-l}{2}\geq\frac{1}{2}\delta$. Thus, when $t=1$, $U_{t,3}$ holds.

We next prove the following result: under the conditions of Lemma 5.2, suppose for each $1\leq s\leq t-1$ (for a fixed $2\leq t\leq T$), the event $U_{s,3}$ holds, then $U_{t,3}$ and $U_{t,4}$ also hold. Let $\Delta\alpha_t=\tilde{\alpha}_t-\alpha^\star$, $\Delta\beta_t=\tilde{\beta}_t-\beta^\star$, and $\gamma_t=\frac{\Delta\alpha_t}{\Delta\beta_t}$ (when $\Delta\beta_t\neq 0$). Note that the following generalized version of the inequality (D.4) holds:

$$\lambda\big((\Delta\alpha_t)^2+(\Delta\beta_t)^2\big)+\sum_{i=1}^{n}\big(\Delta\alpha_t+\Delta\beta_t\hat{p}_i\big)^2+\sum_{s=1}^{t-1}\big(\Delta\alpha_t+\Delta\beta_t p_s\big)^2\leq 2w_{t-1}^2. \tag{D.25}$$

Similar to the proof of Lemma 5.1, we also divide the proof into three cases.

**Case 1**: $\Delta\beta_t=0$. In this case, (D.25) becomes $(\Delta\alpha_t)^2(\lambda+n+t-1)\leq 2w_{t-1}^2$, and

$$||\theta^\star-\tilde{\theta}_t||^2=(\Delta\alpha_t)^2+(\Delta\beta_t)^2=(\Delta\alpha_t)^2\leq\frac{2w_{t-1}^2}{n+t-1}. \tag{D.26}$$

Therefore, combining $\sigma\leq u-l$, $\delta\leq u-l$, and (D.26), we obtain

$$||\theta^\star-\tilde{\theta}_t||^2\leq\frac{4(u-l)^2w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2}.$$

In addition, (D.26) also implies

$$|\bar{p}_{1:n}-p_t|\geq|\bar{p}_{1:n}-\psi(\theta^\star)|-|p_t-\psi(\theta^\star)|\geq|\bar{p}_{1:n}-\psi(\theta^\star)|-\frac{\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{2\beta_{\max}^2}\cdot\frac{\sqrt{2}w_{t-1}}{\sqrt{n+t-1}}\geq\frac{1}{2}\delta,$$

where the second inequality follows from (D.26) and Lipschitz continuity of the function $\psi(\cdot)$, and the last inequality holds since from the assumption of $\delta\geq\frac{\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}}{\beta_{\max}^2}\cdot\frac{T^{1/4}w_T}{n^{1/2}}$, we have

$$\frac{w_{t-1}}{\sqrt{n+t-1}}\leq\frac{w_T}{\sqrt{n}}\leq\frac{\beta_{\max}^2}{\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}}\delta. \tag{D.27}$$

**Case 2**: $\Delta\beta_t \neq 0$, $|\gamma_t| \geq 4u+1$. In this case, we have

$$||\theta^\star - \tilde{\theta}_t||^2 \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{\lambda(1+\gamma_t^2)+\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2} \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{n(\gamma_t+\bar{p}_{1:n})^2} \leq \frac{4w_{t-1}^2}{n}, \quad \text{(D.28)}$$

where the second inequality holds since $\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2 \geq n(\gamma_t+\bar{p}_{1:n})^2$, and the last inequality follows from $1+\gamma_t^2 \leq 2(\gamma_t+\bar{p}_{1:n})^2$, which is easily verified by noting $(\gamma_t+2\bar{p}_{1:n})^2 \geq (|\gamma_t|-2\bar{p}_{1:n})^2 \geq (2\bar{p}_{1:n}+1)^2 \geq 2\bar{p}_{1:n}^2+1$. Then, (D.28) implies

$$||\theta^\star - \tilde{\theta}_t||^2 \leq \frac{8(u-l)^2 w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2},$$

and in addition,

$$|\bar{p}_{1:n}-p_t| \geq |\bar{p}_{1:n}-\psi(\theta^\star)| - |p_t-\psi(\theta^\star)| \geq |\bar{p}_{1:n}-\psi(\theta^\star)| - \frac{\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{2\beta_{\max}^2}\frac{2w_{t-1}}{\sqrt{n}} \geq (1-\frac{\sqrt{2}}{2})\delta,$$

where the last inequality follows from (D.27).

**Case 3**: $\Delta\beta_t \neq 0$, $|\gamma_t| < 4u+1$. Recall the definitions of $C_0$ and $C_1$:

$$C_0 = \frac{l|\beta_{\max}|}{u|\beta_{\min}|}, \quad C_1 = \frac{4(C_0+1)^2}{C_0^2}(1+(4u+1)^2).$$

**Subcase 3.1**: $1+\gamma_t^2 \leq C_1\frac{(\gamma_t+\bar{p}_{1:n})^2}{\delta^2}$. In this subcase, we have

$$||\theta^\star - \tilde{\theta}_t||^2 \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{n(\gamma_t+\bar{p}_{1:n})^2} \leq \frac{2C_1 w_{t-1}^2}{n\delta^2}.$$

From the assumption of $\sigma \leq \delta$, we have

$$||\theta^\star - \tilde{\theta}_t||^2 \leq \frac{4C_1 w_{t-1}^2}{n\delta^2+n\sigma^2} \leq \frac{4C_1 w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2},$$

and in addition,

$$\begin{aligned}
|p_t-\bar{p}_{1:n}| &\geq |\psi(\theta^\star)-\bar{p}_{1:n}| - |p_t-\psi(\theta^\star)| \\
&\geq |\psi(\theta^\star)-\bar{p}_{1:n}| - \frac{\sqrt{\alpha_{\max}^2+\beta_{\max}^2}}{2\beta_{\max}^2}\frac{\sqrt{2C_1}w_{t-1}}{\sqrt{n}|\psi(\theta^\star)-\bar{p}_{1:n}|} \\
&\geq |\psi(\theta^\star)-\bar{p}_{1:n}| - \frac{\sqrt{C_1}}{2T^{1/4}} \\
&\geq \frac{1}{2}\delta,
\end{aligned}$$

where the third inequality follows from (D.27), and in the last inequality, we utilize the assumption of $\delta \geq \sqrt{C_1}T^{-1/4}$.

**Subcase 3.2**: $1+\gamma_t^2 > C_1\frac{(\gamma_t+\bar{p}_{1:n})^2}{\delta^2}$. In this subcase, we have

$$||\theta^\star-\tilde{\theta}_t||^2 \le \frac{2w_{t-1}^2(\gamma_t^2+1)}{\lambda(\gamma_t^2+1)+\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2} \le \frac{2w_{t-1}^2((4u^2+1)^2+1)}{\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2}.$$

To proceed, we establish the following inequality:

$$\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2 \ge n\sigma^2+(n\wedge(t-1))\min\left\{(1-\frac{\sqrt{2}}{2})^2,\frac{C_0^2}{4}\right\}\cdot(\psi(\theta^\star)-\bar{p}_{1:n})^2. \quad \text{(D.29)}$$

Note that $\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2$ is convex in $\gamma_t$ and is minimized at $\gamma_t=-\frac{\sum_{i=1}^n\hat{p}_i+\sum_{s=1}^{t-1}p_s}{n+t-1}$. We have

$$\begin{aligned}\sum_{i=1}^n(\gamma_t+\hat{p}_i)^2+\sum_{s=1}^{t-1}(\gamma_t+p_s)^2 &\ge \sum_{i=1}^n(\hat{p}_i-\frac{\sum_{i=1}^n\hat{p}_i+\sum_{s=1}^{t-1}p_s}{n+t-1})^2+\sum_{s=1}^{t-1}(p_s-\frac{\sum_{i=1}^n\hat{p}_i+\sum_{s=1}^{t-1}p_s}{n+t-1})^2\\ &= \mathrm{Var}((\hat{p}_1,\ldots,\hat{p}_n),(p_1,\ldots,p_{t-1})),\end{aligned}$$

where $((\hat{p}_1,\ldots,\hat{p}_n),(p_1,\ldots,p_{t-1}))\in\mathbb{R}^{(n+t-1)\times 1}$. Define

$$f(p_1,\ldots,p_{t-1}) := \mathrm{Var}((\hat{p}_1,\ldots,\hat{p}_n),(p_1,\ldots,p_{t-1})).$$

Then

$$\begin{aligned}f(p_1,\ldots,p_{t-1}) &= \|((\hat{p}_1,\ldots,\hat{p}_n),(p_1,\ldots,p_{t-1}))\|_2^2-\frac{\left[1_{(n+t-1)\times 1}^\top((\hat{p}_1,\ldots,\hat{p}_n),(p_1,\ldots,p_{t-1}))\right]^2}{(n+t-1)}\\ &= \|(\hat{p}_1,\ldots,\hat{p}_n)\|_2^2+\|(p_1,\ldots,p_{t-1})\|_2^2-\frac{\left[1_{n\times 1}^\top(\hat{p}_1,\ldots,\hat{p}_n)+1_{(t-1)\times 1}^\top(p_1,\ldots,p_{t-1})\right]^2}{n+t-1},\end{aligned}$$

thus

$$\frac{\partial f(p_1,\ldots,p_{t-1})}{\partial(p_1,\ldots,p_{t-1})} = 2(p_1,\ldots,p_{t-1})-2\frac{\left[1_n^\top(\hat{p}_1,\ldots,\hat{p}_n)+1_{t-1}^\top(p_1,\ldots,p_{t-1})\right]}{n+t-1}1_{(t-1)\times 1},$$

$$\frac{\partial^2 f(p_1,\ldots,p_{t-1})}{\partial(p_1,\ldots,p_{t-1})^2} = 2\left(I_{(t-1)}-\frac{1_{(t-1)\times 1}1_{(t-1)\times 1}^\top}{n+t-1}\right)\succeq 0.$$

Therefore, we know that $f(p_1,\ldots,p_{t-1})$ is convex in $(p_1,\ldots,p_{t-1})$ and is minimized at

$$(p_1,\ldots,p_{t-1}) = \bar{p}_{1:n}1_{(t-1)\times 1}.$$

By the Taylor series of $f(p_1,\ldots,p_{t-1})$ at point $\bar{p}_{1:n}1_{(t-1)\times 1}$, we have

$$
\begin{aligned}
&f(p_1,\ldots,p_{t-1}) - f\left(\bar{p}_{1:n}1_{(t-1)\times 1}\right)\\
=&\left((p_1,\ldots,p_{t-1}) - \bar{p}_{1:n}1_{(t-1)\times 1}\right)^\top \left[I_{(t-1)} - \frac{1_{(t-1)\times 1}1_{(t-1)\times 1}^\top}{n+t-1}\right]\left((p_1,\ldots,p_{t-1}) - \bar{p}_{1:n}1_{(t-1)\times 1}\right)\\
=&\|(p_1,\ldots,p_{t-1}) - \bar{p}_{1:n}1_{(t-1)\times 1}\|_2^2 - \frac{\left(\sum_{s=1}^{t-1}(p_s-\bar{p}_{1:n})\right)^2}{n+t-1}\\
=&\sum_{s=1}^{t-1}(p_s-\bar{p}_{1:n})^2 - \frac{\left(\sum_{s=1}^{t-1}(p_s-\bar{p}_{1:n})\right)^2}{n+t-1}\\
\geq&\frac{n}{n+t-1}\sum_{s=1}^{t-1}(p_s-\bar{p}_{1:n})^2\\
\geq&\frac{n(t-1)}{(n+t-1)}\cdot\min\left\{(1-\frac{\sqrt{2}}{2})^2,\frac{C_0^2}{4}\right\}\cdot\delta^2,
\end{aligned}
$$

where the last inequality is by the induction assumption that $U_{s,2} = \left\{|p_s-\bar{p}_{1:n}| \geq \min\left\{1-\frac{\sqrt{2}}{2},\frac{C_0}{2}\right\}\cdot\delta\right\}$ holds for $s=1,\ldots,t-1$. Using also the fact that $f\left(\bar{p}_{1:n}1_{(t-1)\times 1}\right) = \mathrm{Var}(\hat{p}_1,\ldots,\hat{p}_n) = n\sigma^2$, we have

$$
\sum_{i=1}^{n}(\gamma_t+\hat{p}_i)^2 + \sum_{s=1}^{t-1}(\gamma_t+p_s)^2 \geq f(p_1,\ldots,p_{t-1}) \geq n\sigma^2 + (n\wedge(t-1))\min\left\{(1-\frac{\sqrt{2}}{2})^2,\frac{C_0^2}{4}\right\}\cdot\delta^2.
$$

Therefore, we have proven (D.29) and can conclude that

$$
||\tilde{\theta}_t-\theta^\star||^2 \leq 2\max\left\{2(\sqrt{2}+1)^2,\frac{4}{C_0^2}\right\}\cdot((4u+1)^2+1)\cdot\frac{w_{t-1}^2}{(n\wedge(t-1))\delta^2+n\sigma^2}.
$$

Now, it suffices to bound the term $|p_t-\bar{p}_{1:n}|$. We still have (D.8) (which we proved in the single-historical-price setting), i.e., the following inequality:

$$
|\gamma_t+p_t| \geq C_0|\gamma_t+\psi(\theta^\star)|, \tag{D.30}
$$

thus

$$
\begin{aligned}
|p_t-\bar{p}_{1:n}| &\geq |p_t+\gamma_t| - |\gamma_t+\bar{p}_{1:n}|\\
&\geq C_0|\gamma_t+\psi(\theta^\star)| - |\gamma_t+\bar{p}_{1:n}|\\
&\geq C_0(|\psi(\theta^\star)-\bar{p}_{1:n}| - |\gamma_t+\bar{p}_{1:n}|) - |\gamma_t+\bar{p}_{1:n}|\\
&= C_0|\psi(\theta^\star)-\bar{p}_{1:n}| - (C_0+1)|\gamma_t+\bar{p}_{1:n}|\\
&\geq \left(C_0 - (C_0+1)\frac{\sqrt{1+(4u+1)^2}}{\sqrt{C_1}}\right)|\psi(\theta^\star)-\bar{p}_{1:n}|\\
&\geq \frac{C_0}{2}\delta,
\end{aligned}
$$

where the second inequality follows from (D.30), the fourth inequality follows from the assumption of Subcase 3.2, i.e., $1+\gamma_t^2 > C_1\frac{(\gamma_t+\bar{p}_{1:n})^2}{\delta^2}$ and $|\gamma_t| \leq 4u+1$, and the last inequality follows from the

definition of $C_1$.

Therefore, combining the above three cases, we conclude that

$$|\bar{p}_{1:n} - \psi(\theta^\star)| \geq \min\Big\{1 - \frac{\sqrt{2}}{2}, \frac{C_0}{2}\Big\} \cdot \delta,$$
$$||\theta^\star - \tilde{\theta}_t||^2 \leq \max\Big\{8(u-l)^2, 4C_1, 2\max\big\{2(\sqrt{2}+1)^2, \frac{4}{C_0^2}\big\} \cdot ((4u+1)^2+1)\Big\} \cdot \frac{w_{t-1}^2}{(n \wedge (t-1))\delta^2 + n\sigma^2},$$

i.e., $U_{t,3}$ and $U_{t,4}$ hold, which completes the inductive arguments. □

### D.2.3 Proof of Lemma D.4

Since $\theta^\star \in \mathcal{C}_t$ for each $t \in [T]$, and $\tilde{\theta}_t \in \mathcal{C}_t$ for each $2 \leq t \leq T$, the inequality (D.25) still holds. For each $2 \leq t \leq T$, we bound $||\theta^\star - \tilde{\theta}_t||^2$ by considering the following three cases.

**Case 1**: $\Delta\beta_t = 0$. In this case, (D.25) becomes $(\Delta\alpha_t)^2(\lambda + n + t - 1) \leq 2w_{t-1}^2$, and

$$||\theta^\star - \tilde{\theta}_t||^2 = (\Delta\alpha_t)^2 \leq \frac{2w_{t-1}^2}{n} \leq \frac{2(u-l)^2 w_{t-1}^2}{n\sigma^2},$$

where the second inequality holds since $\sigma \leq u - l$.

**Case 2**: $\Delta\beta_t \neq 0$, $|\gamma_t| \geq 4u + 1$. In this case, we have

$$||\theta^\star - \tilde{\theta}_t||^2 \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{\sum_{i=1}^n(\gamma_t + \hat{p}_i)^2} \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{n(\gamma_t + \bar{p}_{1:n})^2} \leq \frac{4w_{t-1}^2}{n} \leq \frac{4(u-l)^2 w_{t-1}^2}{n\sigma^2},$$

where the second inequality holds since $\sum_{i=1}^n(\gamma_t + \hat{p}_i)^2 \geq n(\gamma_t + \bar{p}_{1:n})^2$. and the third inequality follows from $1 + \gamma_t^2 \leq 2(\gamma_t + \bar{p}_{1:n})^2$.

**Case 3**: $\Delta\beta_t \neq 0$, $|\gamma_t| < 4u + 1$. In this case,

$$||\theta^\star - \tilde{\theta}_t||^2 \leq \frac{2w_{t-1}^2(1+\gamma_t^2)}{\sum_{i=1}^n(\gamma_t + \hat{p}_i)^2} \leq \frac{2((4u+1)^2+1)w_{t-1}^2}{\sum_{i=1}^n(\hat{p}_i - \bar{p}_{1:n})^2} = \frac{2((4u+1)^2+1)w_{t-1}^2}{n\sigma^2},$$

where the second inequality holds since $\sum_{i=1}^n(\hat{p}_i - \bar{p}_{1:n})^2 = \min_{x\in\mathbb{R}}(\hat{p}_i + x)^2 \leq \sum_{i=1}^n(\hat{p}_i + \gamma_t)^2$.

Therefore, combining the above three cases, we obtain $||\theta^\star - \tilde{\theta}_t||^2 \leq 2((4u+1)^2+1) \cdot \frac{w_{t-1}^2}{n\sigma^2}$, which completes the proof. □

### D.2.4 Proof of Theorem 5.4

Similar to the proof of Theorem 5.2, we consider normal random noise with standard deviation $R$, and for simplicity, we assume $\xi = \frac{1}{2}$. The proof is divided into two major steps.

**Step 1.** In the first step, we prove the following result: for any pricing policy $\pi$,

$$\sup_{\theta\in\Theta_0(\delta)} R_\theta^\pi(T, n, \sigma, \delta) = \Omega\Big(\sqrt{T} \wedge \frac{T}{\delta^{-2} + n\sigma^2 + (n \wedge T)\delta^2}\Big), \tag{D.31}$$

where $\Theta_0(\delta) = \big\{\theta \in \Theta^\dagger : \psi(\theta) - \bar{p}_{1:n} \in [\frac{\delta}{2}, \delta]\big\}$. When (i) $\delta > \frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$; or (ii) $\delta \leq \frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$ and $n\sigma^2 > \frac{l^2R^2|\beta_{\max}|}{512\beta_{\min}^2 K_0 e}\frac{\sqrt{T}}{(\log T)^{\lambda_0}}$, (D.31) provides the desired lower

bound in Theorem 5.4.

To prove (D.31), it suffices to show that $\sup_{\theta\in\Theta_0(\delta)} R^\pi_\theta(T,n,\sigma,\delta)$ is lower bounded by $\Omega(\sqrt{T}\wedge \frac{T}{\delta^{-2}+n\sigma^2+n\delta^2})$ and $\Omega(\sqrt{T}\wedge \frac{T}{\delta^{-2}+n\sigma^2+T\delta^2})$. The proofs of these two bounds are similar to (D.13) in the proof of Theorem 5.2, and we only highlight the difference here. For the first lower bound $\Omega(\sqrt{T}\wedge \frac{T}{\delta^{-2}+n\sigma^2+n\delta^2})$, by defining a similar prior distribution $q$ as (D.14) and letting $C(\theta)=(\psi(\theta),1)$, we have

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_q[\mathbb{E}^\pi_\theta[(p_t-\psi(\theta))^2]] &\geq \sum_{t=2}^T \frac{R^2\alpha^2_{\min}/(4\beta^2_{\min})}{R^2\mathcal{I}(q)+\sum_{i=1}^n \mathbb{E}_q[(\hat{p}_i-\psi(\theta))^2]+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}^\pi_\theta[(p_s-\psi(\theta))^2]]}\\ &\geq \sum_{t=2}^T \frac{R^2\alpha^2_{\min}/(4\beta^2_{\min})}{R^2\mathcal{I}(q)+2n\sigma^2+2n\delta^2+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}^\pi_\theta[(p_s-\psi(\theta))^2]]}\\ &\geq \frac{(T-1)R^2\alpha^2_{\min}/(4\beta^2_{\min})}{R^2\mathcal{I}(q)+2n\sigma^2+2n\delta^2+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}^\pi_\theta[(p_s-\psi(\theta))^2]]},\end{aligned}$$

where the second inequality follows from $\sum_{i=1}^n(\hat{p}_i-p^\star_\theta)^2\leq 2\sum_{i=1}^n(\hat{p}_i-\bar{p}_{1:n})^2+2n(\bar{p}_{1:n}-p^\star_\theta)^2\leq 2n\sigma^2+2n\delta^2$. Since $\mathcal{I}(q)=\Theta(\delta^{-2})$, the first lower bound $\Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+n\sigma^2+n\delta^2})$ can be proved. For the second lower bound $\Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+n\sigma^2+T\delta^2})$, letting $C(\theta)=(-\bar{p}_{1:n},1)$ and applying the multivariate van Trees inequality to the prior distribution $q$ defined in (D.14), we have

$$\begin{aligned}\mathbb{E}_q\big[\mathbb{E}^\pi_\theta[(p_t-\psi(\theta))^2]\big] &\geq \frac{(\mathbb{E}_q[C(\theta)^\intercal\frac{\partial\psi}{\partial\theta}])^2}{\mathcal{I}(q)+\mathbb{E}_q[C(\theta)^\intercal\mathcal{I}^\pi_{t-1}(\theta)C(\theta)]}\\ &\geq \frac{R^2(\alpha_{\min}+\beta_{\min}\bar{p}_{1:n})^2/(4\beta^2_{\min})}{R^2\mathcal{I}(q)+\sum_{i=1}^n(\hat{p}_i-\bar{p}_{1:n})^2+\sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}^\pi_\theta[(p_s-\bar{p}_{1:n})^2]]}\\ &\geq \frac{R^2(\alpha_{\min}+\beta_{\min}\bar{p}_{1:n})^2/(4\beta^2_{\min})}{R^2\mathcal{I}(q)+n\sigma^2+2(t-1)\delta^2+2\sum_{s=1}^{t-1}\mathbb{E}_{\tilde{q}}[\mathbb{E}^\pi_\theta[(p_s-p^\star_\theta)^2]]},\end{aligned}$$

where the second inequality follows from $(p_s-\bar{p}_{1:n})^2\leq 2(p_s-p^\star_\theta)^2+2(\bar{p}_{1:n}-p^\star_\theta)^2$. Again noting that $\mathcal{I}(q)=\Theta(\delta^{-2})$, we conclude that the regret is lower bounded by $\Omega(\sqrt{T}\wedge\frac{T}{\delta^{-2}+n\sigma^2+T\delta^2})$.

**Step 2.** In this step, we complete the proof by showing that when $\delta\leq\frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$ and $n\sigma^2\leq\frac{l^2R^2|\beta_{\max}|}{512\beta^2_{\min}K_0e}\frac{\sqrt{T}}{(\log T)^{\lambda_0}}$, for any admissible policy $\pi\in\Pi^\circ$, there exists $\theta\in\Theta^\dagger$ satisfying $|\psi(\theta)-\bar{p}_{1:n}|\in[\frac{1}{2}\delta,\frac{3}{2}\delta]$ such that

$$R^\pi_\theta(T)=\Omega\left(\frac{\sqrt{T}}{(\log T)^{\lambda_0}}\right). \tag{D.32}$$

The proof of the above (D.32) is similar to (D.20) in the proof of Theorem 5.2. For completeness, the details are illustrated as follows. We first define two two vectors of demand parameters $\theta_1=(\alpha_1,\beta_1)$ and $\theta_2=(\alpha_2,\beta_2)$ as follows:

$$-\frac{\alpha_1}{2\beta_1}=\bar{p}_{1:n}+\delta,\quad -\frac{\alpha_2}{2\beta_2}=\bar{p}_{1:n}+\delta+\Delta,\quad \alpha_1-\alpha_2+(\beta_1-\beta_2)\bar{p}_{1:n}=0, \tag{D.33}$$

where $\Delta>0$ is to be determined. We consider $P^\pi_1$, $P^\pi_2$ as the two probability measures induced by

the common policy $\pi$ and two demand parameters $\theta_1$ and $\theta_2$ respectively. That is, for each $i = 1, 2$,

$$P_i^\pi(\hat{D}_1, \dots \hat{D}_n, D_1, \dots, D_T) = \prod_{t=1}^{n} \Big( \frac{1}{R} \phi\big( \frac{\hat{D}_t - (\alpha_i + \beta_i \hat{p}_t)}{R} \big) \Big) \cdot \prod_{t=1}^{T} \Big( \frac{1}{R} \phi\big( \frac{D_t - (\alpha_i + \beta_i p_t)}{R} \big) \Big).$$

It is easily verified that the KL divergence between $P_1^\pi$ and $P_2^\pi$ is

$$\begin{aligned} KL(P_1^\pi, P_2^\pi) &= \frac{1}{2R^2} \Big( \sum_{i=1}^{n} \big( (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)\hat{p}_i \big)^2 + \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi \big[ \big( (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2) p_t \big)^2 \big] \Big) \\ &= \frac{(\beta_1 - \beta_2)^2}{2R^2} \Big( \sum_{i=1}^{n} \big( \hat{p}_i - \bar{p}_{1:n} \big)^2 + \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi \big[ \big( p_t - \bar{p}_{1:n} \big)^2 \big] \Big) \\ &= \frac{2\beta_2^2 \Delta^2}{(\bar{p}_{1:n} + 2\delta)^2 R^2} \Big( n\sigma^2 + \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi \big[ \big( p_t - \bar{p}_{1:n} \big)^2 \big] \Big) \\ &\le \frac{2\beta_{\min}^2 \Delta^2}{l^2 R^2} \Big( n\sigma^2 + 2 \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi \big[ \big( p_t - \psi(\theta_1) \big)^2 \big] + 2T\delta^2 \Big), \end{aligned}$$

where the second identity follows from (D.33) and the third identity holds since $(\beta_1 - \beta_2)^2 = \frac{4\beta_2^2 \Delta^2}{(\bar{p}_{1:n} + 2\delta)^2}$ due to (D.33). Therefore, we have

$$R_{\theta_1}^\pi(T) \ge |\beta_{\max}| \cdot \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi [(p_t - \psi(\theta_1))^2] \ge |\beta_{\max}| \cdot \Big( \frac{l^2 R^2}{4\beta_{\min}^2 \Delta^2} KL(P_1^\pi, P_2^\pi) - \frac{n\sigma^2}{2} - T\delta^2 \Big). \quad \text{(D.34)}$$

On the other hand, we have

$$\begin{aligned} \frac{1}{32} e^{-KL(P_1^\pi, P_2^\pi)} \cdot T\Delta^2 &\le \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi [(p_t - \psi(\theta_1))^2] + \sum_{t=1}^{T} \mathbb{E}_{\theta_2}^\pi [(p_t - \psi(\theta_1))^2] \\ &\le \frac{1}{|\beta_{\max}|} 2K_0 \sqrt{T} (\log T)^{\lambda_0}, \end{aligned} \quad \text{(D.35)}$$

where the first inequality follows from Theorem 2.2 in Tsybakov (2009), the second inequality follows from the assumption on the policy $\pi$. Therefore, (D.35) implies

$$KL(P_1^\pi, P_2^\pi) \ge \log \Big( \frac{\sqrt{T} |\beta_{\max}| \Delta^2}{64 K_0 (\log T)^{\lambda_0}} \Big).$$

Thus, by letting $\Delta^2 = \frac{64 K_0 e (\log T)^{\lambda_0}}{|\beta_{\max}| \sqrt{T}}$, from (D.34), the regret can be lower bounded by

$$\begin{aligned} R_{\theta_1}^\pi(T) &\ge |\beta_{\max}| \cdot \Big( \frac{l^2 R^2}{4\beta_{\min}^2 \Delta^2} \log \Big( \frac{\sqrt{T} \Delta^2}{64 K_0 (\log T)^{\lambda_0}} \Big) - \frac{n\sigma^2}{2} - T\delta^2 \Big) \\ &= |\beta_{\max}| \cdot \Big( \frac{l^2 R^2 |\beta_{\max}|}{256 \beta_{\min}^2 K_0 e} \cdot \frac{\sqrt{T}}{(\log T)^{\lambda_0}} - \frac{n\sigma^2}{2} - T\delta^2 \Big) \\ &\ge \frac{l^2 R^2 \beta_{\max}^2}{512 \beta_{\min}^2 K_0 e} \cdot \frac{\sqrt{T}}{(\log T)^{\lambda_0}}, \end{aligned}$$

where the second inequality follows from the definition of $\Delta$, $\delta \le \frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}$ and $n\sigma^2 \le \frac{l^2R^2|\beta_{\max}|}{512\beta_{\min}^2 K_0 e}\frac{\sqrt{T}}{(\log T)^{\lambda_0}}$. Therefore, $R_{\theta_1}^{\pi}(T) = \Omega(\frac{\sqrt{T}}{(\log T)^{\lambda_0}})$.

Combining Step 1 and Step 2, we conclude that for any admissible policy $\pi \in \Pi^{\circ}$, there exists $\theta \in \Theta^{\dagger}$ satisfying $|\psi(\theta) - \bar{p}_{1:n}| \in [(1-\xi)\delta, (1+\xi)\delta]$, such that

$$R_{\theta}^{\pi}(T) = \begin{cases} \Omega\Big(\sqrt{T} \wedge \frac{T}{\delta^{-2}+(n\wedge T)\delta^2+n\sigma^2}\Big), & \text{if } \delta > \frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0}; \\ \Omega\big(T\delta^2 \wedge \frac{T}{n\sigma^2}\big), & \text{if } \delta \le \frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0} \text{ and } n\sigma^2 > \frac{l^2R^2|\beta_{\max}|}{512\beta_{\min}^2 K_0 e}\frac{\sqrt{T}}{(\log T)^{\lambda_0}}; \\ \Omega(\frac{\sqrt{T}}{(\log T)^{\lambda_0}}), & \text{if } \delta \le \frac{lR}{32|\beta_{\min}|}\sqrt{\frac{|\beta_{\max}|}{K_0 e}}T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}\lambda_0} \text{ and } n\sigma^2 \le \frac{l^2R^2|\beta_{\max}|}{512\beta_{\min}^2 K_0 e}\frac{\sqrt{T}}{(\log T)^{\lambda_0}}, \end{cases}$$

which implies Theorem 5.4. □

## D.3 Proof of Proposition 5.1 in Section 5.6

Suppose $\theta^\star \in \mathcal{C}_0$ and $\frac{\min_{\theta\in\mathcal{C}_0}|\psi(\theta)-\bar{p}_{1:n}|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} > K$, we have the following inequalities for each $t \ge 1$:

$$\begin{aligned} |p_t^{\text{myopic}} - \bar{p}_{1:n}| &\ge |\psi(\theta^\star) - \bar{p}_{1:n}| - |\psi(\theta^\star) - p_t^{\text{myopic}}| \ge |\psi(\theta^\star) - \bar{p}_{1:n}| - \max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)| \\ &\ge |\psi(\theta^\star) - \bar{p}_{1:n}| - \frac{1}{K}\min_{\theta\in\mathcal{C}_0}|\psi(\theta)-\bar{p}_{1:n}| \ge (1-\frac{1}{K})\cdot\delta, \end{aligned} \tag{D.36}$$

where the first inequality follows from the triangle inequality, the second inequality holds since $\theta^\star \in \mathcal{C}_0$, $p_t^{\text{myopic}} = \psi(\theta_{t-1}^{\text{LS}})$, and $\theta_{t-1}^{\text{LS}} \in \mathcal{C}_0$ by its definition, and the last inequality holds since $\theta^\star \in \mathcal{C}_0$. That is to say, events $\{U_{t,3} : t \ge 1\}$ defined in Lemma 5.2 are automatically satisfied if ignoring the constant factor under assumptions $\theta \in \mathcal{C}_0$ and $\frac{\min_{\theta\in\mathcal{C}_0}|\psi(\theta)-\bar{p}_{1:n}|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} > K$.

We next bound the estimation error $||\theta^\star - \theta_t^{\text{LS}}||^2$ for each $t \ge 0$. Suppose $\theta^\star \in \mathcal{C}_t$, i.e., $||\theta^\star - \theta_t^{\text{LS}}||_{V_{t,n}}^2 \le w_{t-1}^2$, and therefore, $||\theta^\star - \theta_t^{\text{LS}}||^2 \le \frac{w_{t-1}^2}{\lambda_{\min}(V_{t,n})}$. Then it suffices to bound the minimum eigenvalue of $V_{t,n}$ from below by $\Omega\big((n\wedge t)\delta^2 + n\sigma^2\big)$. Note that

$$\lambda_{\min}(V_{t,n}) = \min_{(x_1,x_2)\in\mathbb{R}^2: x_1^2+x_2^2=1}\left\{\sum_{i=1}^{n}(x_1+\hat{p}_i x_2)^2 + \sum_{s=1}^{t}(x_1+p_s x_2)^2\right\} + \lambda.$$

Let $(x_1^\star, x_2^\star)$ be the optimal solution to the above optimization problem. Then we consider the following cases: $|x_2^\star| \ge \frac{1}{2(1+2u)}$ and $|x_2^\star| < \frac{1}{2(1+2u)}$.

Case 1: $|x_2^\star| \ge \frac{1}{2(1+2u)}$. In this case, we have

$$\begin{aligned} \lambda_{\min}(V_{t,n}) &= \sum_{i=1}^{n}(x_1^\star + \hat{p}_i x_2^\star)^2 + \sum_{s=1}^{t}(x_1^\star + p_s x_2^\star)^2 + \lambda \\ &= \sum_{i=1}^{n}\big(x_1^\star + \bar{p}_{1:n}x_2^\star\big)^2 + (x_2^\star)^2\sum_{i=1}^{n}(\hat{p}_i - \bar{p}_{1:n})^2 + \sum_{s=1}^{t}(x_1^\star + p_s x_2^\star)^2 + \lambda \\ &\ge (x_2^\star)^2\sum_{s=1}^{n\wedge t}(\bar{p}_{1:n} - p_s)^2 + (x_2^\star)^2 n\sigma^2 + \lambda \\ &\ge \frac{1}{4(1+2u)^2}\Big(\big(1-\frac{1}{K}\big)^2\cdot(n\wedge t)\cdot\delta^2 + n\sigma^2\Big) + \lambda, \end{aligned} \tag{D.37}$$

where the first inequality follows from $a^2+b^2 \geq \frac{1}{2}(a-b)^2$, the second inequality follows from the assumption that $|x_2^\star| \geq \frac{1}{2(1+2u)}$, and the last inequality follows from (D.36).

Case 2: $|x_2^\star| < \frac{1}{2(1+2u)}$. In this case, since $(x_1^\star)^2+(x_2^\star)^2=1$, we must have $(x_1^\star)^2 \geq 1-\frac{1}{4(1+2u)^2}$, and therefore,

$$\begin{aligned}\lambda_{\min}(V_{t,n}) &\geq \sum_{i=1}^{n}\Big((x_1^\star)^2+2x_1^\star x_2^\star \hat{p}_i\Big)+\lambda \geq n\big((x_1^\star)^2-\frac{u}{1+2u}\big)+\lambda \\ &\geq n\big(1-\frac{1}{4(1+2u)^2}-\frac{u}{1+2u}\big)+\lambda \geq \frac{1}{2}n+\lambda \\ &\geq \frac{1}{4(u-l)^2}\Big((n\wedge t)\delta^2+n\sigma^2\Big)+\lambda, \end{aligned} \tag{D.38}$$

where the second inequality follows from $2x_1^\star x_2^\star \hat{p}_i \geq -2u|x_2^\star| \geq -\frac{u}{1+2u}$ due to $|x_1^\star| \leq 1$ and $|x_2^\star| \leq \frac{1}{2(1+2u)}$, the third inequality holds since $\frac{1}{4(1+2u)^2}+\frac{u}{1+2u} \leq \frac{1}{2(1+2u)}+\frac{u}{1+2u}=\frac{1}{2}$.

Combining inequalities (D.37) and (D.38), when $\theta \in \mathcal{C}_t$ for each $t \geq 0$, if $\frac{\min_{\theta\in\mathcal{C}_0}|\psi(\theta)-\bar{p}_{1:n}|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} > K$, we have

$$\lambda_{\min}(V_{t,n}) \geq \min\Big\{\frac{1}{4(1+2u)^2}(1-\frac{1}{K})^2, \frac{1}{4}(u-l)^2\Big\}\cdot\Big((n\wedge t)\delta^2+n\sigma^2\Big)+\lambda,$$

and thus,

$$||\theta^\star-\theta_t^{\mathrm{LS}}||^2 \leq \big(\min\{\frac{1}{4(1+2u)^2}(1-\frac{1}{K})^2, \frac{1}{4}(u-l)^2\}\big)^{-1}\frac{w_{t-1}^2}{(n\wedge t)\delta^2+n\sigma^2}.$$

Therefore, the regret of the myopic policy is upper bounded as follows:

$$\begin{aligned}&\sum_{t=1}^{T}\psi(\theta^\star)\big(\alpha^\star+\beta^\star\psi(\theta^\star)\big)-p_t^{\mathrm{myopic}}\big(\alpha^\star+\beta^\star p_t^{\mathrm{myopic}}\big) \\ &\leq |\beta_{\min}|\cdot\sum_{t=1}^{T}\big(\psi(\theta^\star)-\psi(\theta_{t-1}^{\mathrm{LS}})\big)^2 \\ &\leq \frac{|\beta_{\min}|(\alpha_{\max}^2+\beta_{\max}^2)}{4\beta_{\max}^4}\sum_{t=1}^{T}||\theta^\star-\theta_{t-1}^{\mathrm{LS}}||^2 \\ &\leq \frac{|\beta_{\min}|(\alpha_{\max}^2+\beta_{\max}^2)}{4\beta_{\max}^4}\sum_{t=1}^{T}\big(\min\{\frac{1}{4(1+2u)^2}(1-\frac{1}{K})^2, \frac{1}{4}(u-l)^2\}\big)^{-1}\frac{w_{t-1}^2}{(n\wedge t)\delta^2+n\sigma^2} \\ &= \mathcal{O}\Big(\frac{T\log T}{(n\wedge T)\delta^2+n\sigma^2}\Big).\end{aligned}$$

Note that from Lemma D.2, by letting $\epsilon=\frac{1}{T^2}\wedge\frac{1}{n\sigma^2}$, we have with probability $1-\frac{1}{T^2}\wedge\frac{1}{n\sigma^2}$, $\theta\in\mathcal{C}_t$ for all $0\leq t\leq T$. Thus, with probability $1-\frac{1}{T^2}\wedge\frac{1}{n\sigma^2}$, if the condition $\frac{\min_{\theta\in\mathcal{C}_0}|\psi(\theta)-\bar{p}_{1:n}|}{\max_{\theta_1,\theta_2\in\mathcal{C}_0}|\psi(\theta_1)-\psi(\theta_2)|} > K$ holds, the myopic policy achieves the regret $\widetilde{\mathcal{O}}(\frac{T}{(n\wedge T)\delta^2+n\sigma^2})$. □

## D.4 On the Definition of the Optimal Regret

In §5.3 and §5.4, we define the optimal regret as

$$R^\star(T,n,\delta,\sigma) = \inf_{\pi\in\Pi^\circ} \sup_{\substack{\mathcal{D}\in\mathcal{E}(R);\\ \theta\in\Theta^\dagger:|\psi(\theta)-\bar{p}_{1:n}|\in[(1-\xi)\delta,(1+\xi)\delta]}} R_\theta^\pi(T),$$

where the environment class is chosen as $\{\theta\in\Theta^\dagger : |\psi(\theta)-\bar{p}_{1:n}|\in[(1-\xi)\delta,(1+\xi)\delta]\}$. In this section of the appendix, we give some justifications on this definition of the instance-dependent environment class.

### D.4.1 Comparison to the "Worst-Case" Environment Class

One possible way to define the environment class is to allow the demand parameter $\theta\in\Theta^\dagger$ to vary over the entire set $\Theta^\dagger$. This corresponds to the *optimal worst-case regret* (also known as the *minimax regret*):

$$R^{\mathrm{wc}}(T,n,\sigma) = \inf_{\pi\in\Pi^\circ} \sup_{\mathcal{D}\in\mathcal{E}(R),\theta\in\Theta^\dagger} R_\theta^\pi(T).$$

As a byproduct of our results, we can easily characterize the rate of the optimal worst-case regret.

**Corollary D.1.** *Consider the* OPOD *problem. Then*

$$R^{\mathrm{wc}}(T,n,\sigma) = \widetilde{\Theta}(\sqrt{T}\wedge\frac{T}{n\sigma^2}).$$

Corollary D.1 shows that when $n\sigma^2$ is within $\widetilde{\mathcal{O}}(\sqrt{T})$, the optimal worst-case regret is always $\widetilde{\Theta}(\sqrt{T})$, and when $n\sigma^2$ exceeds $\widetilde{\Omega}(\sqrt{T})$, the optimal worst-case regret decays according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$. This demonstrates that the offline data may help to reduce the worst-case regret, but only when they are dispersive enough, i.e., $n\sigma^2\gtrsim\sqrt{T}$. For example, in the single-historical-price setting with $\sigma=0$, even if the seller has infinitely many offline data, i.e., $n=\infty$, the best achievable worst-case regret is still $\widetilde{\Theta}(\sqrt{T})$, and does not improve over the classical setting where there is no offline data. This suggests that the optimal worst-case regret may fail to fully and precisely reflect the value of the offline data (especially when they are not so dispersive), and the goal of achieving the optimal worst-case regret may be too weak. Indeed, the worst case seldom happens in reality and the decision makers are more interested in the actually incurred regret. The offline data thus should play a more powerful role, not only to reduce the regret in the (rare) worst-case scenario, but also to reduce the regret in a per-instance way. The value of the offline data should also be characterized more precisely.

Observing that the definition of the optimal worst-case regret and the choice of the environment class $\Theta^\dagger$ are too conservative, we consider a less conservative environment class by restricting $|\psi(\theta)-\bar{p}_{1:n}|$ to have the same order as $\delta$ (note that our algorithm does not need to known $\delta$). The resulting $\widetilde{\Theta}(\sqrt{T}\wedge\frac{T}{n\sigma^2+(n\wedge T)\delta^2})$ optimal instance-dependent regret significantly improves the $\widetilde{\Theta}(\sqrt{T})$ optimal worst-case regret when $\delta$ is large enough, thus better characterizing the value of offline data. Our results imply that the location of the offline data is an important metric that intrinsically affects the statistical complexity of the OPOD problem. To the best of knowledge, our results provide the first tight and general instance-dependent regret bounds for the dynamic pricing problem with an

unknown linear demand model[63], with the help of offline data.

### D.4.2 Comparison to the "Local" Environment Class

Another possible way to define the instance-dependent regret is to choose the environment class as the set of all the demand parameters $\theta \in \Theta^\dagger$ such that $|\psi(\theta) - \bar{p}_{1:n}|$ exactly equals the generalized distance $\delta$, i.e., $\{\theta \in \Theta^\dagger : |\psi(\theta) - \bar{p}_{1:n}| = \delta\}$. This leads to the following definition of the *local optimal regret*:

$$R^{\mathrm{loc}}(T, n, \delta, \sigma) = \inf_{\pi \in \Pi^\circ} \sup_{\substack{\mathcal{D} \in \mathcal{E}(R);\\ \theta \in \Theta^\dagger : |\psi(\theta) - \bar{p}_{1:n}| = \delta}} R^\pi_\theta(T).$$

With this definition, we can establish the following result on $R^{\mathrm{loc}}(T, n, \sigma, \delta)$ when $\sigma = 0$ and $\delta = \Theta(1)$, whose proof is deferred to Appendix D.4.3.

**Proposition D.1.** *Consider the* OPOD *problem with a single historical price $\hat{p}$. When $\delta = \Theta(1)$, we have*

$$R^{\mathrm{loc}}(T, n, \delta) := R^{\mathrm{loc}}(T, n, \delta, 0) = \begin{cases} \widetilde{\Theta}(\sqrt{T}), & \text{if } n \lesssim \sqrt{T}; \\ \widetilde{\Theta}(\log T), & \text{if } n \gtrsim \sqrt{T}. \end{cases}$$

Note that when $\sqrt{T} \lesssim n \lesssim T$, the local optimal regret $R^{\mathrm{loc}}(T, n, \delta) = \widetilde{\Theta}(\log T)$ is significantly smaller than the optimal instance-dependent regret $R^\star(T, n, \delta) = \widetilde{\Theta}(\frac{T}{n})$. But why does this happen? The caveat is that the rate of $R^{\mathrm{loc}}(T, n, \delta)$ is meaningless in the sense that it cannot be uniformly achieved by any single algorithm! That is to say, if we consider multiple different values of $\delta$, e.g., $\delta = 1, \delta = 1.1, \delta = 1.11, \ldots$, while $R^{\mathrm{loc}}(T, n, 1) = \widetilde{\Theta}(\log T), R^{\mathrm{loc}}(T, n, 1.1) = \widetilde{\Theta}(\log T), R^{\mathrm{loc}}(T, n, 1.11) = \widetilde{\Theta}(\log T), \ldots$, they are actually achieved by *different* algorithms that are specially designed for $\delta = 1, \delta = 1.1, \delta = 1.11, \ldots$ respectively, and there is no algorithm that can achieve $R^{\mathrm{loc}}(T, n, \delta) = \widetilde{\Theta}(\log T)$ for all of $\delta = 1, \delta = 1.1, \delta = 1.11, \ldots$ simultaneously.

To see this, we give a concrete algorithm $\tilde{\pi} \in \Pi^\circ$ that achieves the regret of $\mathcal{O}(\log T)$ for some specific value of $\delta = \delta_0$ but incurs the regret of $\Omega(\sqrt{T})$ for $\delta = \delta_0 + T^{-\frac{1}{4}}$. The algorithm is named as "Speculator($\delta_0$)" and is presented in Algorithm D.1. When $\delta = 1$, and $n = \sqrt{T}$, the Speculator(1) algorithm incurs the regret of $\mathcal{O}(\log T)$ in the first stage and constant regret in the second stage. However, when $\delta = 1 + T^{-\frac{1}{4}}$, the Speculator(1) algorithm must incur the regret of $\Omega(T \times (T^{-\frac{1}{4}})^2) = \Omega(\sqrt{T})$ in the second stage, since with high probability, the algorithm mistakenly charges $\hat{p} + \delta_0$ or $\hat{p} - \delta_0$ for the whole second stage.

In fact, with the above definition of the local optimal regret, any learning algorithm faces the above dilemma, i.e., its regret is not universally optimal when $\delta$ changes, and the reason is as follows. Using KL-divergence arguments, we can show that when $n = \Theta(\sqrt{T})$, for any $\delta = \Theta(1)$ and any policy $\pi$, the sum of the local instance-dependent regrets under $\delta$ and $\delta + \Theta(T^{-\frac{1}{4}})$ is lower bounded

[63] We note that Broder and Rusmevichientong (2012), Keskin and Zeevi (2014) and Qiang and Bayati (2016) provide $\widetilde{\Theta}(\log T)$ regret bounds for this dynamic pricing problem under certain separability assumptions. However, they do not obtain a regret bound that directly depends on the instance parameters in a tight way.

**Algorithm D.1** Speculator($\delta_0$): an algorithm that bets $\delta = \delta_0$

**Input**: specific guess $\delta_0$, historical price $\hat{p}$, offline demand data $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_n$, support of unknown parameters $\Theta^\dagger$, support of feasible price $[l, u]$, length of the selling horizon $T$

1: **while** $t \in [\lfloor\sqrt{T}\rfloor]$ **do**
2: Treat the prices $\hat{p} + \delta_0$ and $\hat{p} - \delta_0$ as two arms, and run the UCB algorithm for the two-armed bandits;
3: Construct the confidence interval $\tilde{C}$ for the optimal price based on the least square regression on both the offline and online data
4: **if** $\hat{p} + \delta_0 \in \tilde{C}$ (or $\hat{p} - \delta_0 \in \tilde{C}$) **then**
5: Charge the price $\hat{p} + \delta_0$ (or $\hat{p} - \delta_0$) when $t = [\lfloor\sqrt{T}\rfloor] + 1, \ldots, T$;
6: **else**
7: Charge the myopic price from the least square estimation when $t = [\lfloor\sqrt{T}\rfloor]+1, \ldots, T$.

by $\Omega(\sqrt{T})$, i.e.,

$$\sup_{\substack{\mathcal{D}\in\mathcal{E}(R);\\ \theta\in\Theta^\dagger:|\psi(\theta)-\hat{p}|=\delta}} R^\pi_\theta(T) + \sup_{\substack{\mathcal{D}\in\mathcal{E}(R);\\ \theta\in\Theta^\dagger:|\psi(\theta)-\hat{p}|=\delta+\Theta(T^{-\frac{1}{4}})}} R^\pi_\theta(T) = \Omega(\sqrt{T}),$$

which implies that for any policy $\pi$, under at least one problem instance, i.e., $\delta$ or $\delta + \Theta(T^{-\frac{1}{4}})$, the regret is greater than $\Omega(\sqrt{T})$. The huge gap between $\Omega(\sqrt{T})$ and $\Theta(\log T)$ implies that when $n = \Theta(\sqrt{T})$, the optimal rate of $R^{\mathrm{loc}}(T, n, \delta)$ defined in Proposition D.1 cannot be achieved by a single learning algorithm for different values of $\delta$. Thus, $R^{\mathrm{loc}}(T, n, \delta)$ fails to be a valid complexity measure for the `OPOD` problem. In fact, the statistical complexity of an online pricing problem heavily relies on the fact that there are infinitely many continuous and "indistinguishable" prices. If we directly define the environment class as $\{\theta \in \Theta^\dagger : |\psi(\theta) - \hat{p}| = \delta\}$, then the resulting $R^{\mathrm{loc}}(T, n, \delta)$ becomes too "sensitive and specific" to two discrete prices $\hat{p} + \delta$ and $\hat{p} - \delta$, leaving chances for an algorithm that "bets $\delta = \delta_0$" to perform "abnormally well" when $\delta$ happens to be $\delta_0$. By contrast, under the definition of the optimal regret $R^\star(T, n, \delta)$ in §5.3 and §5.4, we can design a learning algorithm that uniformly achieves the optimal regret rate for any possible value of $\delta$.

### D.4.3 Proof of Proposition D.1 in Appendix D.4.2

The proof will be divided into proving the regret lower bound and regret upper bound respectively.

**Lower bound: Case 1.** We first prove that when $n \le \frac{R^2 l^2 |\beta_{\max}|}{256\beta^2_{\min} K_0 e \delta^2}\sqrt{T}$, for any admissible policy $\pi \in \Pi^\circ$, and any $\theta \in \Theta^\dagger$ with $\frac{\alpha}{-2\beta} = \hat{p} + \delta$, the regret is lower bounded by $\Omega(\sqrt{T})$. To see this, we construct two problem instances $\theta_1 = (\alpha_1, \beta_1)$ and $\theta_2 = (\alpha_2, \beta_2)$ satisfying the following conditions:

$$-\frac{\alpha_1}{2\beta_1} = \hat{p} + \delta, \quad -\frac{\alpha_2}{2\beta_2} = \hat{p} + \delta + \Delta, \quad (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)(\hat{p} + \delta) = 0.$$

where the value of $\Delta$ is to be specified. That is, the optimal price under the two problem instances is $\hat{p} + \delta$ and $\hat{p} + \delta + \Delta$ respectively, and the two demand functions intersect at the price $\hat{p} + \delta$. Using

similar arguments in inequality (D.21), we have

$$KL(P_1^\pi, P_2^\pi) \le \frac{2\beta_{\min}^2 \Delta^2}{R^2 l^2}\Big(n\delta^2 + \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi[(p_t - \psi(\theta_1))^2]\Big).$$

In addition, by defining the two disjoint intervals $I_1 = [\hat{p}+\delta-\frac{1}{4}\Delta, \hat{p}+\delta+\frac{1}{4}\Delta]$ and $I_2 = [\hat{p}+\delta+\frac{3}{4}\Delta, \hat{p}+\delta+\frac{5}{4}\Delta]$, and using similar arguments to inequality (D.23), we have the following lower bound on the sum of regret under $\theta_1$ and $\theta_2$:

$$R_{\theta_1}^\pi(T) + R_{\theta_2}^\pi(T) \ge \frac{1}{32}|\beta_{\max}| \cdot e^{-KL(P_1^\pi, P_2^\pi)} \cdot T\Delta^2.$$

Since $\pi \in \Pi^\circ$, we further have

$$KL(P_1^\pi, P_2^\pi) \ge \log(\sqrt{T}\Delta^2) + \log\Big(\frac{|\beta_{\max}|}{64K_0}\Big).$$

Thus, by letting $\Delta^2 = \frac{64K_0 e}{\sqrt{T}|\beta_{\max}|}$, we have

$$\begin{aligned}
R_{\theta_1}^\pi(T) &\ge |\beta_{\max}| \sum_{t=1}^{T} \mathbb{E}_{\theta_1}^\pi[(p_t - \psi(\theta_1))^2] \\
&\ge |\beta_{\max}|\Big(\frac{R^2 l^2}{2\beta_{\min}^2 \Delta^2} \cdot \big(\log(\sqrt{T}\Delta^2) + \log\frac{|\beta_{\max}|}{64K_0}\big) - n\delta^2\Big) \\
&= |\beta_{\max}|\Big(\frac{R^2 l^2 |\beta_{\max}|}{128\beta_{\min}^2 K_0 e}\sqrt{T} - n\delta^2\Big) \\
&\ge \frac{R^2 l^2 \beta_{\max}^2}{256\beta_{\min}^2 K_0 e}\sqrt{T},
\end{aligned}$$

where the equation follows from the choice of $\Delta$, and the last inequality holds since $n \le \frac{R^2 l^2 |\beta_{\max}|}{256\beta_{\min}^2 K_0 e \delta^2}\sqrt{T}$.

**Lower bound: Case 2.** We then prove when $n > \frac{R^2 l^2 |\beta_{\max}|}{256\beta_{\min}^2 K_0 e\delta^2}\sqrt{T}$, for any policy $\pi$ (not necessarily in the admissible policy class), $\max_{\theta\in\Theta^\dagger:\psi(\theta)-\hat{p}=\delta} R_\theta^\pi(T) = \Omega(\log T)$. Since $\psi(\theta) = -\frac{\alpha}{2\beta}$, the constraint for $\theta$ becomes $\{(-2\beta(\hat{p}+\delta), \beta) : \beta \in [\beta_{\min} \vee \frac{\alpha_{\max}}{-2(\hat{p}+\delta)}, \beta_{\max} \wedge \frac{\alpha_{\min}}{-2(\hat{p}+\delta)}]\}$. In this case, the problem is reduced to a single-dimensional problem, and it suffices to prove that there exists some $\beta \in [\beta_{\min} \vee \frac{\alpha_{\max}}{-2(\hat{p}+\delta)}, \beta_{\max} \wedge \frac{\alpha_{\min}}{-2(\hat{p}+\delta)}]$, such that $R_\theta^\pi(T) = \Omega(\log T)$, where $\theta = (-2\beta(\hat{p}+\delta), \beta)$.

To this end, we invoke again the van Trees inequality in Lemma (D.3), by letting $C(\theta) = (-\hat{p}\ 1)$, and $q(\cdot) : \mathbb{R} \to \mathbb{R}^+$ be an absolutely continuous density on $\beta \in [\beta_{\min} \vee \frac{\alpha_{\max}}{-2(\hat{p}+\delta)}, \beta_{\max} \wedge \frac{\alpha_{\min}}{-2(\hat{p}+\delta)}]$ with positive value on $(\beta_{\min} \vee \frac{\alpha_{\max}}{-2(\hat{p}+\delta)}, \beta_{\max} \wedge \frac{\alpha_{\min}}{-2(\hat{p}+\delta)})$ and zero on the boundary $\{\beta_{\min} \vee \frac{\alpha_{\max}}{-2(\hat{p}+\delta)}, \beta_{\max} \wedge \frac{\alpha_{\min}}{-2(\hat{p}+\delta)}\}$. In this case, similar to the first lower bound in Step 1 of the proof of Theorem 5.2, we obtain the following inequality for $\theta = (-2\beta(\hat{p}+\delta), \beta)$:

$$\sum_{t=1}^{T} \mathbb{E}_q\big[\mathbb{E}_\theta^\pi[(p_t - \psi(\theta))^2]\big] \ge \sum_{t=2}^{T} \frac{R^2 c_1'}{R^2\mathcal{I}(q) + \sum_{s=1}^{t-1}\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_s - \hat{p})^2]]} \ge \sum_{t=2}^{T} \frac{R^2 c_1'}{R^2\mathcal{I}(q) + (t-1)(u-l)^2},$$

where $c_1' = (\min_{\theta\in\Theta^\dagger:\psi(\theta)=\hat{p}+\delta} \frac{\alpha+\beta\hat{p}}{2\beta^2})^2$, and $\mathcal{I}(q)$ is defined in Lemma (D.3). Since both $c_1'$ and $\mathcal{I}(q)$

are constants (recall that $\delta$ is assumed to be a constant), then we have

$$\max_{\beta \in [\beta_{\min} \vee \frac{\alpha_{\max}}{-2(\hat{p}+\delta)}, \beta_{\max} \wedge \frac{\alpha_{\min}}{-2(\hat{p}+\delta)}]} R_\theta^\pi(T) = \Omega(\log T),$$

which completes the proof.

**Upper bound.** When $n < \sqrt{T}$, from Theorem 5.1, O3FU algorithm is admissible and achieves the regret upper bound $\mathcal{O}(\sqrt{T})$, which matches the lower bound proven in the above Case 1. In the following, we first prove that when $n \geq \sqrt{T}$, Speculator($\delta_0$) achieves the regret upper bound $\mathcal{O}(\sqrt{T})$ for any $\theta \in \Theta^\dagger$, and therefore is admissible. Then we will prove that when $\delta = \delta_0$, Speculator($\delta_0$) achieves the regret upper bound $\mathcal{O}(\log T)$.

When $\theta^\star$ is arbitrary, $\delta = |\psi(\theta^\star) - \hat{p}|$ is also arbitrary and not necessarily equals $\delta_0$, the regret in the first $\sqrt{T}$ periods is $\mathcal{O}(\sqrt{T})$ due to at most a constant loss in each period. In addition, it can be easily verified that the sum of squared dispersion for $n$ offline prices and $\lfloor\sqrt{T}\rfloor$ online prices is lower bounded by $\Omega(\sqrt{T})$. Specifically, from (5.9), for $\hat{p}_1 = \ldots = \hat{p}_n = \hat{p}$ and $p_t \in \{\hat{p} + \delta_0, \hat{p} - \delta_0\}$ for each $t \in [\lfloor\sqrt{T}\rfloor]$, we have

$$J(\hat{p}_1, \ldots, \hat{p}_n, p_1, \ldots, p_{\lfloor\sqrt{T}\rfloor}) \geq J(\hat{p}_1, \ldots, \hat{p}_n) + \frac{n}{n+\sqrt{T}} \sum_{s=1}^{\lfloor\sqrt{T}\rfloor} (p_s - \bar{p}_{1:n})^2 = \frac{n\lfloor\sqrt{T}\rfloor}{n + \lfloor\sqrt{T}\rfloor} \delta_0^2 \gtrsim \sqrt{T}\delta_0^2.$$

Therefore, $\lambda_{\min}(V_{\lfloor\sqrt{T}\rfloor,n}) = \Omega(\sqrt{T})$, and the squared radius of the confidence interval $\tilde{C}$ is at most $\Theta(\frac{1}{\lambda_{\min}(V_{\lfloor\sqrt{T}\rfloor,n})}) = \Theta(\sqrt{T})$. Since the true optimal price lies in $\tilde{C}$ with high probability, it follows that for any price within $\tilde{C}$, its squared deviation from the optimal price in each period $\lfloor\sqrt{T}\rfloor + 1, \ldots, T$ is no more than $\frac{1}{\sqrt{T}}$, and therefore, the cumulative revenue loss in periods $\lfloor\sqrt{T}\rfloor + 1, \ldots, T$ is no more than $\mathcal{O}(\sqrt{T})$.

When $\delta = \delta_0$, from Theorem 5 in Abbasi-Yadkori et al. (2011), the regret of Speculator($\delta_0$) in the first $\lfloor\sqrt{T}\rfloor$ periods is upper bounded by $\widetilde{\mathcal{O}}(\log T)$. In the remaining periods from $\lfloor\sqrt{T}\rfloor + 1$ to $T$, since the optimal price is either $\hat{p} + \delta_0$ or $\hat{p} - \delta_0$, which belongs to the confidence interval $\tilde{C}$ with high probability, by construction, Speculator($\delta_0$) chooses the optimal price from period $\lfloor\sqrt{T}\rfloor + 1$ to $T$ with high probability. Note that the squared length of $\tilde{C}$ is $\Theta(\frac{1}{\sqrt{T}})$, so $\hat{p} + \delta_0$ and $\hat{p} - \delta_0$ cannot belong to $\tilde{C}$ at the same time. In this case, it can be verified that the regret from period $\lfloor\sqrt{T}\rfloor + 1$ to $T$ is upper bounded by $\widetilde{\mathcal{O}}(\log T)$. □

## D.5 Extension to Generalized Linear Model

In this section of the appendix, we discuss the extension of our regret upper bounds to the generalized linear model. For simplicity, we focus on the single-historical-price setting, and leave the discussion on the multiple-historical-price setting to the interested readers. Consider the following demand model:

$$D_t = g(\alpha^\star + \beta^\star p_t) + \varepsilon_t, \tag{D.39}$$

where $g(\cdot)$ is an increasing function whose form is known to the seller (we refer to $g(\cdot)$ as the *link function*), $(\alpha^\star, \beta^\star)$ is the unknown demand parameter in the compact set $\Theta^\dagger$, and $\{\epsilon_t\}_{t\geq 1}$ is a sequence of i.i.d. sub-Gaussian random variables. We also assume that the conditional probability of $D_t$ given $p_t$ is from the exponential family, which is a standard assumption in the literature; see, e.g., Filippi et al. (2010). Since the expected demand function is the composition of the link function $g(\cdot) : \mathbb{R} \to \mathbb{R}$ and the linear function $p \mapsto \alpha^\star + \beta^\star p$, the above equation (D.39) is referred to as the *generalized linear model* (GLM). Similar as before, we let $\theta := (\alpha, \beta)$, $r(p;\theta) := p \cdot g(\alpha + \beta p)$ and $\psi(\theta) := \arg\max_{p\in[\underline{p},\overline{p}]} r(p;\theta)$. The definition of the regret $R^\pi_{\theta^\star}(T)$ for any given policy $\pi$ remains the same.

We make the following assumptions on the optimal price $\psi(\theta)$, the expected revenue $r(p;\theta)$, and the link function $g(\cdot)$.

**Assumption D.1.** *There exist constants $L_0 > 0$, $0 < \lambda_1 < \lambda_2$ and $0 < L_1 < L_2$, such that*

*(a)* $|\psi(\theta_1) - \psi(\theta_2)| \leq L_0 \cdot ||\theta_1 - \theta_2||$ *for any* $\theta_1, \theta_2 \in \Theta^\dagger$;

*(b)* $\lambda_1 \cdot (\psi(\theta) - p)^2 \leq r(\psi(\theta);\theta) - r(p;\theta) \leq \lambda_2 \cdot (\psi(\theta) - p)^2$ *for any* $p \in [\underline{p}, \overline{p}]$ *and* $\theta \in \Theta^\dagger$;

*(c)* $g(x)$ *is twice differentiable in* $\mathcal{X} := \{\alpha + \beta p : (\alpha, \beta) \in \Theta^\dagger, p \in [\underline{p}, \overline{p}]\}$, *with* $L_1 \leq g'(x) \leq L_2$ *for any* $x \in \mathcal{X}$, *and bounded second-order derivative in* $\mathcal{X}$.

Condition (a) requires that the optimal price $\psi(\theta)$ is Lipschitz continuous in $\Theta^\dagger$ with Lipschitz constant $L_0$, which is satisfied if $\psi(\cdot)$ is differentiable and the norm of its gradient is upper bounded. Condition (b) is satisfied if for any $\theta \in \Theta^\dagger$, the optimal price $\psi(\theta)$ is an interior point of $[\underline{p}, \overline{p}]$, and the second-order derivative of $r(p;\theta)$, with respect to $p$, exists, and is lower bounded by $\lambda_1$ and upper bounded by $\lambda_2$. Condition (c) is similar to Assumptions 1 and 2 in Li et al. (2017) on the generalized linear contextual bandit, and our condition is slightly stronger to make sure that our instance-dependent upper bound holds. Note that under condition (c), condition (b) can also be satisfied if for any $\theta \in \Theta^\dagger$, $\psi(\theta)$ is an interior point of $[\underline{p}, \overline{p}]$, and the expected revenue, as a function of the *mean demand*, is concave whose second-order derivative is lower bounded by $\lambda_1$ and upper bounded by $\lambda_2$. Note that the concavity of the expected revenue with respect to the mean demand (instead of the price) is more commonly assumed in the literature of revenue management; see, e.g., Wang et al. (2014). All of conditions (a)-(c) can be satisfied by the commonly used linear model (i.e., $g(x) = x$), logit model (i.e., $g(x) = \frac{e^x}{1+e^x}$), and exponential model (i.e., $g(x) = e^x$).

The algorithm for the generalized linear model (D.39) can be modified from O3FU as follows. Let $\hat{\theta}_t$ be the following maximum likelihood estimator (instead of the least-squares estimator in O3FU):

$$\hat{\theta}_t := \mathop{\arg\max}_{\theta=(\alpha,\beta)\in\Theta^\dagger} n\left(\hat{D}_i \cdot (\alpha + \beta\hat{p}) - m(\alpha + \beta\hat{p})\right) + \sum_{s=1}^{t}\left(D_s \cdot (\alpha + \beta p_s) - m(\alpha + \beta p_s)\right),$$

where $m(\cdot)$ is the function such that $m(\alpha + \beta p) = g'(\alpha + \beta p)$ for any $(\alpha, \beta) \in \Theta^\dagger$ and $p \in [l, u]$. Then we let $(p_t, \tilde{\theta}_t) := \arg\max_{p\in[l,u],\theta=(\alpha,\beta)\in\mathcal{C}_{t-1}} p \cdot g(\alpha + \beta p)$. Besides, for the confidence ellipsoid $\mathcal{C}_{t-1}$, the confidence radius $w_t$ needs to be modified accordingly by applying the high-probability confidence bound in Lemma 3 of Li et al. (2017). We refer to this modified algorithm as O3FU-GLM.

The following proposition establishes a similar regret upper bound for O3FU-GLM to O3FU in Theorem 5.3.

**Proposition D.2.** *Let $\pi$ be O3FU-GLM algorithm for the* OPOD *problem. Then there exists a finite constant $K_5 > 0$ such that for any $T \geq 1$, $n \geq 0$ and $\hat{p} \in [l, u]$, and for any possible value of $\theta^\star \in \Theta^\dagger$, we have*

$$R^\pi_{\theta^\star}(T) \leq K_5\Big(\sqrt{T} \wedge \frac{T \log T}{(n \wedge T)\delta^2}\Big) \cdot \log T.$$

*Proof of Proposition D.2.* Under Assumption D.1, Proposition D.2 can be proven under a similar framework to Theorem 5.1. We next only highlight the main differences and omit the detailed verification.

First, to prove the instance-independent upper bound $\widetilde{\mathcal{O}}(\sqrt{T}\log T)$, we first note the following upper bound on the regret of algorithm O3FU-GLM: when $\theta^\star \in \mathcal{C}_{t-1}$,

$$\psi(\theta^\star) \cdot g(\alpha^\star + \beta^\star \psi(\theta^\star)) - p_t \cdot g(\alpha^\star + \beta^\star p_t) \leq p_t \cdot g(\tilde{\alpha}_t + \tilde{\beta}_t p_t) - p_t \cdot g(\alpha^\star + \beta^\star p_t) \leq u \cdot L_2 |(\tilde{\theta}_t - \theta^\star)^\top x_t|,$$

where $x_t = [1\ p_t]^\top$, the first inequality follows from $\theta^\star \in \mathcal{C}_{t-1}$, $\psi(\theta^\star) \in [l, u]$ and the definition of $(p_t, \tilde{\theta}_t)$, and the second inequality follows from condition (c) of Assumption D.1 and the mean value theorem. With the above inequality, the regret upper bound $\widetilde{\mathcal{O}}(\sqrt{T}\log T)$ can be proven similar to Step 1 of Theorem 5.1. In particular, to bound the probability for the event $\{\theta^\star \in \mathcal{C}_t\}_{t\geq 1}$, Lemma 3 in Li et al. (2017) established for the generalized linear contextual bandit will be useful, and plays a similar role to Theorem 2 in Abbasi-Yadkori et al. (2011) established for the linear contextual bandit, which is applied in our previous proof.

Second, to prove the instance-dependent upper bound $\widetilde{\mathcal{O}}(\frac{T(\log T)^2}{(n\wedge T)\delta^2})$, we first note that the regret of O3FU-GLM is upper bounded by the cumulative estimation error for the true parameter $\theta^\star$ as follows:

$$\sum_{t=2}^{T} \mathbb{E}^\pi_{\theta^\star}\Big[r\big(\psi(\theta^\star); \theta^\star\big) - r\big(p_t; \theta^\star\big)\Big] \leq \lambda_2 \sum_{t=1}^{T} \mathbb{E}^\pi_{\theta^\star}\Big[\big(\psi(\theta^\star) - p_t\big)^2\Big] \leq \lambda_2 L_0 \sum_{t=1}^{T} \mathbb{E}^\pi_{\theta^\star}\big[||\theta^\star - \tilde{\theta}_t||^2\big],$$

where the two inequalities hold due to condition (b) and condition (a) in Assumption D.1 respectively. With the above inequality, it suffices to establish Lemma 5.1 for O3FU-GLM. To this end, we also start from the same inequality to (D.4) in Step 2 of the proof of Lemma 5.1, and discuss the same three cases. For Case 1, Case 2 and Case 3.1, the proof is similar under condition (a) in Assumption D.1 that $\psi(\cdot)$ is Lipschitz continuous. For Case 3.2, the crucial step is to show inequality (D.8), whose proof can be modified by invoking conditions (b) and (c) in Assumption D.1. Specifically, we refine $A_1$, $A_2$, $A_3$ and $A_4$ as

$$\begin{aligned} &A_1 = p_t \cdot g(\tilde{\alpha}_t + \tilde{\beta}_t p_t), \quad A_2 = p_t \cdot g(\alpha^\star + \beta^\star p_t), \\ &A_3 = \psi(\theta^\star) \cdot g(\tilde{\alpha}_t + \tilde{\beta}_t \psi(\theta^\star)), \quad A_4 = \psi(\theta^\star) \cdot g(\alpha^\star + \beta^\star \psi(\theta^\star)), \end{aligned}$$

and inequalities in (D.9) and (D.10) continue to hold, i.e.,

$$A_1 \geq A_3, \quad A_1 \geq A_4 \geq A_2.$$

For the case when $A_3 \geq A_2$, inequality (D.11) will be modified to

$$\begin{aligned}
|\Delta\alpha_t + \Delta\beta_t p_t| &\geq \frac{1}{L_2} \cdot \frac{A_1 - A_2}{p_t} \\
&\geq \frac{1}{L_2} \cdot \frac{|A_4 - A_3|}{p_t} \\
&= \frac{1}{L_2} \cdot \frac{\psi(\theta^\star)}{p_t} \cdot \left| g(\alpha^\star + \beta^\star \psi(\theta^\star)) - g(\tilde{\alpha}_t + \tilde{\beta}_t \psi(\theta^\star)) \right| \\
&\geq \frac{L_1}{L_2} \cdot \frac{\psi(\theta^\star)}{p_t} |\Delta\alpha_t + \Delta\beta_t \psi(\theta^\star)| \\
&\geq \frac{L_1}{L_2} \cdot \frac{l}{u} |\Delta\alpha_t + \Delta\beta_t \psi(\theta^\star)|,
\end{aligned}$$

where the first inequality follows from $|g(x) - g(y)| \leq L_2|x - y|$ guaranteed by condition (c) of Assumption D.1 and the mean value theorem, the second inequality holds since $A_3, A_4 \in [A_2, A_1]$, and the third inequality holds due to $|g(x) - g(y)| \geq L_1|x - y|$ guaranteed by condition (c) of Assumption D.1 and the mean value theorem. For the case when $A_3 < A_2$, inequality (D.12) will be modified to

$$\begin{aligned}
|\Delta\alpha_t + \Delta\beta_t p_t| &\geq \frac{1}{L_2} \cdot \frac{A_1 - A_2}{p_t} \\
&\geq \frac{1}{L_2} \cdot \frac{A_4 - A_2}{p_t} \\
&= \frac{1}{L_2} \cdot \frac{r(\psi(\theta^\star); \theta^\star) - r(p_t; \theta^\star)}{p_t} \\
&\geq \frac{\lambda_1}{L_2} \cdot \frac{(\psi(\theta^\star) - p_t)^2}{p_t} \\
&\geq \frac{\lambda_1}{L_2 \lambda_2 p_t} \cdot |A_1 - A_3| \\
&\geq \frac{\lambda_1}{L_2 \lambda_2 p_t} \cdot |A_4 - A_3| \\
&\geq \frac{\lambda_1 L_1}{L_2 \lambda_2} \cdot \frac{\psi(\theta^\star)}{p_t} \cdot |\Delta\alpha_t + \Delta\beta_t \psi(\theta^\star)| \\
&\geq \frac{\lambda_1 L_1}{L_2 \lambda_2} \cdot \frac{l}{u} \cdot |\Delta\alpha_t + \Delta\beta_t \psi(\theta^\star)|,
\end{aligned}$$

where the third and fourth inequalities follow from condition (b) in Assumption D.1, the fifth inequality follows from the assumption that $A_3 < A_2$, and the sixth inequality follows from condition (c) in Assumption D.1 and the mean value theorem. The remaining analysis for Case 3.2 is similar, whose details are therefore omitted. □

## D.6 Extension to Adaptive Offline Data

In this section of the appendix, we extend our main results to the setting that in the offline stage, the seller's pricing decisions are made adaptively based on the previous price and sales data according to some possibly unknown policy $\hat{\pi}$. Therefore, for each $i = 2, \ldots, n$, $\hat{p}_i$ may depend on the previous data $\hat{p}_1, \hat{D}_1, \ldots, \hat{p}_{i-1}, \hat{D}_{i-1}$.

When the offline data are generated adaptively according to some possibly unknown policy $\hat{\pi}$, the historical price $\hat{p}_i$ is a function of $\hat{p}_1, \hat{D}_1, \ldots, \hat{p}_{i-1}, \hat{D}_{i-1}$, for each $i = 2, \ldots, n$, which contains uncertainty arising from the random noise, and therefore is a random variable. Nevertheless, in many practical scenarios, the seller's primary concern is to understand the effect of this *particular* pricing sequence $\{\hat{p}_1, \ldots, \hat{p}_n\}$ on the online learning process. Thus, we will measure the performance of a learning algorithm via the conditional expected revenue given the realization of $\hat{p}_1, \ldots, \hat{p}_n$, and study the impact of this exact sequence on the online learning process.

Specifically, for any pricing policy $\pi$, let $R^{\pi}_{\theta^\star}(T, \hat{p}_1, \ldots, \hat{p}_n)$ be defined as the conditional regret as follows:

$$R^{\pi}_{\theta^\star}(T, \hat{p}_1, \ldots, \hat{p}_n) = \mathbb{E}^{\pi}_{\theta^\star}\big[Tr^\star(\theta^\star) - \sum_{t=1}^{T} p_t(\alpha^\star + \beta^\star p_t)\big|\hat{p}_1, \ldots, \hat{p}_n\big].$$

For any pricing policy $\pi$, it is said to be admissible if there exists some constant $K_0 > 0$ such that $R^{\pi}_{\theta^\star}(T, \hat{p}_1, \ldots, \hat{p}_n) \le K_0\sqrt{T}\log T$, for any $T \ge 1$, $n \ge 0$, $\theta^\star \in \Theta^\dagger$, and $\hat{p}_1, \ldots, \hat{p}_n \in [l, u]$. Let $\hat{\Pi}^\circ$ be the set of all admissible policies. For notation convenience, we also define $\hat{\delta} = |\bar{p}_{1:n} - \psi(\theta^\star)|$, and $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - \bar{p}_{1:n})^2}$, both of which depend on the realizations of $\hat{p}_1, \ldots, \hat{p}_n$. We provide matching upper and lower bounds on regret in Proposition D.3, which indicates that M-O3FU algorithm remains optimal even for adaptive offline data.

**Proposition D.3.** *Consider the* OPOD *problem with the offline data generated from some possibly unknown policy $\hat{\pi}$.*

*(a) Let $\pi$ be M-O3FU algorithm. For any sample path of historical prices $\hat{p}_1, \ldots \hat{p}_n$, $T \ge 1$, $n \ge 1$, and any possible value of $\theta^\star \in \Theta^\dagger$,*

$$R^{\pi}_{\theta^\star}(T, \hat{p}_1, \ldots, \hat{p}_n) = \begin{cases} \widetilde{\mathcal{O}}\big(T\hat{\delta}^2 + 1\big), & \text{if } \hat{\delta}^2 \lesssim \frac{1}{n\hat{\sigma}^2} \lesssim \frac{1}{\sqrt{T}}; \\ \widetilde{\mathcal{O}}\big(\sqrt{T} \wedge \frac{T}{(n\wedge T)\hat{\delta}^2 + n\hat{\sigma}^2}\big), & \text{otherwise.} \end{cases}$$

*(b) For any pricing policy $\pi$, $T \ge 2$, $n \ge 1$, $\hat{\delta} \in [0, u-l]$, and realization of $\hat{p}_1, \ldots, \hat{p}_n \in [l, u]$,*

$$\sup_{\substack{\mathcal{D}\in\mathcal{E}(R);\\ \theta\in\Theta^\dagger:|\bar{p}_{1:n}-\psi(\theta)|\in[(1-\xi)\hat{\delta},(1+\xi)\hat{\delta}]}} R^{\pi}_{\theta}(T, \hat{p}_1, \ldots, \hat{p}_n) = \widetilde{\Omega}(\sqrt{T} \wedge \frac{T}{\hat{\delta}^{-2} + (n\wedge T)\hat{\delta}^2 + n\hat{\sigma}^2}).$$

*If for any value of $\theta \in \Theta^\dagger$, $\mathbb{E}^{\hat{\pi}}_{\theta}[\hat{\delta}(\theta)] \lesssim T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}}$ and $\mathbb{E}^{\hat{\pi}}_{\theta}[n\hat{\sigma}^2] \lesssim \frac{\sqrt{T}}{\log T}$ (where the expectation is taken over $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$), then for any admissible policy $\pi \in \hat{\Pi}^\circ$, $T \ge 2$, $n \ge 1$, $\theta^\star \in \Theta^\dagger$, $\mathbb{E}^{\hat{\pi}}[R^{\pi}_{\theta}(T, \hat{p}_1, \ldots, \hat{p}_n)] = \widetilde{\Omega}(\sqrt{T})$.*

*Proof of Proposition D.3.* The proof is similar to Theorem 5.3 and Theorem 5.4, and we only highlight the differences and omit detailed analysis.

(a) Similar to the proof of Theorem 5.3, we need to show the two upper bounds $\mathcal{O}(\sqrt{T}\log T)$ and $\mathcal{O}(\frac{T(\log T)^2}{n\hat{\sigma}^2 + (n\wedge T)\hat{\delta}^2})$.

To see the first upper bound, if conditioning on the realization of $\hat{p}_1, \ldots, \hat{p}_t$, the upper bound on $\sum_{t=1}^{T} ||x_t||^2_{V^{-1}_{t-1,n}}$, i.e., (D.24), and the concentration inequality in Lemma D.2 still hold, then

we can apply similar arguments to Step 1 in the proof of Theorem 5.3 to obtain the first upper bound $\mathcal{O}(\sqrt{T}\log T)$. To this end, we notice that the upper bound (D.24) is derived from Lemma D.1, and in the statement of Lemma D.1, the sequence $\{X_t : t \geq 1\}$ and the matrix $V$ can be arbitrary. Therefore, for any given realization of $\hat{p}_1, \ldots, \hat{p}_n$, by letting $V = \lambda I + \sum_{i=1}^{n} x_i x_i^\top$, we have similar upper bound on $\sum_{t=1}^{T} ||x_t||^2_{V_{t-1,n}^{-1}}$. Moreover, the key ingredient to prove Lemma D.2 in Abbasi-Yadkori et al. (2011) is their Theorem 1. For any given realization of $\hat{p}_1, \ldots, \hat{p}_n$, Theorem 1 in Abbasi-Yadkori et al. (2011) continues to hold, and therefore, the bound for the conditional probability given $\hat{p}_1, \ldots \hat{p}_n$ in Lemma D.2 also holds.

To see the second upper bound, it suffices to establish the concentration inequality in Lemma D.2, the sample-path inequality in Lemma 5.2 and Lemma D.4. As discussed above, given realization of $\hat{p}_1, \ldots, \hat{p}_n$, Lemma D.2 continues to hold. In both Lemma 5.2 and Lemma D.4, we conduct the sample-path analysis and treat each quantity as an arbitrary and deterministic number. Therefore, conditioning on the realization of $\hat{p}_1, \ldots, \hat{p}_t$, Lemma 5.2 and Lemma D.4 also continue to hold.

(b) We divide the proof for the lower bound into two steps.

**Lower bound: Step 1.** In this step, similar to (D.31), we prove for any policy $\pi$,

$$\sup_{\theta \in \Theta_0(\hat{\delta}, \hat{p}_1, \ldots, \hat{p}_n)} R_\theta^\pi(T, \hat{p}_1, \ldots, \hat{p}_n) = \Omega\Big(\sqrt{T} \wedge \frac{T}{\hat{\delta}^{-2} + n\hat{\sigma}^2 + (n \wedge T)\hat{\delta}^2}\Big), \tag{D.40}$$

where $\Theta_0(\hat{\delta}, \hat{p}_1, \ldots, \hat{p}_n) = \big\{\theta \in \Theta^\dagger : \psi(\theta) - \bar{p}_{1:n} \in [\frac{1}{2}\hat{\delta}, \hat{\delta}]\big\}$. Here we highlight the dependence of set $\Theta_0$ on $\hat{p}_1, \ldots, \hat{p}_n$.

To see (D.40), we first note that since $l \leq \bar{p}_{1:n} \leq u$, $\Theta_0'(\hat{\delta}) := \{\theta \in \Theta^\dagger : \psi(\theta) - u \geq \frac{1}{2}\hat{\delta},\ \psi(\theta) - l \leq \hat{\delta}\}$ must be a subset of $\Theta_0(\hat{\delta}, \hat{p}_1, \ldots, \hat{p}_n)$. From the definition of $\Theta_0'(\hat{\delta})$, there exist some positive constants $x_0$, $y_0$, $\epsilon$ such that $\Theta_1(\hat{\delta}) := [x_0 - \frac{1}{2}\epsilon\hat{\delta}, x_0 + \frac{3}{2}\epsilon\hat{\delta}] \times [y_0 - \frac{1}{2}\epsilon\hat{\delta}, y_0 + \frac{3}{2}\epsilon\hat{\delta}] \subseteq \Theta_0'(\hat{\delta})$. Then we define a prior distribution $q(\cdot)$ for the unknown parameter $\theta$ on the set $\Theta_1(\hat{\delta})$, whose expression is the same as (D.31). The remaining proof is similar to that of (D.31) as long as when applying the van Trees inequality, we consider the expectation $\mathbb{E}_q[\mathbb{E}_\theta^\pi[(p_t - \psi(\theta))^2 | \hat{p}_1, \ldots, \hat{p}_n]]$ by conditioning on the realization of $\hat{p}_1, \ldots, \hat{p}_n$. In particular, although the Fisher information function $\mathcal{I}(q)$ depends on the historical prices $\hat{p}_1, \ldots, \hat{p}_n$, since the support of $q(\cdot)$ is independent of $\hat{p}_1, \ldots, \hat{p}_n$ and the function $C(\theta)$ and its derivative are bounded, we can verify that $\mathcal{I}(q) = \Theta(\hat{\delta}^{-2})$ with the hidden constant independent of the realization of $\hat{p}_1, \ldots, \hat{p}_n$.

**Lower bound: Step 2.** In this step, we complete the proof by showing that when $\mathbb{E}_\theta^{\hat{\pi}}[\hat{\delta}(\theta)] \lesssim T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}}$ and $\mathbb{E}_\theta^{\hat{\pi}}[n\hat{\sigma}^2] \lesssim \frac{\sqrt{T}}{\log T}$, then for any admissible policy $\pi \in \hat{\Pi}^\circ$,

$$\mathbb{E}_{\hat{p}_1, \ldots, \hat{p}_n}[R_\theta^\pi(T, \hat{p}_1, \ldots, \hat{p}_n)] = \Omega(\frac{\sqrt{T}}{\log T}). \tag{D.41}$$

To show (D.41), for any given realization of $\hat{p}_1, \ldots, \hat{p}_n$, we define two parameters $\theta_1$ and $\theta_2$ satisfying

$$-\frac{\alpha_1}{2\beta_1} = \bar{p}_{1:n} + \hat{\delta}, \quad -\frac{\alpha_2}{2\beta_2} = \bar{p}_{1:n} + \hat{\delta} + \Delta, \quad \alpha_1 - \alpha_2 = -(\beta_1 - \beta_2)\bar{p}_{1:n},$$

where $\Delta > 0$ is to be determined. Then we define two random variables

$$X = (\hat{D}_1, \ldots, \hat{D}_n, D_1, \ldots, D_n, p_1, \ldots, p_n)$$

and $Y = (\hat{p}_1, \ldots, \hat{p}_n)$. For any policy $\pi$, let $\mathbb{P}_i^\pi(X, Y)$ be the joint distribution of $(X, Y)$, $\mathbb{P}_i^\pi(X|Y)$ be the conditional probability measure of $X$ given $Y$, and $\mathbb{P}_i^\pi(X)$ be the marginal distribution of $X$, each of which is associated with the policy $\pi$ and demand parameter $\theta_i$, $i = 1, 2$. Then we have

$$\begin{aligned}
&\mathbb{E}_{Y\sim\mathbb{P}_1^\pi}\big[KL\big(\mathbb{P}_1^\pi(X|Y), \mathbb{P}_2^\pi(X|Y)\big)\big] \\
&\quad\le KL(\mathbb{P}_1^\pi(X, Y), \mathbb{P}_2^\pi(X, Y)) \\
&\quad= \frac{1}{2R^2}\Big(\sum_{i=1}^n \mathbb{E}_{\theta_1}^{\hat{\pi}}\big[\big((\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)\hat{p}_i\big)^2\big] + \sum_{t=1}^T \mathbb{E}_{\theta_1}^{\hat{\pi},\pi}\big[\big((\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)p_t\big)^2\big]\Big) \\
&\quad\le \frac{(\beta_1 - \beta_2)^2}{2R^2}\Big(n\mathbb{E}_{\theta_1}^{\hat{\pi}}[\hat{\sigma}^2] + 2\sum_{t=1}^T \mathbb{E}_{\theta_1}^{\hat{\pi},\pi}\big[\big(p_t - \psi(\theta_1)\big)^2\big] + 2T\mathbb{E}_{\theta_1}^{\hat{\pi}}[\hat{\delta}^2(\theta_1)]\Big),
\end{aligned} \tag{D.42}$$

where the first inequality holds since from the chain rule of KL divergence, $KL\big(\mathbb{P}_1^\pi(X, Y), \mathbb{P}_2^\pi(X, Y)\big) = KL\big(\mathbb{P}_1^\pi(Y), \mathbb{P}_2^\pi(Y)\big) + \mathbb{E}_{Y\sim\mathbb{P}_1^\pi}\big[KL\big(\mathbb{P}_1^\pi(X|Y), \mathbb{P}_2^\pi(X|Y)\big)\big]$, and $KL\big(\mathbb{P}_1^\pi(Y), \mathbb{P}_2^\pi(Y)\big) \ge 0$.

On the other hand, by applying Theorem 2.2 in Tsybakov (2009) and using the fact that $\pi$ is admissible, we have

$$\begin{aligned}
\frac{1}{32}e^{-KL(\mathbb{P}_1^\pi(X|Y), \mathbb{P}_2^\pi(X|Y))} \cdot T\Delta^2 &\le \sum_{t=1}^T \mathbb{E}_{\theta_1}^\pi[(p_t - \psi(\theta_1))^2|\hat{p}_1, \ldots, \hat{p}_n] + \sum_{t=1}^T \mathbb{E}_{\theta_2}^\pi[(p_t - \psi(\theta_1))^2|\hat{p}_1, \ldots, \hat{p}_n] \\
&\le 2K_0\sqrt{T}\log T.
\end{aligned} \tag{D.43}$$

Taking the expectation over $Y$ on both sides of (D.43), we conclude from Jensen's inequality that

$$\mathbb{E}_{Y\sim\mathbb{P}_1^\pi}[KL(\mathbb{P}_1^\pi(X|Y), \mathbb{P}_2^\pi(X|Y))] \ge \log\big(\frac{\sqrt{T}\Delta^2}{64K_0\log T}\big),$$

With inequalities (D.42), the remaining analysis is similar to Step 2 in the proof of Theorem 5.4 and therefore is omitted. □

## D.7 Multi-Armed Bandits with Offline Data

In this section of the appendix, we discuss "MAB with offline data" and show the optimal regret rate exhibits phase transitions by deploying results from Shivaswamy and Joachims (2012) and Gur and Momeni (2022). We also compare the MAB problem with the `OPOD` problem studied in this paper.

Consider a $K$-armed bandit, where the seller chooses arms from set $\{1, 2, \ldots, K\}$ for each period $t \in [T]$. The distribution of reward for each arm $i$ is sub-Gaussian, denoted by $\mathcal{D}_i$, with the mean value $\mu_i$, $i \in [K]$. Let $i^\star$ be the arm with the highest mean reward, i.e., $\mu_{i^\star} = \max\{\mu_i : i \in [K]\}$, $\Delta_i$ be the sub-optimality gap for each arm $i \ne i^\star$, i.e., $\Delta_i = \mu_{i^\star} - \mu_i$, and $\Delta$ be a lower bound such that $0 < \Delta \le \min\{\Delta_i : i \in [K], i \ne i^\star\}$. We denote $\mathcal{S} = (\Delta, \mathcal{D}_1, \ldots, \mathcal{D}_K)$ as the class that includes any possible latent rewards distributions with the lower bound $\Delta$.

We assume that before the start of online learning, there are some pre-existing offline data, which consists of $H_i$ observations of random rewards for each arm $i \in [K]$. The decision maker can use the offline data as well as online data to make online decisions. For any given latent distributions of rewards $\mathcal{D} := (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K)$, we define the regret of any learning policy $\pi$ as the worst-case difference between the expected rewards generated by the optimal clairvoyant policy and the policy $\pi$: $R^\pi(T) = \sup_{\mathcal{S}} \left\{ T\mu_{i^\star} - \mathbb{E}^\pi_{\mathcal{D}}\left[\sum_{t=1}^T \mu_{\pi_t}\right]\right\}$, where the operator $\mathbb{E}^\pi_{\mathcal{D}}[\cdot]$ denotes the expectation induced by the policy $\pi$ and the latent distribution $\mathcal{D}$. The optimal regret is defined as $R^\star(T) = \inf_\pi R^\pi(T)$, which naturally depends on the number of offline observations $H_1, H_2, \ldots, H_K$, and therefore, is also denoted as $R^\star(T, H_1, \ldots, H_K)$.

We first present the upper and lower bounds on the optimal regret for the $K$-armed bandit with offline data in the following proposition, where the regret upper bound is provided in Theorem 2 of Shivaswamy and Joachims (2012), and the regret lower bound is implied from Theorem 1 of Gur and Momeni (2022).

**Proposition D.4.** *There exist positive constants $C_1, C_2, C_3, C_4$ such that the optimal regret satisfies*

$$R^\star(T, H_1, \ldots, H_K) \le \sum_{i=1}^K \Delta_i \Big( \big( \frac{8\log(T+H_i)}{\Delta_i^2} - H_i \big)^+ + C_1 \Big), \tag{D.44}$$

$$R^\star(T, H_1, \ldots, H_K) \ge C_2 \sum_{i=1}^K \Delta \Big( \frac{\log T}{\Delta^2} - C_3 H_i + \frac{1}{\Delta^2} \log \frac{C_4 \Delta^2}{K} \Big)^+. \tag{D.45}$$

Note that the regret lower bound in (D.45) is nontrivial only when $\Delta \gtrsim T^{-\frac{1}{2}}$. Combining (D.44) and (D.45), we discover the following phase transitions for $K$-armed bandits under the assumption of $\Delta = \Omega(T^{-\frac{1}{2}})$: the optimal regret rate in $K$-armed bandits decreases from $\Theta(\frac{\log T}{\Delta})$ to constant when the number of offline observations for each arm exceeds $\Theta(\frac{\log T}{\Delta^2})$. See Figure D-1 for illustration.

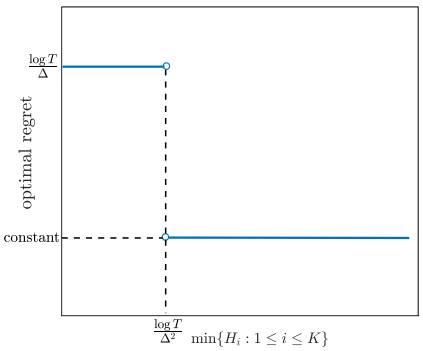


Figure D-1: Phase transition in $K$-armed bandits with offline data when $\Delta = \Omega(T^{-\frac{1}{2}})$.

There are three key differences between the $K$-armed bandit with offline data and the `OPOD` problem considered in this paper. First, in the $K$-armed bandit, the phase transition of the optimal regret only occur when the amount of offline data for *each* arm exceeds the threshold $\Theta(\frac{\log T}{\Delta^2})$; in other words, the offline data need to be *balanced* among different arms. By contrast, in the `OPOD` problem, even when there is only one historical price, i.e., $\sigma = 0$, the optimal regret rate can drop

from $\widetilde{\Theta}(\sqrt{T})$ to $\widetilde{\Theta}(\frac{\log T}{\delta^2})$ as the amount of offline data $n$ increases. The reason is that in $K$-armed bandit, different arms are independent, and the knowledge about the reward of one arm does not help to understand that of another. However, in dynamic pricing, the demands under different prices are connected with each other by the parametric assumption of the linear demand function. Therefore, knowing even one point at the demand curve can lead to a significant decrease in the optimal regret rate (see our Corollary 5.1 when $n = \infty$, or the incumbent price setting in Keskin and Zeevi 2014). Second, in the $K$-armed bandit, the optimal regret rate only shows two phases. In the first phase when the amount of the offline data is small, i.e., $\min\{H_i : i \in [K]\} = O(\frac{\log T}{\Delta^2})$, the optimal regret is always $\Theta(\frac{\log T}{\Delta})$ and the offline data do not help to reduce the optimal regret. In the second phase when the amount of the offline data is large, i.e., $\min\{H_i : i \in [K]\} = \Omega(\frac{\log T}{\Delta^2})$, the optimal regret becomes a constant. In the `OPOD` problem, however, the optimal regret gradually changes as the size of the offline data increases, experiencing different phase transitions depending on the magnitude of $\sigma$ and $\delta$. Third, under the so-called "well-separated" condition in bandits, where the sub-optimality gap $\Delta$ is a constant independent of $T$, the $K$-armed bandit exhibits *weak* phase transition in the sense that the drop of the optimal regret rate is within $\log T$. By comparison, under our well-separated condition , i.e., $\delta$ is a constant independent of $T$, the `OPOD` problem shows *strong* phase transitions in the sense that the drops of the optimal regret rate in multiple phases are measured in $T^\kappa$ for some $\kappa > 0$, which are much more significant even if we ignore the logarithmic factor.

## D.8 Tables in Sections 5.3 and 5.4

Table D.1: Regret upper bound in Theorem 5.1 for the single-historical-price setting.

| **Case 1:** $\delta \gtrsim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ | | | |
|---|---|---|---|
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\delta^2}$ | $\frac{\sqrt{T}\log T}{\delta^2} \lesssim n \lesssim T$ | $n \gtrsim T$ |
| upper bound | $\mathcal{O}(\sqrt{T}\log T)$ | $\mathcal{O}(\frac{T(\log T)^2}{n\delta^2})$ | $\mathcal{O}(\frac{(\log T)^2}{\delta^2})$ |
| **Case 2:** $\delta \lesssim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ | | | |
| offline sample size | $n \ge 0$ | | |
| upper bound | $\mathcal{O}(\sqrt{T}\log T)$ | | |

Table D.2: Regret lower bound in Theorem 5.2 for the single-historical-price setting.

| | | | |
|---|---|---|---|
| **Case 1:** $\delta \gtrsim T^{-\frac{1}{4}}$ | | | |
| offline sample size | $0 < n \lesssim \frac{\sqrt{T}}{\delta^2}$ | $\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim T$ | $n \gtrsim T$ |
| lower bound | $\Omega(\sqrt{T})$ | $\Omega(\frac{T}{n\delta^2} \vee \log T)$ | $\Omega(\frac{1}{\delta^2} \vee \log T)$ |
| **Case 2:** $T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}} \lesssim \delta \lesssim T^{-\frac{1}{4}}$ | | | |
| offline sample size | $n > 0$ | | |
| lower bound | $\Omega(T\delta^2)$ | | |
| **Case 3:** $\delta \lesssim T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}}$ | | | |
| offline sample size | $n > 0$ | | |
| lower bound | $\Omega(\frac{\sqrt{T}}{\log T})$ | | |

Table D.3: Regret upper bound in Theorem 5.3 for the multiple-historical-price setting.

| | | | | |
|---|---|---|---|---|
| **Case 1:** $\delta \gtrsim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ and $\sigma \lesssim \delta$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\delta^2}$ | $\frac{\sqrt{T}\log T}{\delta^2} \lesssim n \lesssim T$ | $T \lesssim n \lesssim \frac{T\delta^2}{\sigma^2}$ | $n \gtrsim \frac{T\delta^2}{\sigma^2}$ |
| upper bound | $\mathcal{O}(\sqrt{T}\log T)$ | $\mathcal{O}(\frac{T(\log T)^2}{n\delta^2})$ | $\mathcal{O}(\frac{(\log T)^2}{\delta^2})$ | $\mathcal{O}(\frac{T(\log T)^2}{n\sigma^2}+1)$ |
| **Case 2:** $\delta \gtrsim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ and $\sigma \gtrsim \delta$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\sigma^2}$ | | $n \gtrsim \frac{\sqrt{T}\log T}{\sigma^2}$ | |
| upper bound | $\mathcal{O}(\sqrt{T}\log T)$ | | $\mathcal{O}(\frac{T(\log T)^2}{n\sigma^2}+1)$ | |
| **Case 3:** $\delta \lesssim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\sigma^2}$ | $\frac{\sqrt{T}\log T}{\sigma^2} \lesssim n \lesssim \frac{(\log T)^2}{\sigma^2\delta^2}$ | $n \gtrsim \frac{(\log T)^2}{\sigma^2\delta^2}$ | |
| upper bound | $\mathcal{O}(\sqrt{T}\log T)$ | $\mathcal{O}(T\delta^2+1)$ | $\mathcal{O}(\frac{T(\log T)^2}{n\sigma^2}+1)$ | |

Table D.4: Regret lower bound in Theorem 5.4 for the multiple-historical-price setting.

| | | | | |
|---|---|---|---|---|
| **Case 1:** $\delta \gtrsim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ and $\sigma \lesssim \delta$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\delta^2}$ | $\frac{\sqrt{T}\log T}{\delta^2} \lesssim n \lesssim T$ | $T \lesssim n \lesssim \frac{T\delta^2}{\sigma^2}$ | $n \gtrsim \frac{T\delta^2}{\sigma^2}$ |
| lower bound | $\Omega(\frac{\sqrt{T}}{\log T})$ | $\Omega(\frac{T}{n\delta^2})$ | $\Omega(\frac{1}{\delta^2})$ | $\Omega(\frac{T}{n\sigma^2})$ |
| **Case 2:** $\delta \gtrsim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ and $\sigma \gtrsim \delta$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\sigma^2}$ | | $n \gtrsim \frac{\sqrt{T}\log T}{\sigma^2}$ | |
| lower bound | $\Omega(\frac{\sqrt{T}}{\log T})$ | | $\Omega(\frac{T}{n\sigma^2})$ | |
| **Case 3:** $T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}} \lesssim \delta \lesssim T^{-\frac{1}{4}}(\log T)^{\frac{1}{2}}$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{1}{\sigma^2\delta^2}$ | | $n \gtrsim \frac{1}{\sigma^2\delta^2}$ | |
| lower bound | $\Omega(\frac{\sqrt{T}}{\log T})$ | | $\Omega(\frac{T}{n\sigma^2})$ | |
| **Case 4:** $\delta \lesssim T^{-\frac{1}{4}}(\log T)^{-\frac{1}{2}}$ | | | | |
| offline sample size | $0 \le n \lesssim \frac{\sqrt{T}\log T}{\sigma^2}$ | $\frac{\sqrt{T}\log T}{\sigma^2} \lesssim n \lesssim \frac{1}{\sigma^2\delta^2}$ | $n \gtrsim \frac{1}{\sigma^2\delta^2}$ | |
| lower bound | $\Omega(\frac{\sqrt{T}}{\log T})$ | $\Omega(T\delta^2)$ | $\Omega(\frac{T}{n\sigma^2})$ | |

# Appendix E

# Supplementary Material for Chapter 6

## E.1 Results on Distribution-Dependent Regret Bounds

For simplicity, we only present the results of distribution-dependent regret bounds for the U-BwSC problem. Extensions to general switching cost structures are analogous to Section 6.4 of the main article.

To achieve tight distribution-dependent regret bounds, we propose the LS-SE2 algorithm and the AdaLS2 algorithm, which are stated in Algorithm E.1 and Algorithm E.2 respectively. Note that the difference between the new algorithms and the original algorithms in Section 6.3 is only on the epoch schedules (which are optimized for distribution-dependent regret).

**Algorithm E.1** Limited-Switch Successive Elimination 2(LS-SE2)

**Input:** Switching budget $S$, number of actions $K$, horizon $T$.

**Initialization:** Compute $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$. Divide the entire time horizon $T$ into $q(S,K)+1$ epochs: $(t_0 : t_1], (t_1 : t_2], \ldots, (t_{q(S,K)} : t_{q(S,K)+1}]$, where the endpoints are defined by $t_0 = 0$ and

$$t_j = \left\lfloor K^{1-\frac{j}{q(S,K)+1}} T^{\frac{j}{q(S,K)+1}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+1.$$

Let $A_1 = [K]$. Let $a_0$ be a random action in $[K]$.

**Policy:** The same as Lines 1 to 10 of Algorithm 6.1.

For any environment $\mathcal{D}$, let $k^\star = \arg\max_{k\in[K]} \mu_k$ denote the optimal action, and $\Delta = \Delta(\mathcal{D}) = \min_{k\neq k^\star} |\mu_{k^\star} - \mu_k| > 0$ denote the gap between the rewards of the optimal action and the best sub-optimal action. We have the following upper and lower bounds on regret.

**Proposition E.1.** *Let $\pi$ be the LS-SE2 policy. There exists an absolute constant $C \geq 0$ such that for all $\mathcal{D}$, for all $K > 1$, $S \geq 0$ and $T \geq K$,*

$$R^\pi_{\mathcal{D}}(K,T) \leq C\left(K^{1-\frac{1}{q(S,K)+1}} \log K\right) \frac{T^{\frac{1}{q(S,K)+1}} \log T}{\Delta},$$

*where $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$.*

**Algorithm E.2** Adaptive Limited-Switch Policy 2 (AdaLS2)

**Input:** Switching budget $S$, number of actions $K$, horizon $T$, tuning parameter $\lambda = 1/2$.

**Initialization:** Compute $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$ and $r(S,K) = (S-1)\%(K-1)$. Define $\widehat{r}(S,K) = \max\{r(S,K) + 1 - q(S,K), 0\}$. Define $T_0^{(1)} = T_0^{(2)=0}$, $t_0^{(1)} = t_0^{(2)} = 0$ and

$$t_j^{(1)} = \left\lfloor (K - \widehat{r}(S,K))^{1-\frac{j}{q(S,K)+1}} T^{\frac{j}{q(S,K)+1}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+1,$$

$$t_j^{(2)} = \left\lfloor K^{1-\frac{j}{q(S,K)+2}} T^{\frac{j}{q(S,K)+2}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+2.$$

Let $A_1 = [K]$. Let $A_1^{(2)}$ be a subset of $A_1$ obtained by uniformly sampling $\widehat{r}(S,K)$ actions from $A_1$ *without replacement* (thus $|A_1^{(2)}| = \widehat{r}(S,K)$). Let $A_1^{(1)} = A_1 \setminus A_1^{(2)}$. Let $a_0$ be a random action in $A_1^{(1)}$.

**Policy:** The same as Lines 1 to 16 of Algorithm 6.2.

**Theorem E.1.** *Let $\pi$ be the AdaLS2 policy. There exists an absolute constant $C \geq 0$ such that for all $\mathcal{D}$, for all $K > 1$, $S \geq 0$ and $T \geq K$,*

$$R_{\mathcal{D}}^{\pi}(K,T) \leq C(\log K \log T) \cdot \max\left\{ \frac{(K - r(S,K))^{2-\frac{1}{q(S,K)+1}}}{K} \frac{T^{\frac{1}{q(S,K)+1}}}{\Delta},\ K^{1-\frac{1}{q(S,K)+1}} \frac{T^{\frac{1}{q(S,K)+1}}}{\Delta} \right\},$$

*where $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$ and $r(S,K) = (S-1)\%(K-1)$.*

**Theorem E.2.** *There exists an absolute constant $C > 0$ such that for all $K > 1, S \geq 0, T \geq 2K$ and for all policy $\pi \in \Pi_S$,*

$$\sup_{\Delta \in [0,1]} \Delta R_{\mathcal{D}}^{\pi}(K,T) \geq \frac{C}{\log T} \cdot \max\left\{ \frac{(K - r(S,K))^{2-\frac{1}{q(S,K)+1}}}{K} T^{\frac{1}{q(S,K)+1}},\ K^{1-\frac{1}{q(S,K)+2}} T^{\frac{1}{q(S,K)+2}} \right\},$$

*where $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$ and $r(S,K) = (S-1)\%(K-1)$.*

Note that the upper bound in Theorem E.1 and the lower bound in Theorem E.2 match in the minimax sense (up to logarithmic factors), which implies that the AdaLS2 algorithm can be considered as near-optimal. We thus characterize the distribution-dependent complexity of the U-BwSC problem. We also note that when $S = \Omega(K \log T)$, both LS-SE2 and AdaLS2 algorithms recover the well-known $\mathcal{O}\left(\frac{K \log T}{\Delta}\right)$ distribution-dependent regret bound of the classical MAB (up to a $\log K$ factor), which is shown to be rate-optimal (Lai and Robbins 1985).

We omit the proofs of above results: the proof of Proposition E.1 resembles the proof of Proposition 6.1 in Appendix E.7, the proof of Theorem E.1 resembles the proof of Theorem 6.1 in Appendix E.8, and the proof of Theorem E.2 resembles the proof of Theorem 6.2 in Appendix E.12. The difference is mainly on the partition of epochs.

Besides results on regret upper and lower bounds, we also establish Corollary E.1, which can be viewed as a counterpart of Corollary 6.2 in Section 6.3.3 of the main article.

**Corollary E.1.** *For any $K > 1$, for any environment $\mathcal{D}$, let $\Delta = \min_{k \in [K], k \neq k^\star} |\mu_{k^\star} - \mu_k|$ denote the gap between the mean rewards of the optimal action and the best sub-optimal action.*

1. *$N(K-1)+1$ switches are necessary and sufficient for uniformly achieving $\widetilde{\mathcal{O}}(KT^{\frac{1}{N+1}}/\Delta)$ distribution-dependent regret for all $\mathcal{D}$ in the $K$-armed MAB ($N \in \mathbb{Z}_{>0}$).*
2. *$\Omega(\frac{K \log T}{\log \log T})$ switches are necessary for uniformly achieving $\widetilde{\mathcal{O}}(K \log T/\Delta)$ distribution-dependent regret for all $\mathcal{D}$ in the $K$-armed MAB.*

## E.2 Rounding Issues of Algorithms

We present a more rigorous version of Algorithm 6.1, which takes care of the rounding issues in Line 4 and Line 7 of Algorithm 6.1. The key idea is to maintain a "after-rounding" epoch schedule $(T_j)_{j=0}^{q(S,K)+1}$ which is slightly different from the original epoch schedule $(t_j)_0^{q(S,K)+1}$; see Algorithm E.3. In the proof of Theorem 6.1, we will directly analyze Algorithm E.3.

We remark that the main algorithm of our article, the AdaLS algorithm (Algorithm 6.2), has *already* taken care of the rounding issues (using a similar idea). The other two algorithms HS-SE and AS-SE omit the rounding issues — they can be easily modified to incorporate the rounding issues, using exactly the same idea as Algorithm E.3.

---

**Algorithm E.3** Limited-Switch Successive Elimination (LS-SE)

---

**Input:** Switching budget $S$, number of actions $K$, horizon $T$.

**Initialization:** Compute $q(S,K) = \left\lfloor \frac{S-1}{K-1} \right\rfloor$. Define $t_0 = 0$ and

$$t_j = \left\lfloor K^{1-\frac{2-2^{-(j-1)}}{2-2^{-q(S,K)}}} T^{\frac{2-2^{-(j-1)}}{2-2^{-q(S,K)}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+1.$$

Let $A_1 = [K]$. Let $a_0$ be a random action in $[K]$. Let $T_0 = 0$.

**Policy:**

1: **for** $l = 1, \ldots, q(S,K)$ **do**
2:  **if** $a_{T_{l-1}} \in A_l$ **then**
3:  **for** $i = a_{T_{l-1}}$ and then $i \in A_l \setminus \{a_{T_{l-1}}\}$ **do** ▷ starting from $i = a_{T_{l-1}}$ is critical
4:  Choose action $i$ for $\left\lfloor \frac{t_l - T_{l-1}}{|A_l|} \right\rfloor$ *consecutive* rounds.
5:  **else**
6:  **for** $i \in A_l$ **do**
7:  Choose action $i$ for $\left\lfloor \frac{t_l - T_{l-1}}{|A_l|} \right\rfloor$ *consecutive* rounds.
8:  Mark the last round as $T_l$, and mark the last chosen action as $a_{T_l}$. ▷ record $T_l$
9:  Elimination: compute $\mathsf{UCB}_i(T_l)$ and $\mathsf{LCB}_i(T_l)$ for all $i \in A_l$ and let ▷ learn from data

$$A_{l+1} = \left\{ i \in A_l \mid \mathsf{UCB}_i(T_l) \geq \max_{j \in A_l} \mathsf{LCB}_j(T_l) \right\}.$$

10: For $l = q(S,K)+1$, find an action $i \in A_l$ that maximizes $\bar{\mu}_i(T_{l-1})$. Keep choosing this action until round $T$. Let $T_{q(S,K)+1} := T$.

---

## E.3 Illustration of AdaLS

We use the example of $S = 2K - 2$ to illustrate how AdaLS utilizes the switching budget more efficiently than LS-SE. In this case, $q(S, K) = 1$ and $r(S, K) = K - 2$. The LS-SE algorithm will observe the data only once throughout the entire horizon, and makes at most $K$ switches.

How does AdaLS behave in this case? At initialization, AdaLS computes $\widehat{r}(S, K) = \max\{r(S, K) + 1 - q(S, K), 0\} = K - 2 + 1 - 1 = K - 2$. The algorithm then randomly splits $[K]$ into two subsets $A_1^{(1)}$ and $A_1^{(2)}$, with $|A_1^{(1)}| = K - \widehat{r}(S, K) = 2$ and $|A_1^{(2)}| = \widehat{r}(S, K) = K - 2$. Then, in the execution of the policy, AdaLS treats the actions in $A_1^{(1)}$ and $A_1^{(2)}$ differently, allowing the actions in $A_1^{(2)}$ to be explored more frequently than the actions in $A_1^{(1)}$. Specifically, in the first epoch, AdaLS explores all actions in $[K]$ and makes $K - 1$ switches; then, in the second epoch, AdaLS first explores all uneliminated actions in $A_1^{(2)}$ (which incurs at most $(K - 2) - 1 = K - 3$ switches), and finally commits to a single action (which incurs at most 1 switch). Note that AdaLS may also incur a switch between the first and second epochs, so its total number of switches is at most $(K - 1) + 1 + (K - 3) + 1 = 2K - 2$. Clearly, compared with LS-SE which only makes $K$ switches, AdaLS makes much better use of the switching budget in this case.

## E.4 Reverse Fano-Type Inequalities and Lower Bound Analysis

This section provides an overview of the methodological contributions associated with our proof of Theorem 6.2. We first introduce the GRF inequality, then present our lower bound approach.

### E.4.1 Reverse Fano-Type Inequalities

*Fano's* inequality is a fundamental information-theoretical tool for developing algorithm-independent impossibility results in statistics and machine learning. In one of its most classical forms, it states that for any sequence of $N \geq 2$ probability measures $\mathbb{P}_1, \dots \mathbb{P}_N$ on the same measurable space $(\Omega, \mathcal{F})$, and any sequence of events $E_1, \cdots, E_N$ forming a partition of $\Omega$, it holds that

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(E_i) \leq \frac{\frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}(\mathbb{P}_i \,\|\, \mathbb{Q}) + \log 2}{\log N}, \tag{E.1}$$

where $\mathbb{Q}$ is an arbitrary measure on $(\Omega, \mathcal{F})$, and $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ stands for the *KL divergence*. Fano's inequality has important consequences for various problems in various fields; see Scarlett and Cevher (2019) for a survey. For example, in multiple hypothesis testing, by considering events of the form $E_i = \{\psi = i\}$ where $\psi : \Omega \to [N]$ is a *test*, (E.1) provides a sharp lower bound on the average error probability $\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi \neq i)$ for any test $\psi$.

Many variants of Fano's inequality have been derived in the literature; see Scarlett and Cevher (2019) and Gerchinovitz et al. (2020) for overviews. However, to our knowledge, existing literature does not provide a reverse version of (E.1), i.e., an inequality that establishes a sharp lower bound on $\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(E_i)$ **for any** $E_1, \dots, E_N$ forming a partition, which corresponds to a sharp upper bound on $\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi \neq i)$ for any test $\psi$ in multiple hypothesis testing. While there are indeed some existing inequalities sometimes referred to as "reverse Fano's inequalities" in the literature (e.g., Chu

and Chueh 1966, Tebbe and Dwyer 1968), and some other related inequalities are implied by the recent work of Gerchinovitz et al. (2020), these inequalities either fail to lower bound $\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(E_i)$ **for any** $E_1,\dots,E_N$ forming a partition, or suffer from sub-optimal dependence on $N$; see Appendix E.11.1 for detailed discussions. We fill this gap by developing a reverse version of (E.1) and significantly generalizing it to a much stronger version, i.e., the GRF inequality; see Proposition E.2. The proof builds on the general framework developed by Gerchinovitz et al. (2020), with some new techniques to obtain better dependence on $N$ via localized versions of Pinsker's inequality; see Appendix E.11.

**Proposition E.2** (Generalized Reverse Fano-Type Inequality)**.** *Let $D(\cdot\,\|\,\cdot)$ be the KL divergence or the reverse KL divergence (see Appendix E.11 for definitions). Let $(\Omega_1,\mathcal{F}_1),\dots,(\Omega_N,\mathcal{F}_N)$ be an arbitrary sequence of measurable spaces. For any $i\in[N]$, let $\mathbb{P}_i$ and $\mathbb{Q}_i$ be arbitrary probability measures on $(\Omega_i,\mathcal{F}_i)$, and $E_i\in\mathcal{F}_i$ be an arbitrary event. We have*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(E_i)\geq\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i)-\sqrt{2\cdot\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i)\cdot\frac{1}{N}\sum_{i=1}^{N}D(\mathbb{P}_i\,\|\,\mathbb{Q}_i)}. \tag{E.2}$$

Proposition E.2 is fairly general and enjoys several advantages: (i) it acts in the reverse direction of (E.1), thus enables new applications, (ii) the events $E_1,\dots,E_N$ do **not** need to form a partition, (iii) the probability measures $\mathbb{P}_1,\dots\mathbb{P}_N$ can be defined on different measurable spaces, (iv) $D(\cdot\,\|\,\cdot)$ can be the *reverse* KL divergence, and (v) the probability measures $\mathbb{Q}_1,\cdots,\mathbb{Q}_N$ can vary with events and do not need to be fixed. All of the above advantages will be utilized in our lower bound analysis.

### E.4.2 The Five-Step Approach to Establish Theorem 6.2

Given any $K>1$, $S\geq 0$ and $T\geq 2K$, we focus on the setting of $\mathcal{D}_k=\mathcal{N}(\mu_k,1)$, $\forall k\in[K]$, where $\mathcal{N}(\mu_k,1)$ denotes the Gaussian distribution with mean $\mu_k$ and variance 1. Since in this setting the underlying environment (i.e., reward distributions) $\mathcal{D}$ is completely specified by a vector $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_K)\in\mathbb{R}^K$, for simplicity, we directly use the vector $\boldsymbol{\mu}$ to represent the environment.

**Notation.** For any environment $\boldsymbol{\mu}$, let $X^t_{\boldsymbol{\mu}}(k)\sim\mathcal{N}(\mu_k,1)$ denote the i.i.d. random reward of each action $k$ at round $t$ ($k\in[K],t\in[T]$). For any policy $\pi\in\Pi_S$, for any environment $\boldsymbol{\mu}$, for any $t\in[T]$, we use $a_t$ and $X^t_{\boldsymbol{\mu}}(a_t)$ to denote the random action selected by and the random reward observed by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$, respectively. Let $\mathbb{P}^{\pi}_{\boldsymbol{\mu}}$ denote the law of the random variables $\big(a_1,X^1_{\boldsymbol{\mu}}(a_1)\big),\dots,\big(a_T,X^T_{\boldsymbol{\mu}}(a_T)\big)$, and let $\mathbb{E}^{\pi}_{\boldsymbol{\mu}}$ be the associated expectation operator. Let $R^{\pi}_{\boldsymbol{\mu}}(T):=T\mu^\star-\mathbb{E}^{\pi}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T}\mu_{a_t}\right]$ be policy $\pi$'s distribution-dependent regret under environment $\boldsymbol{\mu}$.

**Outline.** Since the desired lower bound (6.3) becomes the standard $\widetilde{\Omega}(\sqrt{KT})$ lower bound when $S=\Omega(K\log\log T)$ (see Appendix E.12), we focus on the more interesting case of $S=\mathcal{O}(K\log\log T)$. For any such $S,K,T$, we seek to explicitly construct a family of environments $\Phi$, such that for any $S$-switch policy $\pi\in\Pi_S$, the "average-case regret" $\frac{1}{|\Phi|}\sum_{\boldsymbol{\mu}\in\Phi}R^{\pi}_{\boldsymbol{\mu}}(T)$ is lower bounded by (6.3) — this implies that the worst-case regret $R^{\pi}(T)$ is also lower bounded by (6.3). In our proof, we construct two classes of environments to show the two parts in the "max" of (6.3) respectively. In this section, we focus on the more challenging part — the $\widetilde{\Omega}\left(\frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K}T^{\frac{1}{2-2^{-q(S,K)}}}\right)$ lower bound.

Our lower bound proof program consists of five steps:

1. **R**isky **E**vents
2. **C**ombinatorial arguments and lower bounds under a single environment
3. **A**lternative environments, bad events, and lower bound reductions (we construct $\Phi$ here)
4. **P**robability space changing tricks
5. Applying the GRF inequality

Based on the initials of the first four steps, we call the program RECAP. We provide an overview of each step below. The detailed proof can be found in Appendix E.12.

**Step 1: risky events.** We first define a stopping time $\tau$, which is the first round that the learner's number of switches reaches $S$. We then define a class of *risky events* as follows: for any $k \in [K]$, let

$$E_{1,k}^{(1)} := \Big\{\text{action } k \text{ is not chosen in period } \Big[1 : t_1^{(1)}\Big]\Big\},$$
$$E_{j,k}^{(1)} := \Big\{\text{action } k \text{ is not chosen in period } \Big[t_{j-1}^{(1)} : t_j^{(1)}\Big]\Big\}, \quad \forall j \in [2 : q(S,K)],$$
$$E_{q(S,K)+1,k}^{(1)} := \Big\{\text{action } k \text{ is not chosen in period } \Big[t_{q(S,K)}^{(1)} : \lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor\Big]\Big\},$$
$$E_{q(S,K)+2,k}^{(1)} := \Big\{\tau \le \lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor,\ a_\tau = k,\ \text{action } k \text{ is not chosen in period } \Big[t_{q(S,K)}^{(1)} : \tau - 1\Big]\Big\}.$$

By doing so, we get $(q(S,K)+2)K$ risky events (of the form $E_{j,k}^{(1)}$) in total. Note that the time points $(t_j^{(1)})_{j=1}^{q(S,K)+1}$ are fixed and given in (E.45), which are closely related to but slightly different from the time points $(t_j^{(1)})_{j=1}^{q(S,K)+1}$ defined in Algorithm 6.2; see Footnote 65. Moreover, the events $(E_{q(S,K)+2,k}^{(1)})_{k=1}^{K}$ are defined based on the stopping time $\tau$. Such delicate design based on $\tau$ is novel and crucial; see the remark in Appendix E.12.1.

The risky events characterize some important patterns that are "unavoidable" (in a certain sense, as we shall see in Step 2) for any $S$-switch policy under any environment. They are considered "risky" because, while they may not directly lead to large regret under an arbitrary environment, each of them would lead to larger regret under a specifically chosen environment (i.e.., the *alternative* environment in Step 3) which will be included in our environment class $\Phi$.

**Step 2: combinatorial arguments and lower bounds under a single environment.** In this step, we prove a key result (Lemma E.1) using non-trivial combinatorial and probabilistic arguments. The arguments extensively exploit the properties of the switching constraint.

**Lemma E.1.** *For any $S$-switch policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) \ge K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil = \widetilde{\Omega}\Big(\frac{K - r(S,K)}{K}\Big).$$

Lemma E.1 implies the following fact: under any *single* environment $\boldsymbol{\mu}$, the average probability of the risky events is $\widetilde{\Omega}\Big(\frac{K-r(S,K)}{K}\Big)$. That is, the occurrence of a risky event is "probabilistically unavoidable" for any $S$-switch policy under any single environment. This result precisely characterize the potential weakness of an $S$-switch policy and reveals fundamental properties of the switching constraint.

**Step 3: alternative environments, bad events, and lower bound reductions.** In this step, we construct our environment class $\Phi := \left\{\boldsymbol{\mu}_{j,k}^{(1)} \mid j \in [q(S,K)+2], k \in [K]\right\}$, which consists of $(q(S,K)+2)K$ judiciously chosen *alternative* environments. Each alternative environment $\boldsymbol{\mu}_{j,k}^{(1)}$ is designed to make the risky event $E_{j,k}^{(1)}$ become a *bad event* whose occurrence implies large regret. We can then reduce the task of proving a lower bound on the "average-case regret" $\frac{1}{|\Phi|}\sum_{\boldsymbol{\mu}\in\Phi} R_{\boldsymbol{\mu}}^{\pi}(T)$ to the task of proving a lower bound on the "average-case bad event probability" $\frac{1}{(q(S,K)+2)K}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]} \mathbb{P}_{j,k}^{(1)}\left(E_{j,k}^{(1)}\right)$, where $\mathbb{P}_{j,k}^{(1)} := \mathbb{P}_{\boldsymbol{\mu}_{j,k}^{(1)}}^{\pi}$ denotes the *alternative* measure associated with policy $\pi$ and alternative environment $\boldsymbol{\mu}_{j,k}^{(1)}$.

Specifically, our construction of alternative environments is as follows. Let $\mathbf{0} = (0,\ldots,0) \in \mathbb{R}^K$ be the *reference* environment. For any $j \in [q(S,K+2)]$, define a reward gap

$$\Delta_j^{(1)} := \begin{cases} 1, & \text{if } j = 1, \\ \frac{1}{2(q(S,K)+2)}\sqrt{\frac{K-r(S,K)}{t_{j-1}^{(1)}}}, & \text{if } j \in [2 : q(S,K)+1], \\ -\frac{1}{2(q(S,K)+2)}\sqrt{\frac{K-r(S,K)}{t_{q(S,K)}^{(1)}}}, & \text{if } j = q(S,K)+2. \end{cases}$$

For any $j \in [q(S,K)+2], k \in [K]$, define an *alternative* environment $\boldsymbol{\mu}_{j,k}^{(1)} := \left(\mu_{j,k;1}^{(1)},\ldots,\mu_{j,k;K}^{(1)}\right) \in \mathbb{R}^K$ where

$$\mu_{j,k;i}^{(1)} := \begin{cases} \Delta_j^{(1)}, & \text{if } i = k, \\ 0, & \text{otherwise.} \end{cases}$$

Note that each alternative environment only differs from the reference environment in one coordinate.

In Lemma E.9, we show that the risky event $E_{j,k}^{(1)}$ is indeed a bad event under environment $\boldsymbol{\mu}_{j,k}^{(1)}$, in the sense that its occurrence implies that the regret is larger than a universal quantity $\mathcal{R}_{\mathrm{bad}}(S,K,T) = \widetilde{\Omega}\Big((K-r(S,K))^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}}\Big)$. Thus, in order to prove the desired lower bound on $\frac{1}{|\Phi|}\sum_{\boldsymbol{\mu}\in\Phi} R_{\boldsymbol{\mu}}^{\pi}(T)$, it suffices to prove the following statement:

$$\overline{p^{(1)}} := \frac{1}{(q(S,K)+2)K}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]} \mathbb{P}_{j,k}^{(1)}\left(E_{j,k}^{(1)}\right) = \widetilde{\Omega}\left(\frac{K-r(S,K)}{K}\right), \tag{E.3}$$

**Step 4: probability space changing tricks.** Let $\mathbb{Q} := \mathbb{P}_{\mathbf{0}}^{\pi}$ denote the *reference* measure. By applying Lemma E.1 to the reference environment $\mathbf{0}$, we have

$$\overline{q^{(1)}} := \frac{1}{(q(S,K)+2)K}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]} \mathbb{Q}\left(E_{j,k}^{(1)}\right) = \widetilde{\Omega}\left(\frac{K-r(S,K)}{K}\right),$$

Therefore, in order to show (E.3), it suffices to show that $\overline{p^{(1)}}$ is close to $\overline{q^{(1)}}$. Note that $\overline{p^{(1)}}$ is the average of the sequence $\left\{\mathbb{P}_{j,k}^{(1)}\left(E_{j,k}^{(1)}\right)\right\}$, where a sequence of events $\left\{E_{j,k}^{(1)}\right\}$ are evaluated by a sequence of varying alternative measures $\left\{\mathbb{P}_{j,k}^{(1)}\right\}$, while $\overline{q^{(1)}}$ is the average of the sequence $\left\{\mathbb{Q}\left(E_{j,k}^{(1)}\right)\right\}$, where the same sequence of events $\left\{E_{j,k}^{(1)}\right\}$ are evaluated by a single and fixed reference measure $\mathbb{Q}$. Intuitively, we just need a "change of measure"/information-theoretic argument — if the alternative measures $\left\{\mathbb{P}_{j,k}^{(1)}\right\}$ are "close enough" to the reference measure $\mathbb{Q}$, then $\overline{p^{(1)}}$ is close to $\overline{q^{(1)}}$.

Unfortunately, it turns out that the divergence between $\left\{\mathbb{P}_{j,k}^{(1)}\right\}$ and $\mathbb{Q}$ is too large to make the above argument work. An important reason is that such an argument directly deals with the underlying measures $\left\{\mathbb{P}_{j,k}^{(1)}\right\}$ and $\mathbb{Q}$, thus completely overlooks the special structures of the risky event sequence $\left\{E_{j,k}^{(1)}\right\}$. Therefore, we need to integrate the structural properties of risky events into our argument. We develop *probability space changing tricks* to address this challenge. Specifically, we design two sequences of *artificial* measures $\left\{\mathbb{P}'_{j,k}\right\}$ and $\left\{\mathbb{Q}'_{j,k}\right\}$ based on the structural properties of $\left\{E_{j,k}^{(1)}\right\}$, such that (i) each $\mathbb{P}'_{j,k}$ (resp. $\mathbb{Q}'_{j,k}$) is the *restriction* of $\mathbb{P}_{j,k}^{(1)}$ (resp. $\mathbb{Q}$) to a carefully-chosen $\sigma$-algebra $\mathcal{F}'_{j,k}$ which *tightly* contains $E_{j,k}^{(1)}$, and (ii) the reverse KL divergence between $\mathbb{P}'_{j,k}$ and $\mathbb{Q}'_{j,k}$ is small enough. We can then represent $\overline{p^{(1)}}$ and $\overline{q^{(1)}}$ as the averages of $\left\{\mathbb{P}'_{j,k}(E_{j,k}^{(1)})\right\}$ and $\left\{\mathbb{Q}'_{j,k}(E_{j,k}^{(1)})\right\}$, and bound the difference between $\overline{p^{(1)}}$ and $\overline{q^{(1)}}$ by showing that $\left\{\mathbb{P}'_{j,k}\right\}$ and $\left\{\mathbb{Q}'_{j,k}\right\}$ are "close enough."

**Step 5: applying the GRF inequality.** In the last step, we apply the GRF inequality to provide a tight lower bound on $\overline{p^{(1)}}$ in terms of $\overline{q^{(1)}}$, thus completes the proof of (E.3) (and eventually Theorem 6.2). We remark that we thoroughly utilize the five advantages of the GRF inequality in this step: (i) we need to lower bound $\overline{p^{(1)}}$ rather than lower bound $1-\overline{p^{(1)}}$ (the latter is what classical Fano-type inequalities do), (ii) our events $\left\{E_{j,k}^{(1)}\right\}$ (or their complements) do not form a partition, (iii) our measures $\left\{\mathbb{P}'_{j,k}\right\}$ are defined on different measurable spaces, (iv) we need to use the *reverse* (rather than the standard) KL divergence to evaluate the "closeness" between $\left\{\mathbb{P}'_{j,k}\right\}$ and $\left\{\mathbb{Q}'_{j,k}\right\}$, as we need to fix the reference environment to characterize the policy's behavior, and (v) the artificial reference measures $\left\{\mathbb{Q}'_{j,k}\right\}$ are not fixed.

## E.5 Explanations for Section 6.4.1

This section contains some additional explanations for our results in Section 6.4.1.

### E.5.1 Relaxing the Triangle Inequality Assumption in Section 6.4.1

Consider an arbitrary switching graph $G$ with $K = |G| > 1$. In the following, we show that, even without the triangle inequality assumption, a modified version of the results in Section 6.4.1 still hold.

#### Construction of a New Switching Graph that Satisfies the Triangle Inequality

Assume that the switching costs associated with $G$ do not satisfy the triangle inequality. We then run the Floyd-Warshall algorithm (see Cormen et al. 2009) on $G$ to efficiently find the shortest paths between all pairs of vertices. For any $i, j \in [K]$ such that $i \neq j$, let $p_{i,j} = i \to \cdots \to j$ denote the shortest path between $i$ and $j$, and $c'_{i,j}$ denote the total weight of the shortest path between $i$ and $j$. We construct a new switching graph $G' = (V, E')$ — the vertices in $G'$ are the same as $G$, while the edge between $i$ and $j$ in $G'$ is assigned a weight $c'_{i,j}$, which is the total weight of the shortest path between $i$ and $j$ in $G$. Obviously, $G'$ is a switching graph whose switching costs satisfy the triangle

inequality. Therefore, for BwSC problems defined with $G'$, we can apply the HS-SE policy, and the regret upper and lower bounds in Theorem 6.3 and Theorem 6.4 in Section 6.4.1 hold.

#### Modification of the HS-SE policy

In this part we assume that $K = \mathcal{O}(1)$.

For any G-BwSC problem defined with switching graph $G$ (whose switching costs do not satisfy the triangle inequality) and switching budget $S$, we construct a new switching graph $G'$ according to Appendix E.5.1, and construct a new G-BwSC problem defined with switching graph $G'$ and switching budget $S$. Let $\pi'$ denote the HS-SE policy running on the new G-BwSC problem. Obviously $\pi'$ is a $S$-switching budget policy for the new problem. We construct $\pi$ by modifying $\pi'$, aiming to obtain an $S$-switching-budget policy for the original G-BwSC problem. Let $\pi$ switch (on $G$) following $\pi'$ (on $G'$): every time $\pi'$ switches from $i$ to $j$ on $G'$, let $\pi$ switch according to the path $p_{i,j} = i \to \cdots \to j$ on $G$, visiting each vertex in $p_{i,j}$ once (since in the HS-SE policy, every uneliminated action is chosen for at least $\Omega(T^{1/2})$ consecutive rounds in each epoch, while $p_{i,j}$ contains at most $K = o(\sqrt{T})$ vertices, we know that $\pi'$ is a valid policy). Since the total weight of $p_{i,j}$ is $c'_{i,j}$ and $\pi'$ is an $S$-switching-budget policy for $G'$, we know that $\pi$ is an $S$-switching-budget policy for $G$.

### E.5.2 Computation of the Offline Step in the HS-SE Policy

The HS-SE policy is practical — for any given switching graph $G$, the policy only involves solving the shortest Hamiltonian path problem once, which can be finished *offline*. Thus, the computational complexity of the shortest Hamiltonian path problem does not affect the online decision-making process of the HS-SE policy.

Moreover, under the condition that the switching costs satisfy the triangle inequality, the shortest Hamiltonian path problem can be reduced to the celebrated *metric traveling salesman problem* (metric TSP), see Lawler (1985). This means that we can directly apply many commercial solvers for the TSP to solve (or approximately solve) the shortest Hamiltonian path problem efficiently. The reduction also indicates that any approximation algorithm designed for the metric TSP can be adapted to be an approximation algorithm for the shortest Hamiltonian path problem. In particular, the celebrated Christofides algorithm for the metric TSP (Christofides 1976) can be used to compute a good approximation of $H$ in polynomial time.

## E.6 Additional Numerical Experiments

The experiments in Section 6.5 were conducted for $K = 8$. In this section, we repeat the experiments for $K = 4$ (a smaller $K$) and $K = 16$ (a larger $K$). For $K = 4$, we consider $S \in [4 : 12]$ (nine different switching budgets); for $K = 16$, we consider $S \in \{18, 23, 30, 31, 38, 45, 46\}$ (seven different switching budgets). The choices of $T$ and $\Delta$ and the experimental setup are the same as before.

The numerical results for $K = 4$ are shown in Figure E-1. The numerical results for $K = 16$ are shown in Figure E-2. As we can see, the observations that we made for $K = 8$ in Section 6.5 continue to hold. Therefore, our experiments for $K = 4$ and $K = 16$ verify the insights that we get from the experiments for $K = 8$.

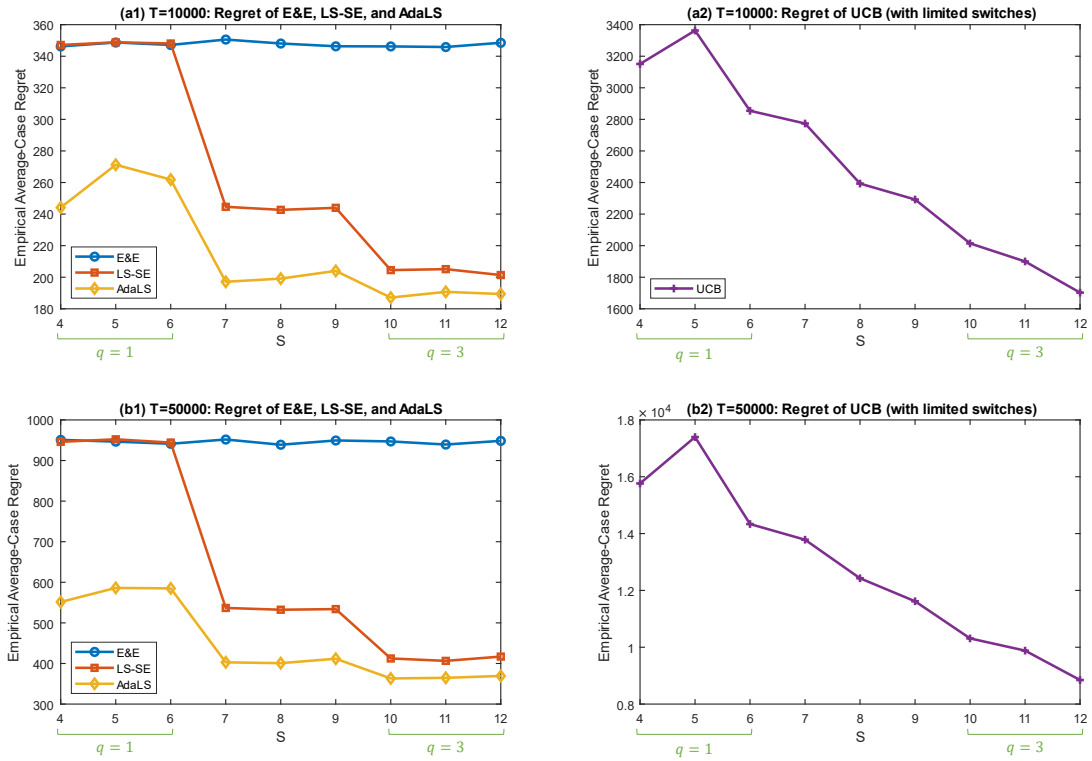


Figure E-1: Empirical average-case regret v.s. the switching budget $S$, under $K = 4$. The regret of UCB has to be plotted separately because it is too large.

Finally, we would like to make two additional comments on our experiment results. First, while our experiments show the significant advantages of AdaLS and LS-SE when $T$ is sufficiently large, we find that AdaLS and LS-SE may not always outperform E&E when we set $T$ to be very small. This is understandable: although AdaLS and LS-SE provably enjoy a smaller regret rate than E&E when $q(S, K) \geq 2$, when $T$ is too small to make the "rate advantage" play a role, the empirical performance of AdaLS and LS-SE may not necessarily be better than E&E.[64] Second, while AdaLS is rate-optimal and performs the best overall in our experiments, as we can observe from Figs. 6-1, E-1 and E-2, its empirical regret can be non-monotone with respect to $S$ — in our experiments, we find that it may (slightly) "over explore" on some instances when $r(S, K)$ becomes larger (possibly because it wants to ensure the optimal worst-case regret rate, which imposes higher requirements on it). Whether there exists a rate-optimal algorithm that (almost) always ensures non-increasing empirical regret as $S$ increases is an interesting open question.

[64] Moreover, when we try to evaluate the magnitude of $T$ and to decide whether it is large enough to make the theoretical rate advantage of certain algorithms show up, we should evaluate $T$ relative to the magnitude of $K$.

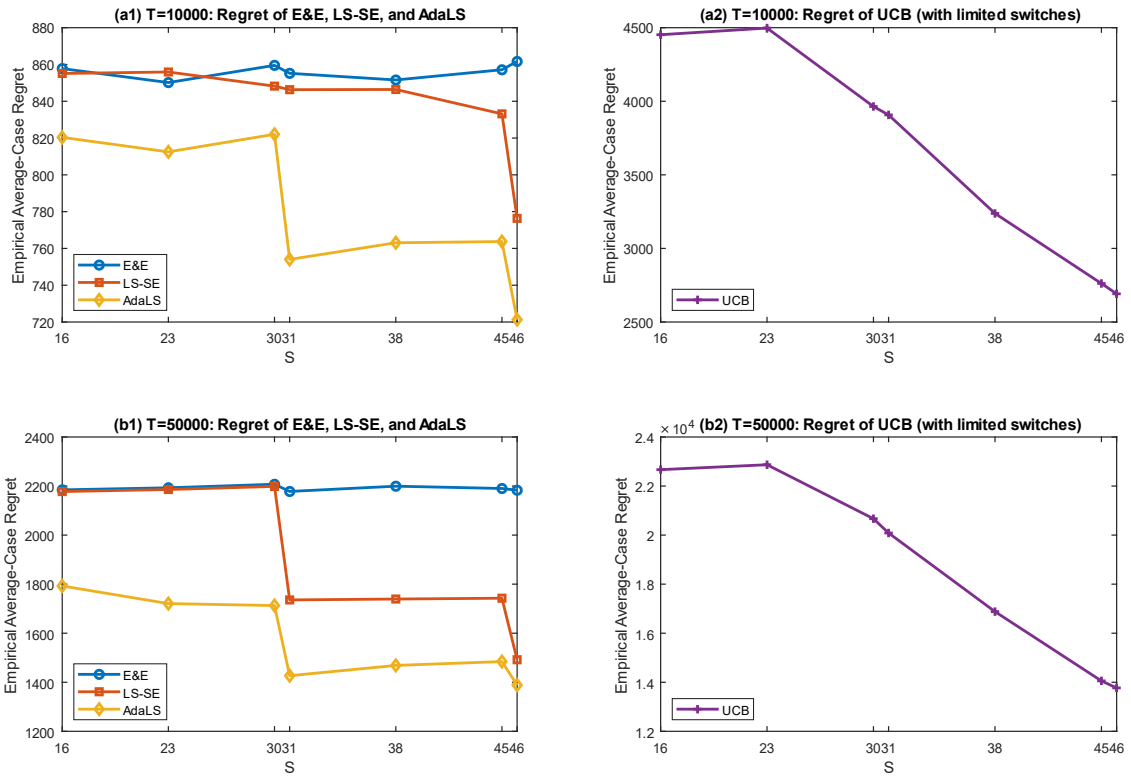


Figure E-2: Empirical average-case regret v.s. the switching budget $S$, under $K = 16$. The regret of UCB has to be plotted separately because it is too large.

## E.7 Proof of Proposition 6.1

Note that our proof is based on the more rigorous version of LS-SE — Algorithm E.3.

### E.7.1 LS-SE is Indeed an $S$-Switch Policy

From round 1 to round $T_1$, LS-SE makes $K-1$ switches. For $1 \le l \le q(S,K)-1$, from round $T_l$ to round $T_{l+1}$:

- If the last action in epoch $l$ remains uneliminated in epoch $l+1$, then it will be the first action in epoch $l+1$, and no switch occurs between round $T_l$ and round $T_l+1$. Since LS-SE makes at most $K-1$ switches within epoch $l+1$, i.e., from round $T_l+1$ to round $T_{l+1}$, it makes at most $0+(K-1)=K-1$ switches from round $T_l$ to round $T_{l+1}$.
- If the last action in epoch $l$ is eliminated before the start of epoch $l+1$, then epoch $l+1$ starts from another uneliminated action, and one switch occurs between round $T_l$ and round $T_l+1$. The elimination implies that $|A_{l+1}| \le K-1$, thus LS-SE makes $|A_{l+1}|-1 \le (K-1)-1 = K-2$ switches within epoch $l+1$, i.e., from round $T_l+1$ to round $T_{l+1}$. Therefore, the LS-SE policy makes at most $1+(K-2)=K-1$ switches from round $T_l$ to round $T_{l+1}$.

From round $T_{q(S,K)}$ to round $T$, since LS-SE does not switch within epoch $q(S,K)+1$, i.e., from round $T_{q(S,K)}+1$ to round $T$, the only possible switch is between round $T_{q(S,K)}$ and $T_{q(S,K)}+1$. Thus LS-SE makes at most 1 switch from round $T_{q(S,K)}$ to round $T$.

Summarizing the above arguments, we find that the LS-SE policy makes at most $q(S,K)(K-1)+1 \le S$ switches from round 1 to round $T$. Thus it is indeed an $S$-switching-budget policy.

### E.7.2 Proof of Upper Bound

Since LS-SE is a $(q(S,K)+1)$-batch policy, existing upper bound analysis for batched MAB (Gao et al. 2019) applies here. Still, we present our upper bound proof for completeness. A difference is that we obtain slightly better dependence on $K$ under the condition $\sup_{i,j\in[K]}|\mu_i-\mu_j| \in [0,1]$.

If $K > T/4$ or $q(S,K)=0$, then the upper bound in Proposition 6.1 becomes $\mathcal{O}(T)$ and is trivial. Therefore, without loss of generality, we assume that $T \ge 4K$ and $q(S,K)>0$.

We start the proof of the upper bound on regret with some definitions. Define the confidence radius as

$$r_i(t) = \sqrt{\frac{6\log T}{N_i(t)}}, \quad \forall i \in [K], t \in [T].$$

The $\mathsf{UCB}_i(t)$ and $\mathsf{LCB}_i(t)$ confidence bounds defined in (6.4) can be expressed as

$$\mathsf{UCB}_i(t) = \bar{\mu}_i(t) + r_i(t), \quad \forall i \in [K], t \in [T],$$

$$\mathsf{LCB}_i(t) = \bar{\mu}_i(t) - r_i(t), \quad \forall i \in [K], t \in [T].$$

Define the *clean event* as

$$\mathcal{E} := \{\forall i \in [K], \forall t \in [T], \quad |\bar{\mu}_i(t) - \mu_i| \le r_t(i)\}.$$

By Hoeffding's inequality for sub-Gaussian variables and a standard union bound argument (see, e.g., Lemma 1.5 in Slivkins 2019), since $T \geq K$, for any policy $\pi$ and any environment $\mathcal{D}$, we always have $\mathbb{P}^{\pi}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \frac{2}{T^3} \cdot T \cdot K \geq 1 - \frac{2}{T}$. Define the *bad event* $\bar{\mathcal{E}}$ as the complement of the clean event.

Let $\pi$ denote the LS-SE policy. First, observe that for any environment $\mathcal{D}$,

$$\begin{aligned} R^{\pi}_{\mathcal{D}}(T) &= \mathbb{E}^{\pi}_{\mathcal{D}}\left[T\mu^{\star} - \sum_{t=1}^{T} \mu_{a_t} \mid \mathcal{E}\right] \mathbb{P}^{\pi}_{\mathcal{D}}(\mathcal{E}) + \mathbb{E}^{\pi}_{\mathcal{D}}\left[T\mu^{\star} - \sum_{t=1}^{T} \mu_{a_t} \mid \bar{\mathcal{E}}\right] \mathbb{P}^{\pi}_{\mathcal{D}}(\bar{\mathcal{E}}) \\ &\leq \mathbb{E}^{\pi}_{\mathcal{D}}\left[T\mu^{\star} - \sum_{t=1}^{T} \mu_{a_t} \mid \mathcal{E}\right] + T \cdot \frac{2}{T} \\ &= \mathbb{E}^{\pi}_{\mathcal{D}}\left[T\mu^{\star} - \sum_{t=1}^{T} \mu_{a_t} \mid \mathcal{E}\right] + 2, \end{aligned} \tag{E.4}$$

so in order to bound $R^{\pi}(T) = \sup_{\mathcal{D}} R^{\pi}_{\mathcal{D}}(T)$, we only need to focus on the clean event.

Consider an arbitrary environment $\mathcal{D}$ and assume the occurrence of the clean event. By the specification of Algorithm E.3, we know that the optimal action $k^{\star} \in A_j$ for all $j \in [q(S,K)+1]$. For any $k \in [K]$, define $\eta_k := \max\{j \in [q(S,K)+1] \mid k \in A_j\}$, i.e., $\eta_k$ the index of the last epoch where action $k$ is uneliminated. Consider any action $k$ such that $\mu_k < \mu_{k^{\star}}$. By the specification of Algorithm E.3, if $\eta_k > 1$, then the confidence intervals of the two actions $k^{\star}$ and $k$ at the end of round $T_{\eta_k - 1}$ must overlap, i.e., $\mathsf{UCB}_k(T_{\eta_k-1}) \geq \mathsf{LCB}_{k^{\star}}(T_{\eta_k-1})$. Therefore,

$$\Delta(k) := \mu_{k^{\star}} - \mu_k \leq 2r_{k^{\star}}(T_{\eta_k-1}) + 2r_k(T_{\eta_k-1}) = 4r_k(T_{\eta_k-1}), \tag{E.5}$$

where the last equality is because $k^{\star}$ and $k$ are chosen for equal times in each epoch $j \in [\eta_k - 1]$, which implies that $N_{k^{\star}}(T_{\eta_k-1}) = N_k(T_{\eta_k-1})$. Since $k$ is never chosen after the $\eta_k$-th epoch, we have $N_k(T_{\eta_k}) = N_k(T)$, and therefore $r_k(T_{\eta_k}) = r_k(T)$.

For any $j \in [q(S,K)]$, since $K \leq \frac{1}{2}\lfloor 2K \rfloor \leq \frac{1}{2}\lfloor K\sqrt{T/K} \rfloor \leq \frac{1}{2}t_j$, we have

$$\begin{aligned} T_j &= T_{j-1} + |A_j| \left\lfloor \frac{t_j - T_{j-1}}{|A_j|} \right\rfloor \\ &\geq T_{j-1} + t_j - T_{j-1} - (|A_j| - 1) \\ &\geq t_j - (K-1) \geq \frac{1}{2} t_j. \end{aligned} \tag{E.6}$$

For any $k \in [K]$, define $R(T;k) := \sum_{t=1}^{T} (\mu^{\star} - \mu_k)\mathbb{1}\{a_t = k\} = \Delta(k)N_k(T)$. For any $k$ such that $\eta_k \in [2 : q(S,K)]$, by (E.5), (E.6), and the specification of Algorithm E.3, conditional on the clean

event $\mathcal{E}$,

$$\begin{aligned}
R(T;k) &= N_k(T)\Delta(k) \\
&\le 4N_k(T_{\eta_k})\sqrt{\frac{6\log T}{N_k(T_{\eta_k-1})}} \\
&\le 4\frac{T_{\eta_k}}{|A_{\eta_k}|}\frac{\sqrt{6\log T}}{\sqrt{T_{\eta_k-1}/K}} \\
&\le 4\sqrt{6K\log T}\frac{1}{|A_{\eta_k}|}\frac{T_{\eta_k}}{\sqrt{T_{\eta_k-1}}} \\
&\le 8\sqrt{3K\log T}\frac{1}{|A_{\eta_k}|}\frac{t_{\eta_k}}{\sqrt{t_{\eta_k-1}}} \\
&= 8\sqrt{6\log T}\frac{1}{|A_{\eta_k}|}K(T/K)^{\frac{1}{2-2^{-q(S,K)}}}.
\end{aligned}$$

For any $k$ such that $\eta_k = 1$, we have $R(T;k) = N_k(T)\Delta(k) \le N_K(T_1) \le \frac{1}{|A_1|}K(T/K)^{\frac{1}{2-2^{-q(S,K)}}}$. Moreover, we have

$$\begin{aligned}
\sum_{k:\eta_k=q(S,K)+1} R(T;k) &= \sum_{k:\eta_k=q(S,K)+1} N_k(T)\Delta(k) \\
&\le T\max_{k:\eta_k=q(S,K)+1}\Delta(k) \\
&\le 4T\max_{k:\eta_k=q(S,K)+1}\sqrt{\frac{6\log T}{N_k(T_{q(S,K)})}} \\
&\le 4T\frac{\sqrt{6\log T}}{\sqrt{T_{q(S,K)}/K}} \\
&\le 8\sqrt{6\log T}K(T/K)^{\frac{1}{2-2^{-q(S,K)}}}.
\end{aligned}$$

Therefore, for any environment $\mathcal{D}$, conditional on the clean event $\mathcal{E}$, we have

$$\begin{aligned}
T\mu^\star - \sum_{t=1}^{T}\mu_{a_t} &= \sum_{k\in[K]} R(T;k) \\
&= \sum_{k:\eta_k=1} R(T;k) + \sum_{k:\eta_k\in[2:q(S,K)]} R(T;k) + \sum_{k:\eta_k=q(S,K)+1} R(T;k) \\
&\le 8\sqrt{6\log T}K(T/K)^{\frac{1}{2-2^{-q(S,K)}}}\left(\sum_{k=1}^{K}\frac{1}{|A_{\eta_k}|}+1\right) \\
&\le 8\sqrt{6\log T}K(T/K)^{\frac{1}{2-2^{-q(S,K)}}}\left(\sum_{j=1}^{K}\frac{1}{j}+1\right) \\
&\le 8\sqrt{6\log T}K(T/K)^{\frac{1}{2-2^{-q(S,K)}}}(\log K+2) \\
&\le 40\sqrt{6}(\log K\log T)K^{1-\frac{1}{2-2^{-q(S,K)}}}T^{\frac{1}{2-2^{-q(S,K)}}}.
\end{aligned}$$

Thus by (E.4) and $R^{\pi}(T) = \sup_{\mathcal{D}} R^{\pi}_{\mathcal{D}}(T)$, we have

$$R^{\pi}(T) \leq 40\sqrt{6}(\log K \log T) K^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}} + 2.$$

□

## E.8 Proof of Theorem 6.1

In this proof, we let the tuning parameter $\lambda = 1/2$ (see Algorithm 6.2). Our proof essentially holds for any $\lambda \in (0, 1)$ being a constant.

### E.8.1 The AdaLS Policy is Indeed an $S$-Switch Policy

According to Section 6.3.2, before the last switch (where we commit to a single action; see Line 16), we allow AdaLS to switch to each action in $A_1^{(1)}$ for at most $q(S, K)$ times, while allowing it to switch to each action in $A_1^{(2)}$ for at most $q(S, K) + 1$ times. Therefore, if $\widehat{r}(S, K) > 0$, then AdaLS will make up to

$$\underbrace{q(S,K)(K - \widehat{r}(S,K))}_{\text{switching to an action in } A_1^{(1)}} + \underbrace{(q(S,K)+1)\widehat{r}(S,K)}_{\text{switching to an action in } A_1^{(2)}} - \underbrace{1}_{\text{the first round}} + \underbrace{1}_{\text{the last switch}} = S$$

switches. See Appendix E.3 for a concrete example when $K = 2S - 2$.

On the other hand, if $\widehat{r}(S, K) = 0$, then AdaLS will behave similar to LS-SE and make up to $q(S, K)(K - 1) + 1 < S$ switches.

To sum up, AdaLS is indeed an $S$-switch policy.

### E.8.2 Proof of Upper Bound

If $K > T/64$, then the upper bound in Theorem 6.1 becomes $\mathcal{O}(T)$ and is trivial. Therefore, without loss of generality, we assume that $T \geq 64K$. If $q(S, K) = 0$, then Algorithm 6.2 will only execute Lines 14 to 16, and it is very easy to show that the regret is upper bounded by

$$\frac{|A_1^{(1)}|}{K} \cdot \mathcal{O}(T) + \frac{|A_2^{(1)}|}{K} \cdot \mathcal{O}(K^{\frac{1}{3}} T^{\frac{2}{3}} \log T) = \mathcal{O}(\log T) \cdot \max\left\{ \frac{K - r(S,K)}{K} T, K^{\frac{1}{3}} T^{\frac{2}{3}} \right\}.$$

Therefore, without loss of generality, we assume that $q(S, K) > 0$ in the proof below.

We start the proof of the upper bound on regret with some definitions. Define the confidence radius as

$$r_i(t) = \sqrt{\frac{6 \log T}{N_i(t)}}, \quad \forall i \in [K], t \in [T].$$

The $\mathtt{UCB}_i(t)$ and $\mathtt{LCB}_i(t)$ confidence bounds defined in (6.4) can be expressed as

$$\mathtt{UCB}_i(t) = \bar{\mu}_i(t) + r_i(t), \quad \forall i \in [K], t \in [T],$$

$$\mathtt{LCB}_i(t) = \bar{\mu}_i(t) - r_i(t), \quad \forall i \in [K], t \in [T].$$

Define the *clean event* as

$$\mathcal{E} := \{\forall i \in [K], \forall t \in [T], \quad |\bar{\mu}_i(t) - \mu_i| \le r_t(i)\}.$$

By Hoeffding's inequality for sub-Gaussian variables and a standard union bound argument (see, e.g., Lemma 1.5 in Slivkins 2019), since $T \ge K$, for any policy $\pi$ and any environment $\mathcal{D}$, we always have $\mathbb{P}^\pi_\mathcal{D}(\mathcal{E}) \ge 1 - \frac{2}{T^3} \cdot T \cdot K \ge 1 - \frac{2}{T}$. Define the *bad event* $\bar{\mathcal{E}}$ as the complement of the clean event.

Let $\pi$ denote the AdaLS policy. First, observe that for any environment $\mathcal{D}$,

$$\begin{aligned} R^\pi_\mathcal{D}(T) &= \mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \mathcal{E}\right] \mathbb{P}^\pi_\mathcal{D}(\mathcal{E}) + \mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \bar{\mathcal{E}}\right] \mathbb{P}^\pi_\mathcal{D}(\bar{\mathcal{E}}) \\ &\le \mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \mathcal{E}\right] + T \cdot \frac{2}{T} \\ &= \mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \mathcal{E}\right] + 2, \end{aligned} \tag{E.7}$$

so in order to bound $R^\pi(T) = \sup_\mathcal{D} R^\pi_\mathcal{D}(T)$, we only need to focus on the clean event.

Define $\mathcal{E}_1 := \left\{k^\star \in A_1^{(1)}\right\}$ and $\mathcal{E}_2 := \left\{k^\star \in A_1^{(2)}\right\}$. Since $A_1^{(1)}$ and $A_1^{(2)}$ are determined by random sampling independent of the clean event $\mathcal{E}$, we have

$$\mathbb{P}^\pi_\mathcal{D}(\mathcal{E}_1 \mid \mathcal{E}) = \mathbb{P}^\pi_\mathcal{D}(\mathcal{E}_1) = \frac{K - \widehat{r}(S,K)}{K},$$

$$\mathbb{P}^\pi_\mathcal{D}(\mathcal{E}_2 \mid \mathcal{E}) = \mathbb{P}^\pi_\mathcal{D}(\mathcal{E}_2) = \frac{\widehat{r}(S,K)}{K}.$$

Thus

$$\mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \mathcal{E}\right] = \frac{K - \widehat{r}(S,K)}{K} \mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \mathcal{E}, \mathcal{E}_1\right] + \frac{\widehat{r}(S,K)}{K} \mathbb{E}^\pi_\mathcal{D}\left[T\mu^\star - \sum_{t=1}^T \mu_{a_t} \mid \mathcal{E}, \mathcal{E}_2\right]. \tag{E.8}$$

Note that we define $\mathbb{E}^\pi_\mathcal{D}(\cdot \mid \mathcal{E}, \mathcal{E}_2) \equiv 0$ if $\widehat{r}(S,K) = 0$.

Define $T^{(1)}_{q(S,K)+1} := T$. For all $j \in [q(S,K)+1]$, "epoch $j$" corresponds to "period $[T^{(1)}_{j-1}+1 : T^{(1)}_j]$". Let $k^\dagger$ denote the action that Algorithm 6.2 chooses in Line 16.

In what follows, we derive several useful inequalities for the time points appeared in Algorithm 6.2. For any $j \in [q(S,K)+2]$, we have

$$\begin{aligned} t^{(2)}_j &= \left\lfloor K^{1 - \frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} T^{\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} \right\rfloor \\ &\ge \left\lfloor K^{1 - \frac{2-2^{2-j}}{2-2^{-q(S,K)}}} T^{\frac{2-2^{2-j}}{2-2^{-q(S,K)}}} \right\rfloor \\ &\ge \left\lfloor (K - \widehat{r}(S,K))^{1 - \frac{2-2^{2-j}}{2-2^{-q(S,K)}}} T^{\frac{2-2^{2-j}}{2-2^{-q(S,K)}}} \right\rfloor = t^{(1)}_{j-1}. \end{aligned} \tag{E.9}$$

By (E.9) and the fact that

$$\forall j \in [q(S,K)], \quad \begin{cases} T_j^{(2)} = T_{j-1}^{(1)} + |A_j^{(2)}| \left\lfloor \frac{t_j^{(2)} - T_{j-1}^{(1)}}{|A_j|} \right\rfloor, \\ T_j^{(1)} = T_j^{(2)} + |A_j^{(1)}| \max\left\{ \left\lfloor \frac{t_j^{(1)}/2 - T_j^{(2)}}{|A_j^{(1)}|} \right\rfloor, \left\lfloor \frac{t_j^{(2)} - T_{j-1}^{(1)}}{|A_j|} \right\rfloor \right\}, \end{cases}$$

we have

$$\forall j \in [q(S,K)], \quad \begin{cases} T_j^{(2)} \le t_j^{(2)}, \\ T_j^{(1)} \le \max\{t_j^{(1)}/2, t_j^{(2)}\} \le t_{j+1}^{(2)}. \end{cases} \tag{E.10}$$

Meanwhile, for any $j \in [q(S,K)]$, since $K \le \frac{1}{8}\lfloor 8K \rfloor \le \frac{1}{8}\lfloor K\sqrt{T/K} \rfloor \le \frac{1}{8} t_j^{(2)}$, we have

$$\begin{aligned} T_j^{(1)} &\ge T_{j-1}^{(1)} + |A_j| \left\lfloor \frac{t_j^{(2)} - T_{j-1}^{(1)}}{|A_j|} \right\rfloor \\ &\ge T_{j-1}^{(1)} + t_j^{(2)} - T_{j-1}^{(1)} - (|A_j| - 1) \\ &\ge t_j^{(2)} - (K-1) \ge \frac{1}{2} t_j^{(2)}. \end{aligned} \tag{E.11}$$

In addition to the above inequalities, we also need a lower bound on $N_k(\cdot)$ for $k \in A_1^{(2)}$, which we provide below.

**Lemma E.2.** *For any $l \in [q(S,K)+1]$, for any $k \in A_l^{(2)}$, we have*

$$N_k(T_l^{(2)}) \ge \frac{t_l^{(2)}}{32K}.$$

*Proof of Lemma E.2.* By (E.10), we have

$$T_{j-1}^{(1)} \le \max\{t_{j-1}^{(1)}/2, t_{j-1}^{(2)}\} \le t_j^{(2)}$$

for all $j \in [q(S,K)+1]$, hence

$$\begin{aligned} n_j^{(2)} &= \left\lfloor \frac{t_j^{(2)} - T_{j-1}^{(1)}}{|A_j|} \right\rfloor \\ &\ge \min\left\{ \left\lfloor \frac{t_j^{(2)} - t_{j-1}^{(1)}/2}{|A_j|} \right\rfloor, \left\lfloor \frac{t_j^{(2)} - t_{j-1}^{(2)}}{|A_j|} \right\rfloor \right\} \\ &\ge \min\left\{ \left\lfloor \frac{t_j^{(2)}/2}{|A_j|} \right\rfloor, \left\lfloor \frac{t_j^{(2)} - t_{j-1}^{(2)}}{|A_j|} \right\rfloor \right\} \end{aligned} \tag{E.12}$$

for all $j \in [q(S,K)+1]$, where the last inequality follows from (E.9). For all $j \le \log_2 \log_2(T/K)$, we

have

$$\begin{aligned}
t_{j-1}^{(2)} &\le K(T/K)^{\frac{2-2^{2-j}}{2-2^{-q(S,K)-1}}} \\
&= K(T/K)^{\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} \cdot (T/K)^{\frac{-2^{1-j}}{2-2^{-q(S,K)-1}}} \\
&\le K(T/K)^{\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} \cdot (T/K)^{-2^{-j}} \\
&\le K(T/K)^{\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} \cdot (T/K)^{-\log_{(T/K)} 2} \\
&= \frac{1}{2} K(T/K)^{\frac{2-2^{1-j}}{2-2^{-q(S,K)-1}}} \\
&\le \frac{1}{2}(t_j^{(2)} + 1). \qquad \text{(E.13)}
\end{aligned}$$

By (E.12) and (E.13), for all $j \le \min\{\log_2 \log_2(T/K), q(S,K)+1\}$, we have

$$\begin{aligned}
n_j^{(2)} &\ge \min\left\{ \left\lfloor \frac{t_j^{(2)}/2}{|A_j|} \right\rfloor, \left\lfloor \frac{t_j^{(2)} - t_{j-1}^{(2)}}{|A_j|} \right\rfloor \right\} \\
&\ge \min\left\{ \left\lfloor \frac{t_j^{(2)}/2}{|A_j|} \right\rfloor, \left\lfloor \frac{(t_j^{(2)} - 1)/2}{|A_j|} \right\rfloor \right\} \\
&\ge \left\lfloor \frac{(t_j^{(2)} - 1)/2}{K} \right\rfloor \\
&\ge \frac{t_j^{(2)} - 1}{4K}, \qquad \text{(E.14)}
\end{aligned}$$

where the last inequality follows from $t_j^{(2)} \ge \lfloor K\sqrt{T/K} \rfloor \ge 8K$.

Fix any $l \in [q(S,K)+1]$ and any $k \in A_l^{(2)}$. If $l \le \log_2 \log_2(T/K)$, then by (E.14), we have

$$N_k(T_l^{(2)}) = \sum_{j \in [l]} n_j^{(2)} \ge n_l^{(2)} \ge \frac{t_l^{(2)} - 1}{4K} \ge \frac{t_l^{(2)}}{8K}.$$

If $l > \log_2 \log_2(T/K)$, then by letting $\widetilde{l} := \lfloor \log_2 \log_2(T/K) \rfloor \ge 2$ and using (E.14), we have

$$N_k(T_l^{(2)}) = \sum_{j \in [l]} n_j^{(2)} \ge n_{\widetilde{l}}^{(2)} \ge \frac{t_{\widetilde{l}}^{(2)} - 1}{4K} \ge \frac{t_l^{(2)}}{32K},$$

where the last inequality follows from

$$\begin{aligned}
t_{\widetilde{l}}^{(2)} &\ge K(T/K)^{\frac{2-2^{1-\widetilde{l}}}{2-2^{-q(S,K)-1}}} - 1 \\
&\ge K(T/K)^{1-2^{-\widetilde{l}}} - 1 \\
&\ge T(T/K)^{-2^{-\widetilde{l}}} - 1 \\
&\ge T(T/K)^{-2^{-\log_2 \log_2(T/K)+1}} - 1 \\
&= T/4 - 1 \ge T/8 + 1 \ge t_l^{(2)}/8 + 1.
\end{aligned}$$

Therefore, no matter $l \le \log_2 \log_2(T/K)$ or $l > \log_2 \log_2(T/K)$, we always have

$$N_k(T_l^{(2)}) \ge \frac{t_l^{(2)}}{32K}.$$

□

To simplify the presentation, we define the following two notations:

$$\mathcal{R}_1 := (K - \widehat{r}(S,K))^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}},$$

$$\mathcal{R}_2 := K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}}.$$

Consider an arbitrary environment $\mathcal{D}$ and assume the occurrence of the clean event $\mathcal{E}$. In what follows, we discuss two cases: $\mathcal{E}_1$ occurs, and $\mathcal{E}_2$ occurs.

**Case 1: Both $\mathcal{E}$ and $\mathcal{E}_1$ occur.** By the specification of Algorithm 6.2, we know that the optimal action $k^\star \in A_l^{(1)} \subset A_l$ for all $l \in [q(S,K)+1]$. Moreover, by Line 3 to Line 12 of Algorithm 6.2, we know that action $k^\star$ is chosen for

$$\max\left\{ \left\lfloor \frac{t_l^{(1)}/2 - T_l^{(2)}}{|A_l^{(1)}|} \right\rfloor, n_l^{(2)} \right\}$$

rounds in each epoch $l \in [q(S,K)]$, which is no less than the number of plays of any other action in epoch $l$. Therefore, for all $k \in [K]$ and $l \in [q(S,K)]$ we have

$$N_{k^\star}(T_l^{(1)}) \ge N_k(T_l^{(1)}), \quad r_{k^\star}(T_l^{(1)}) \le r_k(T_l^{(1)}). \tag{E.15}$$

For any $k \in [K]$, define $\eta_k := \max\{j \in [q(S,K)+1] \mid k \in A_j\}$.

Consider any action $k$ such that $\eta_k > 1$. By Line 12 of Algorithm 6.2, the confidence intervals of the two actions $k^\star$ and $k$ at the end of round $T_{\eta_k - 1}^{(1)}$ must overlap, i.e., $\mathtt{UCB}_k(T_{\eta_k-1}^{(1)}) \ge \mathtt{LCB}_{k^\star}(T_{\eta_k-1}^{(1)})$. Therefore,

$$\Delta(k) := \mu_{k^\star} - \mu_k \le 2r_{k^\star}(T_{\eta_k-1}^{(1)}) + 2r_k(T_{\eta_k-1}^{(1)}) \le 4r_k(T_{\eta_k-1}^{(1)}), \tag{E.16}$$

where the last inequality follows from (E.15). Since $k$ is never chosen after the $\eta_k$-th epoch, we have $N_k(T_{\eta_k}^{(1)}) = N_k(T)$.

Consider the action $k^\dagger$ selected in Line 16 of Algorithm 6.2. By Lines 15 and 16 of Algorithm 6.2, the confidence intervals of the two actions $k^\star$ and $k^\dagger$ at the end of round $T_{q(S,K)+1}^{(2)}$ must overlap, i.e., $\mathtt{UCB}_{k^\dagger}(T_{q(S,K)+1}^{(2)}) \ge \mathtt{LCB}_{k^\star}(T_{q(S,K)+1}^{(2)})$. Therefore,

$$\Delta(k^\dagger) := \mu_{k^\star} - \mu_{k^\dagger} \le 2r_{k^\star}(T_{q(S,K)+1}^{(2)}) + 2r_{k^\dagger}(T_{q(S,K)+1}^{(2)}) \le 4 \max_{k \in A_{q(S,K)+1}} r_k(T_{q(S,K)+1}^{(2)}). \tag{E.17}$$

We now try to prove the following key lemma.

**Lemma E.3.** *Assume both $\mathcal{E}$ and $\mathcal{E}_1$ hold. For any action $k$ such that $\eta_k \in [2 : q(S,K)]$, we have*

$$\frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \le \begin{cases} \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{2}{|A^{(1)}_{\eta_k}|}\max\Big\{\mathcal{R}_1, \frac{32K}{K-\widehat{r}(S,K)}\mathcal{R}_2\Big\}, & \text{if } k \in A^{(1)}_1; \\ \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{8}{|A_{\eta_k}|}\mathcal{R}_2, & \text{if } k \in A^{(2)}_1. \end{cases} \tag{E.18}$$

*Moreover, considering all $k$ such that $\eta_k = q(S,K)+1$ (i.e., $k \in A_{q(S,K)+1}$), we have*

$$\max_{k \in A^{(1)}_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} \le 4\max\Big\{\mathcal{R}_1, \frac{32K}{K-\widehat{r}(S,K)}\mathcal{R}_2\Big\}, \tag{E.19}$$

$$\max_{k \in A^{(2)}_{q(S,K)+1}} \left(\frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}}\right) \le 16\mathcal{R}_2. \tag{E.20}$$

*Proof of Lemma E.3.* We first show (E.18). Fix any action $k$ such that $\eta_k \in [2 : q(S,K)]$.

We first consider the case of $k \in A^{(1)}_1$. By Line 3 to Line 12 of Algorithm 6.2, we know that action $k$ is chosen for

$$\max\left\{\left\lfloor \frac{t^{(1)}_l/2 - T^{(2)}_l}{|A^{(1)}_l|} \right\rfloor, n^{(2)}_l\right\}$$

rounds in each epoch $l \in [\eta_k - 1]$, which is no less than the number of plays of any other action in epoch $l$. This implies

$$N_k(T^{(1)}_l) \ge T^{(1)}_l/K \ge t^{(2)}_l/(2K), \quad \forall l \in [\eta_k - 1], \tag{E.21}$$

where the last inequality follows from (E.11).

If $t^{(1)}_{\eta_k-1}/4 \ge t^{(2)}_{\eta_k-1}$, then we have

$$N_k(T^{(1)}_{\eta_k-1}) \ge \left\lfloor \frac{t^{(1)}_{\eta_k-1}/2 - T^{(2)}_{\eta_k-1}}{|A^{(1)}_{\eta_k-1}|} \right\rfloor \ge \left\lfloor \frac{t^{(1)}_{\eta_k-1}/4}{|A^{(1)}_{\eta_k-1}|} \right\rfloor \ge \left\lfloor \frac{t^{(1)}_{\eta_k-1}}{4(K-\widehat{r}(S,K))} \right\rfloor \ge \frac{t^{(1)}_{\eta_k-1}}{8(K-\widehat{r}(S,K))}, \tag{E.22}$$

where the second inequality utilizes $T^{(2)}_{\eta_k-1} \le t^{(2)}_{\eta_k-1}$ (by (E.10)) and the last inequality follows from

$$K - \widehat{r}(S,K) \le \frac{1}{8}\lfloor (K-\widehat{r}(S,K))\sqrt{T/(K-\widehat{r}(S,K))}\rfloor \le \frac{t^{(1)}_j}{8}, \quad \forall j \in [q(S,K)+1].$$

Hence we have

$$\begin{aligned}
\frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} &\le \frac{N_k(T^{(1)}_{\eta_k-1}) + \max\left\{\left\lfloor \frac{t^{(1)}_{\eta_k}/2 - T^{(2)}_{\eta_k}}{|A^{(1)}_{\eta_k}|}\right\rfloor, n^{(2)}_{\eta_k}\right\}}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&\le \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{1}{|A^{(1)}_{\eta_k}|}\frac{\max\left\{t^{(1)}_{\eta_k}/2, t^{(2)}_{\eta_k}\right\}}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&\le \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{1}{|A^{(1)}_{\eta_k}|}\max\left\{\frac{t^{(1)}_{\eta_k}}{2}\sqrt{\frac{8(K-\widehat{r}(S,K))}{t^{(1)}_{\eta_k-1}}}, t^{(2)}_{\eta_k}\sqrt{\frac{2K}{t^{(2)}_{\eta_k-1}}}\right\} \\
&\qquad\qquad \text{(by (E.21) \& (E.22))} \\
&= \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{1}{|A^{(1)}_{\eta_k}|}\max\left\{t^{(1)}_{\eta_k}\sqrt{\frac{2(K-\widehat{r}(S,K))}{t^{(1)}_{\eta_k-1}}}, t^{(2)}_{\eta_k}\sqrt{\frac{2K}{t^{(2)}_{\eta_k-1}}}\right\} \\
&\le \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{2}{|A^{(1)}_{\eta_k}|}\max\{\mathcal{R}_1, \mathcal{R}_2\}.
\end{aligned}$$

If $t^{(1)}_{\eta_k-1}/4 < t^{(2)}_{\eta_k-1}$, then

$$\left(\frac{T}{K-\widehat{r}(S,K)}\right)^{\frac{2-2^{2-\eta_k}}{2-2^{-q(S,K)}}} \le \frac{2t^{(1)}_{\eta_k-1}}{K-\widehat{r}(S,K)} \le \frac{8t^{(2)}_{\eta_k-1}}{K-\widehat{r}(S,K)} \le \frac{8K}{K-\widehat{r}(S,K)}\left(\frac{T}{K}\right)^{\frac{2-2^{2-\eta_k}}{2-2^{-q(S,K)-1}}},$$

which implies

$$\left(\frac{T}{K-\widehat{r}(S,K)}\right)^{\frac{2-2^{1-\eta_k}}{2-2^{-q(S,K)}}} \le \left(\frac{8K}{K-\widehat{r}(S,K)}\right)^{\frac{2-2^{1-\eta_k}}{2-2^{2-\eta_k}}}\left(\frac{T}{K}\right)^{\frac{2-2^{1-\eta_k}}{2-2^{-q(S,K)-1}}} \le \left(\frac{8K}{K-\widehat{r}(S,K)}\right)^{2}\left(\frac{T}{K}\right)^{\frac{2-2^{1-\eta_k}}{2-2^{-q(S,K)-1}}}.$$

Thus

$$t^{(1)}_{\eta_k} \le (K-\widehat{r}(S,K))\left(\frac{T}{K-\widehat{r}(S,K)}\right)^{\frac{2-2^{1-\eta_k}}{2-2^{-q(S,K)}}} \le \frac{64K}{K-\widehat{r}(S,k)}K\left(\frac{T}{K}\right)^{\frac{2-2^{1-\eta_k}}{2-2^{-q(S,K)-1}}}. \tag{E.23}$$

Hence we have

$$\begin{aligned}
\frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} &\le \frac{N_k(T^{(1)}_{\eta_k-1}) + \max\left\{\left\lfloor \frac{t^{(1)}_{\eta_k}/2 - T^{(2)}_{\eta_k}}{|A^{(1)}_{\eta_k}|}\right\rfloor, n^{(2)}_{\eta_k}\right\}}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&\le \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{1}{|A^{(1)}_{\eta_k}|}\frac{\max\{t^{(1)}_{\eta_k}/2, t^{(2)}_{\eta_k}\}}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&\le \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{1}{|A^{(1)}_{\eta_k}|}\max\{t^{(1)}_{\eta_k}/2, t^{(2)}_{\eta_k}\}\sqrt{\frac{2K}{t^{(2)}_{\eta_k-1}}} && \text{(by (E.21))} \\
&\le \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{1}{|A^{(1)}_{\eta_k}|}\frac{64K}{K-\widehat{r}(S,K)}\mathcal{R}_2. && \text{(by (E.23))}
\end{aligned}$$

We then consider the case of $k \in A^{(2)}_1$. By Line 3 to Line 12 of Algorithm 6.2, we know that

action $k$ is chosen for $n_l^{(2)}$ rounds in each epoch $l \in [\eta_k - 1]$, and $N_k(T^{(1)}_{\eta_k-1}) = N_k(T^{(2)}_{\eta_k-1})$. Since $k \in A^{(1)}_{\eta_k} \subset A^{(1)}_{\eta_k-1}$, by Lemma E.2, we have

$$N_k(T^{(1)}_{\eta_k-1}) = N_k(T^{(2)}_{\eta_k-1}) \geq \frac{t^{(2)}_{\eta_k-1}}{32K}. \tag{E.24}$$

Thus

$$\begin{aligned}
\frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} &\leq \frac{N_k(T^{(1)}_{\eta_k-1}) + n^{(2)}_{\eta_k}}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&= \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{n^{(2)}_{\eta_k}}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&\leq \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{t^{(2)}_{\eta_k}/|A_{\eta_k}|}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}} \\
&\leq \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{t^{(2)}_{\eta_k}/|A_{\eta_k}|}{\sqrt{t^{(2)}_{\eta_k-1}/(32K)}} \qquad \text{(by (E.24))} \\
&\leq \sqrt{N_k(T^{(1)}_{\eta_k-1})} + \frac{8}{|A_{\eta_k}|}\mathcal{R}_2.
\end{aligned}$$

Combing the above three paragraphs, we prove (E.18).

We then show (E.19) and (E.20). Consider $k \in A^{(1)}_{q(S,K)+1}$. If $t^{(1)}_{q(S,K)}/4 \geq t^{(2)}_{q(S,K)}$, then (E.22) holds for $\eta_k = q(S,K)+1$; if $t^{(1)}_{q(S,K)}/4 < t^{(2)}_{q(S,K)}$, then (E.23) and (E.21) hold for $\eta_k = q(S,K)+1$. Thus

$$\begin{aligned}
\max_{k \in A^{(1)}_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} &\leq \max\left\{ T\sqrt{\frac{8(K-\widehat{r}(S,K))}{t^{(1)}_{q(S,K)}}}, \frac{64K}{K-\widehat{r}(S,k)} K \left(\frac{T}{K}\right)^{\frac{2-2^{-q(S,K)}}{2-2^{-q(S,K)-1}}} \sqrt{\frac{2K}{t^{(2)}_{q(S,K)}}} \right\} \\
&\leq 4\max\left\{ \mathcal{R}_1, \frac{32K}{K-\widehat{r}(S,K)}\mathcal{R}_2 \right\}.
\end{aligned}$$

Consider $k \in A^{(2)}_{q(S,K)+1}$. By Lemma E.2, we have

$$N_k(T^{(1)}_{q(S,K)}) = N_k(T^{(2)}_{q(S,K)}) \geq \frac{t^{(2)}_{q(S,K)}}{32K}, \quad N_k(T^{(2)}_{q(S,K)+1}) \geq \frac{t^{(2)}_{q(S,K)+1}}{32K}.$$

Thus we have

$$\begin{aligned}
\max_{k \in A^{(2)}_{q(S,K)+1}} \left( \frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}} \right) &\leq t^{(2)}_{q(S,K)+1}\sqrt{\frac{32K}{t^{(2)}_{q(S,K)}}} + T\sqrt{\frac{32K}{t^{(2)}_{q(S,K)+1}}} \\
&\leq 16\mathcal{R}_2.
\end{aligned}$$

□

For any $k \in [K]$, define $R(T;k) := \sum_{t=1}^T (\mu^\star - \mu_k)\mathbb{1}\{a_t = k\} = \Delta(k)N_k(T)$. Consider any $k$ such

that $\eta_k \in [2 : q(S,K)]$, by (E.16), conditional on the events $\mathcal{E}$ and $\mathcal{E}_1$, we always have

$$R(T;k) = N_k(T)\Delta(k) \le 4N_k(T^{(1)}_{\eta_k})r_k(T^{(1)}_{\eta_k-1}) = 4\sqrt{6\log T}\frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}}$$

Moreover, we have $\sum_{k:\eta_k=1} R(T;k) \le T^{(1)}_1 \le \max\{\mathcal{R}_1/2, \mathcal{R}_2\}$ and

$$\begin{aligned}
&\sum_{k:\eta_k=q(S,K)+1} R(T;k) \\
&= N_{k^\dagger}(T)\Delta(k^\dagger) + \sum_{k\in A_{q(S,K)+1}\setminus\{k^\dagger\}} N_k(T)\Delta(k) \\
&= N_{k^\dagger}(T)\Delta(k^\dagger) + \sum_{k\in A^{(1)}_{q(S,K)+1}\setminus\{k^\dagger\}} N_k(T)\Delta(k) + \sum_{k\in A^{(2)}_{q(S,K)+1}\setminus\{k^\dagger\}} N_k(T^{(2)}_{q(S,K)+1})\Delta(k) \\
&\overset{\text{(i)}}{\le} 4T \max_{k\in A_{q(S,K)+1}} r_k(T^{(2)}_{q(S,K)+1}) + 4T \max_{k\in A^{(1)}_{q(S,K)+1}} r_k(T^{(1)}_{q(S,K)}) + 4T^{(2)}_{q(S,K)+1} \max_{k\in A^{(2)}_{q(S,K)+1}} r_k(T^{(1)}_{q(S,K)}) \\
&\overset{\text{(ii)}}{\le} 4T \max_{k\in A_{q(S,K)+1}} r_k(T^{(2)}_{q(S,K)+1}) + 4T \max_{k\in A^{(1)}_{q(S,K)+1}} r_k(T^{(1)}_{q(S,K)}) + 4t^{(2)}_{q(S,K)+1} \max_{k\in A^{(2)}_{q(S,K)+1}} r_k(T^{(1)}_{q(S,K)}) \\
&= 4\sqrt{6\log T}\left(\max_{k\in A_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}} + \max_{k\in A^{(1)}_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \max_{k\in A^{(2)}_{q(S,K)+1}} \frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_k(T^{(1)}_{q(S,K)})}}\right) \\
&\overset{\text{(iii)}}{\le} 4\sqrt{6\log T}\left(\max_{k\in A^{(1)}_{q(S,K)+1}} \frac{2T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \max_{k\in A^{(2)}_{q(S,K)+1}} \frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \max_{k\in A^{(2)}_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}}\right) \\
&= 4\sqrt{6\log T}\left(\max_{k\in A^{(1)}_{q(S,K)+1}} \frac{2T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \max_{k\in A^{(2)}_{q(S,K)+1}} \left(\frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}}\right)\right),
\end{aligned}$$

where (i) follows from (E.16) and (E.17), (ii) follows from

$$T^{(2)}_{q(S,K)+1} = T^{(1)}_{q(S,K)} + |A^{(2)}_{q(S,K)+1}|\left\lfloor \frac{t^{(2)}_{q(S,K)+1} - T^{(1)}_{q(S,K)}}{|A_{q(S,K)+1}|}\right\rfloor \le t^{(2)}_{q(S,K)+1}$$

(note that (E.10) guarantees $t^{(2)}_{q(S,K)+1} \ge T^{(1)}_{q(S,K)}$), and (iii) follows from

$$\max_{k\in A_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}} = \max\left\{\max_{k\in A^{(1)}_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}}, \max_{k\in A^{(2)}_{q(S,K)+1}} \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}}\right\}.$$

Therefore, for any $\mathcal{D}$, conditional on the events $\mathcal{E}$ and $\mathcal{E}_1$, we have

$$\begin{aligned}
T\mu^\star - \sum_{t=1}^{T} \mu_{a_t} &= \sum_{k\in[K]} R(T;k) = \sum_{k:\eta_k=1} R(T;k) + \sum_{k:\eta_k\in[2:q(S,K)]} R(T;k) + \sum_{k:\eta_k=q(S,K)+1} R(T;k) \\
&\le \max\{\mathcal{R}_1/2, \mathcal{R}_2\} + 4\sqrt{6\log T}\left(\sum_{k:\eta_k\in[2:q(S,K)]} \frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_k(T^{(1)}_{\eta_k-1})}}\right) \\
&\quad + 4\sqrt{6\log T}\left(\max_{k\in A^{(1)}_{q(S,K)+1}} \frac{2T}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \max_{k\in A^{(2)}_{q(S,K)+1}}\left(\frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_k(T^{(1)}_{q(S,K)})}} + \frac{T}{\sqrt{N_k(T^{(2)}_{q(S,K)+1})}}\right)\right).
\end{aligned}$$

Combining the above inequality with Lemma E.3 and $\sum_{k:\eta_k>1}\sqrt{N_k(T^{(1)}_{\eta_k-1})} \le \sqrt{KT}$, we have

$$\mathbb{E}^{\pi}_{\mathcal{D}}\left[T\mu^\star - \sum_{t=1}^{T}\mu_{a_t} \mid \mathcal{E}, \mathcal{E}_1\right] \le \mathcal{O}(\log K\sqrt{\log T})\cdot\max\left\{\mathcal{R}_1,\ \frac{K}{K-\widehat{r}(S,K)}\mathcal{R}_2\right\}. \tag{E.25}$$

**Case 2: Both $\mathcal{E}$ and $\mathcal{E}_2$ occur.** By the specification of Algorithm 6.2, we know that the optimal action $k^\star \in A^{(2)}_l \subset A_l$ for all $l \in [q(S,K)+1]$, and $k^\star \in A_{q(S,K)+2}$. Moreover, by Line 3 to Line 12 of Algorithm 6.2, we know that action $k^\star$ is chosen for $n^{(2)}_l$ rounds in each epoch $l \in [q(S,K)]$, which is no greater than the number of plays of any other action chosen in epoch $l$. Therefore, for all $l \in [q(S,K)]$ and $k \in A_l$ and we have

$$N_{k^\star}(T^{(1)}_l) \le N_k(T^{(1)}_l), \quad r_{k^\star}(T^{(1)}_l) \ge r_k(T^{(1)}_l). \tag{E.26}$$

For any $k \in [K]$, define $\eta_k := \max\{j \in [q(S,K)+1] \mid k \in A_j\}$.

Consider any action $k$ such that $\eta_k > 1$. By Line 12 of Algorithm 6.2, the confidence intervals of the two actions $k^\star$ and $k$ at the end of round $T^{(1)}_{\eta_k-1}$ must overlap, i.e., $\mathtt{UCB}_k(T^{(1)}_{\eta_k-1}) \ge \mathtt{LCB}_{k^\star}(T^{(1)}_{\eta_k-1})$. Therefore,

$$\Delta(k) := \mu_{k^\star} - \mu_k \le 2r_{k^\star}(T^{(1)}_{\eta_k-1}) + 2r_k(T^{(1)}_{\eta_k-1}) \le 4r_{k^\star}(T^{(1)}_{\eta_k-1}), \tag{E.27}$$

where the last inequality follows from (E.26). Since $k$ is never chosen after the $\eta_k$-th epoch, we have $N_k(T^{(1)}_{\eta_k}) = N_k(T)$.

Consider any action $k$ such that $N_k(T^{(1)}_{\eta_k}) > N_k(T^{(2)}_{\eta_k})$. If $\eta_k < q(S,K)+1$, then $k \in A^{(1)}_{\eta_k}$. By Lines 6 and 10 of Algorithm 6.2, the confidence intervals of the two actions $k^\star$ and $k$ at the end of round $T^{(1)}_{\eta_k,k}$ must overlap, i.e., $\mathtt{UCB}_k(T^{(1)}_{\eta_k,k}) \ge \mathtt{LCB}_{k^\star}(T^{(1)}_{\eta_k,k})$. Therefore,

$$\Delta(k) := \mu_{k^\star} - \mu_k \le 2r_{k^\star}(T^{(1)}_{\eta_k,k}) + 2r_k(T^{(1)}_{\eta_k,k}) \le 4r_{k^\star}(T^{(2)}_{\eta_k}), \tag{E.28}$$

where the last inequality follows from

$$N_{k^\star}(T^{(2)}_{\eta_k}) = N_{k^\star}(T^{(1)}_{\eta_k,k}) = N_{k^\star}(T^{(1)}_{\eta_k-1}) + n^{(2)}_{\eta_k} \le N_k(T^{(1)}_{\eta_k-1}) + n^{(2)}_{\eta_k} = N_k(T^{(1)}_{\eta_k,k}).$$

If $\eta_k = q(S,K)+1$, then $k = k^\dagger$. By Lines 15 and 16 of Algorithm 6.2, we have $k^\star \in A^{(2)}_{q(S,K)+2} \ne \varnothing$ and hence $k^\dagger \in A_{q(S,K)+2} = A^{(2)}_{q(S,K)+2}$. Moreover, the confidence intervals of the two actions $k^\star$ and

$k^\dagger$ at the end of round $T^{(2)}_{q(S,K)+1}$ must overlap, i.e., $\mathtt{UCB}_{k^\dagger}(T^{(2)}_{q(S,K)+1}) \geq \mathtt{LCB}_{k^\star}(T^{(2)}_{q(S,K)+1})$. Therefore,

$$\Delta(k^\dagger) := \mu_{k^\star} - \mu_{k^\dagger} \leq 2r_{k^\star}(T^{(2)}_{q(S,K)+1}) + 2r_{k^\dagger}(T^{(2)}_{q(S,K)+1}) \leq 4r_{k^\star}(T^{(2)}_{q(S,K)+1}), \tag{E.29}$$

where the last inequality follows from $k^\star, k^\dagger \in A^{(2)}_{q(S,K)+1}$ and

$$N_{k^\star}(T^{(2)}_{q(S,K)+1}) = N_{k^\star}(T^{(1)}_{q(S,K)}) + n^{(2)}_{q(S,K)+1} \leq N_{k^\dagger}(T^{(1)}_{q(S,K)}) + n^{(2)}_{q(S,K)+1} = N_{k^\dagger}(T^{(2)}_{q(S,K)+1}).$$

We now try to prove the following key lemma.

**Lemma E.4.** *Assume both $\mathcal{E}$ and $\mathcal{E}_2$ hold. For any action $k$ such that $\eta_k \in [2 : q(S,K)]$, we have*

$$\frac{N_k(T^{(2)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(1)}_{\eta_k-1})}} + \frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(2)}_{\eta_k})}} \leq \begin{cases} \frac{24}{|A^{(1)}_{\eta_k}|}\mathcal{R}_2, & \text{if } k \in A^{(1)}_1; \\ \frac{32}{|A_{\eta_k}|}\mathcal{R}_2, & \text{if } k \in A^{(2)}_1. \end{cases} \tag{E.30}$$

*Moreover, considering all $k$ such that $\eta_k = q(S,K)+1$ (i.e., $k \in A_{q(S,K)+1}$), we have*

$$\sum_{k \in A_{q(S,K)+1}} \frac{N_k(T^{(2)}_{q(S,K)+1})}{\sqrt{N_{k^\star}(T^{(1)}_{q(S,K)})}} + \frac{N_k(T)}{\sqrt{N_{k^\star}(T^{(2)}_{q(S,K)+1})}} \leq 16\mathcal{R}_2. \tag{E.31}$$

*Proof of Lemma E.4.* We first show (E.30). Fix any action $k$ such that $\eta_k \in [2 : q(S,K)]$.

We first consider the case of $k \in A^{(1)}_1$. By Line 3 to Line 12 of Algorithm 6.2, we know that every action in $A^{(1)}_{\eta_k}$ (including action $k$) is chosen for

$$\max\left\{ \left\lfloor \frac{t^{(1)}_l/2 - T^{(2)}_l}{|A^{(1)}_l|} \right\rfloor, n^{(2)}_l \right\}$$

rounds in each epoch $l \in [\eta_k - 1]$. This implies

$$N_k(T^{(2)}_{\eta_k}) = N_k(T^{(1)}_{\eta_k-1}) \leq T^{(1)}_{\eta_k-1}/|A^{(1)}_{\eta_k}| \leq t^{(2)}_{\eta_k}/|A^{(1)}_{\eta_k}|, \tag{E.32}$$

where the last inequality follows from (E.10). Moreover, we have

$$N_k(T^{(1)}_{\eta_k}) \leq N_k(T^{(2)}_{\eta_k}) + \max\left\{ \left\lfloor \frac{t^{(1)}_{\eta_k}/2 - T^{(2)}_{\eta_k}}{|A^{(1)}_{\eta_k}|} \right\rfloor, n^{(2)}_{\eta_k} \right\} \leq N_k(T^{(2)}_{\eta_k}) + \frac{\max\{t^{(1)}_{\eta_k}/2, t^{(2)}_{\eta_k}\}}{|A^{(1)}_{\eta_k}|} \leq \frac{2t^{(2)}_{\eta_k+1}}{|A^{(1)}_{\eta_k}|},$$

where the last inequality follows from (E.32) and (E.9). By $k^\star \in A^{(2)}_{q(S,K)+1} \subset A^{(2)}_{\eta_k}$ and Lemma E.2, we have

$$N_{k^\star}(T^{(1)}_{\eta_k-1}) = N_{k^\star}(T^{(2)}_{\eta_k-1}) \geq \frac{t^{(2)}_{\eta_k-1}}{32K}, \quad N_k(T^{(2)}_{\eta_k}) \geq \frac{t^{(2)}_{\eta_k}}{32K}.$$

Therefore, we have

$$\frac{N_k(T^{(2)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(1)}_{\eta_k-1})}} + \frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(2)}_{\eta_k})}} \leq \frac{t^{(2)}_{\eta_k}}{|A^{(1)}_{\eta_k}|}\sqrt{\frac{32K}{t^{(2)}_{\eta_k-1}}} + \frac{2t^{(2)}_{\eta_k+1}}{|A^{(1)}_{\eta_k}|}\sqrt{\frac{32K}{t^{(2)}_{\eta_k}}} \leq \frac{24}{|A^{(1)}_{\eta_k}|}\mathcal{R}_2.$$

We then consider the case of $k \in A_1^{(2)}$. By Line 3 to Line 12 of Algorithm 6.2, we know that action $k$ is chosen for $n_l^{(2)}$ rounds in each epoch $l \in [\eta_k - 1]$, while every action in $A_{\eta_k}$ is chosen for at least $n_l^{(2)}$ rounds in each epoch $l \in [\eta_k - 1]$. This implies

$$N_k(T^{(1)}_{\eta_k-1}) \le T^{(1)}_{\eta_k-1}/|A_{\eta_k}| \le t^{(2)}_{\eta_k}/|A_{\eta_k}|, \tag{E.33}$$

where the last inequality follows from (E.10). Moreover, we have

$$N_k(T^{(1)}_{\eta_k}) = N_k(T^{(2)}_{\eta_k}) \le N_k(T^{(1)}_{\eta_k-1}) + n^{(2)}_{\eta_k} \le N_k(T^{(1)}_{\eta_k-1}) + \frac{t^{(2)}_{\eta_k}}{|A_{\eta_k}|} \le \frac{2t^{(2)}_{\eta_k}}{|A_{\eta_k}|},$$

where the first equality utilizes $\eta_k \le q(S,K)$, and the last inequality follows from (E.33). By $k^\star \in A^{(2)}_{q(S,K)+1} \subset A^{(2)}_{\eta_k}$ and Lemma E.2, we have

$$N_{k^\star}(T^{(1)}_{\eta_k-1}) = N_{k^\star}(T^{(2)}_{\eta_k-1}) \ge \frac{t^{(2)}_{\eta_k-1}}{32K}, \quad N_k(T^{(2)}_{\eta_k}) \ge \frac{t^{(2)}_{\eta_k}}{32K}.$$

Therefore, we have

$$\frac{N_k(T^{(2)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(1)}_{\eta_k-1})}} + \frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(2)}_{\eta_k})}} \le \frac{2t^{(2)}_{\eta_k}}{|A_{\eta_k}|}\sqrt{\frac{32K}{t^{(2)}_{\eta_k-1}}} + \frac{2t^{(2)}_{\eta_k}}{|A_{\eta_k}|}\sqrt{\frac{32K}{t^{(2)}_{\eta_k}}} \le \frac{32}{|A^{(1)}_{\eta_k}|}\mathcal{R}_2.$$

We then show (E.31). We have

$$T^{(2)}_{q(S,K)+1} = T^{(1)}_{q(S,K)} + |A^{(2)}_{q(S,K)+1}|\left\lfloor \frac{t^{(2)}_{q(S,K)+1} - T^{(1)}_{q(S,K)}}{|A_{q(S,K)+1}|} \right\rfloor \le t^{(2)}_{q(S,K)+1}$$

(note that (E.10) guarantees $t^{(2)}_{q(S,K)+1} \ge T^{(1)}_{q(S,K)}$). By $k^\star \in A^{(2)}_{q(S,K)+1} \subset A^{(2)}_{q(S,K)}$ and Lemma E.2, we have

$$N_{k^\star}(T^{(1)}_{q(S,K)}) = N_{k^\star}(T^{(2)}_{q(S,K)}) \ge \frac{t^{(2)}_{q(S,K)}}{32K}, \quad N_k(T^{(2)}_{q(S,K)+1}) \ge \frac{t^{(2)}_{q(S,K)+1}}{32K}.$$

Therefore, we have

$$\begin{aligned}
\sum_{k\in A_{q(S,K)+1}} \frac{N_k(T^{(2)}_{q(S,K)+1})}{\sqrt{N_{k^\star}(T^{(1)}_{q(S,K)})}} + \frac{N_k(T)}{\sqrt{N_{k^\star}(T^{(2)}_{q(S,K)+1})}} &\le \frac{T^{(2)}_{q(S,K)+1}}{\sqrt{N_{k^\star}(T^{(1)}_{q(S,K)})}} + \frac{T}{\sqrt{N_{k^\star}(T^{(2)}_{q(S,K)+1})}} \\
&\le \frac{t^{(2)}_{q(S,K)+1}}{\sqrt{N_{k^\star}(T^{(1)}_{q(S,K)})}} + \frac{T}{\sqrt{N_{k^\star}(T^{(2)}_{q(S,K)+1})}} \\
&\le t^{(2)}_{q(S,K)+1}\sqrt{\frac{32K}{t^{(2)}_{q(S,K)}}} + T\sqrt{\frac{32K}{t^{(2)}_{q(S,K)+1}}} \\
&\le 16\mathcal{R}_2.
\end{aligned}$$

□

For any $k \in [K]$, define $\mathcal{R}(T;k) := \sum_{t=1}^{T}(\mu^\star - \mu_k)\mathbb{1}\{a_t = k\} = \Delta(k)N_k(T)$. If $\eta_k > 1$, then conditional on the events $\mathcal{E}$ and $\mathcal{E}_2$, we always have

$$\begin{aligned}
R(T;k) &= N_k(T)\Delta(k) = N_k(T^{(1)}_{\eta_k})\Delta(k) \\
&\le N_k(T^{(2)}_{\eta_k}) \cdot 4r_{k^\star}(T^{(1)}_{\eta_k-1}) + (N_k(T^{(1)}_{\eta_k}) - N_k(T^{(2)}_{\eta_k})) \cdot \Delta(k) && \text{(by (E.27))} \\
&\le N_k(T^{(2)}_{\eta_k}) \cdot 4r_{k^\star}(T^{(1)}_{\eta_k-1}) + (N_k(T^{(1)}_{\eta_k}) - N_k(T^{(2)}_{\eta_k})) \cdot 4r_{k^\star}(T^{(2)}_{\eta_k}) && \text{(by (E.28) \& (E.29))} \\
&\le N_k(T^{(2)}_{\eta_k})4\sqrt{\frac{6\log T}{N_{k^\star}(T^{(1)}_{\eta_k-1})}} + N_k(T^{(1)}_{\eta_k})4\sqrt{\frac{6\log T}{N_{k^\star}(T^{(2)}_{\eta_k})}} \\
&\le 4\sqrt{6\log T}\left(\frac{N_k(T^{(2)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(1)}_{\eta_k-1})}} + \frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(2)}_{\eta_k})}}\right).
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\sum_{k:\eta_k=1} R(T;k) &\le T^{(2)}_1 + \sum_{k:\eta_k=1}(N_k(T^{(1)}_1) - N_k(T^{(2)}_1)) \cdot \Delta(k) \\
&\le \mathcal{R}_2 + \sum_{k:\eta_k=1}(N_k(T^{(1)}_1) - N_k(T^{(2)}_1)) \cdot 4r_{k^\star}(T^{(2)}_1) && \text{(by (E.28))} \\
&\le \mathcal{R}_2 + 4\sqrt{6\log T}\sum_{k:\eta_k=1}\frac{N_k(T^{(1)}_1)}{\sqrt{N_{k^\star}(T^{(2)}_1)}} \\
&\le \mathcal{R}_2 + 4\sqrt{6\log T}\frac{T^{(1)}_1}{\sqrt{N_{k^\star}(T^{(2)}_1)}} \\
&\le \mathcal{R}_2 + 4\sqrt{6\log T}T^{(1)}_1\sqrt{\frac{2K}{t^{(2)}_1}} && \text{(by } N_{k^\star}(T^{(2)}_1) = n^{(2)}_1 \le \tfrac{t^{(2)}_1}{2K}\text{)} \\
&\le \mathcal{R}_2 + 8\sqrt{6\log T}\mathcal{R}_2.
\end{aligned}$$

Therefore, for any $\mathcal{D}$, conditional on the events $\mathcal{E}$ and $\mathcal{E}_2$, we have

$$\begin{aligned}
T\mu^\star - \sum_{t=1}^{T}\mu_{a_t} &= \sum_{k\in[K]} R(T;k) = \sum_{k:\eta_k=1} R(T;k) + \sum_{k:\eta_k>1} R(T;k) \\
&\le \mathcal{R}_2 + 8\sqrt{6\log T}\mathcal{R}_2 + 4\sqrt{6\log T}\left(\sum_{k:\eta_k>1}\left(\frac{N_k(T^{(2)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(1)}_{\eta_k-1})}} + \frac{N_k(T^{(1)}_{\eta_k})}{\sqrt{N_{k^\star}(T^{(2)}_{\eta_k})}}\right)\right).
\end{aligned}$$

Combining the above inequality with Lemma E.4, we have

$$\mathbb{E}^\pi_{\mathcal{D}}\left[T\mu^\star - \sum_{t=1}^{T}\mu_{a_t} \mid \mathcal{E}, \mathcal{E}_2\right] \le \mathcal{O}(\log K\sqrt{\log T})\mathcal{R}_2. \tag{E.34}$$

**Putting everything together.** Combining Eqs. (E.8), (E.25) and (E.34), we have

$$R^\pi(T) \le \mathcal{O}(\log K\sqrt{\log T}) \cdot \max\left\{\frac{K - \widehat{r}(S,K)}{K}\mathcal{R}_1, \mathcal{R}_2\right\}. \tag{E.35}$$

Since $\widehat{r}(S,K) \in [r(S,K)+1-q(S,K), r(S,K)]$ (note that $q(S,K) \geq 1$), we have

$$1 \leq \frac{K-\widehat{r}(S,K)}{K-r(S,K)} \leq \frac{K-r(S,K)-1+q(S,K)}{K-r(S,K)} = 1 + \frac{q(S,K)-1}{K-r(S,K)} \leq q(S,K).$$

If $q(S,K) \leq \log_2 \log_2(T)$, then (E.35) implies the upper bound in Theorem 6.1 because

$$\begin{aligned}
\frac{K-\widehat{r}(S,K)}{K}\mathcal{R}_1 &\leq \left(\frac{K-\widehat{r}(S,K)}{K-r(S,K)}\right)^2 \cdot \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}} \\
&\leq (q(S,K))^2 \cdot \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}} \\
&\leq 10\sqrt{\log_2 T} \cdot \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}}
\end{aligned}$$

and $\mathcal{R}_2 = K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}}$. If $q(S,K) \geq \log_2 \log_2(T)$, then the right-hand side of (E.35) is $\mathcal{O}(\log K \sqrt{\log T}) \cdot \sqrt{KT}$ because

$$\mathcal{R}_1 \leq (K-\widehat{r}(S,K))(T/(K-\widehat{r}(S,K)))^{\frac{1}{2-2^{-\log_2 \log_2(T)}}} \leq \sqrt{2(K-\widehat{r}(S,K))T} \leq \sqrt{2KT},$$

$$\mathcal{R}_2 \leq K(T/K)^{\frac{1}{2-2^{-\log_2 \log_2(T)-1}}} \leq \sqrt{2KT},$$

which also implies the upper bound in Theorem 6.1. Therefore, we finish the proof of Theorem 6.1. □

## E.9 Proof of Theorem 6.3

Consider an arbitrary switching graph $G$ whose switching costs satisfy the triangle inequality. Recall that $H$ is the total weight of the shortest Hamiltonian path in $G$.

### E.9.1 The HS-SE Policy is Indeed an $S$-Switching-Budget Policy

From round 1 to round $t_1$, HS-SE incurs $H$ switching cost.

For $1 \leq l \leq q'(S,G)-1$, from round $t_l$ to round $t_{l+1}$, no matter whether $l$ is odd or even, no matter whether the last action in epoch $l$ is eliminated before the start of epoch $l+1$ or not, by the switching order (determined by the shortest Hamiltonian path of $G$) and the triangle inequality, HS-SE always incurs at most $H$ switching cost.

From round $t_{q'(S,G)}$ to round $T$, since HS-SE does not switch within epoch $q'(S,G)+1$, i.e., from round $t_{q'(S,G)}+1$ to round $T$, the only possible switch is between round $t_{q'(S,G)}$ and $t_{q'(S,G)}+1$. Thus HS-SE incurs at most $\max_{i,j\in[k]} c_{i,j}$ switching cost from round $t_{q'(S,G)}$ to round $T$.

Summarizing the above arguments, we find that HS-SE incurs at most $q'(S,G)H + \max_{i,j\in[k]} c_{i,j} \leq S$ switching cost from round 1 to round $T$. Thus it is indeed an $S$-switching-budget policy.

### E.9.2 Proof of Upper Bound

The proof is essentially the same as Appendix E.7.2, with $q(S,K)$ replaced by $q'(S,G)$. □

## E.10 Proof of the Upper Bound in Theorem 6.5

Consider an arbitrary $\boldsymbol{c} \in \mathbb{R}_{\geq 0}^K$. Recall that $i_K \in \arg\max_{i\in[K]} c_i$, $i_1 \in \arg\max_{i\in[K]\setminus\{i_1\}} c_i$, $c^{(1)} = \max_{i\in[K]} = c_{i_K}$, $c^{(2)} = \max_{i\neq i_K} c_i = c_{i_1}$, and $\Sigma = \sum_{i=1}^K c_i$.

### E.10.1 The AS-SE Policy is Indeed an $S$-Switching-Budget Policy

From round 1 to round $T$, by the switching order specified in Algorithm 6.4, AS-SE departs from action $i_K$ for at most $\left\lceil \frac{q(S,\boldsymbol{c})}{2} \right\rceil$ times, departs from action $i_1$ for at most $\left\lfloor \frac{q(S,\boldsymbol{c})}{2} \right\rfloor + 1$ times, and departs from every other action for at most $q(S,\boldsymbol{c})$ times. The total switching cost is no larger than

$$\left\lceil \frac{q(S,\boldsymbol{c})}{2} \right\rceil c^{(1)} + \left( \left\lfloor \frac{q(S,\boldsymbol{c})}{2} \right\rfloor + 1 \right) c^{(2)} + q(S,\boldsymbol{c}) \sum_{i\in[K]\setminus\{i_1,i_K\}} c_i. \tag{E.36}$$

If $q(S,\boldsymbol{c}) = \max\left\{1 + 2\left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor, 2\left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor\right\} = 1 + 2\left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor$, then (E.36) is equal to

$$\begin{aligned}
&\left(1 + \left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor\right)\left(c^{(1)} + c^{(2)}\right) + \left(1 + 2\left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor\right)\left(\Sigma - c^{(1)} - c^{(2)}\right) \\
&= \Sigma + \left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor \left(2\Sigma - c^{(1)} - c^{(2)}\right) \\
&\leq \Sigma + S - \Sigma = S.
\end{aligned}$$

If $q(S,\boldsymbol{c}) = \max\left\{1 + 2\left\lfloor \frac{S-\Sigma}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor, 2\left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor\right\} = 2\left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor$, then (E.36) is equal to

$$\begin{aligned}
&\left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor \left(c^{(1)} + c^{(2)}\right) + c^{(2)} + 2\left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor \left(\Sigma - c^{(1)} - c^{(2)}\right) \\
&= c^{(2)} + \left\lfloor \frac{S-c^{(2)}}{2\Sigma - c^{(1)} - c^{(2)}} \right\rfloor \left(2\Sigma - c^{(1)} - c^{(2)}\right) \\
&\leq c^{(2)} + S - c^{(2)} = S.
\end{aligned}$$

Therefore, AS-SE is indeed an $S$-switching-budget policy.

### E.10.2 Proof of Upper Bound

The proof is essentially the same as Appendix E.7.2, with $q(S,K)$ replaced by $q(S,\boldsymbol{c})$. □

## E.11 Information-Theoretic Tools

In this section, we introduce our information-theoretic tools.

For any two probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on the same measurable space $(\Omega, \mathcal{F})$, let $D_{\mathrm{TV}}(\mathbb{P}\|\mathbb{Q}) := \sup_{E\in\mathcal{F}}|\mathbb{P}(E) - \mathbb{Q}(E)|$ denote the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$. We write $\mathbb{P} << \mathbb{Q}$ to indicate that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, and

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) := \begin{cases} \int_\Omega \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \text{if } \mathbb{P} << \mathbb{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

be the *Kullback-Leibler (KL) divergence* between $\mathbb{P}$ and $\mathbb{Q}$. Furthermore, let

$$D_{\mathrm{re}}(\mathbb{P} \,\|\, \mathbb{Q}) := D_{\mathrm{KL}}(\mathbb{Q} \,\|\, \mathbb{P})$$

be the *reverse KL divergence* between $\mathbb{P}$ and $\mathbb{Q}$. For any $p, q \in [0,1]$, let

$$d_{\mathrm{TV}}(p \,\|\, q) := D_{\mathrm{TV}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)),$$

$$d_{\mathrm{KL}}(p \,\|\, q) := D_{\mathrm{TV}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q}),$$

$$d_{\mathrm{re}}(p \,\|\, q) := D_{\mathrm{re}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) = q\log(\frac{q}{p}) + (1-q)\log(\frac{1-q}{1-p}),$$

where $\mathrm{Ber}(p)$ stands for the Bernoulli distribution with mean $p$. More generally, for any divergence denoted by $D(\cdot \,\|\, \cdot)$, let

$$d(p \,\|\, q) := D(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)).$$

### E.11.1 Reverse Fano-Type Inequalities

We first introduce a basic version of the reverse Fano-type inequality below.

**Proposition E.3** (Reverse Fano-Type Inequality)**.** *Let $D$ be the KL divergence or the reverse KL divergence. Let $\mathbb{P}_1, \ldots, \mathbb{P}_N$ and $\mathbb{Q}$ be arbitrary probability measures on a common measurable space $(\Omega, \mathcal{F})$. For any measurable function $\psi : \Omega \mapsto [N]$, we have*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(\psi = i) \geq \frac{1}{N} - \frac{1}{N}\sqrt{2\left(1 - \frac{1}{N}\right)\sum_{i=1}^N D(\mathbb{P}_i \,\|\, \mathbb{Q})}. \tag{E.37}$$

*More generally, let $\mathbb{Q}_1, \ldots, \mathbb{Q}_N$ be arbitrary probability measures on $(\Omega, \mathcal{F})$. For any sequence of events $E_1, \ldots, E_N \in \mathcal{F}$ (not necessarily disjoint), if $\bar{q} := \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(E_i) \in [0, \frac{1}{2}]$, then*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(E_i) \geq \bar{q} - \sqrt{2\bar{q}(1-\bar{q})\frac{1}{N}\sum_{i=1}^N D(\mathbb{P}_i \,\|\, \mathbb{Q}_i)}. \tag{E.38}$$

To the best of our knowledge, both (E.37) and (E.38) are new. Note that the classical Fano's

inequality (E.1) provides a lower bound on the minimum "average error probability"

$$\inf_{\psi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi \neq i).$$

Our inequality (E.37) provides a sharp upper bound on the maximum "average error probability"

$$\sup_{\psi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi \neq i),$$

thus can be viewed as a reverse version of the classical Fano's inequality. Our inequality (E.38) further generalizes (E.37) to arbitrary events.

*Remark 1.* While there are some existing inequalities sometimes referred to as "reverse Fano's inequalities" in the literature (e.g., Chu and Chueh 1966, Tebbe and Dwyer 1968), they are very different from (E.37), as all of them only provide an upper bound on $\inf_\psi \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi \neq i)$ rather than $\sup_\psi \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi \neq i)$, i.e., their upper bound only holds for the minimax test and does not hold for an *arbitrary* test $\phi$. For this reason, we call the inequalities in Proposition E.3 "reverse **Fano-type** inequalities" to distinguish them from the existing "reverse **Fano's** inequalities" in the literature.

*Remark 2.* Gerchinovitz et al. (2020) provide a very general framework to derive Fano-type inequalities, and their results imply two related inequalities. In the setting of (E.37), their results in their Section 4.2 imply

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(\psi = i) \geq \frac{1}{N} - \sqrt{\frac{1}{N \log N} \sum_{i=1}^{N} D(\mathbb{P}_i \parallel \mathbb{Q})},$$

which is worse than our (E.37) by a $\frac{1}{\sqrt{N}}$ factor — importantly, this worse result cannot help us to obtain a tight lower bound for U-BwSC whenever $K - r(S, K) = o(K)$. In the setting of (E.38), Gerchinovitz et al. (2020) show that

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(E_i) \geq 1 - \frac{\frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}(\mathbb{P}_i \parallel \mathbb{Q}) + \log 2}{\log\left(\frac{1}{1-\bar{q}}\right)}.$$

However, this bound becomes meaningless under our condition $\bar{q} \in [0, \frac{1}{2}]$. In other words, this bound is only useful for proving lower bounds for "high-probability events" rather than "low-probability events" — the latter is required in the proof of U-BwSC lower bounds.

### E.11.2 Generalized Reverse Fano-Type Inequalities

In this subsection, we introduce a more general version of the inequalities in Proposition E.3, which enjoy several advantages as described in Appendix E.4. We then give the proof.

**Proposition E.4** (Generalized Reverse Fano-Type Inequality)**.** *Let $D$ be the KL divergence or the reverse KL divergence. Let $(\Omega_1, \mathcal{F}_1), \ldots, (\Omega_N, \mathcal{F}_N)$ be an arbitrary sequence of measurable spaces.*

*For any $i \in [N]$, let $\mathbb{P}_i$ and $\mathbb{Q}_i$ be arbitrary probability measures on $(\Omega_i, \mathcal{F}_i)$, and $E_i \in \mathcal{F}_i$ be an arbitrary event. We have*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(E_i) \geq \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i) - \sqrt{2\cdot\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i)\cdot\frac{1}{N}\sum_{i=1}^{N}D(\mathbb{P}_i \,\|\, \mathbb{Q}_i)}. \tag{E.39}$$

*Moreover, if $\bar{q} := \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i) \in [0, \frac{1}{2}]$, then we have a slightly tighter bound*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(E_i) \geq \bar{q} - \sqrt{2\bar{q}(1-\bar{q})\cdot\frac{1}{N}\sum_{i=1}^{N}D(\mathbb{P}_i \,\|\, \mathbb{Q}_i)}. \tag{E.40}$$

*Proof of Proposition E.4.* Our proof builds on a two-step procedure established by Gerchinovitz et al. (2020), with a few key modifications in the second step to obtain sharper one-sided inequalities.

Our first step is a reduction to Bernoulli distributions. By the joint convexity of general $f$-divergences, we have

$$d\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(E_i) \,\|\, \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i)\right) \leq \frac{1}{N}\sum_{i=1}^{N}d(\mathbb{P}_i(E_i) \,\|\, \mathbb{Q}_i(E_i)) = \frac{1}{N}\sum_{i=1}^{N}d(\mathbb{P}_i(E_i) \,\|\, \mathbb{Q}_i(E_i)).$$

Note that the above inequality holds even if $\mathbb{P}_1, \ldots, \mathbb{P}_N$ are on different measurable spaces. For all $i \in [N]$, since $\mathrm{Ber}(\mathbb{P}_i(E_i))$ (resp., $\mathrm{Ber}(\mathbb{Q}_i(E_i))$) is the law of $\mathbb{1}_{E_i}$ under $\mathbb{P}_i$ (resp., $\mathbb{Q}_i$), using the data-processing inequality for $f$-divergences (see, e.g., Lemma 1 in Gerchinovitz et al. 2020), we have

$$d(\mathbb{P}_i(E_i) \,\|\, \mathbb{Q}_i(E_i)) = D(\mathrm{Ber}(\mathbb{P}_i(E_i)) \,\|\, \mathrm{Ber}(\mathbb{Q}_i(E_i))) \leq D(\mathbb{P}_i' \,\|\, \mathbb{Q}_i').$$

Thus we have

$$d\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(E_i) \,\|\, \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i(E_i)\right) \leq \frac{1}{N}\sum_{i=1}^{N}D(\mathbb{P}_i \,\|\, \mathbb{Q}_i).$$

Let $\bar{p} := \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(A_i)$, we have

$$d_f(\bar{p} \,\|\, \bar{q}) \leq \frac{1}{N}\sum_{i=1}^{N}D(\mathbb{P}_i \,\|\, \mathbb{Q}_i). \tag{E.41}$$

In the second step, we lower bound $d(\bar{p} \,\|\, \bar{q})$ to extract a lower bound on $\bar{p}$. When $D$ is restricted to be the KL divergence or the reverse KL divergence, Lemma E.5 and Lemma E.6, we have

$$d(\bar{p} \,\|\, \bar{q}) \geq \frac{(\bar{p}-\bar{q})^2}{2\bar{q}}$$

for all $q \in [0, 1)$ and

$$d(\bar{p} \,\|\, \bar{q}) \geq \frac{(\bar{p}-\bar{q})^2}{2\bar{q}(1-\bar{q})}$$

for all $q \in [0, \frac{1}{2}]$. Note that Lemma E.5 and Lemma E.6 provide "localized" versions of the Pinsker's inequality that substantially improves over existing "global" variants of the Pinsker's inequality by

exploiting the one-sided condition $\bar{p} \le \bar{q}$ (see the remarks after Lemma E.5 and Lemma E.6). Such improvement is critical for our second step and enables us to obtain tight one-sided inequalities about $\bar{p}$ (with improved dependence on $\bar{q}$), which we describe below:

- If $\bar{p} \notin [0, \bar{q}]$, then $\bar{p} \ge \bar{q}$.
- If $\bar{p} \in [0, \bar{q}]$, then we have
$$d(\bar{p} \parallel \bar{q}) \ge \frac{(\bar{p} - \bar{q})^2}{2\bar{q}},$$
which implies $\bar{p} \ge \bar{q} - \sqrt{2\bar{q} d(\bar{p} \parallel \bar{q})}$.

Therefore, no matter $\bar{p} \in [0, \bar{q}]$ or not, we always have

$$\bar{p} \ge \bar{q} - \sqrt{2\bar{q} d_f(\bar{p} \parallel \bar{q})}. \tag{E.42}$$

The above inequality can be improved to $\bar{p} \ge \bar{q} - \sqrt{2\bar{q}(1-\bar{q}) d(\bar{p} \parallel \bar{q})}$ when $\bar{q} \in [0, \frac{1}{2}]$.

By (E.41), we prove (E.39). □

### E.11.3 Localized Pinsker's Inequalities

**Lemma E.5** (Localized Pinsker's Inequality)**.** *If $0 \le p \le q \le \frac{1}{2}$ or $\frac{1}{2} \le q \le p \le 1$, then*

$$d_{\mathrm{KL}}(p \parallel q) \ge \frac{(p-q)^2}{2q(1-q)} \quad \text{and} \quad d_{\mathrm{KL}}(q \parallel p) \ge \frac{(p-q)^2}{2q(1-q)}.$$

*Proof of Lemma E.5.* If $p = q = 0$ or $p = q = 1$, then $d_{\mathrm{KL}}(p \parallel q) = d_{\mathrm{KL}}(q \parallel p) = 0 = \frac{(p-q)^2}{2q(1-q)}$. In the rest of the proof, we fix $q \in (0, 1)$.

We first define

$$g(x) := \mathrm{kl}(x, q) - \frac{(x-q)^2}{2q(1-q)} = x \log \frac{x}{q} + (1-x) \log \frac{1-x}{1-q} - \frac{(x-q)^2}{2q(1-q)}$$

for all $x \in [0, 1]$. We have

$$g'(x) = \log\left(\frac{x}{1-x}\frac{1-q}{q}\right) - \frac{x-q}{q(1-q)},$$
$$g''(x) = \frac{1}{x(1-x)} - \frac{1}{q(1-q)}.$$

We discuss two cases:

- If $q \le \frac{1}{2}$, then $g''(x) \ge 0$ for all $x \in (0, q]$. Furthermore, since $g'(q) = 0$, we have $g'(x) \le 0$ for all $x \in (0, q)$, which implies that $g(x) \ge g(q) = 0$ for all $x \in [0, q]$. Thus $g(p) = \mathrm{kl}(p, q) - \frac{(p-q)^2}{2q(1-q)} \ge 0$ when $0 \le p \le q \le \frac{1}{2}$.
- If $q \ge \frac{1}{2}$, then $g''(x) \ge 0$ for all $x \in [q, 1)$. Furthermore, since $g'(q) = 0$, we have $g'(x) \ge 0$ for all $x \in (q, 1)$, which implies that $g(x) \ge g(q) = 0$ for all $x \in [q, 1]$. Thus $g(p) = \mathrm{kl}(p, q) - \frac{(p-q)^2}{2q(1-q)} \ge 0$ when $\frac{1}{2} \le q \le p \le 1$.

We then define

$$g(x) := \mathrm{kl}(q, x) - \frac{(x-q)^2}{2q(1-q)} = q \log \frac{q}{x} + (1-q) \log \frac{1-q}{1-x} - \frac{(x-q)^2}{2q(1-q)}$$

for all $x \in (0,1)$. We have

$$g'(x) = \frac{x-q}{x(1-x)} - \frac{x-q}{q(1-q)} = \frac{(x+q-1)(x-q)^2}{x(1-x)q(1-q)}.$$

We discuss two cases:

- If $q \le \frac{1}{2}$, then $x+q-1 \le 0$ and $g'(x) \le 0$ for all $x \in (0,q]$, which implies that $g(x) \ge g(q) = 0$ for all $x \in (0,q]$. Thus $g(p) = \mathrm{kl}(q,p) - \frac{(p-q)^2}{q(1-q)} \ge 0$ when $0 \le p \le q \le \frac{1}{2}$.
- If $q \ge \frac{1}{2}$, then $x+q-1 \ge 0$ and $g'(x) \ge 0$ for all $x \in [q,1)$, which implies that $g(x) \ge g(q) = 0$ for all $x \in (q,1]$. Thus $g(p) = \mathrm{kl}(q,p) - \frac{(p-q)^2}{q(1-q)} \ge 0$ when $\frac{1}{2} \le q \le p \le 1$.

To sum up, if $0 \le p \le q \le \frac{1}{2}$ or $\frac{1}{2} \le q \le p \le 1$, then $\mathrm{kl}(p,q) \ge \frac{(p-q)^2}{2q(1-q)}$ and $\mathrm{kl}(q,p) \ge \frac{(p-q)^2}{2q(1-q)}$. □

**Lemma E.6** (Localized Pinsker's Inequality, version 2)**.** *If $0 \le p \le q \le 1$, then*

$$d_{\mathrm{KL}}(p \parallel q) \ge \frac{(p-q)^2}{2q} \quad \text{and} \quad d_{\mathrm{KL}}(q \parallel p) \ge \frac{(p-q)^2}{2q}.$$

Lemma E.6 is a corollary of Lemma A.2 in Talebi Mazraeh Shahi (2017). It has a slightly worse constant compared with Lemma E.5, but holds for a more general range of $\bar{q}$.

*Remark.* The classical Pinsker's inequality (see, e.g., Lemma 2.5 in Tsybakov 2009) for Bernoulli distributions states that

$$d_{\mathrm{KL}}(p \parallel q) \ge 2(d_{\mathrm{TV}}(p \parallel q))^2 = 2(p-q)^2$$

for all $p, q \in [0,1]$. Many improvements and generalizations of the Pinsker's inequality have been obtained in the literature, including the "refined Pinsker's inequality" by Ordentlich and Weinberger (2005), which states that

$$d_{\mathrm{KL}}(p \parallel q) \ge \frac{\log((1-q)/q)}{1-2q}(p-q)^2$$

for all $p, q \in [0,1]$. The above bounds become substantially weaker than the $\frac{(p-q)^2}{2q(1-q)}$ bound in Lemma E.5 and the $\frac{(p-q)^2}{2q}$ bound in Lemma E.6 as $q$ gets closer to 0, i.e., they lose an $\widetilde{O}(1/q)$ factor when $q \to 0$. In fact, all variants of the Pinsker's inequality that seek to establish a *global* bound which holds for all $p, q \in [0,1]$ must lose such a huge factor compared with Lemmas E.5 and E.6, thus are loose for our purpose (note that Lemmas E.5 and E.6 critically utilizes the *one-sided* condition: $p \le q$). It is also worth noting that Lemmas E.5 and E.6 hold for not only the KL divergence $d_{\mathrm{KL}}(p \parallel q)$ but also the *reverse* KL divergence $d_{\mathrm{re}}(p \parallel q)$. This feasture is crucial for us to prove Proposition E.4 and establish tight lower bounds on the regret of the BwSC problem.

## E.12 Proof of Theorem 6.2

For an overview of the proof, see Appendix E.4.2.

Given any $K > 1$, $S \geq 0$ and $T \geq 2K$, we focus on the setting of $\mathcal{D}_k = \mathcal{N}(\mu_k, 1)$ $(\forall k \in [K])$, as this is sufficient for us to prove the desired lower bound. Note that now the underlying environment (i.e., latent distributions) $\mathcal{D}$ can be completely determined by a vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K) \in \mathbb{R}^K$. For simplicity, in this proof we will directly use the vector $\boldsymbol{\mu}$ to represent the environment.

For any environment $\boldsymbol{\mu}$, let $X_{\boldsymbol{\mu}}^t(k) \sim \mathcal{N}(\mu_k, 1)$ denote the i.i.d. random reward of each action $k$ at round $t$ $(k \in [K], t \in [T])$. For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, for any $t \in [T]$, we use $a_t$ to denote the random action selected by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$, and use $X_{\boldsymbol{\mu}}^t(a_t)$ to denote the random reward observed by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$. Let $\mathcal{H}_t := \big((a_1, X_{\boldsymbol{\mu}}^1(a_1)), \ldots, (a_t, X_{\boldsymbol{\mu}}^t(a_t))\big)$ be the history (of actions and observations) up to round $t$ (inclusive), whose value lies in $\Omega_t := ([K] \times \mathbb{R})^t$. Let $\mathcal{F}_t := \mathcal{B}(\Omega_t)$ be the Borel $\sigma$-algebra on $\Omega_t$. Let $\mathbb{P}_{\boldsymbol{\mu}}^\pi$ be the probability measure induced by (i.e., the law of) $\mathcal{H}_T$, and $\mathbb{E}_{\boldsymbol{\mu}}^\pi$ be the associated expectation operator. Let $R_{\boldsymbol{\mu}}^\pi(T) := T\mu^\star - \mathbb{E}_{\boldsymbol{\mu}}^\pi\left[\sum_{t=1}^T \mu_{a_t}\right]$ be policy $\pi$'s distribution-dependent regret under environment $\boldsymbol{\mu}$.

We argue that in our proof, we only need to consider the case of $q(S,K) + 2 \leq \log_2 \log_2(T/K)$. Suppose $q(S,K) + 2 > \log_2 \log_2(T/K)$, then we have

$$\begin{aligned}
K^{1-\frac{1}{2-2^{-q(S,K)}}} T^{\frac{1}{2-2^{-q(S,K)}}} &= K(T/K)^{\frac{1}{2-2^{-q(S,K)}}} \\
&= K(T/K)^{\frac{1}{2}} (T/K)^{\frac{2^{-q(S,K)-1}}{2-2^{-q(S,K)}}} \\
&\leq \sqrt{KT}(T/K)^{2^{-q(S,K)-1}} \\
&< \sqrt{KT}(T/K)^{2^{-\log_2 \log_2(T/K)+1}} \\
&= \sqrt{KT}(T/K)^{2\log_{T/K}(2)} = 4\sqrt{KT},
\end{aligned}$$

thus the lower bound in Theorem 6.2 becomes $\Omega(\sqrt{KT}/\log T)$ and can be directly obtained by applying the well-known $\Omega(\sqrt{KT})$ lower bound of the classical MAB (see, e.g., Lattimore and Szepesvári 2020). Therefore, the really non-trivial case of Theorem 6.2 is the case of $q(S,K)+2 \leq \log_2 \log_2(T/K)$, and we focus on this case in the rest of our proof.

Our goal is to explicitly construct a family of environments $\Phi$, such that for any $S$-switching-budget policy $\pi \in \Pi_S$, the "average-case regret" $\frac{1}{|\Phi|}\sum_{\boldsymbol{\mu}\in\Phi} R_{\boldsymbol{\mu}}^\pi(T)$ is lower bounded by both

$$\widetilde{\Omega}\left(\frac{(K - r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}}\right) \tag{E.43}$$

and

$$\widetilde{\Omega}\Big(K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}}\Big). \tag{E.44}$$

Since the worst-case regret $R^\pi(T)$ is no less than the "average-case regret" $\frac{1}{|\Phi|}\sum_{\boldsymbol{\mu}\in\Phi} R_{\boldsymbol{\mu}}^\pi(T)$, the above goal directly implies Theorem 6.2. In our proof, we construct two classes of environments $\Phi_1$ and $\Phi_2$ to show the lower bounds (E.43) and (E.44) respectively.

For notational simplicity, we **redefine** the sequence $(t_j^{(1)})_{j=0}^{q(S,G)+1}$ as $t_0^{(1)} = 0$ and

$$t_j^{(1)} = \left\lfloor (K - r(S,K))^{1-\frac{2-2^{1-j}}{2-2^{-q(S,K)}}} T^{\frac{2-2^{1-j}}{2-2^{-q(S,K)}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S,K)+1. \tag{E.45}$$

Note that the above definition is slightly different from the definition of $(t_j^{(1)})_{j=0}^{q(S,K)+1}$ in Algorithm 6.2 (specifically, $\widehat{r}(S,K)$ is replaced by $r(S,K)$) — we only use the above definition in this proof, for the purpose of making the analysis cleaner.[65] Meanwhile, we keep the definition of the sequence $(t_j^{(2)})_{j=0}^{q(S,K)+2}$ the same as the original definition of $(t_j^{(2)})_{j=0}^{q(S,K)+2}$ in Algorithm 6.2.

Our lower bound proof program consists of five steps:

1. **R**isky **E**vents
2. **C**ombinatorial arguments and lower bounds under a single environment
3. **A**lternative environments, bad events, and lower bound reductions
4. **P**robability space changing tricks
5. Applying the GRF inequality

Based on the initials of the first four steps, we call the program RECAP. We present the five steps in the five subsections below.

### E.12.1 Definitions of Risky Events

For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we make some key definitions below.

1. For any $n_1, n_2 \in [T]$, we define a random variable $S(n_1, n_2)$ to be the total switching cost incurred in period $[n_1 : n_2]$ (note that if there is a switch happening between round $n_1 - 1$ and round $n_1$, or between round $n_2$ and round $n_2 + 1$, we do not count its cost in $S(n_1, n_2)$).

2. Second, we define a stopping time

$$\tau := \min\{t \in [T] : S(1:t) = S\}$$

if the set is non-empty and $\tau = \infty$ otherwise. That is, $\tau$ is the first round that the learner's total switching cost reaches $S$.

3. We define a class of *risky* events as follows: for any $k \in [K]$, let
$E_{1,k}^{(1)} := \left\{\text{action } k \text{ is not chosen in period } \left[1 : t_1^{(1)}\right]\right\}$,
$E_{j,k}^{(1)} := \left\{\text{action } k \text{ is not chosen in period } \left[t_{j-1}^{(1)} : t_j^{(1)}\right]\right\}, \quad \forall j \in [2 : q(S,K)]$,
$E_{q(S,K)+1,k}^{(1)} := \left\{\text{action } k \text{ is not chosen in period } \left[t_{q(S,K)}^{(1)} : \lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor\right]\right\}$,
$E_{q(S,K)+2,k}^{(1)} := \left\{\tau \le \lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor,\, a_\tau = k,\, \text{action } k \text{ is not chosen in period } \left[t_{q(S,K)}^{(1)} : \tau - 1\right]\right\}$.
By doing so, we get $(q(S,K)+2)K$ risky events (of the form $E_{j,k}^{(1)}$) in total. Note that the time points $(t_j^{(1)})_{j=1}^{q(S,K)+1}$ are fixed and given in (E.45), and the events $(E_{q(S,K)+2,k}^{(1)})_{k\in[K]}$ are defined based on the stopping time $\tau$. We refer to the above class of risky events as "the first class of risky events," and will use them to prove the lower bound (E.43).

4. We then define another class of risky events: for any $k \in [K]$, let
$E_{1,k}^{(2)} := \left\{\text{action } k \text{ is not chosen in period } \left[1 : t_1^{(2)}\right]\right\}$,

[65] The quantity $\widehat{r}(S,K)$ is introduced in Algorithm 6.2 (as a "proxy" of $r(S,K)$) to help us control the total number of switches more conveniently. In the lower bound analysis, directly dealing with $r(S,K)$ (rather than the algorithmic proxy $\widehat{r}(S,K)$) is better because $r(S,K)$ is the quantity that actually appears in the regret bound.

$E_{j,k}^{(2)} := \Big\{\text{action } k \text{ is not chosen in period } \Big[t_{j-1}^{(2)} : t_j^{(2)}\Big]\Big\}, \quad \forall j \in [2 : q(S,K)+2]$.
By doing so, we get $(q(S,K)+2)K$ risky events (of the form $E_{j,k}^{(2)}$) in total. Note that the time points $(t_j^{(2)})_{j=1}^{q(S,K)+1}$ are fixed and given in Algorithm 6.2. We refer to the above class of risky events as "the second class of risky events," and will use them to prove the lower bound (E.44).

*Remark.* Note that in the first class of risky events, the events $(E_{q(S,K)+2,k}^{(1)})_{k\in[K]}$ are defined based on the stopping time $\tau$. Such delicate design is crucial for our analysis — the importance should become clear quite soon. In particular, if we do not define the events $(E_{q(S,K)+2,k}^{(1)})_{k\in[K]}$ in this step, but only define the events $(E_{q(S,K)+2,k}^{(1)})_{j\in[q(S,K)+1],k\in[K]}$ (which do not involve $\tau$), then in the next step, we can only show

$$\sum_{j\in[q(S,K)+1]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) \geq K - 1 - \left\lfloor \frac{S}{q(S,K)+1} \right\rfloor$$

for Lemma E.7, which is unfortunately a meaningless result when $S\%(K-1) = 0$ (i.e., when $S$ is at the end of a "phase" defined in Section 6.3.4), as the right-hand side $K - 1 - \left\lfloor \frac{S}{q(S,K)+1} \right\rfloor$ becomes 0 when $S\%(K-1) = 0$ — this issue will eventually prevent us from making the floor function $\left\lfloor \frac{S-1}{K-1} \right\rfloor$ appear in the lower bound. By contrast, by defining the events $(E_{q(S,K)+2,k}^{(1)})_{k\in[K]}$ based on $\tau$, we are able to show

$$\sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) \geq K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil$$

for Lemma E.7. Importantly, the right-hand side is always no less than 1 — this property will play a fundamental role in subsequent analysis.

### E.12.2 Combinatorial Arguments and Lower Bounds for Risky Events (Under a Single Environment)

The main purpose of this subsection is to prove the following two lemmas (Lemma E.7 and Lemma E.8) using (non-trivial) combinatorial (and probabilistic) arguments. The arguments extensively exploit the properties of the switching constraint.

**Lemma E.7.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) \geq K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil.$$

**Lemma E.8.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(2)}\Big) \geq K - 1 - \left\lfloor \frac{S}{q(S,K)+2} \right\rfloor.$$

Lemma E.7 and Lemma E.8 lead to the following corollary, which will be utilized in subsequent subsections.

**Corollary E.2.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) \geq \frac{K-r(S,K)}{2(q(S,K)+2)^2K},$$

$$\frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(2)}\Big) \geq \frac{1}{2(q(S,K)+2)^2}.$$

Corollary E.2 tells us the following fact: under any *single* environment $\boldsymbol{\mu}$, the average probability of the first class of risky events is $\widetilde{\Omega}\Big(\frac{K-r(S,K)}{K}\Big)$, and the average probability of the second class of risky events is $\widetilde{\Omega}(1)$.

In the rest of this subsection, we provide proofs for Lemma E.7, Lemma E.8 and Corollary E.2.

*Proof of Lemma E.7.* For any $j \in [q(S,K)]$, we have

$$\begin{aligned}
\sum_{k\in[K]} \mathbb{1}\Big\{E_{j,k}^{(1)}\Big\} &= \text{number of actions that are not chosen in period } \Big[t_{j-1}^{(1)} : t_j^{(1)}\Big] \\
&\geq K - 1 - S\Big(t_{j-1}^{(1)} : t_j^{(1)}\Big) \\
&\geq \left(K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil\right) \mathbb{1}\left\{S\Big(t_{j-1}^{(1)} : t_j^{(1)}\Big) < \frac{S}{q(S,K)+1}\right\}
\end{aligned}$$

almost surely. Thus for any $j \in [q(S,K)]$, we have

$$\begin{aligned}
\sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) &= \sum_{k\in[K]} \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\Big[\mathbb{1}\Big\{E_{j,k}^{(1)}\Big\}\Big] \\
&= \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\sum_{k\in[K]} \mathbb{1}\Big\{E_{j,k}^{(1)}\Big\}\right] \\
&\geq \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\left(K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil\right) \mathbb{1}\left\{S\Big(t_{j-1}^{(1)} : t_j^{(1)}\Big) < \frac{S}{q(S,K)+1}\right\}\right] \\
&= \left(K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil\right) \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\mathbb{1}\left\{S\Big(t_{j-1}^{(1)} : t_j^{(1)}\Big) < \frac{S}{q(S,K)+1}\right\}\right]. \qquad \text{(E.46)}
\end{aligned}$$

Summing (E.46) over $j \in [q(S,K)]$, we have

$$\begin{aligned}
\sum_{j\in[q(S,K)]} \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big) &\geq \left(K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil\right) \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\sum_{j\in[q(S,K)]} \mathbb{1}\left\{S\Big(t_{j-1}^{(1)} : t_j^{(1)}\Big) < \frac{S}{q(S,K)+1}\right\}\right] \\
&\overset{\text{(i)}}{\geq} \left(K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil\right) \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\left(1 - \mathbb{1}\left\{S\Big(1 : t_{q(S,K)}^{(1)}\Big) \geq \frac{q(S,K)}{q(S,K)+1} S\right\}\right)\right].
\end{aligned}$$

Note that (i) follows from

$$\sum_{j\in[q(S,K)]} \mathbb{1}\left\{S\left(t_{j-1}^{(1)}:t_j^{(1)}\right) < \frac{S}{q(S,K)+1}\right\} \geq \mathbb{1}\left\{\bigcup_{j\in[q(S,K)]}\left\{S\left(t_{j-1}^{(1)}:t_j^{(1)}\right) < \frac{S}{q(S,K)+1}\right\}\right\}$$
$$\overset{\text{(ii)}}{\geq} \mathbb{1}\left\{S\left(1:t_{q(S,K)}^{(1)}\right) < \frac{q(S,K)}{q(S,K)+1}S\right\}$$
$$= 1 - \mathbb{1}\left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\},$$

where (ii) follows from the pigeonhole principle.

Now we define

$$E_{\sim,k}^{(1)} := \left\{\text{action } k \text{ is not among the first } \left\lceil \frac{S}{q(S,K)+1}\right\rceil \text{ (different) actions chosen in period } \left[t_{q(S,K)}^{(1)}:T\right]\right\}.$$

If $\left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}$ happens, then by the switching constraint $S(1:T)\leq S$, we have $S\left(t_{q(S,K)}^{(1)}:T\right) \leq \frac{S}{q(S,K)+1}$. If we further assume $E_{\sim,k}^{(1)}$ happens, then by $S\left(t_{q(S,K)}^{(1)}:T\right) \leq \frac{S}{q(S,K)+1}$, either $E_{q(S,K)+1,k}^{(1)} = \left\{\text{action } k \text{ is not chosen in period } \left[t_{q(S,K)}^{(1)}:\lfloor (t_{q(S,K)}^{(1)}+T)/2\rfloor\right]\right\}$ happens, or

$$E_{q(S,K)+2,k}^{(1)} = \left\{\tau \leq \lfloor (t_{q(S,K)}^{(1)}+T)/2\rfloor,\ a_\tau = k,\ \text{action } k \text{ is not chosen in period } \left[t_{q(S,K)}^{(1)}:\tau-1\right]\right\}$$

happens. Therefore, we know that

$$E_{q(S,K)+1,k}^{(1)} \cup E_{q(S,K)+2,k}^{(1)} \ \supset\ E_{\sim,k}^{(1)} \cap \left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}$$

This implies that

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\left(E_{q(S,K)+1,k}^{(1)} \cup E_{q(S,K)+2,k}^{(1)}\right)\\
&\geq \sum_{k\in[K]} \mathbb{P}_{\boldsymbol{\mu}}^{\pi}\left(E_{\sim,k}^{(1)} \cap \left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}\right)\\
&= \sum_{k\in[K]} \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\mathbb{1}\left\{E_{\sim,k}^{(1)}\right\}\mathbb{1}\left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}\right]\\
&= \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\sum_{k\in[K]} \mathbb{1}\left\{E_{\sim,k}^{(1)}\right\}\mathbb{1}\left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}\right]\\
&\overset{\text{(iii)}}{\geq} \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\left(K - \left\lceil \frac{S}{q(S,K)+1}\right\rceil\right)\mathbb{1}\left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}\right]\\
&= \left(K - \left\lceil \frac{S}{q(S,K)+1}\right\rceil\right)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\mathbb{1}\left\{S\left(1:t_{q(S,K)}^{(1)}\right) \geq \frac{q(S,K)}{q(S,K)+1}S\right\}\right],
\end{aligned}$$

where (iii) follow from the definition of $E_{\sim,k}^{(1)}$.

Combining the above two paragraphs, we have

$$\begin{aligned}
&\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big)\\
&\geq \sum_{j\in[q(S,K)]}\sum_{k\in[K]}\mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(1)}\Big)+\sum_{k\in[K]}\mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{q(S,K)+1,k}^{(1)}\cup E_{q(S,K)+2,k}^{(1)}\Big)\\
&\geq \left(K-\left\lceil\frac{S}{q(S,K)+1}\right\rceil\right)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\left(1-\mathbb{1}\left\{S\Big(1:t_{q(S,K)}^{(1)}\Big)\geq\frac{q(S,K)}{q(S,K)+1}S\right\}\right)\right]\\
&\quad+\left(K-\left\lceil\frac{S}{q(S,K)+1}\right\rceil\right)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\mathbb{1}\left\{S\Big(1:t_{q(S,K)}^{(1)}\Big)\geq\frac{q(S,K)}{q(S,K)+1}S\right\}\right]\\
&=K-\left\lceil\frac{S}{q(S,K)+1}\right\rceil.
\end{aligned}$$

□

*Proof of Lemma E.8.* For any $j\in[q(S,K)]$, we have

$$\begin{aligned}
\sum_{k\in[K]}\mathbb{1}\left\{E_{j,k}^{(2)}\right\}&=\text{number of actions that are not chosen in period }\left[t_{j-1}^{(2)}:t_{j}^{(2)}\right]\\
&\geq K-1-S\Big(t_{j-1}^{(2)}:t_{j}^{(2)}\Big)\\
&\geq\left(K-1-\left\lfloor\frac{S}{q(S,K)+2}\right\rfloor\right)\mathbb{1}\left\{S\Big(t_{j-1}^{(2)}:t_{j}^{(2)}\Big)\leq\frac{S}{q(S,K)+2}\right\}
\end{aligned}$$

almost surely. Thus for any $j\in[q(S,K)+2]$, we have

$$\begin{aligned}
\sum_{k\in[K]}\mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(2)}\Big)&=\sum_{k\in[K]}\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\mathbb{1}\left\{E_{j,k}^{(2)}\right\}\right]\\
&=\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\sum_{k\in[K]}\mathbb{1}\left\{E_{j,k}^{(2)}\right\}\right]\\
&\geq\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\left(K-1-\left\lfloor\frac{S}{q(S,K)+2}\right\rfloor\right)\mathbb{1}\left\{S\Big(t_{j-1}^{(2)}:t_{j}^{(2)}\Big)\leq\frac{S}{q(S,K)+2}\right\}\right]\\
&=\left(K-1-\left\lfloor\frac{S}{q(S,K)+2}\right\rfloor\right)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\mathbb{1}\left\{S\Big(t_{j-1}^{(2)}:t_{j}^{(2)}\Big)\leq\frac{S}{q(S,K)+2}\right\}\right]
\end{aligned}\tag{E.47}$$

Summing (E.47) over $j\in[q(S,K)+2]$, we have

$$\begin{aligned}
&\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}_{\boldsymbol{\mu}}^{\pi}\Big(E_{j,k}^{(2)}\Big)\\
&\geq\left(K-1-\left\lfloor\frac{S}{q(S,K)+2}\right\rfloor\right)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\sum_{j\in[q(S,K)+2]}\mathbb{1}\left\{S\Big(t_{j-1}^{(2)}:t_{j}^{(2)}\Big)\leq\frac{S}{q(S,K)+2}\right\}\right]\\
&\overset{\text{(i)}}{\geq}\left(K-1-\left\lfloor\frac{S}{q(S,K)+2}\right\rfloor\right)\mathbb{E}_{\boldsymbol{\mu}}^{\pi}[\mathbb{1}\{S(1:T)\leq S\}]\\
&\overset{\text{(ii)}}{=}K-1-\left\lfloor\frac{S}{q(S,K)+2}\right\rfloor,
\end{aligned}$$

where (i) follows from the pigeonhole principle and (ii) follows from the switching constraint. □

*Proof of Corollary E.2.* Since $K \geq 2$ and $0 \leq r(S,K) \leq K-2$, we have

$$\begin{aligned}
K - \left\lceil \frac{S}{q(S,K)+1} \right\rceil &= K - \left\lceil (K-1) - \frac{K-2-r(S,K)}{q(S,K)+1} \right\rceil \\
&\geq K - \min\left\{ K-1, \left\lceil (K-1) - \frac{K-2-r(S,K)}{q(S,K)+1} + 1 \right\rceil \right\} \\
&= \max\left\{ 1, \frac{K-2-r(S,K)}{q(S,K)+1} \right\} \\
&\geq \frac{K-r(S,K)}{2(q(S,K)+2)}
\end{aligned}$$

and

$$\begin{aligned}
K-1-\left\lfloor \frac{S}{q(S,K)+2} \right\rfloor &= K-1-\left\lfloor \frac{(K-1)q(S,K)+r(S,K)+1}{q(S,K)+2} \right\rfloor \\
&\geq K-1-\left\lceil \frac{(K-1)q(S,K)+r(S,K)+1}{q(S,K)+2} \right\rceil \\
&\geq \frac{2(K-1)-r(S,K)-1}{q(S,K)+2} \\
&\geq \frac{K-1}{q(S,K)+2} \\
&\geq \frac{K}{2(q(S,K)+2)}.
\end{aligned}$$

Combining the above inequalities with Lemma E.7 and Lemma E.8, we obtain Corollary E.2. □

### E.12.3 Alternative Environments, Bad Events, and Lower Bound Reductions

In the rest of the proof, we fix an arbitrary policy $\pi \in \Pi_S$.

In this subsection, we define the following concepts: (i) reference environment & reference measure, (ii) alternative environments & alternative measures, and (iii) bad events. Based on these definitions, we explicitly construct two classes of environments $\Phi_1$ and $\Phi_2$, and reduce the task of proving lower bounds on the "average-case regret" $\frac{1}{|\Phi_1|}\sum_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T)$ and $\frac{1}{|\Phi_2|}\sum_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T)$ to the task of proving lower bounds on the "average-case bad event probability" $\frac{1}{(q(S,K)+2)K}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}^{(1)}_{j,k}\left(E^{(1)}_{j,k}\right)$ and $\frac{1}{(q(S,K)+2)K}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}^{(2)}_{j,k}\left(E^{(2)}_{j,k}\right)$, respectively.

Let $\mathbf{0} = (0,\ldots,0) \in \mathbb{R}^K$ be the *reference* environment. Let $\mathbb{Q} := \mathbb{P}^{\pi}_{\mathbf{0}}$ denote the *reference* measure.

### Results Associated with the First Class of Risky Events

For any $j \in [q(S, K+2)]$, define a reward gap

$$\Delta_j^{(1)} := \begin{cases} 1, & \text{if } j = 1, \\ \frac{1}{2(q(S,K)+2)} \sqrt{\frac{K - r(S,K)}{t_{j-1}^{(1)}}}, & \text{if } j \in [2 : q(S,K)+1], \\ -\frac{1}{2(q(S,K)+2)} \sqrt{\frac{K - r(S,K)}{t_{q(S,K)}^{(1)}}}, & \text{if } j = q(S,K)+2. \end{cases}$$

Note that $\left|\Delta_j^{(1)}\right| \in [0, 1]$ for all $j \in [q(S, K) + 2]$.

For any $j \in [q(S,K)+2], k \in [K]$, define an *alternative* environment $\boldsymbol{\mu}_{j,k}^{(1)} := \left(\mu_{j,k;1}^{(1)}, \ldots, \mu_{j,k;K}^{(1)}\right) \in \mathbb{R}^K$ where

$$\mu_{j,k;i}^{(1)} := \begin{cases} \Delta_j^{(1)}, & \text{if } i = k, \\ 0, & \text{otherwise.} \end{cases}$$

Note that each alternative environment $\boldsymbol{\mu}_{j,k}^{(1)}$ only differs from the reference environment in terms of the mean reward of action $k$.

For any $j \in [q(S, K) + 2], k \in [K]$, let $\mathbb{P}_{j,k}^{(1)} := \mathbb{P}_{\boldsymbol{\mu}_{j,k}^{(1)}}^{\pi}$ denote the *alternative* measure associated with the alternative environment $\boldsymbol{\mu}_{j,k}^{(1)}$.

We explicitly construct a class of environments $\Phi_1 := \left\{\boldsymbol{\mu}_{j,k}^{(1)} \mid j \in [q(S, K) + 2], k \in [K]\right\}$.

For any $j \in [q(S, K) + 2], k \in [K]$, under environment $\boldsymbol{\mu}_{j,k}^{(1)}$, the risky event $E_{j,k}^{(1)}$ becomes a *bad event*[66] whose occurrence would lead to large regret. Specifically:

- Suppose $j = 1$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{1,k}^{(1)}$, choosing any action other than $k$ for one round incurs at least a $\Delta_1^{(1)}$ term in the policy's regret, and the occurrence of $E_{1,k}^{(1)} = \left\{\text{action } k \text{ is not chosen in period } \left[1 : t_1^{(1)}\right]\right\}$ incurs at least a $t_1^{(1)}\Delta_1^{(1)}$ term in the policy's regret.

- Suppose $j \in [2 : q(S, K)]$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{j,k}^{(1)}$, choosing any action other than $k$ for one round incurs at least a $\Delta_j^{(1)}$ term in the policy's regret, and the occurrence of $E_{j,k}^{(1)} = \left\{\text{action } k \text{ is not chosen in period } \left[t_{j-1}^{(1)} : t_j^{(1)}\right]\right\}$ incurs at least a $\left(t_j^{(1)} - t_{j-1}^{(1)} + 1\right)\Delta_j^{(1)}$ term in the policy's regret.

- Suppose $j = q(S, K) + 1$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{q(S,K)+1,k}^{(1)}$, choosing any action other than $k$ for one round incurs at least a $\Delta_{q(S,K)+1}^{(1)}$ term in the policy's regret, and the occurrence of

  $$E_{q(S,K)+1,k}^{(1)} = \left\{\text{action } k \text{ is not chosen in period } \left[t_{q(S,K)}^{(1)} : \lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor\right]\right\}$$

  incurs at least a $\left(\lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor - t_{q(S,K)}^{(1)} + 1\right)\Delta_{q(S,K)+1}^{(1)}$ term in the policy's regret.

[66] In our language, we call $E_{j,k}^{(1)}$ a *risky event* for any environment, but a *bad event* only for environment $\boldsymbol{\mu}_{j,k}^{(1)}$.

- Suppose $j = q(S,K)+2$. Since action $k$ is the worst action under environment $\boldsymbol{\mu}^{(1)}_{q(S,K)+2,k}$, choosing action $k$ for one round incurs at least a $-\Delta^{(1)}_{q(S,K)+2}$ term in the policy's regret. Furthermore, since the occurrence of $E^{(1)}_{q(S,K)+2,k}$ implies the occurrence of

$$\Big\{\text{action } k \text{ is chosen in every round in } [\lfloor (t^{(1)}_{q(S,K)} + T)/2 \rfloor : T]\Big\}$$

(because of the switching constraint), it incurs at least a $-\Big(T - \lfloor (t^{(1)}_{q(S,K)} + T)/2 \rfloor + 1\Big)\Delta^{(1)}_{q(S,K)+2}$ term in the policy's regret.

The above arguments lead to Lemma E.9.

**Lemma E.9** (From risky events to bad events)**.** *For any $j \in [q(S,K)+2], k \in [K]$, under environment $\boldsymbol{\mu}^{(1)}_{j,k}$, the risky event $E^{(1)}_{j,k}$ becomes a* bad event *in the sense that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}\left[T\mu^{(1)}_{j,k;k} - \sum_{t=1}^{T}\mu^{(1)}_{j,k;a_t} \mid E^{(1)}_{j,k}\right] \geq \mathcal{R}_{\mathrm{bad}}(S,K,T),$$

*where*

$$\mathcal{R}_{\mathrm{bad}}(S,K,T) := \frac{(K - r(S,K))}{8(q(S,K)+2)}\left(\frac{T}{K - r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}$$

*is a universal lower bound on the "distribution-dependent regret conditional on the bad event."*

*Proof of Lemma E.9.* By the arguments in the previous paragraph, we have

$$\begin{aligned}
&\mathbb{E}^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}\left[T\mu^{(1)}_{j,k;k} - \sum_{t=1}^{T}\mu^{(1)}_{j,k;a_t} \mid E^{(1)}_{j,k}\right] \\
&\geq \mathbb{E}^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}\left[\Big(t^{(1)}_{j} - t^{(2)}_{j-1} + 1\Big)\mu^{(1)}_{j,k;k} - \sum_{t \in \left[t^{(1)}_{j-1} : t^{(1)}_{j}\right]}\mu^{(1)}_{j,k;a_t} \mid E^{(1)}_{j,k}\right] \\
&\geq \begin{cases} t^{(1)}_1\Delta^{(1)}_1, & \text{if } j = 1, \\ \Big(t^{(1)}_j - t^{(1)}_{j-1} + 1\Big)\Delta^{(1)}_j, & \text{if } j \in [2 : q(S,K)], \\ \Big(\lfloor (t^{(1)}_{q(S,K)} + T)/2 \rfloor - t^{(1)}_{q(S,K)} + 1\Big)\Delta^{(1)}_{q(S,K)+1}, & \text{if } j = q(S,K)+1, \\ -\Big(T - \lfloor (t^{(1)}_{q(S,K)} + T)/2 \rfloor + 1\Big)\Delta^{(1)}_{q(S,K)+2}, & \text{if } j = q(S,K)+2, \end{cases} \\
&\geq \frac{(K - r(S,K))^{1 - \frac{1}{2-2^{-q(S,K)}}}}{8(q(S,K)+2)} T^{\frac{1}{2-2^{-q(S,K)}}},
\end{aligned}$$

where the last inequality follows from the following inequalities:

$$t^{(1)}_1\Delta^{(1)}_1 = t^{(1)}_1 \geq \frac{(K - r(S,K))^{1 - \frac{1}{2-2^{-q(S,K)}}}}{(q(S,K)+2)} T^{\frac{1}{2-2^{-q(S,K)}}},$$

$$
\begin{aligned}
&\Big(t_j^{(1)} - t_{j-1}^{(1)} + 1\Big)\Delta_j^{(1)} \\
&\geq (K - r(S,K))\left(\left(\frac{T}{K-r(S,K)}\right)^{\frac{2-2^{1-j}}{2-2^{-q(S,K)}}} - \left(\frac{T}{K-r(S,K)}\right)^{\frac{2-2^{2-j}}{2-2^{-q(S,K)}}}\right)\Delta_j^{(1)} \\
&\geq \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\left(\frac{T}{K-r(S,K)}\right)^{\frac{2-2^{1-j}}{2-2^{-q(S,K)}}} - \left(\frac{T}{K-r(S,K)}\right)^{\frac{2-2^{2-j}}{2-2^{-q(S,K)}}}\right)\left(\frac{T}{K-r(S,K)}\right)^{-\frac{1-2^{1-j}}{2-2^{-q(S,K)}}} \\
&= \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}} - \left(\frac{T}{K-r(S,K)}\right)^{\frac{1-2^{1-j}}{2-2^{-q(S,K)}}}\right) \\
&= \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}\left(1 - \left(\frac{T}{K-r(S,K)}\right)^{\frac{-2^{1-j}}{2-2^{-q(S,K)}}}\right) \\
&\geq \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}\left(1 - \left(\frac{T}{K-r(S,K)}\right)^{\frac{-2^{-q(S,K)}}{2-2^{-q(S,K)}}}\right) \\
&\geq \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}\left(1 - \left(\frac{T}{K-r(S,K)}\right)^{-2^{-q(S,K)-1}}\right) \\
&\overset{\text{(i)}}{\geq} \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}\left(1 - \left(\frac{T}{K-r(S,K)}\right)^{-\frac{1}{\log_2(T/K)}}\right) \\
&\geq \frac{(K-r(S,K))}{2(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}\Big(1 - (T/K)^{-\frac{1}{\log_2(T/K)}}\Big) \\
&= \frac{(K-r(S,K))}{4(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}, \quad \forall j \in [2 : q(S,K)+1],
\end{aligned}
$$

$$
\begin{aligned}
\Big(\lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor - t_{q(S,K)}^{(1)} + 1\Big)\Delta_{q(S,K)+1}^{(1)} &\geq \frac{1}{2}\Big(t_{q(S,K)+1}^{(1)} - t_{q(S,K)}^{(1)} + 1\Big)\Delta_{q(S,K)+1}^{(1)} \\
&\geq \frac{(K-r(S,K))}{8(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}},
\end{aligned}
$$

$$
\begin{aligned}
-\Big(T - \lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor + 1\Big)\Delta_{q(S,K)+2}^{(1)} &\geq \Big(\lfloor (t_{q(S,K)}^{(1)} + T)/2 \rfloor - t_{q(S,K)}^{(1)} + 1\Big)\Delta_{q(S,K)+1}^{(1)} \\
&\geq \frac{(K-r(S,K))}{8(q(S,K)+2)}\left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}}.
\end{aligned}
$$

Note that in (i) we utilize the fact that $q(S,K) + 1 \leq \log_2 \log_2(T/K)$. □

Based on Lemma E.9, we can reduce the task of proving a lower bound on the policy's (distribution-dependent) regret $R^{\pi}_{\boldsymbol{\mu}_{j,k}^{(1)}}(T)$ to the task of proving a lower bound on the bad event probability $\mathbb{P}_{j,k}^{(1)}\Big(E_{j,k}^{(1)}\Big)$. Consequently, we can reduce the task of proving a lower bound on the policy's "average-case regret" $\frac{1}{|\Phi_1|}\sum_{\boldsymbol{\mu}\in\Phi_1} R^{\pi}_{\boldsymbol{\mu}}(T)$ to the task of proving a lower bound on the "average-case bad event probability" $\frac{1}{(q(S,K)+2)K}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}_{j,k}^{(1)}\Big(E_{j,k}^{(1)}\Big)$.

**Lemma E.10** (Reducing regret lower bounds to bad event probability lower bounds)**.** *For any*

$j \in [q(S,K)+2], k \in [K]$, *we have*

$$R^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}(T) \geq \mathcal{R}_{\mathrm{bad}}(S,K,T) \cdot \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big).$$

*As a result, we have*

$$R^{\pi}(T) \geq \frac{1}{|\Phi_1|}\sum_{\boldsymbol{\mu}\in\Phi_1} R^{\pi}_{\boldsymbol{\mu}}(T) \geq \mathcal{R}_{\mathrm{bad}}(S,K,T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big).$$

*Proof of Lemma E.10.* For any $j \in [q(S,K)+2], k \in [K]$, by Lemma E.9, we have

$$\begin{aligned}
R^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}(T) &= \mathbb{E}^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}\left[T\mu^{(1)}_{j,k;k} - \sum_{t=1}^{T}\mu^{(1)}_{j,k;a_t}\right] \\
&\geq \mathbb{E}^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}\left[T\mu^{(1)}_{j,k;k} - \sum_{t=1}^{T}\mu^{(1)}_{j,k;a_t} \mid E^{(1)}_{j,k}\right] \cdot \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big) \\
&\geq \mathcal{R}_{\mathrm{bad}}(S,K,T) \cdot \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big),
\end{aligned}$$

and hence

$$\begin{aligned}
R^{\pi}(T) &= \sup_{\mathcal{D}} R^{\pi}_{\mathcal{D}}(T) \\
&\geq \sup_{\boldsymbol{\mu}\in\Phi_1} R^{\pi}_{\boldsymbol{\mu}}(T) \\
&\geq \frac{1}{|\Phi_1|}\sum_{\boldsymbol{\mu}\in\Phi_1} R^{\pi}_{\boldsymbol{\mu}}(T) \\
&\geq \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} R^{\pi}_{\boldsymbol{\mu}^{(1)}_{j,k}}(T) \\
&\geq \mathcal{R}_{\mathrm{bad}}(S,K,T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big).
\end{aligned}$$

□

**Results Associated with the Second Class of Risky Events**

For any $j \in [q(S,K+2)]$, define a reward gap

$$\Delta^{(2)}_j := \begin{cases} 1, & \text{if } j = 1, \\ \frac{1}{2(q(S,K)+2)}\sqrt{\frac{K}{t^{(2)}_{j-1}}}, & \text{if } j \in [2 : q(S,K)+2]. \end{cases}$$

Note that $\left|\Delta^{(2)}_j\right| \in [0,1]$ for all $j \in [q(S,K)+2]$.

For any $j \in [q(S,K)+2], k \in [K]$, define an *alternative* environment $\boldsymbol{\mu}^{(2)}_{j,k} := \left(\mu^{(2)}_{j,k;1}, \ldots, \mu^{(2)}_{j,k;K}\right) \in$

$\mathbb{R}^K$ where

$$\mu^{(2)}_{j,k;i} := \begin{cases} \Delta^{(2)}_j, & \text{if } i = k, \\ 0, & \text{otherwise.} \end{cases}$$

Note that each alternative environment $\boldsymbol{\mu}^{(1)}_{j,k}$ only differs from the reference environment in terms of the mean reward of action $k$.

For any $j \in [q(S,K)+2], k \in [K]$, let $\mathbb{P}^{(2)}_{j,k} := \mathbb{P}^{\pi}_{\boldsymbol{\mu}^{(2)}_{j,k}}$ denote the *alternative* measure associated with the alternative environment $\boldsymbol{\mu}^{(2)}_{j,k}$.

We explicitly construct a class of environments $\Phi_2 := \left\{ \boldsymbol{\mu}^{(2)}_{j,k} \mid j \in [q(S,K)+2], k \in [K] \right\}$.

For any $j \in [q(S,K)+2], k \in [K]$, under environment $\boldsymbol{\mu}^{(2)}_{j,k}$, the risky event $E^{(2)}_{j,k}$ becomes a *bad event* whose occurrence would lead to large regret. Similar to our analysis in Appendix E.12.3, we have the following two lemmas.

**Lemma E.11** (From risky events to bad events)**.** *For any $j \in [q(S,K)+2], k \in [K]$, under environment $\boldsymbol{\mu}^{(2)}_{j,k}$, the risky event $E^{(2)}_{j,k}$ becomes a* bad event *in the sense that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\mu}^{(2)}_{j,k}} \left[ T\mu^{(2)}_{j,k;k} - \sum_{t=1}^{T} \mu^{(2)}_{j,k;a_t} \mid E^{(2)}_{j,k} \right] \geq \mathcal{R}_{\mathrm{bad2}}(S,K,T),$$

*where*

$$\mathcal{R}_{\mathrm{bad2}}(S,K,T) := \frac{K}{4(q(S,K)+2)} \left( \frac{T}{K} \right)^{\frac{1}{2-2^{-q(S,K)-1}}}$$

*is a universal lower bound on the "distribution-dependent regret conditional on the bad event."*

**Lemma E.12** (Reducing regret lower bounds to bad event probability lower bounds)**.** *For any $j \in [q(S,K)+2], k \in [K]$, we have*

$$R^{\pi}_{\boldsymbol{\mu}^{(2)}_{j,k}}(T) \geq \mathcal{R}_{\mathrm{bad2}}(S,K,T) \cdot \mathbb{P}^{(2)}_{j,k}\Big(E^{(2)}_{j,k}\Big).$$

*As a result, we have*

$$R^{\pi}(T) \geq \frac{1}{|\Phi_2|} \sum_{\boldsymbol{\mu} \in \Phi_2} R^{\pi}_{\boldsymbol{\mu}}(T) \geq \mathcal{R}_{\mathrm{bad2}}(S,K,T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} \mathbb{P}^{(2)}_{j,k}\Big(E^{(2)}_{j,k}\Big).$$

### E.12.4 Probability Space Changing Tricks

Lemma E.10 and Lemma E.12 indicate that, in order to prove the desired lower bound on the regret $R^{\pi}(T)$, it suffices to prove the following two statements:

$$\overline{p^{(1)}} := \frac{1}{(q(S,K)+2)K} \sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big) = \widetilde{\Omega}\left( \frac{K - r(S,K)}{K} \right), \tag{E.48}$$

$$\overline{p^{(2)}} := \frac{1}{(q(S,K)+2)K} \sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} \mathbb{P}^{(2)}_{j,k}\Big(E^{(2)}_{j,k}\Big) = \widetilde{\Omega}(1). \tag{E.49}$$

That is, we only need to establish tight lower bounds on the average probability $\overline{p^{(1)}}$ and the average probability $\overline{p^{(2)}}$.

By Corollary E.2, we have

$$\overline{q^{(1)}} := \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{Q}\Big(E_{j,k}^{(1)}\Big) = \widetilde{\Omega}\bigg(\frac{K-r(S,K)}{K}\bigg),$$

$$\overline{q^{(2)}} := \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{Q}\Big(E_{j,k}^{(2)}\Big) = \widetilde{\Omega}(1).$$

Therefore, in order to show (E.48), it suffices to show that $\overline{p^{(1)}}$ is close to $\overline{q^{(1)}}$; in order to show (E.49), it suffices to show that $\overline{p^{(2)}}$ is close to $\overline{q^{(2)}}$.

Let us first focus on the relationship between $\overline{p^{(1)}}$ and $\overline{q^{(1)}}$. Note that $\overline{p^{(1)}}$ is the average of the sequence $\Big\{\mathbb{P}_{j,k}^{(1)}\Big(E_{j,k}^{(1)}\Big)\Big\}$[67] (where a sequence of events $\Big\{E_{j,k}^{(1)}\Big\}$ are evaluated by a sequence of varying alternative measures $\Big\{\mathbb{P}_{j,k}^{(1)}\Big\}$), while $\overline{q^{(1)}}$ is the average of the sequence $\Big\{\mathbb{Q}\Big(E_{j,k}^{(1)}\Big)\Big\}$ (where the same sequence of events $\Big\{E_{j,k}^{(1)}\Big\}$ are evaluated by a single and fixed reference measure $\mathbb{Q}$). Intuitively, we just need a "change of measure"/information-theoretic argument — if the alternative measures $\Big\{\mathbb{P}_{j,k}^{(1)}\Big\}$ are "close enough" to the reference measure $\mathbb{Q}$, then $\overline{p^{(1)}}$ is close to $\overline{q^{(1)}}$.

Unfortunately, it turns out that the divergence between $\Big\{\mathbb{P}_{j,k}^{(1)}\Big\}$ and $\mathbb{Q}$ is too large to make the above argument work. An important reason is that such an argument directly deals with the underlying measures $\Big\{\mathbb{P}_{j,k}^{(1)}\Big\}$ and $\mathbb{Q}$, thus completely overlooks the special structures of the risky event sequence $\Big\{E_{j,k}^{(1)}\Big\}$. Therefore, if we want to show that $\overline{p^{(1)}}$ is close to $\overline{q^{(1)}}$, we need to integrate the structural properties of risky events into our argument.

The same challenge exists when we want to show that $\overline{p^{(2)}}$ is close to $\overline{q^{(2)}}$.

We develop *probability space changing tricks* to address this challenge. See below.

### Results Associated with the First Class of Risky Events

We start with some key structural properties of the risky event sequence $\Big\{E_{j,k}^{(1)}\Big\}$.

- For any $k \in [K]$, the occurrence of the event

$$E_{1,k}^{(1)} = \Big\{\text{action } k \text{ is not chosen in period } \Big[1 : t_1^{(1)}\Big]\Big\}$$

  is independent of the random variables $\big(X_{\boldsymbol{\mu}}^t(k)\big)_{t\in[1:t_1^{(1)}]}$ and the random variables $\big(X_{\boldsymbol{\mu}}^t(i)\big)_{t\in[t_1^{(1)}+1:T],i\in[K]}$.

- For any $j \in [2 : q(S,K)]$, $k \in [K]$, the occurrence of the event

$$E_{j,k}^{(1)} = \Big\{\text{action } k \text{ is not chosen in period } \Big[t_{j-1}^{(1)} : t_j^{(1)}\Big]\Big\}$$

  is independent of the random variables $\big(X_{\boldsymbol{\mu}}^t(k)\big)_{t\in[t_{j-1}^{(1)}:t_j^{(1)}]}$ and the random variables $\big(X_{\boldsymbol{\mu}}^t(i)\big)_{t\in[t_j^{(1)}+1:T],i\in[K]}$.

[67] We use the notation $\Big\{\mathbb{P}_{j,k}^{(1)}\Big(E_{j,k}^{(1)}\Big)\Big\}$ to represent the sequence $\Big(\mathbb{P}_{j,k}^{(1)}\Big(E_{j,k}^{(1)}\Big)\Big)_{j\in[q(S,K)+2],k\in[K]}$. In general, we use $\{a_{j,k}\}$ to represent a sequence $(a_{j,k})_{j\in[q(S,K)+2],k\in[K]}$.

- For any $k \in [K]$, the occurrence of the event

$$E^{(1)}_{q(S,K)+1,k} = \left\{\text{action } k \text{ is not chosen in period } \left[t^{(1)}_{q(S,K)} : \lfloor (t^{(1)}_{q(S,K)} + T)/2 \rfloor\right]\right\}$$

is independent of the random variables $\left(X^t_{\boldsymbol{\mu}}(k)\right)_{t \in [t^{(1)}_{q(S,K)} : \lfloor (t^{(1)}_{q(S,K)}+T)/2 \rfloor]}$ and the random variables $\left(X^t_{\boldsymbol{\mu}}(i)\right)_{t \in [\lfloor (t^{(1)}_{q(S,K)}+T)/2 \rfloor + 1 : T], i \in [K]}$.

- For any $k \in [K]$, the occurrence of the event

$$E^{(1)}_{q(S,K)+2,k} := \left\{\tau \le \lfloor (t^{(1)}_{q(S,K)} + T)/2 \rfloor,\ a_\tau = k,\ \text{action } k \text{ is not chosen in period } \left[t^{(1)}_{q(S,K)} : \tau - 1\right]\right\}$$

is independent of the random variables $\left(X^t_{\boldsymbol{\mu}}(k)\right)_{t \in [t^{(1)}_{q(S,K)} : \lfloor (t^{(1)}_{q(S,K)}+T)/2 \rfloor]}$ and the random variables $\left(X^t_{\boldsymbol{\mu}}(i)\right)_{t \in [\lfloor (t^{(1)}_{q(S,K)}+T)/2 \rfloor + 1 : T], i \in [K]}$. (Note that this property crucially relies on the delicate design of $E^{(1)}_{q(S,K)+2,k}$.)

The above properties indicate that, when we want to represent the probability of a risky event $E^{(1)}_{j,k}$ under an environment $\boldsymbol{\mu}$, we do not need to use the "full" measure induced by $\mathcal{H}_T$ (i.e., $\mathbb{P}^\pi_{\boldsymbol{\mu}}$) or the "natural" measure induced by $\mathcal{H}_{t^{(1)}_j}$. Instead, we can use a "ristricted" measure which deliberately "ignores" certain reward information associated with action $k$ — thanks to the structural property of $E^{(1)}_{j,k}$, such ignorance would not affect the measure's well-definedness and value on $E^{(1)}_{j,k}$. Moreover, since the alternative environment $\boldsymbol{\mu}^{(1)}_{j,k}$ only differs from the reference environment in terms of the mean reward of action $k$, such ignorance can help make the measures associated with the two environments "closer." We can then establish tighter bounds on the distance between $\overline{p^{(1)}}$ and $\overline{q^{(1)}}$.

Motivated by the above idea, we design two sequences of *artificial* measures $\left\{\mathbb{P}'_{j,k}\right\}$ and $\left\{\mathbb{Q}'_{j,k}\right\}$ as follows.

**Artificial measures $\left\{\mathbb{P}'_{j,k}\right\}$.** For any $j \in [2 : q(S,K)]$, $k \in [K]$, let $\mathbb{P}'_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\boldsymbol{\mu}^{(1)}_{j,k}}(a_t)\right)_{t \in [1 : t^{(1)}_{j-1} - 1]}, \left(a_t, X^t_{\boldsymbol{\mu}^{(1)}_{j,k}}(a_t)\mathbb{1}\{a_t \neq k\}\right)_{t \in [t^{(1)}_{j-1} : t^{(1)}_j]}\right). \tag{E.50}$$

For $j = 1$, for any $k \in [K]$, let $\mathbb{P}'_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\boldsymbol{\mu}^{(1)}_{j,k}}(a_t)\mathbb{1}\{a_t \neq k\}\right)_{t \in [1 : t^{(1)}_j]}\right).$$

For $j \in \{q(S,K)+1, q(S,K)+2\}$, let $\mathbb{P}'_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\boldsymbol{\mu}^{(1)}_{j,k}}(a_t)\right)_{t \in [1 : t^{(1)}_{q(S,K)} - 1]}, \left(a_t, X^t_{\boldsymbol{\mu}^{(1)}_{j,k}}(a_t)\mathbb{1}\{a_t \neq k\}\right)_{t \in [t^{(1)}_{q(S,K)} : \lfloor (t^{(1)}_{q(S,K)}+T)/2 \rfloor]}\right).$$

**Artificial reference measures $\left\{\mathbb{Q}'_{j,k}\right\}$.** For any $j \in [2 : q(S,K)]$, $k \in [K]$, let $\mathbb{Q}'_{j,k}$ be the

probability measure induced by the joint random variable

$$\Big(\big(a_t, X_{\mathbf{0}}^t(a_t)\big)_{t\in[1:t_{j-1}^{(1)}-1]}, \big(a_t, X_{\mathbf{0}}^t(a_t)\mathbb{1}\{a_t\neq k\}\big)_{t\in[t_{j-1}^{(1)}:t_j^{(1)}]}\Big).$$

For $j=1$, for any $k\in[K]$, let $\mathbb{Q}'_{j,k}$ be the probability measure induced by the joint random variable

$$\Big(\big(a_t, X_{\mathbf{0}}^t(a_t)\mathbb{1}\{a_t\neq k\}\big)_{t\in[1:t_j^{(1)}]}\Big).$$

For $j\in\{q(S,K)+1, q(S,K)+2\}$, let $\mathbb{Q}'_{j,k}$ be the probability measure induced by the joint random variable

$$\Big(\big(a_t, X_{\mathbf{0}}^t(a_t)\big)_{t\in[1:t_{q(S,K)}^{(1)}-1]}, \big(a_t, X_{\mathbf{0}}^t(a_t)\mathbb{1}\{a_t\neq k\}\big)_{t\in[t_{q(S,K)}^{(1)}:\lfloor(t_{q(S,K)}^{(1)}+T)/2\rfloor]}\Big).$$

Let us provide some explanations on the above definitions of artificial measures. For brevity, we focus on the case of $j\in[2:q(S,K)]$. Note that (E.50) is *not* the total data collected by $\pi$ under environment $\boldsymbol{\mu}_{j,k}^{(1)}$ during period $\left[1:t_j^{(1)}\right]$, which should be

$$\mathcal{H}_{t_j^{(1)}} = \left(\left(a_t, X_{\boldsymbol{\mu}_{j,k}^{(1)}}^t(a_t)\right)_{t\in[1:t_j^{(1)}]}\right);$$

instead, (E.50) is a censored variant of $\mathcal{H}_{t_j^{(1)}}$, with all the reward observations associated with action $k$ during period $\left[1:t_j^{(1)}\right]$ being "ignored." Consequently, $\mathbb{P}'_{j,k}$ is *neither* a measure on $(\Omega_T,\mathcal{F}_T)$ *nor* a measure on $\left(\Omega_{t_j^{(1)}},\mathcal{F}_{t_j^{(1)}}\right)$; instead, it is a measure on a "restricted" measurable space $\left(\Omega'_{j,k},\mathcal{F}'_{j,k}\right)$, where

$$\Omega'_{j,k} := \Omega_{t_j^{(1)}} \setminus \left(\Omega_{t_{j-1}^{(1)}-1}\times\left(\{k\}\times(\mathbb{R}\setminus\{0\})\right)^{t_j^{(1)}-t_{j-1}^{(1)}+1}\right),\quad \mathcal{F}'_{j,k} := \mathcal{B}\big(\Omega'_{j,k}\big).$$

Since $\left(\Omega'_{j,k},\mathcal{F}'_{j,k}\right)\subset\left(\Omega_{t_j^{(1)}},\mathcal{F}_{t_j^{(1)}}\right)\subset(\Omega_T,\mathcal{F}_T)$, the artificial measure $\mathbb{P}'_{j,k}$ can be seen as the restriction of $\mathbb{P}_{j,k}^{(1)}$ to a "much smaller" measurable space which still keeps $E_{j,k}^{(1)}$ measurable.[68] Similarly, the artificial reference measure $\mathbb{Q}'_{j,k}$ is the restriction of the reference measure $\mathbb{Q}$ to the same measurable space $\left(\Omega'_{j,k},\mathcal{F}'_{j,k}\right)$. Such restrictions guarantee two nice properties:

1. $\mathbb{P}'_{j,k}(E_{j,k}^{(1)}) = \mathbb{P}_{j,k}^{(1)}(E_{j,k}^{(1)})$ and $\mathbb{Q}'_{j,k}(E_{j,k}^{(1)}) = \mathbb{Q}(E_{j,k}^{(1)})$ for all $j,k$.

2. Since the alternative environment $\boldsymbol{\mu}_{j,k}^{(1)}$ only differs from the reference environment $\mathbf{0}$ in terms of the mean reward of action $k$, the divergence between $\mathbb{P}'_{j,k}$ and $\mathbb{Q}'_{j,k}$ becomes much smaller on the new measurable space $\left(\Omega'_{j,k},\mathcal{F}'_{j,k}\right)$, compared with the divergence between $\mathbb{P}_{j,k}^{(1)}$ and $\mathbb{Q}$ on the original measurable space $(\Omega_T,\mathcal{F}_T)$.

We now use the policy's behavior under the fixed reference environment $\mathbf{0}$ to bound the *reverse* KL divergence between $\mathbb{P}'_{j,k}$ and $\mathbb{Q}'_{j,k}$. Using a standard "divergence decomposition" lemma (see, e.g.,

[68] In measure theory, $\mathcal{F}'_{j,k}$ is called a sub-$\sigma$-algebra, and $\mathbb{P}'_{j,k}$ is called a *restricted measure*.

Lemma 15.1 of Lattimore and Szepesvári 2020), we have

$$D_{\mathrm{re}}\big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\big) = D_{\mathrm{KL}}\big(\mathbb{Q}'_{j,k} \parallel \mathbb{P}'_{j,k}\big) = \mathbb{E}^{\pi}_{\mathbf{0}}\left[\sum_{t=1}^{t^{(1)}_{j-1}-1} \frac{\left[\Delta^{(1)}_j\right]^2}{2} \mathbb{1}\{a_t = k\}\right] = \frac{\left[\Delta^{(1)}_j\right]^2}{2} \mathbb{E}^{\pi}_{\mathbf{0}}\left[\sum_{t=1}^{t^{(1)}_{j-1}-1} \mathbb{1}\{a_t = k\}\right]$$

for all $j \in [2 : q(S,K)]$ and $k \in [K]$. This implies that

$$\sum_{k=1}^{K} D_{\mathrm{re}}\big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\big) = \sum_{k=1}^{K} \frac{\left[\Delta^{(1)}_j\right]^2}{2} \mathbb{E}^{\pi}_{\mathbf{0}}\left[\sum_{t=1}^{t^{(1)}_{j-1}-1} \mathbb{1}\{a_t = k\}\right] = \frac{\left[\Delta^{(1)}_j\right]^2}{2}\left(t^{(1)}_{j-1} - 1\right)$$

for all $j \in [2 : q(S,K)]$. Similarly, we have $\sum_{k=1}^{K} D_{\mathrm{re}}\Big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\Big) = 0$ for $j = 0$ and

$$\sum_{k=1}^{K} D_{\mathrm{re}}\big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\big) = \sum_{k=1}^{K} \frac{\left[\Delta^{(1)}_j\right]^2}{2} \mathbb{E}^{\pi}_{\mathbf{0}}\left[\sum_{t=1}^{t^{(1)}_{q(S,K)}-1} \mathbb{1}\{a_t = k\}\right] = \frac{\left[\Delta^{(1)}_j\right]^2}{2}\left(t^{(1)}_{q(S,K)} - 1\right)$$

for $j \in \{q(S,K)+1, q(S,K)+2\}$. Combining with the definitions of $\Delta^{(1)}_j$ in Appendix E.12.3, we have

$$\sum_{k=1}^{K} D_{\mathrm{re}}\big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\big) \le \frac{K - r(S,K)}{8(q(S,K)+2)^2}$$

for all $j \in [q(S,K)+2]$. Therefore, we have the following lemma.

**Lemma E.13.** *It holds that*

$$\sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} D_{\mathrm{re}}\big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\big) \le \frac{K - r(S,K)}{8(q(S,K)+2)}.$$

*Combined with Corollary E.2, this implies*

$$\frac{1}{(q(S,K)+2)K} \sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} D_{\mathrm{re}}\big(\mathbb{P}'_{j,k} \parallel \mathbb{Q}'_{j,k}\big) \le \frac{K - r(S,K)}{8(q(S,K)+2)^2 K} \le \frac{\overline{q^{(1)}}}{4}.$$

### Results Associated with the Second Class of Risky Events

Similar to Appendix E.12.4, we design two sequences of *artificial* measures $\left\{\mathbb{P}''_{j,k}\right\}$ and $\left\{\mathbb{Q}''_{j,k}\right\}$ as follows.

**Artificial measures $\left\{\mathbb{P}''_{j,k}\right\}$.** For any $j \in [2 : q(S,K)+1]$, $k \in [K]$, let $\mathbb{P}''_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\boldsymbol{\mu}^{(1)}_{j,k}}(a_t)\right)_{t \in [1 : t^{(2)}_{j-1}-1]}, \left(a_t, X^t_{\boldsymbol{\mu}^{(2)}_{j,k}}(a_t)\mathbb{1}\{a_t \neq k\}\right)_{t \in [t^{(2)}_{j-1} : t^{(2)}_{j}]}\right).$$

For $j=1$, for any $k\in[K]$, let $\mathbb{P}''_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\boldsymbol{\mu}^{(2)}_{j,k}}(a_t)\mathbb{1}\{a_t\neq k\}\right)_{t\in[1:t_j^{(2)}]}\right).$$

**Artificial reference measures $\left\{\mathbb{Q}''_{j,k}\right\}$.** For any $j\in[2:q(S,K)]$, $k\in[K]$, let $\mathbb{Q}''_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\mathbf{0}}(a_t)\right)_{t\in[1:t_{j-1}^{(2)}-1]}, \left(a_t, X^t_{\mathbf{0}}(a_t)\mathbb{1}\{a_t\neq k\}\right)_{t\in[t_{j-1}^{(2)}:t_j^{(2)}]}\right).$$

For $j=1$, for any $k\in[K]$, let $\mathbb{Q}''_{j,k}$ be the probability measure induced by the joint random variable

$$\left(\left(a_t, X^t_{\mathbf{0}}(a_t)\mathbb{1}\{a_t\neq k\}\right)_{t\in[1:t_j^{(2)}]}\right).$$

Similar to Lemma E.13, we have the following lemma.

**Lemma E.14.** *It holds that*

$$\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]} D_{\mathrm{re}}\big(\mathbb{P}''_{j,k}\,\|\,\mathbb{Q}'_{j,k}\big)\leq\frac{K}{8(q(S,K)+2)}.$$

*Combined with Corollary E.2, this implies*

$$\frac{1}{\overline{(q(S,K)+2)K}}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]} D_{\mathrm{re}}\big(\mathbb{P}''_{j,k}\,\|\,\mathbb{Q}''_{j,k}\big)\leq\frac{1}{8(q(S,K)+2)^2}\leq\frac{\overline{q^{(2)}}}{4}.$$

### E.12.5 Applying the GRF Inequality

In this part, we apply the GRF inequality to show (E.48) and (E.49), and complete the proof of Theorem 6.2.

We first represent $\overline{p^{(1)}}$ using $\{\mathbb{P}'_{j,k}\}$ and represent $\overline{q^{(1)}}$ using $\{\mathbb{Q}'_{j,k}\}$. Since $\mathbb{P}'_{j,k}(E^{(1)}_{j,k})=\mathbb{P}^{(1)}_{j,k}(E^{(1)}_{j,k})$ and $\mathbb{Q}'_{j,k}(E^{(1)}_{j,k})=\mathbb{Q}(E^{(1)}_{j,k})$ hold for all $j,k$, we have

$$\overline{p^{(1)}}=\frac{1}{\overline{(q(S,K)+2)K}}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big)=\frac{1}{\overline{(q(S,K)+2)K}}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{P}'_{j,k}\Big(E^{(1)}_{j,k}\Big),$$

$$\overline{q^{(1)}}=\frac{1}{\overline{(q(S,K)+2)K}}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{Q}\Big(E^{(2)}_{j,k}\Big)=\frac{1}{\overline{(q(S,K)+2)K}}\sum_{j\in[q(S,K)+2]}\sum_{k\in[K]}\mathbb{Q}'_{j,k}\Big(E^{(1)}_{j,k}\Big).$$

By the GRF inequality (Proposition E.4), we have

$$\begin{aligned}
\overline{p^{(1)}} &\geq \overline{q^{(1)}} - \sqrt{2\overline{q^{(1)}} \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} D_{\mathrm{re}}\Big(\mathbb{P}'_{j,k} \,\|\, \mathbb{Q}'_{j,k}\Big)} \\
&\overset{\text{(i)}}{\geq} \overline{q^{(1)}} - \sqrt{2\overline{q^{(1)}} \frac{\overline{q^{(1)}}}{4}} \\
&= \frac{2-\sqrt{2}}{2} \overline{q^{(1)}} \\
&\overset{\text{(ii)}}{\geq} \frac{2-\sqrt{2}}{4} \frac{K - r(S,K)}{(q(S,K)+2)^2 K}
\end{aligned} \tag{E.51}$$

and

$$\begin{aligned}
\overline{p^{(2)}} &\geq \overline{q^{(2)}} - \sqrt{2\overline{q^{(2)}} \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} D_{\mathrm{re}}\Big(\mathbb{P}''_{j,k} \,\|\, \mathbb{Q}''_{j,k}\Big)} \\
&\overset{\text{(iii)}}{\geq} \overline{q^{(2)}} - \sqrt{2\overline{q^{(2)}} \frac{\overline{q^{(2)}}}{4}} \\
&= \frac{2-\sqrt{2}}{2} \overline{q^{(2)}} \\
&\overset{\text{(iv)}}{\geq} \frac{2-\sqrt{2}}{4} \frac{1}{(q(S,K)+2)^2},
\end{aligned} \tag{E.52}$$

where (i) follows from Lemma E.13, (ii) follows from Corollary E.2, (iii) follows from Lemma E.14, and (iv) follows from Corollary E.2. We thus prove (E.48) and (E.49).

Now we plug (E.51) and (E.52) into Lemma E.10 and Lemma E.12. We have

$$\begin{aligned}
R^{\pi}(T) &\geq \frac{1}{|\Phi_1|} \sum_{\boldsymbol{\mu}\in\Phi_1} R^{\pi}_{\boldsymbol{\mu}}(T) \\
&\geq \mathcal{R}_{\mathrm{bad}}(S,K,T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}^{(1)}_{j,k}\Big(E^{(1)}_{j,k}\Big) \\
&\geq \mathcal{R}_{\mathrm{bad}}(S,K,T) \cdot \frac{2-\sqrt{2}}{4} \frac{K-r(S,K)}{(q(S,K)+2)^2 K} \\
&= \frac{(K-r(S,K))}{8(q(S,K)+2)} \left(\frac{T}{K-r(S,K)}\right)^{\frac{1}{2-2^{-q(S,K)}}} \cdot \frac{2-\sqrt{2}}{4} \frac{K-r(S,K)}{(q(S,K)+2)^2 K} \\
&= \frac{2-\sqrt{2}}{32(q(S,K)+2)^3} \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}} \\
&\geq \frac{2-\sqrt{2}}{32(\log_2\log_2(T/K))^3} \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}} \\
&= \frac{2-\sqrt{2}}{160\log_2(T/K)} \frac{(K-r(S,K))^{2-\frac{1}{2-2^{-q(S,K)}}}}{K} T^{\frac{1}{2-2^{-q(S,K)}}}
\end{aligned}$$

and

$$
\begin{aligned}
R^{\pi}(T) &\geq \frac{1}{|\Phi_2|} \sum_{\boldsymbol{\mu} \in \Phi_2} R^{\pi}_{\boldsymbol{\mu}}(T) \\
&\geq \mathcal{R}_{\mathrm{bad2}}(S, K, T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j \in [q(S,K)+2]} \sum_{k \in [K]} \mathbb{P}^{(2)}_{j,k}\left(E^{(2)}_{j,k}\right) \\
&\geq \mathcal{R}_{\mathrm{bad2}}(S, K, T) \cdot \frac{2-\sqrt{2}}{4} \frac{1}{(q(S,K)+2)^2} \\
&= \frac{K}{4(q(S,K)+2)} \left(\frac{T}{K}\right)^{\frac{1}{2-2^{-q(S,K)-1}}} \cdot \frac{2-\sqrt{2}}{4} \frac{1}{(q(S,K)+2)^2} \\
&= \frac{2-\sqrt{2}}{16(q(S,K)+2)^3} K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}} \\
&\geq \frac{2-\sqrt{2}}{16(\log_2 \log_2(T/K))^3} K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}} \\
&\geq \frac{2-\sqrt{2}}{80 \log_2(T/K)} K^{1-\frac{1}{2-2^{-q(S,K)-1}}} T^{\frac{1}{2-2^{-q(S,K)-1}}}.
\end{aligned}
$$

We thus complete the proof of Theorem 6.2. □

## E.13 Proof of Theorem 6.4

The proof of Theorem 6.4 builds on the proof of Theorem 6.2. In this proof, we emphasize the differences, which are mainly in Appendices E.13.1 to E.13.3, corresponding to the first three steps of the RECAP method.

Given any $K = |G| > 1$, $S \geq 0$ and $T \geq 2K$, we focus on the setting of $\mathcal{D}_k = \mathcal{N}(\mu_k, 1)$ $(\forall k \in [K])$, as this is sufficient for us to prove the desired lower bound. For simplicity, in this proof we will directly use the vector $\boldsymbol{\mu}$ to represent the environment.

For any environment $\boldsymbol{\mu}$, let $X^t_{\boldsymbol{\mu}}(k) \sim \mathcal{N}(\mu_k, 1)$ denote the i.i.d. random reward of each action $k$ at round $t$ $(k \in [K], t \in [T])$. For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, for any $t \in [T]$, we use $a_t$ to denote the random action selected by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$, and use $X^t_{\boldsymbol{\mu}}(a_t)$ to denote the random reward observed by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$. Let $\mathcal{H}_t := \big((a_1, X^1_{\boldsymbol{\mu}}(a_1)), \ldots, (a_t, X^t_{\boldsymbol{\mu}}(a_t))\big)$ be the history (of actions and observations) up to round $t$ (inclusive), whose value lies in $\Omega_t := ([K] \times \mathbb{R})^t$. Let $\mathcal{F}_t := \mathcal{B}(\Omega_t)$ be the Borel $\sigma$-algebra on $\Omega_t$. Let $\mathbb{P}^{\pi}_{\boldsymbol{\mu}}$ be the probability measure induced by (i.e., the law of) $\mathcal{H}_T$, and $\mathbb{E}^{\pi}_{\boldsymbol{\mu}}$ be the associated expectation operator. Let $R^{\pi}_{\boldsymbol{\mu}}(T) := T\mu^{\star} - \mathbb{E}^{\pi}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} \mu_{a_t}\right]$ be policy $\pi$'s distribution-dependent regret under environment $\boldsymbol{\mu}$.

Similar to Appendix E.12, we argue that in our proof, we only need to consider the case of $q''(S,G) + 2 \leq \log_2 \log_2(T)$. Suppose $q''(S,G) + 2 > \log_2 \log_2(T)$, then we have

$$
T^{\frac{1}{2-2^{-q''(S,G)}}} \leq 4\sqrt{T},
$$

thus the lower bound in Theorem 6.4 becomes $\Omega(\sqrt{T}/(K \log T))$ and can be directly obtained by applying the well-known $\Omega(\sqrt{KT})$ lower bound of the classical MAB (see, e.g., Lattimore and Szepesvári

2020). Therefore, the really non-trivial case of Theorem 6.4 is the case of $q''(S,G)+2 \le \log_2\log_2(T)$, and we focus on this case in the rest of our proof.

Our goal is to explicitly construct a family of environments $\Phi$, such that for any $S$-switching-budget policy $\pi \in \Pi_S$, the "worst-case regret" $\max_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T)$ is lower bounded by

$$\Omega\left(\frac{1}{K\log T}T^{\frac{1}{2-2^{-q''(S,G)}}}\right)$$

Since the worst-case regret $R^{\pi}(T)$ is no less than the $\max_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T)$, the above goal directly implies Theorem 6.4.

Without loss of generality, we assume that $1 \in \arg\max_{i\in[K]} \min_{j\neq i} c_{i,j}$. For notational simplicity, we **redefine** the sequence $(t_j)_{j=0}^{q''(S,G)+1}$ as $t_0 = 0$ and

$$t_j = \left\lfloor T^{\frac{2-2^{-(i-1)}}{2-2^{-q''(S,G)}}} \right\rfloor, \quad \forall j = 1,\ldots,q''(S,G)+1. \tag{E.53}$$

Note that the above definition is different from the definition of $(t_j)_{j=0}^{q'(S,G)+1}$ in Algorithm 6.3 — we only use the above definition in this proof, for the purpose of simplifying notations.

### E.13.1 Definitions of Risky Events

For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we make some key definitions below.

1. For any $n_1, n_2 \in [T]$, we define a random variable $S(n_1,n_2)$ to be the total switching cost incurred in period $[n_1 : n_2]$ (note that if there is a switch happening between round $n_1-1$ and round $n_1$, or between round $n_2$ and round $n_2+1$, we do not count its cost in $S(n_1,n_2)$).

2. Second, we define a stopping time

$$\tau := \min\Big\{t \in [T] : \text{all of the actions in } [K] \text{ are chosen in period } [t_{q''(S,G)} : \tau]\Big\}$$

if the set is non-empty and $\tau = \infty$ otherwise. Note that this definition is different from the definition used in Appendix E.12.1.

3. We define a class of *risky* events as follows: for any $k \in [K]$, let

$E_{1,k} := \{\text{action } k \text{ is not chosen in period } [1 : t_1]\}$,

$E_{j,k} := \{\text{action } k \text{ is not chosen in period } [t_{j-1} : t_j]\}, \quad \forall j \in [2 : q''(S,G)]$,

$E_{q''(S,G)+1,k} := \Big\{\text{action } k \text{ is not chosen in period } \big[t_{q''(S,G)} : \lfloor (t_{q''(S,G)}+T)/2 \rfloor\big]\Big\}$,

$E_{q''(S,G)+2,k} := \Big\{\tau \le \lfloor (t_{q''(S,G)}+T)/2 \rfloor,\, a_\tau = k,\, S(1 : t_{q''(S,G)}) \ge q''(S,G)H\Big\}$.

By doing so, we get $(q''(S,G)+2)K$ risky events (of the form $E_{j,k}$) in total. Note that the time points $(t_j)_{j=1}^{q''(S,G)+1}$ are fixed and given in (E.53), and the events $(E_{q''(S,G)+2,k})_{k\in[K]}$ are defined based on the stopping time $\tau$.

### E.13.2 Combinatorial Arguments and Lower Bounds for Risky Events (Under a Single Environment)

The main purpose of this subsection is to prove the following lemma using (non-trivial) combinatorial (and probabilistic) arguments. Compared with the second step in the proof of Theorem 6.2, we develop new techniques here to deal with the general switching cost structure, while paying less attention to the order of $K$.

**Lemma E.15.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\sum_{j\in[q''(S,G)+2]} \sum_{k\in[K]} \mathbb{P}^{\pi}_{\boldsymbol{\mu}}(E_{j,k}) \geq 1.$$

Lemma E.15 leads to the following corollary, which will be utilized in subsequent subsections.

**Corollary E.3.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\frac{1}{(q''(S,G)+2)K} \sum_{j\in[q''(S,G)+2]} \sum_{k\in[K]} \mathbb{P}^{\pi}_{\boldsymbol{\mu}}(E_{j,k}) \geq \frac{1}{(q''(S,G)+2)K},$$

Corollary E.3 tells us the following fact: under any *single* environment $\boldsymbol{\mu}$, the average probability of the risky events is $\widetilde{\Omega}\big(\frac{1}{K}\big)$.

In the rest of this subsection, we provide a proof for Lemma E.15.

*Proof of Lemma E.15.* Since $H$ is the total weight of the shortest Hamiltonian path of $G$, for any $j \in [q''(S,G)]$, we have

$$\begin{aligned}
\sum_{k\in[K]} \mathbb{1}\{E_{j,k}\} &= \text{number of actions that are not chosen in period } [t_{j-1}:t_j] \\
&\geq \mathbb{1}\{\text{not all actions are chosen in period } [t_{j-1}:t_j]\} \\
&\geq \mathbb{1}\{S(t_{j-1}:t_j) < H\}
\end{aligned}$$

almost surely. Thus for any $j \in [q''(S,G)]$, we have

$$\begin{aligned}
\sum_{k\in[K]} \mathbb{P}^{\pi}_{\boldsymbol{\mu}}(E_{j,k}) &= \sum_{k\in[K]} \mathbb{E}^{\pi}_{\boldsymbol{\mu}}[\mathbb{1}\{E_{j,k}\}] \\
&= \mathbb{E}^{\pi}_{\boldsymbol{\mu}}\left[\sum_{k\in[K]} \mathbb{1}\{E_{j,k}\}\right] \\
&\geq \mathbb{E}^{\pi}_{\boldsymbol{\mu}}[\mathbb{1}\{S(t_{j-1}:t_j) < H\}]. \qquad \text{(E.54)}
\end{aligned}$$

Summing (E.54) over $j \in [q''(S,G)]$, we have

$$\begin{aligned}
\sum_{j\in[q''(S,G)]} \sum_{k\in[K]} \mathbb{P}^{\pi}_{\boldsymbol{\mu}}(E_{j,k}) &\geq \mathbb{E}^{\pi}_{\boldsymbol{\mu}}\left[\sum_{j\in[q''(S,G)]} \mathbb{1}\{S(t_{j-1}:t_j) < H\}\right] \\
&\overset{\text{(i)}}{\geq} \mathbb{E}^{\pi}_{\boldsymbol{\mu}}\big[\big(1 - \mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big) \geq q''(S,G)H\big\}\big)\big].
\end{aligned}$$

Note that (i) follows from

$$\begin{aligned}\sum_{j\in[q''(S,G)]} \mathbb{1}\{S(t_{j-1}:t_j)<H\} &\geq \mathbb{1}\left\{\bigcup_{j\in[q''(S,G)]}\{S(t_{j-1}:t_j)<H\}\right\}\\ &\overset{\text{(ii)}}{\geq} \mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)<q''(S,G)H\big\}\\ &= 1-\mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\},\end{aligned}$$

where (ii) follows from the pigeonhole principle.

Now we define

$$E_{\sim,k} := \big\{\text{action } k \text{ is not among the first } K-1 \text{ (different) actions chosen in period } \big[t_{q''(S,G)}:T\big]\big\}.$$

If both $\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}$ and $E_{\sim,k}$ happen, then since $\tau$ is the first time that all actions of $[K]$ have been chosen after round $t_{q''(S,G)}$, we know that either

$$E_{q''(S,G)+1,k} = \big\{\text{action } k \text{ is not chosen in period } \big[t_{q''(S,G)}:\lfloor(t_{q''(S,G)}+T)/2\rfloor\big]\big\}$$

happens, or

$$E_{q''(S,G)+2,k} = \big\{\tau\leq\lfloor(t_{q''(S,G)}+T)/2\rfloor,\, a_\tau=k,\, S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}$$

happens. Therefore, we know that

$$E_{q''(S,G)+1,k}\cup E_{q''(S,G)+2,k} \;\supset\; E_{\sim,k}\cap\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}.$$

This implies that

$$\begin{aligned}\sum_{k\in[K]}\mathbb{P}^\pi_{\boldsymbol{\mu}}\big(E_{q''(S,G)+1,k}\cup E_{q''(S,G)+2,k}\big) &\geq \sum_{k\in[K]}\mathbb{P}^\pi_{\boldsymbol{\mu}}\big(E_{\sim,k}\cap\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}\big)\\ &= \sum_{k\in[K]}\mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\mathbb{1}\{E_{\sim,k}\}\mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}\big]\\ &= \mathbb{E}^\pi_{\boldsymbol{\mu}}\left[\sum_{k\in[K]}\mathbb{1}\{E_{\sim,k}\}\mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}\right]\\ &\overset{\text{(iii)}}{\geq} \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}\big],\end{aligned}$$

where (iii) follow from the definition of $E_{\sim,k}$.

Combining the above two paragraphs, we have

$$\begin{aligned}\sum_{j\in[q''(S,G)+2]}\sum_{k\in[K]}\mathbb{P}^{\pi}_{\boldsymbol{\mu}}(E_{j,k}) &\geq \sum_{j\in[q''(S,G)]}\sum_{k\in[K]}\mathbb{P}^{\pi}_{\boldsymbol{\mu}}(E_{j,k}) + \sum_{k\in[K]}\mathbb{P}^{\pi}_{\boldsymbol{\mu}}\big(E_{q''(S,G)+1,k}\cup E_{q''(S,G)+2,k}\big)\\ &\geq \mathbb{E}^{\pi}_{\boldsymbol{\mu}}\big[\big(1-\mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}\big)\big]\\ &\quad+\mathbb{E}^{\pi}_{\boldsymbol{\mu}}\big[\mathbb{1}\big\{S\big(1:t_{q''(S,G)}\big)\geq q''(S,G)H\big\}\big]\\ &=1.\end{aligned}$$

□

### E.13.3 Alternative Environments, Bad Events, and Lower Bound Reductions

Compared with the proof of Theorem 6.2, the main difference in this step is that we define the reference and alternative environments in a new way. In particular, we make action 1 the *unique* optimal action in the reference environment, and let all the alternative environments in the form of $\boldsymbol{\mu}_{q''(S,G)+2,k}$ $(k\neq 1)$ be the same as the reference environments. Since any switch from or to action 1 incurs a cost at least $\max_i \min_{j\neq i} c_{i,j}$, such new techniques help us to make the quantity $\max_i \min_{j\neq i} c_{i,j}$ appear in the lower bound.

In the rest of the proof, we fix an arbitrary policy $\pi\in\Pi_S$.

For any $j\in[q''(S,G)+2]$, define a reward gap

$$\Delta_j := \begin{cases}1, & \text{if } j=1,\\ \frac{1}{2(q''(S,G)+2)}\sqrt{\frac{1}{t_{j-1}}}, & \text{if } j\in[2:q''(S,G)+1],\\ -\frac{1}{2(q''(S,G)+2)}\sqrt{\frac{1}{t_{q''(S,G)}}}, & \text{if } j=q''(S,G)+2.\end{cases}$$

Note that $|\Delta_j|\in[0,1]$ for all $j\in[q''(S,G)+2]$.

Let $\alpha=(\frac{\Delta_{q''(S,G)+1}}{2},0,0,\ldots,0)\in\mathbb{R}^K$ be the *reference* environment. Let $\mathbb{Q}:=\mathbb{P}^{\pi}_{\alpha}$ denote the *reference* measure.

For any $j\in[q''(S,G)+1], k\in[K]$, define an *alternative* environment $\boldsymbol{\mu}_{j,k}:=(\mu_{j,k;1},\ldots,\mu_{j,k;K})\in\mathbb{R}^K$ where

$$\mu_{j,k;i}:=\begin{cases}\alpha_i+\Delta_j, & \text{if } i=k,\\ \alpha_i, & \text{otherwise.}\end{cases}$$

Note that each alternative environment $\boldsymbol{\mu}_{j,k}$ define above only differs from the reference environment in terms of the mean reward of action $k$.

For $j=q''(S,G)+2$, for any $k\neq 1$, define an *alternative* environment $\boldsymbol{\mu}_{q''(S,G)+2,k}:=\alpha$ which is the same as the reference environment. For $j=q''(S,G)+2$ and $k=1$, define an *alternative* environment $\boldsymbol{\mu}_{q''(S,G)+2,1}:=\big(\mu_{q''(S,G)+2,1;1},\ldots,\mu_{q''(S,G)+2,1;K}\big)\in\mathbb{R}^K$ where

$$\mu_{q''(S,G)+2,1;i}:=\begin{cases}\alpha_i+\Delta_j, & \text{if } i=1,\\ \alpha_i, & \text{otherwise.}\end{cases}$$

Note that the above definitions are different from those in the proof of Theorem 6.2.

For any $j \in [q''(S,G)+2], k \in [K]$, let $\mathbb{P}_{j,k} := \mathbb{P}^{\pi}_{\boldsymbol{\mu}_{j,k}}$ denote the *alternative* measure associated with the alternative environment $\boldsymbol{\mu}_{j,k}$.

We explicitly construct a class of environments $\Phi := \{\boldsymbol{\mu}_{j,k} \mid j \in [q''(S,G)+2], k \in [K]\}$.

For any $j \in [q''(S,G)+2], k \in [K]$, under environment $\boldsymbol{\mu}_{j,k}$, the risky event $E_{j,k}$ becomes a *bad event* whose occurrence would lead to large regret. Specifically:

- Suppose $j = 1$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{1,k}$, choosing any action other than $k$ for one round incurs at least a $\Delta_1 - \frac{\Delta_{q''(S,G)+1}}{2} \geq \frac{\Delta_1}{2}$ term in the policy's regret, and the occurrence of $E_{1,k} = \{\text{action } k \text{ is not chosen in period } [1 : t_1]\}$ incurs at least a $t_1\Delta_1/2$ term in the policy's regret.

- Suppose $j \in [2 : q''(S,G)]$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{j,k}$, choosing any action other than $k$ for one round incurs at least a $\Delta_j - \frac{\Delta_{q''(S,G)+1}}{2} \geq \frac{\Delta_j}{2}$ term in the policy's regret, and the occurrence of $E_{j,k} = \{\text{action } k \text{ is not chosen in period } [t_{j-1} : t_j]\}$ incurs at least a $(t_j - t_{j-1} + 1)\Delta_j/2$ term in the policy's regret.

- Suppose $j = q''(S,G)+1$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{q''(S,G)+1,k}$, choosing any action other than $k$ for one round incurs at least a $\Delta_{q''(S,G)+1} - \frac{\Delta_{q''(S,G)+1}}{2} \geq \frac{\Delta_{q''(S,G)+1}}{2}$ term in the policy's regret, and the occurrence of $E_{q''(S,G)+1,k} = \big\{\text{action } k \text{ is not chosen in period } \big[t_{q''(S,G)} : \lfloor (t_{q''(S,G)} + T)/2 \rfloor\big]\big\}$ incurs at least a

  $$\big(\lfloor (t_{q''(S,G)} + T)/2 \rfloor - t_{q''(S,G)} + 1\big)\Delta_{q''(S,G)+1}/2$$

  term in the policy's regret.

- Suppose $j = q''(S,G)+2$ and $k = 1$. Since action 1 is the worst action under environment $\boldsymbol{\mu}_{q''(S,G)+2,1}$, choosing action 1 for one round incurs at least a $-\Delta_{q''(S,G)+2} - \frac{\Delta_{q''(S,G)+1}}{2} = -\frac{\Delta_{q''(S,G)+2}}{2}$ term in the policy's regret. Furthermore, by the switching constraint $S(1:T) \leq S < (q''(S,G)+1)H + \min_{j \neq 1} c_{1,j}$, the occurrence of $E_{q''(S,G)+2,1}$ implies the occurrence of $\{\text{no switch happens after round } \tau\}$ (as the remaining switching budget does not allow for switching from action 1), thus essentially implies the occurrence of $\big\{\text{action 1 is chosen in every round in period } [\lfloor (t_{q''(S,G)} + T)/2 \rfloor : T]\big\}$. As a result, the occurrence of $E_{q''(S,G)+2,1}$ incurs at least a $-\big(T - \lfloor (t_{q''(S,G)} + T)/2 \rfloor + 1\big)\Delta_{q''(S,G)+2}/2$ term in the policy's regret.

- Suppose $j = q''(S,G)+2$ and $k \neq 1$. Since action 1 is the unique optimal action under environment $\boldsymbol{\mu}_{q''(S,G)+2,k} = \alpha$, choosing action $k \neq 1$ for one round incurs at least a $\frac{\Delta_{q''(S,G)+1}}{2}$ term in the policy's regret. Furthermore, by the switching constraint $S(1:T) \leq S < (q''(S,G)+1)H + \min_{j \neq 1} c_{j,1}$ (we utilize the symmetry of switching costs here), the occurrence of $E_{q''(S,G)+2,1}$ implies the occurrence of $\{\text{action 1 is never chosen after round } \tau\}$ (as the remaining switching budget does not allow for switching to action 1), thus essentially implies the occurrence of $\big\{\text{action 1 is not chosen in period } [\lfloor (t_{q''(S,G)} + T)/2 \rfloor : T]\big\}$. As a result, the occurrence of $E_{q''(S,G)+2,1}$ incurs at least a $\big(T - \lfloor (t_{q''(S,G)} + T)/2 \rfloor + 1\big)\Delta_{q''(S,G)+1}/2 = -\big(T - \lfloor (t_{q''(S,G)} + T)/2 \rfloor + 1\big)\Delta_{q''(S,G)+2}/2$ term in the policy's regret.

Note that the arguments in the last two bullets are very different from what we have done in the proof of Theorem 6.2. The above arguments lead to Lemma E.16.

**Lemma E.16** (From risky events to bad events)**.** *For any $j \in [q''(S,G)+2], k \in [K]$, under environment $\boldsymbol{\mu}_{j,k}$, the risky event $E_{j,k}$ becomes a* bad event *in the sense that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\mu}_{j,k}}\left[T\mu_{j,k;k} - \sum_{t=1}^{T}\mu_{j,k;a_t} \mid E_{j,k}\right] \geq \mathcal{R}_{\mathrm{bad}}(S,G,T),$$

*where*

$$\mathcal{R}_{\mathrm{bad}}(S,G,T) := \frac{1}{16(q''(S,G)+2)}T^{\frac{1}{2-2^{-q''(S,G)}}}$$

*is a universal lower bound on the "distribution-dependent regret conditional on the bad event."*

*Proof of Lemma E.16.* By the arguments in the previous paragraph, we have

$$\begin{aligned}
&\mathbb{E}^{\pi}_{\boldsymbol{\mu}_{j,k}}\left[T\mu_{j,k;k} - \sum_{t=1}^{T}\mu_{j,k;a_t} \mid E_{j,k}\right] \\
&\geq \mathbb{E}^{\pi}_{\boldsymbol{\mu}_{j,k}}\left[\Big(t_j - t^{(2)}_{j-1}+1\Big)\mu_{j,k;k} - \sum_{t\in[t_{j-1}:t_j]}\mu_{j,k;a_t} \mid E_{j,k}\right] \\
&\geq \begin{cases} t_1\Delta_1/2, & \text{if } j=1, \\ (t_j - t_{j-1}+1)\Delta_j/2, & \text{if } j\in[2:q''(S,G)], \\ \big(\lfloor (t_{q''(S,G)}+T)/2\rfloor - t_{q''(S,G)}+1\big)\Delta_{q''(S,G)+1}/2, & \text{if } j = q''(S,G)+1, \\ -\big(T - \lfloor (t_{q''(S,G)}+T)/2\rfloor+1\big)\Delta_{q''(S,G)+2}/2, & \text{if } j = q''(S,G)+2, \end{cases} \\
&\geq \frac{1}{16(q''(S,G)+2)}T^{\frac{1}{2-2^{-q''(S,G)}}},
\end{aligned}$$

where the last inequality follows from the same algebra presented in the proof of Lemma E.9. □

Based on Lemma E.16, we can reduce the task of proving a lower bound on the policy's (distribution-dependent) regret $R^{\pi}_{\boldsymbol{\mu}_{j,k}}(T)$ to the task of proving a lower bound on the bad event probability $\mathbb{P}_{j,k}(E_{j,k})$. Consequently, we can reduce the task of proving a lower bound on $\sup_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T)$ to the task of proving a lower bound on the "average-case bad event probability"

$$\frac{1}{(q''(S,G)+2)K}\sum_{j\in[q''(S,G)+2]}\sum_{k\in[K]}\mathbb{P}_{j,k}(E_{j,k}).$$

**Lemma E.17** (Reducing regret lower bounds to bad event probability lower bounds)**.** *For any $j \in [q''(S,G)+2], k\in[K]$, we have*

$$R^{\pi}_{\boldsymbol{\mu}_{j,k}}(T) \geq \mathcal{R}_{\mathrm{bad}}(S,G,T)\cdot\mathbb{P}_{j,k}(E_{j,k}).$$

*As a result, we have*

$$R^{\pi}(T) \geq \sup_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T) \geq \mathcal{R}_{\mathrm{bad}}(S,K,T)\cdot\frac{1}{(q''(S,G)+2)K}\sum_{j\in[q''(S,G)+2]}\sum_{k\in[K]}\mathbb{P}_{j,k}(E_{j,k}).$$

*Proof of Lemma E.17.* The proof is almost the same as the proof of Lemma E.10. □

### E.13.4 Probability Space Changing Tricks

Lemma E.17 indicates that, in order to prove the desired lower bound on the regret $R^{\pi}(T)$, it suffices to prove the following statement:

$$\bar{p} := \frac{1}{(q''(S,G)+2)K} \sum_{j\in[q''(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{j,k}(E_{j,k}) = \widetilde{\Omega}\bigg(\frac{1}{K}\bigg), \tag{E.55}$$

That is, we only need to establish tight lower bounds on the average probability $\bar{p}$.

By Corollary E.3, we have

$$\bar{q} := \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{Q}(E_{j,k}) \geq \frac{1}{(q(S,K)+2)K} = \widetilde{\Omega}\bigg(\frac{1}{K}\bigg),$$

Therefore, in order to show (E.55), it suffices to show that $\bar{p}$ is close to $\bar{q}$. Similar to the proof of Theorem 6.2, we apply the probability space changing tricks. The arguments are very similar to the arguments in Appendix E.12.4 and are omitted here.

### E.13.5 Applying the GRF Inequality

Similar to Appendix E.12.5, by applying the GRF inequality, we can show that

$$\begin{aligned} \bar{p} &\geq \bar{q} - \sqrt{2\bar{q}\frac{\bar{q}}{4}} \\ &= \frac{2-\sqrt{2}}{2}\bar{q}. \end{aligned} \tag{E.56}$$

Now we plug (E.56) into Lemma E.17. We have

$$\begin{aligned} R^{\pi}(T) &\geq \sup_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T) \\ &\geq \mathcal{R}_{\mathrm{bad}}(S,G,T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{j,k}(E_{j,k}) \\ &\geq \mathcal{R}_{\mathrm{bad}}(S,G,T) \cdot \frac{2-\sqrt{2}}{2}\bar{q} \\ &= \frac{1}{16(q''(S,G)+2)} T^{\frac{1}{2-2^{-q''(S,G)}}} \cdot \frac{2-\sqrt{2}}{2} \frac{1}{(q(S,K)+2)K} \\ &= \Omega\bigg(\frac{1}{K\log T} T^{\frac{1}{2-2^{-q''(S,G)}}}\bigg). \end{aligned}$$

We thus complete the proof of Theorem 6.4. □

## E.14 Proof of the Lower Bound in Theorem 6.5

The proof of the lower bound builds on the proof of Theorem 6.2. In this proof, we emphasize the differences, which are mainly in Appendices E.14.1 and E.14.2, corresponding to the first two steps of the RECAP method.

Given any $K > 1$, $S \geq 0$ and $T \geq 2K$, we focus on the setting of $\mathcal{D}_k = \mathcal{N}(\mu_k, 1)$ $(\forall k \in [K])$, as this is sufficient for us to prove the desired lower bound. For simplicity, in this proof we will directly use the vector $\boldsymbol{\mu}$ to represent the environment.

For any environment $\boldsymbol{\mu}$, let $X_{\boldsymbol{\mu}}^t(k) \sim \mathcal{N}(\mu_k, 1)$ denote the i.i.d. random reward of each action $k$ at round $t$ $(k \in [K], t \in [T])$. For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, for any $t \in [T]$, we use $a_t$ to denote the random action selected by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$, and use $X_{\boldsymbol{\mu}}^t(a_t)$ to denote the random reward observed by policy $\pi$ at round $t$ under environment $\boldsymbol{\mu}$. Let $\mathcal{H}_t := \big((a_1, X_{\boldsymbol{\mu}}^1(a_1)), \ldots, (a_t, X_{\boldsymbol{\mu}}^t(a_t))\big)$ be the history (of actions and observations) up to round $t$ (inclusive), whose value lies in $\Omega_t := ([K] \times \mathbb{R})^t$. Let $\mathcal{F}_t := \mathcal{B}(\Omega_t)$ be the Borel $\sigma$-algebra on $\Omega_t$. Let $\mathbb{P}_{\boldsymbol{\mu}}^{\pi}$ be the probability measure induced by (i.e., the law of) $\mathcal{H}_T$, and $\mathbb{E}_{\boldsymbol{\mu}}^{\pi}$ be the associated expectation operator. Let $R_{\boldsymbol{\mu}}^{\pi}(T) := T\mu^\star - \mathbb{E}_{\boldsymbol{\mu}}^{\pi}\left[\sum_{t=1}^T \mu_{a_t}\right]$ be policy $\pi$'s distribution-dependent regret under environment $\boldsymbol{\mu}$.

Similar to Appendix E.12, we argue that in our proof, we only need to consider the case of $q(S, \boldsymbol{c}) + 2 \leq \log_2 \log_2(T)$. Suppose $q(S, \boldsymbol{c}) + 2 > \log_2 \log_2(T)$, then we have

$$T^{\frac{1}{2-2^{-q(S,\boldsymbol{c})}}} \leq 4\sqrt{T},$$

thus the lower bound in Theorem 6.5 becomes $\widetilde{\Omega}(\sqrt{T})$ and can be directly obtained by applying the well-known $\Omega(\sqrt{KT})$ lower bound of the classical MAB (see, e.g., Lattimore and Szepesvári 2020). Therefore, the really non-trivial case of Theorem 6.5 is the case of $q(S, \boldsymbol{c}) + 2 \leq \log_2 \log_2(T)$, and we focus on this case in the rest of our proof.

Our goal is to explicitly construct a family of environments $\Phi$, such that for any $S$-switching-budget policy $\pi \in \Pi_S$, the "worst-case regret" $\max_{\boldsymbol{\mu} \in \Phi} R_{\boldsymbol{\mu}}^{\pi}(T)$ is lower bounded by

$$\Omega\left(\frac{1}{K \log T} T^{\frac{1}{2-2^{-q(S,\boldsymbol{c})}}}\right)$$

Since the worst-case regret $R^{\pi}(T)$ is no less than the $\max_{\boldsymbol{\mu} \in \Phi} R_{\boldsymbol{\mu}}^{\pi}(T)$, the above goal directly implies the lower bound in Theorem 6.5.

Without loss of generality, we assume that $1 \in \arg\max_{i \in [K]} \min_{j \neq i} c_{i,j}$. For notational simplicity, we **redefine** the sequence $(t_j)_{j=0}^{q(S,\boldsymbol{c})+1}$ as $t_0 = 0$ and

$$t_j = \left\lfloor T^{\frac{2-2^{-(j-1)}}{2-2^{-q(S,\boldsymbol{c})}}} \right\rfloor, \quad \forall j = 1, \ldots, q(S, \boldsymbol{c}) + 1. \tag{E.57}$$

Note that the above definition is different from the definition of $(t_j)_{j=0}^{q(S,\boldsymbol{c})+1}$ in Algorithm 6.4 — we only use the above definition in this proof, for the purpose of simplifying notations.

### E.14.1 Definitions of Risky Events

For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we make some key definitions below.

1. For any $n_1, n_2 \in [T]$, we define a random variable $S(n_1, n_2)$ to be the total switching cost incurred in period $[n_1 : n_2]$ (note that if there is a switch happening between round $n_1 - 1$ and round $n_1$, or between round $n_2$ and round $n_2 + 1$, we do not count its cost in $S(n_1, n_2)$).

2. Second, we define a stopping time

$$\tau := \min\big\{t \in [T] : \text{all of the actions in } [K] \text{ are chosen in period } [t_{q(S,\boldsymbol{c})} : \tau]\big\}$$

if the set is non-empty and $\tau = \infty$ otherwise. Note that this definition is different from the definition used in Appendix E.12.1.

3. We define a class of *risky* events as follows: for any $k \in [K]$, let

$E_{1,k} := \{\text{action } k \text{ is not chosen in period } [1 : t_1]\}$,

$E_{j,k} := \{\text{action } k \text{ is not chosen in period } [t_{j-1} : t_j]\}, \quad \forall j \in [2 : q(S, \boldsymbol{c})]$,

$E_{q(S,\boldsymbol{c})+1,k} := \big\{\text{action } k \text{ is not chosen in period } \big[t_{q(S,\boldsymbol{c})} : \lfloor (t_{q(S,\boldsymbol{c})} + T)/2 \rfloor\big]\big\}$,

$E_{q(S,\boldsymbol{c})+2,k} := \big\{\tau \le \lfloor (t_{q(S,\boldsymbol{c})} + T)/2 \rfloor, a_\tau = k, S(1 : t_{q(S,\boldsymbol{c})}) > S - \Sigma\big\}$.

By doing so, we get $(q(S, \boldsymbol{c}) + 2)K$ risky events (of the form $E_{j,k}$) in total. Note that the time points $(t_j)_{j=1}^{q(S,\boldsymbol{c})+1}$ are fixed and given in (E.57), and the events $(E_{q(S,\boldsymbol{c})+2,k})_{k\in[K]}$ are defined based on the stopping time $\tau$.

### E.14.2 Combinatorial Arguments and Lower Bounds for Risky Events (Under a Single Environment)

The main purpose of this subsection is to prove the following lemma using (non-trivial) combinatorial (and probabilistic) arguments. Compared with the second step in the proof of Theorem 6.2, we develop new techniques here to deal with the departure cost structure, while paying less attention to the order of $K$.

**Lemma E.18.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\sum_{j\in[q(S,\boldsymbol{c})+2]} \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) \ge 1.$$

Lemma E.18 leads to the following corollary, which will be utilized in subsequent subsections.

**Corollary E.4.** *For any policy $\pi \in \Pi_S$, for any environment $\boldsymbol{\mu}$, we have*

$$\frac{1}{(q(S,\boldsymbol{c})+2)K} \sum_{j\in[q(S,\boldsymbol{c})+2]} \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) \ge \frac{1}{(q(S,\boldsymbol{c})+2)K},$$

Corollary E.4 tells us the following fact: under any *single* environment $\boldsymbol{\mu}$, the average probability of the risky events is $\widetilde{\Omega}\big(\frac{1}{K}\big)$.

In the rest of this subsection, we provide a proof for Lemma E.18.

*Proof of Lemma E.18.* We discuss two cases: $q(S, \boldsymbol{c})$ is odd, and $q(S, \boldsymbol{c})$ is even. Suppose that $q(S, \boldsymbol{c})$ is odd. For any $j \in [q(S, \boldsymbol{c})]$, we have

$$\begin{aligned}\sum_{k\in[K]} \mathbb{1}\{E_{j,k}\} &= \text{number of actions that are not chosen in period } [t_{j-1} : t_j] \\ &\geq \mathbb{1}\{\text{not all actions are chosen in period } [t_{j-1} : t_j]\} \\ &= 1 - \mathbb{1}\{\text{all actions are chosen in period } [t_{j-1} : t_j]\}\end{aligned}$$

almost surely. Thus for any $j \in \{1, 3, \cdots, q(S, \boldsymbol{c}) - 2\}$, we have

$$\begin{aligned}&\sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) + \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j+1,k}) \\ &= \sum_{k\in[K]} \mathbb{E}^\pi_{\boldsymbol{\mu}}[\mathbb{1}\{E_{j,k}\}] + \sum_{k\in[K]} \mathbb{E}^\pi_{\boldsymbol{\mu}}[\mathbb{1}\{E_{j+1,k}\}] \\ &= \mathbb{E}^\pi_{\boldsymbol{\mu}}\left[\sum_{k\in[K]} \mathbb{1}\{E_{j,k}\}\right] + \mathbb{E}^\pi_{\boldsymbol{\mu}}\left[\sum_{k\in[K]} \mathbb{1}\{E_{j+1,k}\}\right] \\ &\geq 1 - \mathbb{1}\{\text{all actions are chosen in period } [t_{j-1} : t_j]\} \\ &\quad + 1 - \mathbb{1}\{\text{all actions are chosen in period } [t_j : t_{j+1}]\} \\ &\geq 1 - \mathbb{1}\{\text{all actions are chosen in both period } [t_{j-1} : t_j] \text{ and period } [t_j : t_{j+1}]\} \\ &= 1 - \mathbb{1}\Big\{S(t_{j-1} : t_{j+1}) \geq 2\Sigma - c^{(1)} - c^{(2)}\Big\} \\ &= \mathbb{1}\Big\{S(t_{j-1} : t_{j+1}) < 2\Sigma - c^{(1)} - c^{(2)}\Big\}.\end{aligned}$$

Moreover, for $j = q(S, \boldsymbol{c})$, we can show that

$$\sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}\big(E_{q(S,\boldsymbol{c}),k}\big) \geq 1 - \mathbb{1}\Big\{S(t_{j-1} : t_j) \geq \Sigma - c^{(1)}\Big\} = \mathbb{1}\Big\{S(t_{j-1} : t_j) < \Sigma - c^{(1)}\Big\}.$$

Therefore, by the pigeonhole principle and the definition of $q(S, \boldsymbol{c})$, we have

$$\begin{aligned}&\sum_{j\in[q(S,\boldsymbol{c})]} \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) \\ &\geq \mathbb{E}^\pi_{\boldsymbol{\mu}}\left[\sum_{j\in\{1,3,\cdots,q(S,\boldsymbol{c})-2\}} \mathbb{1}\Big\{S(t_{j-1} : t_{j+1}) < 2\Sigma - c^{(1)} - c^{(2)}\Big\} + \mathbb{1}\Big\{S\big(t_{q(S,\boldsymbol{c})-1} : t_{q(S,\boldsymbol{c})}\big) < \Sigma - c^{(1)}\Big\}\right] \\ &\geq \mathbb{E}^\pi_{\boldsymbol{\mu}}\left[\left(1 - \mathbb{1}\Big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) \geq q(S, \boldsymbol{c})\Sigma - \left\lceil \frac{q(S,\boldsymbol{c})}{2} \right\rceil c^{(1)} - \left\lfloor \frac{q(S,\boldsymbol{c})}{2} \right\rfloor c^{(2)}\Big\}\right)\right] \\ &\geq \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\big(1 - \mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big)\big].\end{aligned}$$

Suppose that $q(S, \boldsymbol{c})$ is even. Then using similar arguments, we can still show that

$$\sum_{j\in[q(S,\boldsymbol{c})]} \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) \geq \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\big(1 - \mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big)\big].$$

Therefore, the above inequality always holds.

Now we define

$$E_{\sim,k} := \big\{\text{action } k \text{ is not among the first } K-1 \text{ (different) actions chosen in period } \big[t_{q(S,\boldsymbol{c})} : T\big]\big\}.$$

If both $\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}$ and $E_{\sim,k}$ happen, then since $\tau$ is the first time that all actions of $[K]$ have been chosen after round $t_{q(S,\boldsymbol{c})}$, we know that either

$$E_{q(S,\boldsymbol{c})+1,k} = \big\{\text{action } k \text{ is not chosen in period } \big[t_{q(S,\boldsymbol{c})} : \lfloor (t_{q(S,\boldsymbol{c})} + T)/2 \rfloor\big]\big\}$$

happens, or

$$E_{q(S,\boldsymbol{c})+2,k} = \big\{\tau \le \lfloor (t_{q(S,\boldsymbol{c})} + T)/2 \rfloor,\, a_\tau = k,\, S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}$$

happens. Therefore, we know that

$$E_{q(S,\boldsymbol{c})+1,k} \cup E_{q(S,\boldsymbol{c})+2,k} \;\supset\; E_{\sim,k} \cap \big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}.$$

This implies that

$$\begin{aligned}
\sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}\big(E_{q(S,\boldsymbol{c})+1,k} \cup E_{q(S,\boldsymbol{c})+2,k}\big) &\ge \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}\big(E_{\sim,k} \cap \big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big) \\
&= \sum_{k\in[K]} \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\mathbb{1}\{E_{\sim,k}\}\mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big] \\
&= \mathbb{E}^\pi_{\boldsymbol{\mu}}\left[\sum_{k\in[K]} \mathbb{1}\{E_{\sim,k}\}\mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\right] \\
&\overset{\text{(iii)}}{\ge} \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big],
\end{aligned}$$

where (iii) follow from the definition of $E_{\sim,k}$.

Combining the above two paragraphs, we have

$$\begin{aligned}
\sum_{j\in[q(S,\boldsymbol{c})+2]} \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) &\ge \sum_{j\in[q(S,\boldsymbol{c})]} \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}(E_{j,k}) + \sum_{k\in[K]} \mathbb{P}^\pi_{\boldsymbol{\mu}}\big(E_{q(S,\boldsymbol{c})+1,k} \cup E_{q(S,\boldsymbol{c})+2,k}\big) \\
&\ge \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\big(1 - \mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big)\big] \\
&\quad + \mathbb{E}^\pi_{\boldsymbol{\mu}}\big[\mathbb{1}\big\{S\big(1 : t_{q(S,\boldsymbol{c})}\big) > S - \Sigma\big\}\big] \\
&= 1.
\end{aligned}$$

□

### E.14.3 Alternative Environments, Bad Events, and Lower Bound Reductions

In the rest of the proof, we fix an arbitrary policy $\pi \in \Pi_S$.

For any $j \in [q(S, \boldsymbol{c}) + 2]$, define a reward gap

$$\Delta_j := \begin{cases} 1, & \text{if } j = 1, \\ \frac{1}{2(q(S,\boldsymbol{c})+2)} \sqrt{\frac{1}{t_{j-1}}}, & \text{if } j \in [2 : q(S, \boldsymbol{c}) + 1], \\ -\frac{1}{2(q(S,\boldsymbol{c})+2)} \sqrt{\frac{1}{t_{q(S,\boldsymbol{c})}}}, & \text{if } j = q(S, \boldsymbol{c}) + 2. \end{cases}$$

Note that $|\Delta_j| \in [0, 1]$ for all $j \in [q(S, \boldsymbol{c}) + 2]$.

Let $\boldsymbol{\mu} = (0, \ldots, 0) \in \mathbb{R}^K$ be the *reference* environment. Let $\mathbb{Q} := \mathbb{P}^{\pi}_{\mathbf{0}}$ denote the *reference* measure.

For any $j \in [q(S, \boldsymbol{c}) + 2], k \in [K]$, define an *alternative* environment $\boldsymbol{\mu}_{j,k} := (\mu_{j,k;1}, \ldots, \mu_{j,k;K}) \in \mathbb{R}^K$ where

$$\mu_{j,k;i} := \begin{cases} 0 + \Delta_j, & \text{if } i = k, \\ 0, & \text{otherwise.} \end{cases}$$

Note that each alternative environment $\boldsymbol{\mu}_{j,k}$ only differs from the reference environment in terms of the mean reward of action $k$.

For any $j \in [q(S, \boldsymbol{c}) + 2], k \in [K]$, let $\mathbb{P}_{j,k} := \mathbb{P}^{\pi}_{\boldsymbol{\mu}_{j,k}}$ denote the *alternative* measure associated with the alternative environment $\boldsymbol{\mu}_{j,k}$.

We explicitly construct a class of environments $\Phi := \{\boldsymbol{\mu}_{j,k} \mid j \in [q(S, \boldsymbol{c}) + 2], k \in [K]\}$.

For any $j \in [q(S, \boldsymbol{c}) + 2], k \in [K]$, under environment $\boldsymbol{\mu}_{j,k}$, the risky event $E_{j,k}$ becomes a *bad event* whose occurrence would lead to large regret. Specifically:

- Suppose $j = 1$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{1,k}$, choosing any action other than $k$ for one round incurs at least a $\Delta_1$ term in the policy's regret, and the occurrence of $E_{1,k} = \{\text{action } k \text{ is not chosen in period } [1 : t_1]\}$ incurs at least a $t_1\Delta_1$ term in the policy's regret.

- Suppose $j \in [2 : q(S, \boldsymbol{c})]$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{j,k}$, choosing any action other than $k$ for one round incurs at least a $\Delta_j$ term in the policy's regret, and the occurrence of $E_{j,k} = \{\text{action } k \text{ is not chosen in period } [t_{j-1} : t_j]\}$ incurs at least a $(t_j - t_{j-1} + 1)\Delta_j$ term in the policy's regret.

- Suppose $j = q(S, \boldsymbol{c}) + 1$. Since action $k$ is the unique optimal action under environment $\boldsymbol{\mu}_{q(S,\boldsymbol{c})+1,k}$, choosing any action other than $k$ for one round incurs at least a $\Delta_{q(S,\boldsymbol{c})+1}$ term in the policy's regret, and the occurrence of

  $$E_{q(S,\boldsymbol{c})+1,k} = \big\{\text{action } k \text{ is not chosen in period } \big[t_{q(S,\boldsymbol{c})} : \lfloor (t_{q(S,\boldsymbol{c})} + T)/2 \rfloor\big]\big\}$$

  incurs at least a $\big(\lfloor (t_{q(S,\boldsymbol{c})} + T)/2 \rfloor - t_{q(S,\boldsymbol{c})} + 1\big)\Delta_{q(S,\boldsymbol{c})+1}$ term in the policy's regret.

- Suppose $j = q(S, \boldsymbol{c}) + 2$. Since action $k$ is the worst action under environment $\boldsymbol{\mu}_{q(S,\boldsymbol{c})+2,k}$, choosing action $k$ for one round incurs at least a $-\Delta_{q(S,\boldsymbol{c})+2}$ term in the policy's regret. Furthermore, by the switching constraint $S(1 : T) \le S < S(1 : t_{q(S,\boldsymbol{c})}) + \Sigma$, the occurrence of $E_{q(S,\boldsymbol{c})+2,k}$ implies the occurrence of $\{\text{no switch happens after round } \tau\}$ (as the remaining switching budget does not allow for switching from action $k$), thus essentially implies the

occurrence of $\Big\{$action $k$ is chosen in every round in $[\lfloor (t^{(1)}_{q(S,\boldsymbol{c})} + T)/2 \rfloor : T]\Big\}$. As a result, the occurrence of $E_{q(S,\boldsymbol{c})+2,k}$ incurs at least a $-\Big(T - \lfloor (t^{(1)}_{q(S,\boldsymbol{c})} + T)/2 \rfloor + 1\Big)\Delta_{q(S,\boldsymbol{c})+2}$ term in the policy's regret.

The above arguments lead to Lemma E.19.

**Lemma E.19** (From risky events to bad events)**.** *For any $j \in [q(S,\boldsymbol{c})+2], k \in [K]$, under environment $\boldsymbol{\mu}_{j,k}$, the risky event $E_{j,k}$ becomes a* bad event *in the sense that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\mu}_{j,k}}\left[T\mu_{j,k;k} - \sum_{t=1}^{T}\mu_{j,k;a_t} \mid E_{j,k}\right] \geq \mathcal{R}_{\mathrm{bad}}(S,G,T),$$

*where*

$$\mathcal{R}_{\mathrm{bad}}(S,G,T) := \frac{1}{8(q(S,\boldsymbol{c})+2)}T^{\frac{1}{2-2^{-q(S,\boldsymbol{c})}}}$$

*is a universal lower bound on the "distribution-dependent regret conditional on the bad event."*

*Proof of Lemma E.19.* The proof is almost the same as the proof of Lemma E.9. □

Based on Lemma E.19, we can reduce the task of proving a lower bound on the policy's (distribution-dependent) regret $R^{\pi}_{\boldsymbol{\mu}_{j,k}}(T)$ to the task of proving a lower bound on the bad event probability $\mathbb{P}_{j,k}(E_{j,k})$. Consequently, we can reduce the task of proving a lower bound on $\sup_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T)$ to the task of proving a lower bound on the "average-case bad event probability"

$$\frac{1}{(q(S,\boldsymbol{c})+2)K}\sum_{j\in[q(S,\boldsymbol{c})+2]}\sum_{k\in[K]}\mathbb{P}_{j,k}(E_{j,k}).$$

**Lemma E.20** (Reducing regret lower bounds to bad event probability lower bounds)**.** *For any $j \in [q(S,\boldsymbol{c})+2], k \in [K]$, we have*

$$R^{\pi}_{\boldsymbol{\mu}_{j,k}}(T) \geq \mathcal{R}_{\mathrm{bad}}(S,G,T)\cdot\mathbb{P}_{j,k}(E_{j,k}).$$

*As a result, we have*

$$R^{\pi}(T) \geq \sup_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T) \geq \mathcal{R}_{\mathrm{bad}}(S,K,T)\cdot\frac{1}{(q(S,\boldsymbol{c})+2)K}\sum_{j\in[q(S,\boldsymbol{c})+2]}\sum_{k\in[K]}\mathbb{P}_{j,k}(E_{j,k}).$$

*Proof of Lemma E.20.* The proof is almost the same as the proof of Lemma E.10. □

### E.14.4 Probability Space Changing Tricks

Lemma E.20 indicates that, in order to prove the desired lower bound on the regret $R^{\pi}(T)$, it suffices to prove the following statement:

$$\bar{p} := \frac{1}{(q(S,\boldsymbol{c})+2)K}\sum_{j\in[q''(S,K)+2]}\sum_{k\in[K]}\mathbb{P}_{j,k}(E_{j,k}) = \widetilde{\Omega}\left(\frac{1}{K}\right), \tag{E.58}$$

That is, we only need to establish tight lower bounds on the average probability $\bar{p}$.

By Corollary E.4, we have

$$\bar{q} := \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{Q}(E_{j,k}) \geq \frac{1}{(q(S,K)+2)K} = \widetilde{\Omega}\bigg(\frac{1}{K}\bigg),$$

Therefore, in order to show (E.58), it suffices to show that $\bar{p}$ is close to $\bar{q}$. Similar to the proof of Theorem 6.2, we apply the probability space changing tricks. The arguments are very similar to the arguments in Appendix E.12.4 and are omitted here.

### E.14.5 Applying the GRF Inequality

Similar to Appendix E.12.5, by applying the GRF inequality, we can show that

$$\begin{aligned} \bar{p} &\geq \bar{q} - \sqrt{2\bar{q}\frac{\bar{q}}{4}} \\ &= \frac{2-\sqrt{2}}{2}\bar{q}. \end{aligned} \tag{E.59}$$

Now we plug (E.59) into Lemma E.20. We have

$$\begin{aligned} R^{\pi}(T) &\geq \sup_{\boldsymbol{\mu}\in\Phi} R^{\pi}_{\boldsymbol{\mu}}(T) \\ &\geq \mathcal{R}_{\mathrm{bad}}(S,G,T) \cdot \frac{1}{(q(S,K)+2)K} \sum_{j\in[q(S,K)+2]} \sum_{k\in[K]} \mathbb{P}_{j,k}(E_{j,k}) \\ &\geq \mathcal{R}_{\mathrm{bad}}(S,G,T) \cdot \frac{2-\sqrt{2}}{2}\bar{q} \\ &= \frac{1}{8(q(S,\boldsymbol{c})+2)} T^{\frac{1}{2-2^{-q(S,\boldsymbol{c})}}} \cdot \frac{2-\sqrt{2}}{2} \frac{1}{(q(S,K)+2)K} \\ &= \Omega\bigg(\frac{1}{K\log T} T^{\frac{1}{2-2^{-q(S,\boldsymbol{c})}}}\bigg). \end{aligned}$$

We thus complete the proof of the lower bound. □

# Appendix F

# Supplementary Material for Chapter 7

## F.1 Extended Models in the Bandits with Knapsacks Setup

In this section, we introduce the bandits with knapsacks (BwK) problem and its variant, the bandits with knapsacks under limited switches (BwK-LS) problem. We explain the relations between BNRM v.s. BwK, and the relations between BNRM-LS v.s. BwK-LS in Section F.1.1.

The BwK problem is a general learning framework introduced in Badanidiyuru et al. (2018) and has been later on extensively studied in the machine learning literature. It generalizes the classical MAB problem and includes the BNRM problem as a special case. See Badanidiyuru et al. (2018) for a review of relevant literature.

### BwK Setup

Similar to the BNRM problem, let there be a discrete, finite time horizon with $T$ periods. Time starts from period 1 and ends in period $T$. Unlike BNRM, there is no "product" nor "consumption matrix" in BwK. Let there be $d$ different resources, each endowed with finite initial capacity $B_i$, $\forall i \in [d]$.

In each period $t$, the decision maker pulls one arm from a finite set of $K$ distinct arms, which we denote using $z_t \in [K]$. Each time an arm $k \in [K]$ is pulled, a random reward $R_k \in [0, R_{\mathsf{max}}]$ is received at a random cost $C_{i,k} \in [0, C_{\mathsf{max}}]$ of each resource $i$, which we denote using the cost vector $\boldsymbol{C}_k \in [0, C_{\mathsf{max}}]^d$. The distributions of both the random reward and the random cost vector are fixed but unknown to the decision maker, and have to be sequentially learned over time. The decision maker stops at the earlier time when one or more resource constraint is violated, or when the time horizon ends. We use $\mathcal{I} = (T, \boldsymbol{B}, K, d; \boldsymbol{R}, \boldsymbol{C})$ to stand for one instance of the BwK problem.

### Regime for Regret Analysis

Similar to the BNRM problem, we derive non-asymptotic bounds on the regret of policies in terms of the number of time periods $T$. Again, we adopt the following regret analysis regime: there exists an arbitrary positive constant $\underline{b} > 0$, such that $B_{\mathsf{min}} \geq \underline{b}T$.

Following the literature, we assume $C_{\mathsf{max}}, R_{\mathsf{max}}$ are all absolute constants that do not depend on $T$ or $\boldsymbol{B}$. The other parameters $K$ and $d$ do not depend on $T$ or $\boldsymbol{B}$, either. Yet we write out our regret

bounds' exact dependence on $K$ and $d$ in our main theorems and all the proofs. Obtaining regret upper and lower bounds that are tight in the orders of $K$ and $d$ is an interesting future direction.

#### New Constraint to BwK

We model the business constraint of limited changes between arms as a hard constraint, and define the BwK-LS problems as the BwK problems with an extra constraint of limited switches. Specifically, on top of the initial resource capacities, the decision maker is initially endowed with a fixed number of switching budget $s$, to change the arm from one to another. When two consecutive arm pulls are different, i.e., $z_t \neq z_{t+1}$, one unit of switching budget is consumed. When there is no switching budget remaining, the decision maker cannot change the arm anymore, and has to keep pulling the last arm pulled.

There are other ways to model the business constraint of limited switches, but all are beyond the scope of this paper; see Section F.1.2 for more discussions. We can view the BwK problems as the BwK-LS problems under an infinite switching budget. Since a limited switching budget restricts the family of admissible policies, any admissible algorithm for the BwK-LS problem is also an admissible algorithm for the BwK problem.

### F.1.1 Comparison of the Models

The BNRM problem and the BwK problem are closely related. The distinct price vectors in the BNRM problem corresponds to the distinct arms in the BwK problem. For each $k \in [K]$, the revenue of price vector $\boldsymbol{p}_k$, $\sum_{j=1}^{n} Q_{j,k} p_{j,k}$, corresponds to the reward of arm $k$, $R_k$. And for each resource $i \in [d]$, the consumption of price vector $\boldsymbol{p}_k$, $\sum_{j=1}^{n} Q_{j,k} a_{ij}$, corresponds to the cost of arm $k$, $C_{i,k}$. The BwK problem is more general than the BNRM problem, in the sense that for any fixed arm $k \in [K]$, the reward $R_k$ and any cost $C_{i,k}$ can have an arbitrary and unknown relationship. But in the BNRM problem, the revenue $\sum_{j=1}^{n} Q_{j,k} p_{j,k}$ and any consumption $\sum_{j=1}^{n} Q_{j,k} a_{ij}$ are correlated through the random demand $Q_{j,k}$ (intuitively, the revenue earned is proportional to the consumption of resources). The BNRM problem can thus be understood as a special case of the BwK problem with a specific reward-cost structure.

Due to this reason, after adding the switching constraint, the BwK-LS problem is more general than the BNRM-LS problem. Consequently, establishing an upper bound on regret for BwK-LS is more challenging than establishing an upper bound on regret for BNRM-LS; meanwhile, establishing a lower bound on regret for BNRM-LS is more challenging than establishing a lower bound on regret for BwK-LS, as one has to construct hard problem instances without breaking the specific structure of BNRM-LS (this can be highly non-trivial, as illustrated in Section 7.1.2).

### F.1.2 Related Modeling Components

We survey other related modeling components that have appeared in the literature, including the stopping criterion, performance metric, and the modeling of limited switches.

### Stopping Criterion

At each point in time, as long as the remaining inventory for any resource is zero, the selling horizon stops. This stopping criterion is standard in the BNRM and BwK literature, see Besbes and Zeevi (2012), Badanidiyuru et al. (2018). We refer to this stopping criterion as the "ungenerous" stopping criterion.

There is a second stopping criterion that is common in the revenue management literature when the stochastic distribution is known. This setup assumes time horizon never stops. Even if some resources are stocked-out, the decision maker continues to generate revenue from products that does not use the stocked-out resources. Either the admissible policy eliminates the possibility to sell stocked-out resources (Gallego and Van Ryzin 1997, Rusmevichientong et al. 2020), or the realized demand in each period is simply the minimum between the remaining inventory and the generated demand (Ma et al. 2021). We refer to this stopping criterion as the "generous" stopping criterion.

Following each trajectory of randomness, the ungenerous stopping criterion stops earlier than the generous stopping criterion, hence the regret is larger.

### Modeling Limited Switches

We model the switching budget as a hard constraint that cannot be violated, which is common in the literature. Cheung et al. (2017) considers a dynamic pricing model where the demand function is unknown but belongs to a known finite set, and a pricing policy makes limited number of price changes. Chen and Chao (2019) studies a multi-period stochastic joint inventory replenishment and pricing problem with unknown demand and limited price changes. Chen et al. (2020) considers the dynamic pricing and inventory control problem in the face of censored demand. Simchi-Levi and Xu (2023) considers the stochastic MAB problem with a general switching constraint. All the above papers adopt the same modeling approach, yet none of the above papers considers the existence of non-replenishable resource constraints.

There is another prevalent way of modeling, which models the business constraint as incurring switching costs; see Agrawal et al. (1988, 1990), Cesa-Bianchi et al. (2013), and Jun (2004) for a survey. Most papers from this stream of research penalize switching costs into the objective function of the reward calculation. That is, the objective is to minimize the sum of the regret incurred, plus the switching costs. Since we treat the switching budget as a hard constraint that can never be violated, we face a more challenging learning task.

## F.2 Bandits with Knapsacks Under Limited Switches

In this section, we study the BwK-LS problem, introduce an efficient algorithm, and provide matching upper and lower bounds of the optimal regret.

### Admissible Policies and Clairvoyant Policies

Recall that in Section 7.4, we distinguish between a BNRM *instance* $\mathcal{I} = (T, \boldsymbol{B}, K, d, n, P, A; \boldsymbol{Q})$ and a BNRM *problem* $\mathcal{P} = (T, \boldsymbol{B}, K, d, n, P, A)$ based on whether the underlying demand distributions $\boldsymbol{Q}$

are specified or not. In this section, we distinguish between a BwK *instance* $\mathcal{I} = (T, \boldsymbol{B}, K, d; \boldsymbol{R}, \boldsymbol{C})$ and a BwK *problem* $\mathcal{P} = (T, \boldsymbol{B}, K, d)$ based on whether the underlying reward and cost distributions $\boldsymbol{R}, \boldsymbol{C}$ are specified or not. Consider a BwK problem $\mathcal{P} = (T, \boldsymbol{B}, K, d)$. Let $\phi$ denote any non-anticipating learning policy; specifically, $\phi$ consists of a sequence of (possibly randomized) decision rules $(\phi^t)_{t\in[T]}$, where each $\phi^t$ establishes a probability kernel acting from the space of historical actions and observations in periods $1, \ldots, t-1$ to the space of actions at period $t$. For any $s \in \mathbb{N}$, let $\Phi[s]$ be the set of learning policies that change actions for no more than $s$ times almost surely under all possible distributions $\boldsymbol{R}, \boldsymbol{C}$. For any $s, s' \in \mathbb{N}$ such that $s \leq s'$, $\Phi[s] \subseteq \Phi[s']$. Let $\Phi[\infty]$ be the set of all learning policies with an infinite switching budget. Let $\mathsf{Rev}_{\boldsymbol{R},\boldsymbol{C}}(\phi)$ be the expected reward that a learning policy $\phi$ generates under distributions $\boldsymbol{R}, \boldsymbol{C}$.

As we have defined in Section 7.2, $\pi$ refers to a *clairvoyant* policy with full distributional information about the true distributions $\boldsymbol{R}, \boldsymbol{C}$. For any $s \in \mathbb{N}$, let $\Pi_{\boldsymbol{R},\boldsymbol{C}}[s]$ be the set of clairvoyant policies that change actions for no more than $s$ times under the true distributions $\boldsymbol{R}, \boldsymbol{C}$. For any $s, s' \in \mathbb{N}$ such that $s \leq s'$, $\Pi_{\boldsymbol{R},\boldsymbol{C}}[s] \subseteq \Pi_{\boldsymbol{R},\boldsymbol{C}}[s']$. Let $\Pi_{\boldsymbol{R},\boldsymbol{C}}[\infty]$ be the set of clairvoyant policies with an infinite switching budget. Let $\mathsf{Rev}_{\boldsymbol{R},\boldsymbol{C}}(\pi)$ be the expected revenue that a clairvoyant policy $\pi \in \Pi_{\boldsymbol{R},\boldsymbol{C}}$ generates under distributions $\boldsymbol{R}, \boldsymbol{C}$. Let $\pi^\star_{\boldsymbol{R},\boldsymbol{C}}[s] \in \arg\sup_{\pi\in\Pi_{\boldsymbol{R},\boldsymbol{C}}[s]} \mathsf{Rev}(\pi)$ be one optimal clairvoyant policy with switching budget $s$, and $\pi^\star_{\boldsymbol{R},\boldsymbol{C}}$ be one of the optimal dynamic policies with an infinite switching budget (i.e., without a switching constraint).

### Performance Metrics

The performance of an $s$-switch learning policy $\phi \in \Phi[s]$ is measured against the performance of the optimal $s$-switch clairvoyant policy $\pi^\star_{\boldsymbol{R},\boldsymbol{C}}[s]$. Specifically, for any BNRM problem $\mathcal{P}$ and switching budget $s$, we define the *$s$-switch regret* of a learning policy $\phi \in \Phi[s]$ as the worst-case difference between the expected revenue of the optimal $s$-switch clairvoyant policy $\pi^\star_{\boldsymbol{R},\boldsymbol{C}}[s]$ and the expected revenue of policy $\phi$:

$$R^\phi_s(T) = \sup_{\boldsymbol{R},\boldsymbol{C}} \left\{ \mathsf{Rev}_{\boldsymbol{R},\boldsymbol{C}}(\pi^\star_{\boldsymbol{R},\boldsymbol{C}}[s]) - \mathsf{Rev}_{\boldsymbol{R},\boldsymbol{C}}(\phi) \right\}.$$

We also measure the performance of policy $\phi$ against the performance of the optimal unlimited-switch clairvoyant policy $\pi^\star_{\boldsymbol{R},\boldsymbol{C}}$. Specifically, we define the *overall regret* of a learning policy $\phi \in \Phi[s]$ as the worst-case difference between the expected revenue of the optimal unlimited-switch clairvoyant policy $\pi^\star_{\boldsymbol{R},\boldsymbol{C}}$ and the expected revenue of the policy $\phi$:

$$R^\phi(T) = \sup_{\boldsymbol{R},\boldsymbol{C}} \left\{ \mathsf{Rev}_{\boldsymbol{R},\boldsymbol{C}}(\pi^\star_{\boldsymbol{R},\boldsymbol{C}}) - \mathsf{Rev}_{\boldsymbol{R},\boldsymbol{C}}(\phi) \right\}.$$

Intuitively, the $s$-switch regret $R^\phi_s(T)$ measures the "informational revenue loss" due to not knowing the underlying distributions, while the overall regret $R^\phi(T)$ measures the "overall revenue loss" due to not knowing the underlying distributions and not being able to switch freely. Clearly, the overall regret $R^\phi(T)$ is always larger than the $s$-switch regret $R^\phi_s(T)$. Interestingly, as we will show later, for all $s$, $R^\phi(T)$ and $R^\phi_s(T)$ are always in the same order in terms of the dependence on $T$.

### F.2.1 Upper Bound

In this subsection, we describe the Limited-Switch Learning via Two-Stage Linear Programming (LS-2SLP) algorithm as Algorithm F.1 in the BwK-LS setup. Note that, Algorithm 7.2 should not be directly viewed as a special case of Algorithm F.1, as it utilizes BNRM's feedback structure and can have much better empirical performance in the BNRM-LS setup. We analyze the performance of Algorithm F.1 as follows. The proof can be found in the full version of our paper (Simchi-Levi et al. 2019).

**Theorem F.1.** *Let $\phi$ be the LS-2SLP policy as suggested by Algorithm F.1. Let $\underline{b} > 0$ be an arbitrary constant. For any BwK problem $\mathcal{P} = (T, \boldsymbol{B}, K, d)$ with $T \geq 1, d \geq 0, K > d+1$ and $B_{\min}/T \geq \underline{b}$, $\phi$ is guaranteed to be a $s$-switch learning policy, and the regret incurred by $\phi$ satisfies*

$$R_s^{\phi}(T) \leq R^{\phi}(T) \leq \left(\max\{cn/\underline{b}, c'\} \cdot \sqrt{d \log[dKT]} K^{1-\frac{1}{2-2^{-\nu(s,d)}}} \log T\right) \cdot T^{\frac{1}{2-2^{-\nu(s,d)}}},$$

*where $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$, and $c, c' > 0$ are some absolute constants.*

### F.2.2 Lower Bound

As we have discussed in Section F.1.1, the lower bound construction in Theorem 7.4 for the BNRM-LS problem directly applies to the BwK-LS problem. So the lower bound result holds.

**Corollary F.1.** *Let $\underline{b} > 0$ be an arbitrary constant. For any $T \geq 1, d \geq 0, K \geq 2(d+1)$ and $\boldsymbol{B}$ such that $B_{\min}/T \geq \underline{b}$, for the BwK problem $\mathcal{P} = (T, \boldsymbol{B}, K, d)$, for any switching budget $s \geq 0$ and any $\phi \in \Phi[s]$,*

$$R^{\phi}(T) \geq R_s^{\phi}(T) \geq \left(\min\{c\underline{b}, c'\} \cdot (d+1)^{-3} K^{-\frac{3}{2}-\frac{1}{2-2^{-\nu(s,d)}}} (\log T)^{-\frac{5}{2}}\right) \cdot T^{\frac{1}{2-2^{-\nu(s,d)}}},$$

*where $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$, and $c, c' > 0$ are some absolute constants.*

**Algorithm F.1** Limited-Switch Learning via Two-Stage Linear Programming (LS-2SLP) for BwK-LS

**Input:** Problem parameters $(T, \boldsymbol{B}, K, d)$; switching budget $s$; discounting factor $\gamma$.

**Initialization:** Calculate $\nu(s,d) = \left\lfloor \frac{s-d-1}{K-1} \right\rfloor$. Define $t_0 = 0$ and

$$t_l = \left\lfloor K^{1-\frac{2-2^{-(l-1)}}{2-2^{-\nu(s,d)}}} T^{\frac{2-2^{-(l-1)}}{2-2^{-\nu(s,d)}}} \right\rfloor, \quad \forall l = 1, \ldots, \nu(s,d)+1.$$

Set $\gamma = 1 - 3\frac{C_{\mathsf{max}}\sqrt{d \log[dKT]} \log T}{B_{\mathsf{min}}} t_1$.

**Notation:** Let $T_l$ denote the ending period of epoch $l$ (which will be determined by the algorithm). Let $z_t$ denote the algorithm's action at period $t$. Let $z_0 \in [K]$ be a random action.

**Policy:**

1: **for** epoch $l = 1, \ldots, \nu(s,d)$ **do**

2: **if** $l = 1$ **then**

3: Set $T_0 = L_k^{\mathsf{rew}}(0) = L_{i,k}^{\mathsf{cost}}(0) = 0$ and $U_k^{\mathsf{rew}}(0) = U_{i,k}^{\mathsf{cost}}(0) = \infty$, $\forall i \in [d], \forall k \in [K]$.

4: **else**

5: Let $n_k(T_{l-1})$ be the total number of periods that action $k$ is chosen, up to period $T_{l-1}$. Calculate $\bar{c}_{i,k}(T_{l-1})$ to be the empirical average consumption of resource $i$ by selecting arm $k$, up to period $T_{l-1}$; Calculate $\bar{r}_k(T_{l-1})$ to be the empirical average reward by selecting arm $k$, up to period $T_{l-1}$. Calculate confidence radius $\nabla_k(T_{l-1}) = \sqrt{\frac{\log[(d+1)KT]}{n_k(T_{l-1})}}$ and

$$\begin{cases} U_k^{\mathsf{rew}}(T_{l-1}) = \min\left\{\bar{r}_{j,k}(T_{l-1}) + R_{\mathsf{max}}\nabla_k(T_{l-1}), U_k^{\mathsf{rew}}(T_{l-2})\right\}, \\ L_k^{\mathsf{rew}}(T_{l-1}) = \max\left\{\bar{r}_{j,k}(T_{l-1}) - R_{\mathsf{max}}\nabla_k(T_{l-1}), L_k^{\mathsf{rew}}(T_{l-2})\right\}, \end{cases} \quad \forall k \in [K],$$

$$\begin{cases} U_{i,k}^{\mathsf{cost}}(T_{l-1}) = \min\left\{\bar{c}_{i,k}(T_{l-1}) + C_{\mathsf{max}}\nabla_k(T_{l-1}), U_{i,k}^{\mathsf{cost}}(T_{l-2})\right\}, \\ L_{i,k}^{\mathsf{cost}}(T_{l-1}) = \max\left\{\bar{c}_{i,k}(T_{l-1}) - C_{\mathsf{max}}\nabla_k(T_{l-1}), L_{i,k}^{\mathsf{cost}}(T_{l-2})\right\}, \end{cases} \quad \forall i \in [d], \forall k \in [K].$$

6: Solve the first-stage pessimistic LP:

$$\begin{aligned} \mathsf{J}_l^{\mathsf{PES}} = \max_{(x_1,\ldots,x_K)} & \sum_{k\in[K]} L_k^{\mathsf{rew}}(T_{l-1}) x_k \\ \text{s.t.} & \sum_{k\in[K]} U_{i,k}^{\mathsf{cost}}(T_{l-1}) x_k \le B_i && \forall i \in [d] \\ & \sum_{k\in[K]} x_k \le T \\ & x_k \ge 0 && \forall k \in [K] \end{aligned}$$

7: For each $j \in [K]$, solve the second-stage exploration LP:

$$\begin{aligned} \boldsymbol{x}^{l,j} = \arg \max_{(x_1,\ldots,x_K)} \; & x_j \\ \text{s.t.} \sum_{k\in[K]} & U_k^{\mathsf{rew}}(T_{l-1}) x_k \geq \mathsf{J}_l^{\mathsf{PES}} \\ \sum_{k\in[K]} & L_{i,k}^{\mathsf{cost}}(T_{l-1}) x_k \leq B_i \qquad \forall i \in [d] \\ \sum_{k\in[K]} & x_k \leq T \\ & x_k \geq 0 \qquad \forall\, k \in [K] \end{aligned}$$

8: For all $k \in [K]$, let $N_k^l = \frac{(t_l - t_{l-1})}{T} \sum_{j=1}^K \frac{1}{K} (\boldsymbol{x}^{l,j})_k$. Let $z_{T_{l-1}+1} = z_{T_{l-1}}$. Starting from this arm, Select each arm $k$ for $\gamma N_k^l$ *consecutive* periods, $k \in [K]$ (we overlook the rounding issues here, which are easy to fix in regret analysis). Stop the algorithm once time horizon is met or one of the resources is exhausted.

9: End epoch $l$. Mark the last period in epoch $l$ as $T_l$.

10: For epoch $\nu(s,d)+1$ (the last epoch), calculate $\bar{c}_{i,k}(T_{\nu(s,d)})$ to be the empirical average consumption of resource $i$ by selecting arm $k$, up to period $T_{\nu(s,d)}$; Calculate $\bar{r}_k(T_{\nu(s,d)})$ to be the empirical average reward by selecting arm $k$, up to period $T_{\nu(s,d)}$. Solve the following deterministic LP

$$\begin{aligned} \max_{(x_1,\ldots,x_K)} \sum_{k\in[K]} & \bar{r}_{j,k}(T_{\nu(s,d)})\; x_k \\ \text{s.t.} \sum_{k\in[K]} & \bar{r}_{j,k}(T_{\nu(s,d)})\; x_k \leq B_i \qquad \forall\; i \in [d] \\ \sum_{k\in[K]} & x_k \leq T \\ & x_k \geq 0 \qquad \forall\; k \in [K], \end{aligned}$$

and find an optimal solution with the least number of non-zero variables, $\boldsymbol{x}_{\tilde{\boldsymbol{q}}}^\star$. Let $N_k^{\nu(s,d)+1} = \frac{(T - t_{\nu(s,d)})}{T} (\boldsymbol{x}_{\tilde{\boldsymbol{q}}}^\star)_k$ for all $k \in [K]$. First let $z_{T_{\nu(s,d)}+1} = z_{T_{\nu(s,d)}}$. Start from this arm, choose each arm $k$ for $\gamma N_k^{\nu(s,d)+1}$ *consecutive* periods, $k \in [K]$ (we overlook the rounding issues here, which are easy to fix in regret analysis). Stop the algorithm once time horizon is met or one of the resources is exhausted. End epoch $\nu(s,d)+1$.

# Bibliography


Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Naoki Abe and Philip Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, pages 3–11, 1999.

Naoki Abe, Alan Biermann, and Philip Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Daniel Adelman. Dynamic bid prices in revenue management. *Operations Research*, 55(4):647–661, 2007.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E Schapire. Contextual bandit learning with predictable rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

Alekh Agarwal, Sham M Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*, pages 20095–20107, 2020a.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114, 2020b.

Rajeev Agrawal, Manjunath V Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

Rajeev Agrawal, Manjunath V Hedge, and Demosthenis Teneketzis. Multi-armed bandit problems with multiple plays and switching cost. *Stochastics and Stochastic Reports*, 29(4):437–459, 1990.

Shipra Agrawal. Recent advances in multiarmed bandits for sequential decision making. In *Operations Research & Management Science in the Age of Analytics*, pages 167–188. INFORMS, 2019.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *ACM Conference on Economics and Computation*, pages 989–1006, 2014.

Shipra Agrawal and Nikhil R Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems*, pages 3458–3467, 2016.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. In *Conference on Learning Theory*, pages 76–78, 2017.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Hyun-Soo Ahn, Christopher Ryan, Joline Uichanco, and Mengzhenyu Zhang. On the performance of certainty-equivalent pricing. *Available at SSRN 3502478*, 2021.

Kenneth S Alexander. The central limit theorem for weighted empirical processes indexed by sets. *Journal of Multivariate Analysis*, 22(2):313–339, 1987.

Jason M Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. *Mathematics of Operations Research*, 46(1):179–203, 2021.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Alessandro Arlotto and Itai Gurvich. Uniformly bounded regret in the multisecretary problem. *Stochastic Systems*, 9(3):231–260, 2019.

Manjari Asawa and Demosthenis Teneketzis. Multi-armed bandits with switching penalties. *IEEE Transactions on Automatic Control*, 41(3):328–348, 1996.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633, 2007.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM*, 65(3):1–55, 2018.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2-3):111–139, 2010.

Santiago R Balseiro, David B Brown, and Chen Chen. Dynamic pricing of relocating resources in large networks. *Management Science*, 67(7):4075–4094, 2021.

Gah-Yi Ban and N Bora Keskin. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 67(9):5549–5568, 2021.

Jeffrey S Banks and Rangarajan K Sundaram. Switching costs and the gittins index. *Econometrica: Journal of the Econometric Society*, pages 687–694, 1994.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881, 2022.

Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *Annals of Applied Probability*, 25(2):753–822, 2015.

Dirk Bergemann and Juuso Välimäki. Stationary multi-choice bandit problems. *Journal of Economic dynamics and Control*, 25(10):1585–1594, 2001.

Dimitri P Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 1995.

Omar Besbes and Assaf Zeevi. Blind network revenue management. *Operations Research*, 60(6): 1537–1550, 2012.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Gabriel Bitran and René Caldentey. An overview of pricing models for revenue management. *Manufacturing & Service Operations Management*, 5(3):203–229, 2003.

Djallel Bouneffouf, Srinivasan Parthasarathy, Horst Samulowitz, and Martin Wistub. Optimal exploitation of clustering and history information in multi-armed bandit. *arXiv preprint arXiv:1906.03979*, 2019.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Revised Lectures*, pages 169–207, 2004.

Robert L Bray and Ioannis Stamatopoulos. Menu costs and the bullwhip effect: Supply chain implications of dynamic pricing. *Operations Research*, 70(2):748–765, 2022.

Monica Brezzi and Tze Leung Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87–108, 2002.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Jinzhi Bu, David Simchi-Levi, and Yunzong Xu. Online pricing with offline data: Phase transition and inverse square law. *Management Science*, 68(12):8568–8588, 2022.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

Pornpawee Bumpensanti and He Wang. A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Science*, 66(7):2993–3009, 2020.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, 2020.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, pages 1160–1168, 2013.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, 2014.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

Boxiao Chen and Xiuli Chao. Parametric demand learning with limited price explorations in a backlog stochastic inventory system. *IISE Transactions*, 51(6):605–613, 2019.

Boxiao Chen, Xiuli Chao, and Yining Wang. Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Operations Research*, 68(5):1445–1456, 2020.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.

Qi Chen, Stefanus Jasin, and Izak Duenyas. Real-time dynamic pricing with minimal and flexible price adjustment. *Management Science*, 62(8):2437–2455, 2015.

Qi Chen, Stefanus Jasin, and Izak 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, 2019.

Yiwei Chen and Cong Shi. Network revenue management with online inverse batch gradient descent method. *Available at SSRN*, 2019.

Wang Chi Cheung, David Simchi-Levi, and He Wang. Dynamic pricing and demand learning with limited price experimentation. *Operations Research*, 65(6):1722–1731, 2017.

Nicos Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. Technical report, Carnegie Mellon University Management Sciences Research Group, 1976.

J T Chu and J C Chueh. Inequalities between information measures and error probability. *Journal of the Franklin Institute*, 282(2):121–125, 1966.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

William L Cooper. Asymptotic behavior of an allocation policy for revenue management. *Operations Research*, 50(4):720–727, 2002.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 2009.

Jose Correa, Paul Dütting, Felix Fischer, and Kevin Schewior. Prophet inequalities for independent and identically distributed random variables from an unknown distribution. *Mathematics of Operations Research*, 47(2):1287–1309, 2022.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. In *Advances in neural information processing systems*, pages 1422–1432, 2018.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Annual ACM symposium on Theory of computing*, pages 459–467, 2014.

Arnoud V den Boer. Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of Operations Research*, 39(3):863–888, 2014.

Arnoud 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.

Arnoud V den Boer and N Bora Keskin. Dynamic pricing with demand learning and reference effects. *Management Science*, 68(10):7112–7130, 2022.

Arnoud V den Boer and Bert Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783, 2013.

Cyril Domb. *Phase Transitions and Critical Phenomena*, volume 19. Elsevier, 2000.

Kefan Dong, Yingkai Li, Qin Zhang, and Yuan Zhou. Multinomial logit bandit with low switching cost. In *International Conference on Machine Learning*, pages 2607–2615, 2020.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674, 2019a.

Simon S Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. *arXiv preprint arXiv:1901.09018*, 2019b.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020a.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic q-learning with function approximation in deterministic systems: near-optimal bounds on approximation error and sample complexity. In *Advances in Neural Information Processing Systems*, pages 22327–22337, 2020b.

Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, page 169–178, 2011.

Wedad Elmaghraby and Pınar Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, 49(10): 1287–1309, 2003.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Fei Feng, Ruosong Wang, Wotao Yin, Simon S Du, and Lin F Yang. Provably efficient exploration for RL with unsupervised learning. *arXiv preprint arXiv:2003.06898*, 2020.

Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations Research*, 66(6):1586–1602, 2018.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

Dylan J Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan J Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548, 2018.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Eric Friedman. Active learning for smooth problems. In *Conference on Learning Theory*, 2009.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2019.

Guillermo Gallego and Garrett Van Ryzin. A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research*, 45(1):24–41, 1997.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, pages 503–513, 2019.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Sebastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. Fano's inequality for random variables. *Statistical Science*, 35(2):178–201, 2020.

Richard D Gill and Boris Y Levit. Applications of the van trees inequality: A bayesian cramér-rao bound. *Bernoulli*, 1:59, 2001.

Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Annals of Probability*, 34(3):1143–1216, 2006.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings*, 1995.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li wei H. Lehman, Matthieu Komorowski, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 2019.

Sudipto Guha and Kamesh Munagala. Multi-armed bandits with metric switching costs. In *International Colloquium on Automata, Languages, and Programming*, pages 496–507. Springer, 2009.

Sudipto Guha and Kamesh Munagala. Approximation algorithms for bayesian multi-armed bandit problems. *arXiv preprint arXiv:1306.3525*, 2013.

Yonatan Gur and Ahmadreza Momeni. Adaptive sequential experiments with unknown information arrival processes. *Manufacturing & Service Operations Management*, 24(5):2666–2684, 2022.

MohammadTaghi Hajiaghayi, Robert Kleinberg, and Tuomas Sandholm. Automated online mechanism design and prophet inequalities. In *AAAI Conference on Artificial Intelligence*, pages 58–65, 2007.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *International Conference on Machine Learning*, pages 353–360, 2007.

Steve Hanneke. Rates of convergence in active learning. *Annals of Statistics*, 39(1):333–361, 2011.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. *arXiv preprint arXiv:1910.06996*, 2019.

J Michael Harrison, N Bora Keskin, and Assaf Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Annual ACM Symposium on Theory of Computing*, pages 128–141, 2016.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Chih-Wei Hsu, Branislav Kveton, Ofer Meshi, Mladenov Martin, and Csaba Szepesvári. Empirical bayes regret minimization. *arXiv preprint arXiv:1904.02664*, 2019.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010, 2020.

Nicole Immorlica, Karthik Abinav Sankararaman, Robert E Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 202–219, 2019.

Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media, 2012.

Stefanus Jasin. Reoptimization and self-adjusting price control for network revenue management. *Operations Research*, 62(5):1168–1178, 2014.

Su Jia, Andrew Li, and R Ravi. Markdown pricing under unknown demand. *Available at SSRN 3861379*, 2021.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems*, pages 2747–2758, 2020.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Steffen Jørgensen, Sihem Taboubi, and Georges Zaccour. Retail promotions with negative brand image effects: Is cooperation possible? *European Journal of Operational Research*, 150(2):395–405, 2003.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Online learning for changing environments using coin betting. *Electronic Journal of Statistics*, 11(2):5282–5310, 2017.

Tackseung Jun. A survey on the bandit problem with switching costs. *De Economist*, 152(4):513–541, 2004.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, 2018.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *International Conference on Robotics and Automation*, 2019.

Dara Kerr. Detest uber's surge pricing? some drivers don't like it either, Aug 2015. URL `https://www.cnet.com/tech/tech-industry/detest-ubers-surge-pricing-some-drivers-dont-like-it-either/`.

N Bora Keskin and Assaf Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.

N Bora Keskin and Assaf Zeevi. Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, 42(2):277–307, 2016.

Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Tomer Koren, Roi Livni, and Yishay Mansour. Bandits with movement costs and adaptive pricing. In *Conference on Learning Theory*, pages 1242–1268. PMLR, 2017.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924, 2017.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20:1–50, 2019.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*, pages 5805–5814, 2021.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 2019.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, pages 817–824, 2008.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, 2019.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(1):765–796, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670, 2020.

Eugene L Lawler. The traveling salesman problem: a guided tour of combinatorial optimization. *Wiley-Interscience Series in Discrete Mathematics*, 1985.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *International Journal of Robotics Research*, 2018.

Daniel Levy, Shantanu Dutta, Mark Bergen, and Robert Venable. Price adjustment at multiproduct retailers. *Managerial and Decision Economics*, 19(2):81–120, 1998.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *International World Wide Web Conference*, pages 661–670, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080, 2017.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174, 2019.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

Qian Liu and Garrett Van Ryzin. On the choice-based linear programming model for network revenue management. *Manufacturing & Service Operations Management*, 10(2):288–310, 2008.

Will Ma, David Simchi-Levi, and Jinglong Zhao. Dynamic pricing (and assortment) under a static calendar. *Management Science*, 67(4):2292–2313, 2021.

Yuhang Ma, Paat Rusmevichientong, Mika Sumida, and Huseyin Topaloglu. An approximation algorithm for network revenue management under nonstationary arrivals. *Operations Research*, 68 (3):834–855, 2020.

Constantinos Maglaras and Joern Meissner. Dynamic pricing strategies for multiproduct revenue management problems. *Manufacturing & Service Operations Management*, 8(2):136–148, 2006.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27 (6):1808–1829, 1999.

H Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Conference on Learning Theory*, 2009.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Sentao Miao and Xiuli Chao. Dynamic joint assortment and pricing optimization with demand learning. *Manufacturing & Service Operations Management*, 2020.

Sentao Miao and Yining Wang. Network revenue management with nonparametric demand learning:\sqrt{T}-regret and polynomial dimension dependency. *Available at SSRN 3948140*, 2021.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. *arXiv preprint arXiv:1911.05815*, 2019.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 2003.

Rémi Munos. Performance bounds in $\ell_p$-norm for approximate value iteration. *SIAM Journal on Control and Optimization*, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 2008.

Mila Nambiar, David Simchi-Levi, and He Wang. Dynamic learning and pricing with model misspecification. *Management Science*, 2019.

Serguei Netessine. Dynamic pricing of inventory/capacity with infrequent price changes. *European Journal of Operational Research*, 174(1):553–580, 2006.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 2021.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8874–8882, 2018.

Erik Ordentlich and Marcelo J Weinberger. A distribution dependent refinement of pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840, 2005.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 2008.

Georgia Perakis and Divya Singhvi. Dynamic pricing with unknown non-parametric demand and limited price changes. *Available at SSRN 3336949*, 2019.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Annals of Statistics*, 41(2):693–721, 2013.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *Annals of Statistics*, 44(2):660–681, 2016.

Robert Phillips. *Pricing and Revenue Optimization*. Stanford University Press, 2005.

Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. In *International Conference on Robotics and Automation*, 2016.

Yury Polyanskiy. Lecture 1 of Information Theoretic Methods in Statistics and Computer Science. `https://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf`, 2020. [Online; accessed Nov/6/2021].

Stephen Portnoy. Asymptotic behavior of $m$-estimators of $p$ regression parameters when $p^2/n$ is large. i. consistency. *Annals of Statistics*, 12(4):1298–1309, 1984.

Stephen Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, pages 356–366, 1988.

Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2011.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

John A Rice. *Mathematical Statistics and Data Analysis*. Cengage Learning, 2006.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. In *Conference on Learning Theory*, 2010.

Stéphane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *International Conference on Machine Learning*, 2012.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Paat Rusmevichientong, Mika Sumida, and Huseyin Topaloglu. Dynamic assortment optimization for reusable products with random usage durations. *Management Science*, 2020.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Advances in bandits with knapsacks. *arXiv preprint arXiv:2002.00253*, 2020.

Jonathan Scarlett and Volkan Cevher. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

Noam Scheiber. How uber uses psychological tricks to push its drivers' buttons, Apr 2017. URL `https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html`.

Rajat Sen, Alexander Rakhlin, Lexing Ying, Rahul Kidambi, Dean Foster, Daniel N Hill, and Inderjit S Dhillon. Top-k extreme contextual bandits with arm hierarchy. In *International Conference on Machine Learning*, pages 9422–9433, 2021.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054, 2012.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *arXiv preprint arXiv:2003.12699*, 2020.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

David Simchi-Levi and Yunzong Xu. Phase transitions in bandits with switching constraints. *Management Science*, forthcoming, 2023.

David Simchi-Levi, Phillip Kaminsky, and Edith Simchi-Levi. *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies*. Tata McGraw-Hill Education, 2008.

David Simchi-Levi, Yunzong Xu, and Jinglong Zhao. Blind network revenue management and bandits with knapsacks under limited switches. *Available at SSRN 3479477*, 2019.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Aleksandrs Slivkins and Jennifer Wortman Vaughan. Online decision making in crowdsourcing markets: Theoretical challenges. *ACM SIGecom Exchanges*, 12(2):4–23, 2014.

Ioannis Stamatopoulos, Achal Bassamboo, and Antonio Moreno. The effects of menu costs on retail performance: Evidence from adoption of the electronic shelf label technology. *Management Science*, 2020.

Mohammad Sadegh Talebi Mazraeh Shahi. *Minimizing Regret in Combinatorial Bandits and Reinforcement Learning*. PhD thesis, KTH Royal Institute of Technology, 2017.

Kalyan Talluri and Garrett Van Ryzin. An analysis of bid-price controls for network revenue management. *Management Science*, 44(11-part-1):1577–1593, 1998.

Kalyan Talluri and Garrett Van Ryzin. *The Theory and Practice of Revenue Management*, volume 68. Springer Science & Business Media, 2006.

Terence Tao. *An Introduction to Measure Theory*, volume 126. American Mathematical Society, 2011.

D L Tebbe and S J Dwyer. Uncertainty and the probability of error. *IEEE Transactions on Information Theory*, 14(3):516–518, 1968.

Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

Ambuj Tewari and Susan Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

Huseyin Topaloglu. Using lagrangian relaxation to compute capacity-dependent bid prices in network revenue management. *Operations Research*, 57(3):637–649, 2009.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 1996.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 1997.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, 2020.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv:2102.02981*, 2021.

Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 2018.

Nicolas Verzelen and Fanny Villers. Goodness-of-fit tests for high-dimensional Gaussian linear models. *Annals of Statistics*, 2010.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Ruosong Wang, Dean Foster, and Sham M Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2020a.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020b.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham M Kakade. Instabilities of offline RL with pre-trained neural representation. *International Conference on Machine Learning*, 2021a.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yuanhao Wang, Ruosong Wang, and Sham M Kakade. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. *arXiv:2103.12690*, 2021b.

Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, pages 4300–4354, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, 2021.

Zheng Wen and Benjamin Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, 2021.

Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020a.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020b.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Li Ye, Yishi Lin, Hong Xie, and John Lui. Combining offline causal inference and online bandit learning for data driven decisions. *arXiv preprint arXiv:2001.05699*, 2020.

Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. In *International Conference on Healthcare Informatics*, 2019.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, 2021.

Mark J Zbaracki, Mark Ritson, Daniel Levy, Shantanu Dutta, and Mark Bergen. Managerial and customer costs of price adjustment: direct evidence from industrial markets. *Review of Economics and Statistics*, 86(2):514–533, 2004.

Nikita Zhivotovskiy and Steve Hanneke. Localization of VC classes: Beyond local Rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 18–33, 2016.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502, 2020.