Environment Adaptive Regret Analysis in Bandit Problems(Dissertation_全文)

Tsuchiya, Taira

Environment Adaptive Regret Analysis in Bandit Problems
（バンディット問題における環境適応的リグレット解析）

Taira Tsuchiya
土屋 平

A Doctor Thesis
博士論文

ABSTRACT

We live in an uncertain world, intermittently making decisions based on the knowledge we have accumulated so far. In such a world, how to make decisions to mitigate regrets is probably one of the most pressing concerns. In today's information-oriented society, such expectations are not limited to people, but also to computers that are programmed to maximize gains. Given such universal *sequential decision-making problems under uncertainty*, what decision-making procedures would be desirable in order to minimize regret?

A *bandit problem* is one of the most fundamental models in the field of statistical machine learning for dealing with such a sequential decision-making problem under uncertainty. In this problem, the learner sequentially selects an action from a given set of actions under the assumption the learner incurs and observes the loss for the selected action, and the goal of the learner is to minimize the cumulative loss. Instead of the cumulative loss, it is typical to consider the regret, which is the difference between the cumulative loss incurred by the learner and that of an optimal action. The most basic model of the bandit problem, the multi-armed bandit problem, was invented in the 1930s with clinical trial applications in mind. This model has been studied separately in two very different streams: the *stochastic regime* in which losses are obtained according to a certain distribution, and the *adversarial regime* in which arbitrary bounded loss vectors are given. Later, around 2010, with the rapid development of Internet technology and its accompanying web advertising, the bandit problem and its algorithms have been extensively investigated, broadening their scope of applications.

In the course of this development, it became clear that the simple multi-armed bandit problem, which does not take into account the correlation between actions and the information available at each time, cannot effectively make decisions in highly complex real-world problems. To address this issue, a new framework, commonly called *structured bandits* that appropriately incorporates the structure of the problem under investigation, has been actively investigated.

Adaptivity of algorithms is one of the most fundamental keywords in statistical machine learning, playing an important role in improving the performance of bandit algorithms. While bandit algorithms have historically been designed for a certain worst-case scenario, recent adaptive algorithms are based on the idea that performance can be improved by making decisions adaptively to an underlying structure and loss sequences.

An important example of adaptability is *adaptation to loss distributions in the stochastic regime*. While most existing algorithms aim to attain a favorable regret upper bound assuming that the underlying distribution is the most unfavorable one to the learner, such distribution is rarely encountered in real-world problems. Hence, an adaptive action selection procedure with respect to the distribution behind losses is expected to yield better performance than an algorithm that aims to improve performance for the worst-case distribution.

A second important example of adaptivity is *data dependence in the adversarial regime*. In the adversarial regime, as well as in the stochastic regime, a large number of algorithms have been developed that perform well for a loss sequence that is very annoying for the learner. However, losses in real problems have good properties. For example, a loss vector has almost the same value as in the previous time, or losses for the optimal action are close to zero. Therefore, it can be expected that an algorithm that works adaptively to such quantities should perform well in practice.

A third important example of adaptivity is *best-of-both-worlds*. When employing a bandit algorithm in a real-world problem, it is often the case that the underlying environment is unknown. Moreover, it is known that algorithms for the stochastic regime suffer from a linear regret in the adversarial regime, and classical algorithms for the adversarial regime perform much worse in the stochastic regime than algorithms for the stochastic regime. Therefore, it is desirable to achieve optimality in both environments without knowing the underlying environment, and an algorithm with such a property is called a best-of-both-worlds algorithm.

Despite the importance of such adaptability in applying the bandit algorithm to real-world problems, it has not been sufficiently investigated in the structured bandit problem, which has expanded its field of application significantly in recent years. In this dissertation, to overcome this limitation, we aim to realize the three adaptive properties in several structured bandit problems. We deal with the following problems: online learning with full information, multi-armed bandits, combinatorial semi-bandits, and partial monitoring. The organization of this dissertation is as follows.

In Chapter 1, we review the history of the bandit problem and then provide a concise background and contributions of this dissertation.

In Chapter 2, we present the fundamentals of the bandit problem, its algorithms, and the detailed background of this dissertation. First, representative results for the stochastic and adversarial regimes are presented, and in particular, the theoretical framework of the algorithm that plays a major role in this dissertation is detailed. We then review the history of adaptivity in the bandit problem, and in particular, the best-of-both-worlds property. Finally, we introduce some important examples of structured bandits and their concrete applications.

Chapters 3 through 6 are the main results of this dissertation.

In Chapter 3, we consider partial monitoring in the stochastic regime. Partial monitoring is the setting in which the learner can observe only abstract feedback symbols instead of losses and includes many important problems as special cases, such as the multi-armed bandit problem, dynamic pricing, and efficient label prediction. To further the applicability of partial monitoring, we aim to construct an algorithm that is adaptive to distribution and empirically powerful. We design an algorithm based on Thompson sampling, known for its splendid empirical performance in some bandit problems, and show that the algorithm can achieve a logarithmic distribution-dependent regret bound.

In Chapter 4, we continue to focus on partial monitoring and investigate best-of-both-worlds adaptivity for it. Existing best-of-both-worlds algorithms have been constructed only for relatively simple bandit problems, and it has been left unclear whether best-of-both-worlds can be achieved in a setting where only abstract feedback can be observed. To obtain a best-of-both-world algorithm for partial monitoring, we base our study on the follow-the-regularized-leader framework, which was first established as an algorithmic framework for the adversarial regime and emerged as a representative approach to constructing best-of-both-worlds algorithms. To apply this framework to partial monitoring, we advance the theory of adaptive learning rate and that of determining the action selection probability by optimization, which is a technique for maintaining the stability of the follow-the-regularized-leader. This development allows us to realize best-of-both-worlds algorithms for partial monitoring.

In Chapter 5, we further develop the follow-the-regularized-leader, a powerful framework for achieving adaptivity. Looking at the history of the adaptive learning rate of follow-the-regularized-leader, one can see that it depends on only one of the terms that appear in a regret upper bound, either the stability term or the term related to the strength of regularization. However, the induced property by adapting to each term is different from each other. Based on this observation, we establish an adaptive learning rate framework that depends on each of these terms simultaneously. We demonstrate that this allows us to simultaneously achieve best-of-both-worlds and certain data-dependent bounds for multi-armed bandits and partial monitoring, whereas existing approaches can achieve only one of those adaptivities.

In Chapter 6, we construct an adaptive algorithm for the combinatorial semi-bandit problem in which the action set involves a combinatorial structure. Existing regret bounds of the best-of-both-worlds algorithms in the stochastic regime are called optimal, only focusing on the dependence on the mean of the loss distribution. However, the bounds have a gap with the true optimality that is determined not only by the mean but also by the higher-order information of the distribution. To fill this gap and improve the performance of the best-of-both-worlds algorithms in the stochastic regime by exploiting the higher-order information of an underlying distribution, we develop an adaptive learning rate in the follow-the-regularized-leader that is adaptive to variances of the underlying distribution. This allows us to derive a regret upper bound that depends on the action variance in a stochastic regime and several data-dependent regret upper bounds in the adversarial regime.

Finally, in Chapter 7, we summarize the contributions of this dissertation and discuss future research challenges in sequential decision-making problems.

論文要旨

人はこれまで得られた知識を基に，断続的に意思決定を繰り返しながら不確実な世界を生きている．どのように意思決定をすれば後悔を減らすことができるかは，大きな関心事の一つだろう．情報化が進んだ現代社会では，人だけでなく，利益を最大化するようプログラムされた計算機にもそのような期待が寄せられる．このように普遍的に存在する**不確実性のもとでの逐次的意思決定問題**において，後悔を最小限に抑えるにはどのような意思決定の手続きが望ましいだろうか．

バンディット問題は，統計的機械学習の分野において，不確実性のもとでの逐次的意思決定問題を扱う最も基礎的なモデルの 1 つである．この問題では，学習者が複数の行動の選択肢から 1 つを選択し，選択した行動についてのみ損失を観測するという枠組みのもとで，累積損失を最小化することを目指す．ここでは，累積損失の代わりに，学習者が被った累積損失と最適な行動の被る累積損失の差である**リグレット**を評価指標として考えることが一般的である．バンディット問題の最も基礎的なモデルである多腕バンディット問題は，1930 年代に治験への応用を念頭に考案された．このモデルは，損失がある分布に従って得られる**確率的な環境**と任意の有界な値が与えられる**敵対的な環境**の大きく異なる潮流の中で別々に研究がなされてきた．その後，2010 年ごろからインターネット技術の進展とそれに伴うウェブ広告の急速な発展により，バンディット問題とそのアルゴリズムはその活躍の場を大きく広げ盛んに研究されてきた．

この発展の過程で，行動の選択肢間の相関や，各時刻で得られる情報を利用することができない単純な多腕バンディット問題の枠組みでは，高度に複雑化した実問題で効果的に意思決定を行うことができないことが明らかになった．この問題を解決するために，一般に**構造化バンディット問題**と呼ばれる，対象とする問題の構造を適切に取り込んだ枠組みが盛んに研究されてきた．

アルゴリズムの**適応性**は統計的機械学習において最も重要なキーワードの 1 つである．バンディット問題においても，適応性はアルゴリズムの性能向上に重要な役割を果たす．伝統的なバンディットアルゴリズムはある種の最悪ケースを考慮した設計がなされてきた一方で，適応的アルゴリズムは背後の構造や損失系列に対して適応的に意思決定することで性能を改善できないだろうかという思想に基づく．

一つ目の適応性の重要な例が，確率的環境における**損失分布に対する適応性**である．既存の多くのアルゴリズムは，背後の分布が学習者にとって最も都合が悪いものであるとした上で良いリグレット上界を達成することを目標としていたが，実問題ではそのような性質の分布が登場することは稀である．そこで背後の分布に対して，適応的に行動を選択を行うことができれば，悪い分布に対しての性能改善を目指すアルゴリズムより良好な性能が得られると考えられる．

二つ目の適応性の重要な例が，敵対的環境における**データ依存性**である．敵対的環境に対しても確率的な環境と同様に学習者にとって非常に厄介な損失系列が与えられたもとでも良好に動作するアルゴリズムが構築されてきた．しかし実問題に登場する損失は，例えば前時刻とほとんど変わらない損失値を持っていたり，最適な行動の損失値は限りなくゼロに近いなど性質が良い場合が多く存在する．このような問題の性質の良さを定量的に表す量に対して適応的に動作するアルゴリズムは実用上も良い性能を達成すると考えられる．

三つ目の重要な適応性の例が，確率的環境と敵対的環境の両方で単一のアルゴリズムで

同時に最適性を達成することを目指す**両環境最適性**である．実問題においてバンディット問題を利用するとき，背後の環境が確率的であるか敵対的であるかは未知であることが多い．また，確率的環境のためのアルゴリズムは敵対的環境においては線形リグレットを被ることが多く，一方で伝統的な敵対的環境のためのアルゴリズムは確率的環境において非常に性能が悪い．そこで，背後の環境を知らずして両環境における最適性を達成することが望まれ，このような性質を持つアルゴリズムは両環境最適であると呼ばれる．

このような適応性は実問題におけるバンディットアルゴリズムの適用において非常に重要であるにもかかわらず，近年活躍の場を大きく広げている構造化バンディット問題においては十分研究がなされていない．そこで本博士論文では，この限界を克服しバンディット問題の利用可能性を向上させるべく，上で挙げた三つの適応性を複数の構造化バンディット問題において実現することを目指す．具体的には，我々は全情報が得られるオンライン学習，多腕バンディット問題，組合せ半バンディット問題，部分観測問題の四つの問題を扱う．以下に本博士論文の構成を示す．

第一章では，バンディット問題の歴史と本博士論文の背景を簡潔に振り返り，本博士論文の貢献を議論する．

第二章では，バンディット問題の基礎的な理論とそのアルゴリズム，および本博士論文の詳細な背景をまとめる．はじめに確率的環境と敵対的環境の代表的な結果を紹介し，特に本博士論文で主要な役割を果たすアルゴリズムの理論的な枠組みについて詳述する．そして，バンディット問題における適応性，特に両環境最適性の歴史を振り返り，最後に構造化バンディット問題の重要な例とその具体的な応用例を紹介する．

第三章から第六章の四つの章は，本博士論文の主結果である．

第三章では，確率的環境における部分観測問題を考える．部分観測問題は，損失が直接観測される代わりに，抽象的なフィードバック記号のみが観測される設定であり，多腕バンディット問題や動的価格設定，効率的ラベル予測など非常に多くの重要な問題を特別なケースとして含む．そこで，このような複雑な構造を持つ逐次的意思決定問題において，分布に対して適応的であり，経験的に強力に動作するアルゴリズムの構築を目指す．具体的には，基本的なバンディット問題においてその経験的な性能の良さで知られるトンプソンサンプリングアルゴリズムをこの設定に対して設計し，背後のフィードバックの分布に適応的な対数オーダーのリグレットを達成できることを明らかにする．

第四章では，引き続き部分観測問題を対象とし，両環境最適性に焦点を当てる．これまでの両環境最適なアルゴリズムは，比較的単純なバンディット問題のみに対して構築されており，抽象的なフィードバックのみしか観測できない設定においても両環境最適性を達成できるかは不明であった．そこで部分観測問題における両環境最適なアルゴリズムの構築をするために，follow-the-regularized-leader の枠組みを基礎とする．この枠組みは，もともと敵対的環境を念頭に発展し，近年における両環境最適アルゴリズムを実現するための代表的な枠組みである．これを部分観測問題に適用するために，follow-the-regularized-leader の安定性を維持する技術である最適化によって行動探索割合の決定理論および適応的学習率の技術を発展させる．ここで発展させた技術を活用し，部分観測問題において両環境最適性を達成できることを明らかにする．

第五章では，適応性獲得のための強力な枠組みである follow-the-regularized-leader をさらに発展させる．Follow-the-regularized-leader の適応的学習率の歴史を紐解くと，これらはいずれもリグレット上界に現れる安定性の項あるいは正則化の強さに関する項どちらか

一方のみに依存していることがわかり，これらの依存性によって実現可能な適応性の種類は異なる．この観察をもとに，それぞれの構成要素に同時に依存するような適応的学習率の枠組みを確立する．これによって，多腕バンディット問題と部分観測問題において，両環境最適性およびあるデータ依存の上界を同時に達成できることを示す．

第六章では，行動の選択肢が組合せ構造を持つ設定である組合せ半バンディット問題における適応的アルゴリズムの構築を行う．既存の両環境最適なアルゴリズムの確率的環境における上界は，損失分布の期待値のみを考慮した意味での最適性であり，分布の高次の情報から定まる真の最適性との乖離があった．そこで，この乖離を埋め，分布の高次の情報を利用することで両環境最適なアルゴリズムの性能を向上させるために，follow-the-regularized-leader において背後の分布の分散に適応的に動作する学習率を設計する．これにより，確率的環境においては行動の分散に依存したリグレット上界を，敵対的環境においては複数のデータ依存リグレット上界を導出する．

最後に，第七章では，本博士論文における貢献をまとめる．また，今後の逐次的意思決定問題，バンディット問題における研究課題について述べる．

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Professor Junya Honda for his unwavering guidance and support over the past four years. We began to work together in 2020 when I was a second-year master's student at the University of Tokyo. His guidance persisted even after our transfer to Kyoto University, and I am thankful for his enduring support in countless matters. I have been inspired by his amazing attitude toward working and his speed to instantly switch contexts between work and hobbies. His work ethic is truly admirable and has been a great source of inspiration to me. I was incredibly fortunate to have been under his guidance, and my Ph.D. program would not have been the same without him.

I would also like to express my sincere thanks to Professor Toshiyuki Tanaka and Professor Hisashi Kashima, who served on my dissertation committee. Their insightful feedback and constructive criticism played a crucial role in improving my dissertation.

My heartfelt thanks also go to Professor Hidetoshi Shimodaira, the principal investigator of the lab in Kyoto. His endless enthusiasm for research, ability to listen keenly and ask incisive questions on topics beyond his area of expertise, and youthful spirit of enjoying research are truly attractive.

My research life was incredibly enriched in the lab in Tokyo to which I belonged during the first half year of the doctoral course and the master's course. I would like to deeply thank Professor Masashi Sugiyama, who is the principal investigator of the lab in Tokyo. He provided me with a vast amount of invaluable guidance and perspective on research. More than that, he fostered an exceptional research environment with passionate and diverse students, and he afforded numerous opportunities for external research presentations and engagements. My sincere gratitude is extended to Professor Issei Sato, Professor Tetsunori Kobayashi, Professor Tetsuji Ogawa, and Professor Naohiro Tawara. They furnished me with the crucial foundational skills that have played a significant role in my doctoral research. I would like to deeply acknowledge them.

In the lab in Tokyo, I acquired a great deal of knowledge and met many friends I still keep in touch with. Nontawat Charoenphakdee spent a vast amount of time talking with me from the time I joined the lab. I have learned so much from him through research collaboration, and we spent numerous hours in having meals (and a few karaokes) in Tokyo and Kyoto, and enjoyed walks along the Kamo River. Yuko Kuroki always brought joy and motivation to work in the lab, and I admire her attitude toward her research. Conversations with Han Bao over meals or desks were always insightful, and I learned a vast amount from him. He later moved to Kyoto and became a precious friend, making my life there enjoyable. Sharing a hotel room with him in Bangalore in July 2023 and engaging in numerous conversations stands out as a particularly enjoyable memory from my doctoral life. Jongyeong Lee, a fellow bandit researcher and a good friend, often joined me for meals in Tokyo. Masahiro Fujisawa, a valued colleague in the Ph.D. program, shared many meals and discussions that inspired my doctoral life. Takuya Shimada, a friend since our undergraduate days, stayed in touch even after I began my doctoral course. His kindness in taking lectures and having a phone meeting with me and his constant endeavor to take on new challenges was truly inspirational. Hideaki Imamura and Takuo

Kaneko, with whom I started a seminar reading mathematics and statistics books in the summer of 2020, provided not only a lot of learning but also some of the most enjoyable times during my Ph.D. program. I would also like to express my gratitude to them for taking me out for meals and caring for me. Takuo also joined me in participating in Kaggle competitions under the guise of studying reinforcement learning, which was a lot of fun. I am indebted to numerous lab members who I cannot fully acknowledge here for their kindness and support.

Kyoto is one of the best places in the world to spend my life, especially in spring and fall. Despite most of my time there coinciding with the COVID-19 pandemic, I still had precious opportunities for interactions. I am grateful to Kenta Kihara, Masahiro Naito, Hiroaki Yamagiwa, Momose Oyama, Ryoma Hashimoto, and Yusuke Takase. In the final year of my doctoral program, when the research lab regained its original form, I had the opportunity to have many means with them at numerous restaurants near the university. These occasions provided an enjoyable moment, which greatly served as a source of motivation in advancing my research. Junta Fujinaga was an undergraduate when I was a doctoral first-year student, and we did bandit research together. His vitality was amazing, and it was good to have conversations with him. My gratitude also goes to Naomi Nakagami and many other university administrative staff. My thanks extend to all other lab members.

My doctoral research would not have been possible without the invaluable support of collaborators. I received plenty of advice and inspiration from Dr. Shinji Ito, who has an amazing passion for research and is very kind. Without him, my doctoral course would not have been the same. Engaging in research discussions while strolling through the vibrant streets of London and exploring the city on a bicycle with him in July 2022, stands out as some of the most thrilling moments throughout my doctoral studies. Also, when I stayed at NEC in 2023, I received much help from Kei Takemura, Dr. Tatsuya Matsuoka, and many other researchers. I would like to express my gratitude to them. I am also deeply indebted to Professor Junpei Komiyama, and I would be happy to work on joint research again.

I had significant exchanges with France during my doctoral program. Dorian Baudry and Charles Riou stayed in the lab in Kyoto for about two months in June 2022. We toured many Kyoto tourist spots together, which I had never visited, walking a distance greater than I ever had before, which was an exhilarating experience. This taught me the importance of stepping away from research to exercise on weekends, which in turn greatly benefited my research progress. I am thankful to Professor Odalric-Ambrym Maillard and Professor Emilie Kaufmann, who hosted me at INRIA in Lille, France, in 2023. The experience of living abroad and conducting research was invaluable, and I met many new friends.

My heartfelt thanks go out to the machine learning community and the people in it. The Information-based Induction Sciences (IBIS) conference, in particular, has always been a great source of inspiration and motivation, offering stimulating encounters with everyone I meet there.

I would also like to express my appreciation for the financial support I received. My research progress was greatly facilitated by the Japan Society for the Promotion of Science (JSPS) Fellowship Program, KAKENHI Grant Number JP21J21272, Japan, and the Japan Science and Technology Agency (JST) ACT-X Program, Grant Number JPMJAX210E, Japan, and UTokyo Toyota-Dwango AI Scholarship. They have been definitely instrumental in driving my research forward.

Last but not least, my deepest thanks go to my family for their financial and unending support, and encouragement. Without them, I may not have even had the opportunity to pursue my studies at the university.

Contents

Chapter 1

Introduction

We human beings live in an uncertain world, making decisions intermittently, consciously or unconsciously, based on prior existing knowledge. Decision-making is not always deductive. We often make decisions by inductively extracting knowledge that is approximately plausible from the limited information that we have received in the past. It is not only humans who make decisions. In today's highly information-oriented world, people program computers to make desirable decisions. In this way, decision-making problems under uncertainty are universal in the real world. But, can they be formulated mathematically in a form that is useful to people? Moreover, to what extent can we make decisions "adaptively" to empirically available information? This dissertation aims to give partial answers to these questions using the framework of the bandit problem, the most representative mathematical model of sequential decision-making problems under uncertainty.

1.1 Statistical Machine Learning and Sequential Decision Making under Uncertainty

What is the ideal way to make decisions based on the finite amount of experience available so far? *Statistical machine learning* is a research field that aims to realize this goal in mathematically grounded ways. There are three major frameworks in machine learning: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the problem of finding a function that takes an input as its argument and outputs a label as accurately as possible, given a finite number of inputs and labels obtained according to certain rules. In general, it is common to evaluate the performance of the function with respect to new inputs. For example, consider the problem of predicting whether a new input image is a dog or a cat, given images of dogs and cats and labels indicating which one an image is. This problem is called (binary) classification and is an example of supervised learning. The problem of dividing a given set of inputs into multiple sets is called clustering, and is an example of unsupervised learning. Reinforcement learning, on the other hand, has characteristics that differ significantly from these frameworks.

Reinforcement learning is a general framework of *sequential decision-making under uncertainty*. Since there are various settings in reinforcement learning, let us consider one fundamental formulation here. In this framework, there is an agent and an environment, and we are initially given the set of states $\mathcal{S}$ in which the agent goes through transitions and the set of actions $\mathcal{A}$ ($|\mathcal{A}| < \infty$) that the agent can take. For each episode $t \in \{1, \dots, T\}$, the environment has a function $r_t \colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ that determines the *reward* according to the agent's state and action pairs, and a transition function $P_t \colon \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ that gives the next state. In each episode, the agent sequentially determines what action to take in each state based on policy $\pi \colon \mathcal{S} \to \mathcal{P}_{|\mathcal{A}|}$, receives a next state and reward from the environment, and then moves to the next state. At the end of each episode, using the experience of the

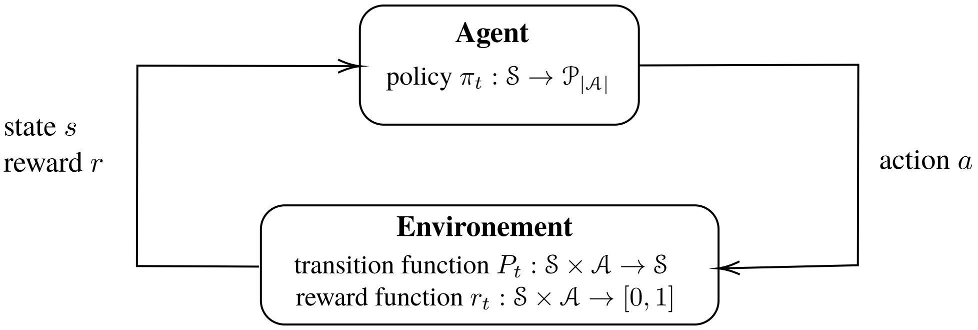

Agent
policy $\pi_t : \mathcal{S} \to \mathcal{P}_{|\mathcal{A}|}$
state s
reward r
action a
Environement
transition function $P_t : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$
reward function $r_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$

Figure 1.1: Overview of reinforcement learning.

previous episodes $\{1, \dots, \tau - 1\}$, policy $\pi \colon \mathcal{S} \to \mathcal{P}_{|\mathcal{A}|}$, the function that determines what action to take in each state, is updated. Here, $\mathcal{P}_k$ is the $(k-1)$-dimensional probability simplex. The overview of reinforcement learning is given in Figure 1.1. For example, consider the problem where there are 3×3 integer lattice points each of which has an unknown reward, and the agent moves from the start (lower left) to the goal (upper right), choosing whether to move up or right at each lattice point, so as to obtain the reward as much as possible. Then, one can see that this problem can be described as a reinforcement learning problem by appropriately defining the functions given above. To summarize, the framework of reinforcement learning deals with sequentially obtained data and is thus considerably different from the framework of supervised learning and unsupervised learning, in which we make inferences based on a finite amount of data given in advance.

This reinforcement learning framework is developing at an unprecedented rate, especially in its applications. In Othello, Go, and the computer game Atari, which have been frequently used for assessing performance of reinforcement learning algorithms since the early days of the field, it is now possible to achieve performance that far surpasses even that of human professionals. Other applications include robotics (Degrave et al., 2022; Ibarz et al., 2021), machine translation (Wu et al., 2018), and even magnetic control of tokamak plasmas (Degrave et al., 2022). Furthermore, at the time of writing this dissertation (April 6th, 2023), ChatGPT, a large-scale language model using reinforcement learning from human feedback, demonstrated capabilities that could replace search engines that have been the most basic means of obtaining information. As such, the reinforcement learning framework as a sequential decision-making model has been developed as a modern scientific and significantly powerful technology.

Bandit problems belong to the framework of reinforcement learning. In fact, the problem corresponds to a reinforcement learning problem where there is only one state, $|\mathcal{S}| = 1$, and the number of action choices within each episode is one. Despite its seemingly simple setup, the bandit problem can model a tremendous number of important real-world problems. Because of its simplicity, its theory and algorithms are supported by a very strong mathematical foundation, and theoretical advances in the bandit problem are often strongly associated with its numerical performance. Informally, the most basic model of the bandit problem, the *multi-armed bandit* problem, is a problem in which k slot machines are presented and played for a total amount of T times to maximize the cumulative reward. Formally, k arms (or actions) are given at the beginning, and the *learner* (corresponding to the agent in reinforcement learning) selects one arm $A_t \in [k]$ from these arms. At the same time, the underlying *environment* (or adversary) determines losses (negative reward) $\ell_t = (\ell_{t1}, \dots, \ell_{tk})^\top \in [0, 1]^k$. The learner then suffers a loss ℓ_{tA_t} of the chosen arm and can observe only that loss. One of the most important characteristics of the bandit problem is that the learner can only observe losses of the selected arm, and thus there is an *exploration and exploitation tradeoff*: either to select an arm that currently appears optimal or to select an arm that has the potential to become truly optimal. The goal of the learner is to minimize the cumulative loss. Equivalently, the learner aims to minimize a *regret*, which is a shifted metric of the cumulative loss, namely, the expectation of a difference between the cumulative loss of the selected arms and the cumulative loss of the optimal arm. If the regret of an algorithm is $o(T)$, then the loss difference from the optimal choice suffered per round approaches 0 as $T \to \infty$, meaning that the algorithm is "learning".

The origin of the bandit problem dates back to the 1930s when it was invented with its use in clinical trials in mind (Thompson, 1933). Here, the experimental conditions, such as drugs and dosing regimen, correspond to the arms, the number of patients corresponds to the total round of samples, and the result corresponds to the loss for each condition. Since then, the bandit problem has been developed in two different streams: a *stochastic regime* (stochastic environment) and an *adversarial regime* (adversarial environment). The stochastic regime is a setting in which losses (ℓ_t) are independently and identically distributed (i.i.d.) by a certain distribution decided by the environment, whereas the adversarial regime is a setting where (ℓ_t) can be arbitrarily determined. The achievable regrets in these regimes are rather different. For multi-armed bandits with loss vectors ℓ_t in $[0, 1]^k$, it is possible to achieve $O(\log T)$ and $O(\sqrt{T})$ regrets for the stochastic and adversarial regimes, respectively, both of which are known to be optimal. While a number of algorithms have been developed that can achieve these optimal bounds, their underlying ideas for designing algorithms are rather different (Lai and Robbins, 1985; Auer et al., 2002b,a), which we will see in the next chapter. An optimal algorithm for the stochastic regime has the advantage of achieving a very small regret of $O(\log T)$ if the underlying environment can be regarded as stochastic. In contrast, an optimal algorithm for the adversarial regime can achieve a certain degree of regret even when the underlying environment is highly unpredictable or can abruptly change. Thereafter, the bandit problem has been extensively studied since around 2010 due to the rapid advances in Internet technology and the accompanying rapid development of web advertising.

Over the course of its development, it has become clear that the framework of the vanilla multi-armed bandits introduced so far is not sufficient to apply it to real-world problems. For example, consider the problem where users visit a website at regular intervals and we recommend products that they would like to purchase. If we apply the vanilla multi-armed bandits framework to this problem, each product is considered as an independent arm, and the product with the highest estimated probability of purchase is recommended. However, this model totally ignores the features of users and products in its formulation and does not fully exploit the structure of the problem. Other important realistic problems include the problem setup where the losses cannot be directly observed, such as the spam/ham decision problem for e-mails, and the problem setup in which the set of actions has a combinatorial structure, such as the problem of transporting data over a network graph, both of which cannot be handled by the vanilla multi-armed bandits. To address this issue, a framework has been investigated that captures the highly complex real-world problems in a structured manner: *structured bandits*. The Structured bandit is a class of problem that incorporates a specific structure into the vanilla multi-armed bandit problem, it includes partial monitoring, combinatorial semi-bandits, and linear bandits, which will be discussed in the following, as an instance. In this dissertation, we focus on this concept that has become very important in recent years. A further detailed background mentioned here is provided in Chapter 2.

1.2 Major Challenges in Bandit Problems

Adaptivity is an extremely important keyword in the field of statistical machine learning. Efforts to adaptively exploit the underlying structure of a given problem to improve task performance can be found in a wide range of setups. These include Lasso (Tibshirani, 1996) that exploits the sparsity of features in linear regression to improve prediction performance and interpretability, and matrix completion (Candes and Plan, 2010) that exploits the low-rank property of matrices to complement missing values.

Such adaptivity also plays an important role in the bandit problem to improve the performance of algorithms. Adaptivity in the bandit problem is based on the principle that if the sequence of losses has a certain benign structure, the algorithm should be adaptive to it and achieve a small regret, which will be illustrated below. However, there has been little amount of studies aiming to construct sufficiently adaptive algorithms for structured bandits, which have been rapidly advancing in recent years. In this dissertation, we consider how far we can go in pursuit of adaptivity mainly for structured bandits.

Environment Adaptivity in Bandits What kind of adaptivity can be considered in the bandit problem? As mentioned above, it is known that in the bandit problem, given a distribution of losses behind the stochastic regime, it is possible to achieve a *distribution-dependent bound* of $O(\log T)$. This type of bounds has been well studied in relatively simple settings such as multi-armed bandits (Lai and Robbins, 1985; Burnetas and Katehakis, 1996). However, such a distribution-dependent bound has not been sufficiently studied for structured bandits with complex structures, while a minimax optimal bound of $O(\sqrt{T})$, which is optimal for the worst-case distribution, has been extensively investigated (Auer, 2002; Abbasi-Yadkori et al., 2011; Chu et al., 2011; Bartók et al., 2011). Such distribution-dependent bounds that fully exploit the information of the underlying distribution can be regarded as a certain kind of algorithmic adaptivity.

Another example of adaptivity is a *data-dependent bound* in the adversarial regime. As mentioned above, the regret achievable in the adversarial regime is $O(\sqrt{T})$. However, this bound is a very pessimistic result because one needs to consider all possibilities of losses in the adversarial regime, and we rarely encounter such a worst-case scenario in real-world problems. In a practical scenario, a loss vector ℓ_t at each time may often be similar to that at the previous time, *i.e.,* $\ell_{ta} \simeq \ell_{t+1,a}$ for all $a \in [k]$ (Chiang et al., 2013; Wei and Luo, 2018), or there could be a setting where the optimal arm a^* suffers a loss 0 most of the time, *i.e.,* $\ell_{ta^*} = 0$ for almost all $t \in [T]$ (Allenberg et al., 2006). Such settings are clearly benign compared to the most pessimistic cases. In fact, it is known that we can achieve regret upper bounds that depend on such quantities $\sum_{t=2}^{T} \|\ell_t - \ell_{t-1}\|$ or $\min_{a \in [k]} \sum_{t=1}^{T} \ell_{ta}$, and regret upper bounds that depend on such quantities are called data-dependent bounds (Allenberg et al., 2006; Hazan and Kale, 2011; Wei and Luo, 2018; Bubeck et al., 2018).

Another very important example of adaptivity we investigate is the simultaneous optimality in stochastic and adversarial regimes. As mentioned above, bandit games are classified into stochastic and adversarial regimes dependent on how losses are generated, and in the case of multi-armed bandits, $O(\log T)$ for the stochastic regime and $O(\sqrt{T})$ for the adversarial regime are the best achievable bounds. However, in many real-world problems, we often have no prior knowledge on whether the underlying environment is stochastic or adversarial. Existing algorithms specialized for the stochastic regime suffer a linear regret in the adversarial regime, while the classical algorithms intended for the adversarial regime cannot achieve good performance in the stochastic regime. Hence, the following question arises. *Can we establish a single algorithm that is optimal in both stochastic and adversarial regimes?* Algorithms with such adaptivity are called *best-of-both-worlds* algorithms and have been actively studied in recent years (Bubeck and Slivkins 2012; Seldin and Slivkins 2014; Auer and Chiang 2016; Seldin and Lugosi 2017; Wei and Luo 2018; Zimmert and Seldin 2019, to name a few). What about the intermediate environments between these two? Many of the real-world problems belong to environments that are neither completely stochastic nor adversarial. Therefore, it is desirable to construct an algorithm that achieves good performance in a setting in which losses are essentially stochastic and some adversarial noise is added to the observed losses or feedback. Such an environment is called *stochastic regime with adversarial corruptions* (Lykouris et al., 2018). Despite the practical importance of developing best-of-both-worlds algorithms with a favorable guarantee in the stochastic regime with adversarial corruptions, most existing best-of-both-worlds algorithms have only been considered in relatively simple settings such as the vanilla multi-armed bandits and have not been very well studied in structured bandits. The precise definitions of the various concepts that have appeared so far are given in Chapter 2.

1.3 Contributions of this Dissertation

This dissertation primarily explores adaptivity in structured bandits. The major research results of this dissertation are as follows:

> In several structured bandit problems, we can establish bandit algorithms with further adaptive guarantees by exploiting properties behind the problem and an underlying environment through environment-adaptive regret analysis.

The contributions of each chapter are briefly presented in the following.

1.3.1 Chapter 3: Thompson Sampling for Partial Monitoring

Motivation We begin with partial monitoring, a very generic instance of structured bandits dealing with limited feedback (Rustichini, 1999; Piccolboni and Schindelhauer, 2001). Partial monitoring is attracting broad interest because it includes a wide range of problems such as the multi-armed bandit problem (Lai and Robbins, 1985; Auer et al., 2002b,a), a linear optimization problem with full or bandit feedback (Zinkevich, 2003; Dani et al., 2008), dynamic pricing (Kleinberg and Leighton, 2003), and label efficient prediction (Cesa-Bianchi et al., 2005).

Several studies have investigated partial monitoring in the stochasitc regime (Bartók et al., 2012; Vanchinathan et al., 2014; Komiyama et al., 2015a). However, all of these algorithms only have a guarantee with respect to a worst-case feedback distribution or need to solve the optimization problem every round to deal with limited feedback, both of which result in poor performance in a realistic number of rounds. Therefore, we aim to design and analyze Thompson sampling, which is generally considered to have the best performance in the stochastic regime, for partial monitoring, with an $O(\log T)$ distribution-dependent bound.

Contribution Using the accept-reject sampling, we propose a new Thompson-sampling-based algorithm for PM, which is equipped with a numerical scheme to obtain a posterior sample from the complicated posterior distribution. We derive an $O(\log T)$ distribution-dependent regret bound for the proposed algorithm on locally observable games for a linearized variant of the problem. This is the first regret bound for Thompson sampling on locally observable games. Finally, we compare the performance of the proposed algorithm with existing algorithms in numerical experiments and show that the proposed algorithm outperforms existing algorithms.

1.3.2 Chapter 4: Best of Both Worlds Algorithms for Partial Monitoring

Motivation While Thomson sampling is a very strong algorithm in the stochastic regime, it is known that the algorithm suffers linear regret even in a slightly stochastic regime. Recall that, as discussed in Section 1.2, it is not often the case that feedback is generated in a completely stochastic manner. Hence, it is desirable to build an algorithm that performs well not only in the stochastic regime but also in the adversarial regime and in the stochastic regime with adversarial corruptions. Can we construct such a best-of-both-worlds algorithm for partial monitoring, which is a very complex structured bandit?

Contribution We resolve this question affirmatively by establishing new best-of-both-worlds algorithms for partial monitoring. Our algorithm is based on the follow-the-regularized-leader framework, which was originally developed in the context of online optimization and recently adopted to achieve a best-of-both-world guarantee in the vanilla multi-armed bandits. We rely on two recent theoretical advances: a framework of exploration by optimization, a method for enhancing the stability of algorithms, which is important in partial monitoring, and a method for adjusting the learning rate in follow-the-regularized-leader when dealing with indirect feedback. We show that for easy partial monitoring games, the regret is $O((\log T)^2)$ in the stochastic regime and $O(\sqrt{T \log T})$ in the adversarial regime. Moreover, we show that for hard partial monitoring games, the regret is $O((\log T)^2)$ in the stochastic regime and $O((T \log T)^{2/3})$ in the adversarial regime, both of which are nearly-optimal in their class of games. We also provide regret bounds for the stochastic regime with adversarial corruptions.

1.3.3 Chapter 5: Stability-penalty-adaptive Follow-the-regularized-leader: Sparsity, Game dependency, Best of both worlds

Motivation As we see in Chapters 4 and 6, follow-the-regularized-leader is a powerful tool for exploiting loss adaptivity, which is made possible by properly designing the learning rate, which is a component of the regularizer in follow-the-regularized-leader, based on the observations observed so far. Still, looking at the history of adaptive learning rates, we notice that they are designed to depend on only one of the two components that appear in the regret upper bound of follow-the-regularized-leader. Then, what if we could construct the learning rate so that it depends on these components simultaneously? Would it enhance the adaptivity of follow-the-regularized-leader?

Contribution We answer this question affirmatively by developing a generic adaptive learning rate that jointly depends on the two components. This result yields algorithms that have best-of-both-worlds guarantees and data-dependent bounds simultaneously. In particular, leveraging the new adaptive learning rate framework, we establish the first best-of-both-worlds algorithm with a sparsity-dependent bound that becomes small when the underlying losses are sparse. Additionally, we explore partial monitoring and demonstrate that the proposed learning rate framework allows us to achieve both a best-of-both-worlds guarantee and a game-dependent bound that becomes small when the essential difficulty of the underlying problem is easier than the worst-case game, which is often the case in the practical scenario.

1.3.4 Chapter 6: Further Adaptive Algorithms for Combinatorial Semi-bandits

Motivation In this chapter, we consider combinatorial semi-bandits as one example of structured bandits. Combinatorial semi-bandits include practically important problems such as the online shortest path problem and online advertisement placement, and thus it is important to develop adaptive algorithms for this structured bandit problem. Suppose that in the ads allocation problem and the online shortest path problem, the losses are generated in a stochastic manner. In the ads allocation problem, the user click rate is very small, and in the shortest path problem, there is basically no significant change in the time required to move from a start to a goal. Hence, variances of the distributions of arms in each problem are considered to be very small. Can we exploit the smallness of variances by constructing an algorithm with a variance-dependent bound? Meanwhile, in the adversarial regime, we also consider what kind of data-dependent bound can be achieved while guaranteeing the best-of-both-worlds performance.

Contribution We show that all of these can be accomplished simultaneously with a single algorithm. In particular, we establish best-of-both-worlds algorithms based on follow-the-regularized-leader as done in Chapters 4 and 5, establishing a new regularizer and its learning rate leading to the desired properties. In the stochastic regime, we prove a variance-dependent regret bound dependent with a tight suboptimality gap. Additionally, in the adversarial regime, we show that the same algorithm simultaneously achieves various data-dependent regret bounds. We also numerically test the proposed algorithm and confirm its superior or competitive performance over existing algorithms, including Thompson sampling under most settings.

1.4 Organization of this Dissertation

The organization of this dissertation is summarized as follows. In Chapter 2, we introduce the fundamentals of the bandit problem and its algorithm and then provide a detailed background of the dissertation. We also detail the follow-the-regularized-leader and structured bandits, which are central subjects of this dissertation. Chapters 3 and 4 focus on partial monitoring, a very complex sequential decision-making problem. In particular, in Chapter 3, we construct a numerically high-performance algorithm that achieves a data-dependent regret bound in a stochastic regime based on Thompson sampling. In Chapter 4, we construct best-of-both-worlds algorithms for partial monitoring based on the follow-the-regularized-leader framework. In Chapter 5, we establish a generic adaptive learning rate framework for follow-the-regularized-leader, which enables us to achieve simultaneous adaptivity of best-of-both-worlds and data-dependent bounds in multi-armed bandits and partial monitoring. In Chapter 6, we target combinatorial semi-bandits and construct an algorithm that simultaneously achieves best-of-both-worlds, several important data-dependent bounds, and bounds adaptive to the arm variance in a stochastic regime. Chapters 3 through 6, which provide the major contributions of this dissertation, provide sufficient information to be read on their own. Finally, Chapter 7 concludes this dissertation. The organization of this dissertation can be found in Figure 1.2.

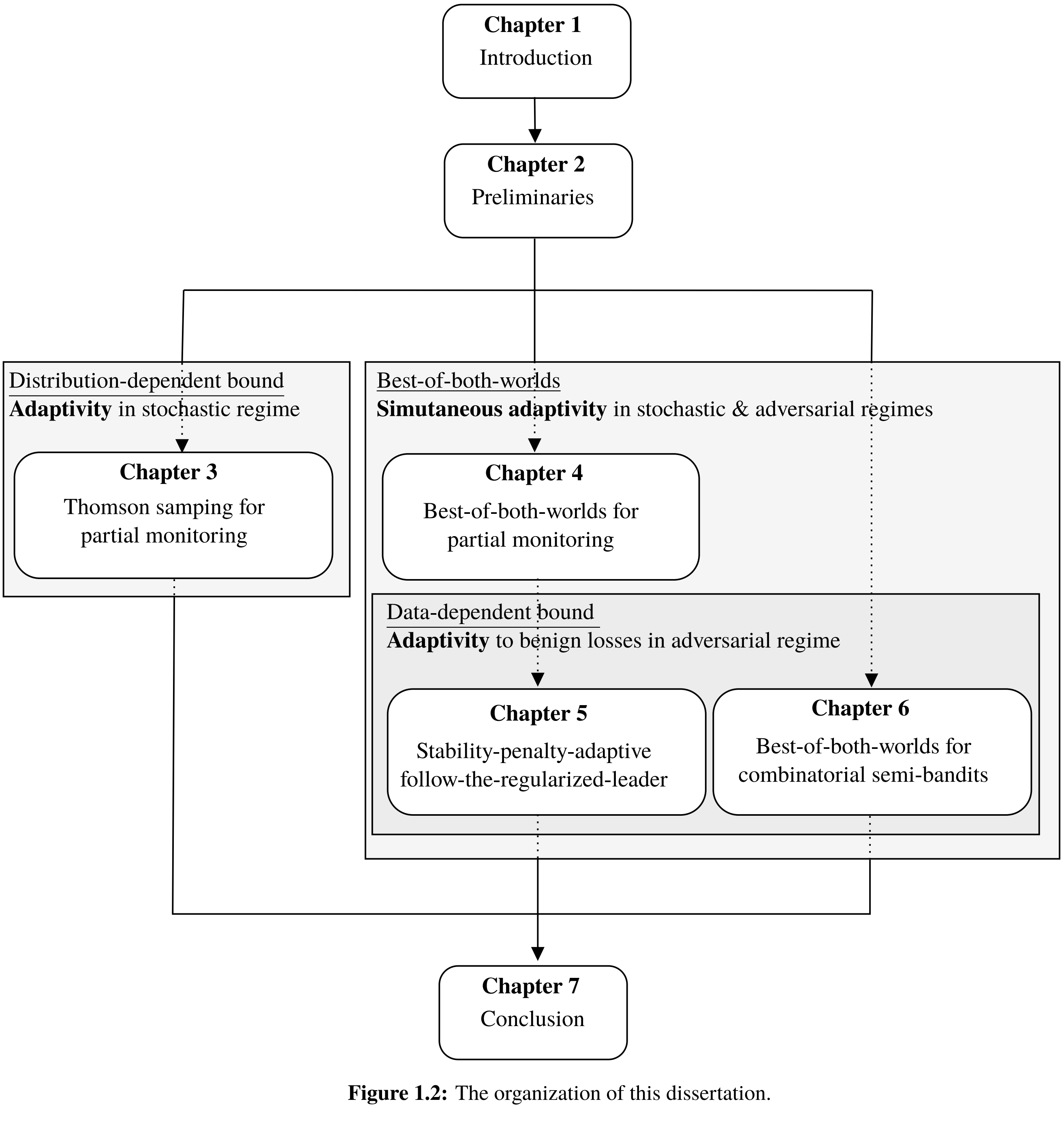

Chapter 1
Introduction
Chapter 2
Preliminaries
Distribution-dependent bound
Adaptivity in stochastic regime
Chapter 3
Thomson samping for partial monitoring
Best-of-both-worlds
Simutaneous adaptivity in stochastic & adversarial regimes
Chapter 4
Best-of-both-worlds for partial monitoring
Data-dependent bound
Adaptivity to benign losses in adversarial regime
Chapter 5
Stability-penalty-adaptive follow-the-regularized-leader
Chapter 6
Best-of-both-worlds for combinatorial semi-bandits
Chapter 7
Conclusion

Figure 1.2: The organization of this dissertation.

Chapter 2

Preliminaries: Foundations of Bandit Problems and its Algorithms

In this chapter, we begin with formulating the most fundamental model in the bandit problem, the multi-armed bandits. In the literature, the bandit problem has been studied in two regimes: the stochastic regime and the adversarial regime. First, we introduce the stochastic regime, the achievable lower bounds, and representative algorithms. Next, we present the definition of the adversarial regime, its achievable lower bounds, and its seminal algorithms. Furthermore, we review the definition and background of the best-of-both-worlds property, which has been developed rapidly in recent years. Finally, a formulation of structured bandits, which is a model incorporating a structure into multi-armed bandits, is presented, as well as specific applications of each formulation.

2.1 Notation

Let $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{N}$, and $\mathbb{Z}$ be the set of all real numbers, the set of all non-negative real numbers, the set of all natural numbers, and the set of all integers, respectively. For $n \in \mathbb{N}$, define $[n] = \{1, \dots, n\}$. For an event A, we define the indicator function $\mathbb{1}[A]$ to take the value 1 if A occurs and 0 otherwise. Let $\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ be the Euclidian, ℓ_1-, and ℓ_∞-norms for a vector x respectively, and let $\|x\|_A = \sqrt{x^\top A x}$ for a positive semidefinite matrix $A \succeq 0$. Let $\|A\|_\infty = \max_{i,j} |A_{ij}|$ be the maximum norm for a matrix A. Let $\mathcal{P}_k = \{p \in [0,1]^k : \|p\|_1 = 1\}$ be the $(k-1)$-dimensional probability simplex. A vector $e_a \in \{0,1\}^k$ is the a-th orthonormal basis of $\mathbb{R}^k$. $\mathbf{1}$ is the all-one vector. Let $H(p) = \sum_{a=1}^k p_a \log(1/p_a)$ be the Shannon entropy for a probability vector $p \in \mathcal{P}_k$. Let $D(p\|q) = \sum_{a=1}^k p_a \log(p_a/q_a)$ be the Kullback-Leibler divergence of p from q. For a convex function $\psi \colon \mathbb{R}^k \to \mathbb{R}$, let $D_\psi \colon \mathbb{R}^k \times \mathrm{dom}(\psi) \to \mathbb{R}_+$ be the Bregman divergence induced by ψ, *i.e.*, $D_\psi(p,q) = \psi(p) - \psi(q) - \langle \nabla\psi(q), p - q \rangle$.

2.2 Formulations of Multi-armed Bandits

This section formulates multi-armed bandits, which are one of the fundamental models for all bandit models. Informally, the multi-armed bandit problem is a problem in which k slot machines, called an arm or action, are presented and played for a total amount of T times to minimize the cumulative loss (or equivalently maximize the cumulative reward.) Formally, let $T \in \mathbb{N}$ be horizon that is the number of time the learner can select arm. In multi-armed bandits with k-arms, at each round $t \in [T]$, the environment determines the loss vector $\ell_t = (\ell_{t1}, \ell_{t2}, \dots, \ell_{tk})^\top \in [0,1]^k$.[1] The learner then chooses an arm $A_t \in [k]$

[1] In stochastic bandit problems, which will be introduced later, we sometimes consider the whole real number $\mathbb{R}^k$ as a domain of the loss, but this dissertation only deals with bounded losses.

Algorithm 2.1: The procedure of multi-armed bandits

```
input: horizon $T$, number of arms $k$
for $t = 1, 2, \ldots, T$ do
    The environment determines the loss vector
      $\ell_t = (\ell_{t1}, \ell_{t2}, \ldots, \ell_{tk})^\top \in [0, 1]^k$.
    The learner chooses an arm $A_t \in [k] := \{1, \ldots, k\}$ without knowing $\ell_t$.
    The learner incurs and observes the loss $\ell_{tA_t}$ for the chosen arm.
```

without knowing ℓ_t. After that, the learner observes only the loss ℓ_{tA_t} for the chosen arm. This procedure of the multi-armed bandit problem is given in Algorithm 2.1.

The performance of the learner (or algorithm) π is evaluated by the *regret*[2] Reg_T, which is the difference between the cumulative loss of the learner and the single optimal arm, that is,

$$\mathsf{Reg}_T^\pi = \mathbb{E}\left[\sum_{t=1}^{T}(\ell_{tA_t} - \ell_{ta^*})\right] \quad \text{for} \quad a^* = \operatorname*{arg\,min}_{a \in [k]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{ta}\right], \tag{2.1}$$

where the expectation is taken with respect to the internal randomness of the algorithm and the randomness of the loss vectors $(\ell_t)_{t=1}^T$. We omit the superscript π when it is clear from the context. If the regret of an algorithm is $o(T)$, then the loss difference from the optimal choice suffered per round approaches 0, meaning that the algorithm is learning. Table 2.1 summarizes the notation used for multi-armed bandits.

One of the most important characteristics of the multi-armed bandit problem is that we can only observe the loss for the taken arm. This characteristic causes the challenge of exploration and exploitation tradeoffs: on the one hand, we want to select the arm with the fewest losses incurred so far (*exploitation*), and on the other hand, we prefer to select the arm with less information in order to look for an arm that can possibly have a smaller loss (*exploitation*). These are not always possible at the same time, and it is necessary to balance them effectively when selecting an arm.

Historically, two major regimes have been considered, stochastic and adversarial, depending on the procedure by which loss vectors are selected. In the following sections, we will introduce the definitions and achievable regrets of each regime and some fundamental algorithms for each regime. Since most of our contributions are based on algorithms

Table 2.1: Notation for multi-armed bandits

Symbol	Meaning
$k \in \mathbb{N}$	number of arms
$T \in \mathbb{N}$	horizon
$\ell_t = (\ell_{t1}, \ldots \ell_{tk})^\top \in [0, 1]^k$	loss vector at round t
$A_t \in [k]$	taken action at round t
$a^* \in [k]$	optimal arm
ν^*	underlying distribution of arms
$\mu = \mathbb{E}_{\ell_t \sim \nu^*}[\ell_t] \in \mathbb{R}^k$	loss mean
$\Delta_a \in (0, 1]$	suboptimality gap for suboptimal arm $a \neq a^*$
$\Delta_{\min} = \min_{a \neq a^*} \Delta_a$	minimum suboptimality gap
Reg_T^π	(pseudo-)regret of algorithm π with horizon T

[2] There are several definitions of a regret, and the regret here is sometimes called pseudo-regret. We will only focus on this definition of a regret in this dissertation.

originally considered only for the adversarial regime, we will provide a more detailed analysis of algorithms in the adversarial regime.

2.3 Stochastic Regime and its Algorithms

In this section, we formally define the stochastic regime for multi-armed bandits and then discuss achievable regret bounds. We also briefly introduce the well-known algorithms, the UCB algorithm and Thompson sampling, which are known to be optimal in the stochastic regime.

2.3.1 Definitions and Achievable Regret for Stochastic Regime

In the *stochastic regime*, a sequence of loss vector $\ell_1, \ell_2, \dots$ follows an unknown distribution ν^* with mean $\mu \in [0,1]^k$ in an i.i.d. manner. Define the minimum suboptimality gap by $\Delta_{\min} = \min_{a \neq a^*} \Delta_a$ for $\Delta_a = \mathbb{E}_{\ell_t \sim \nu^*}[(\ell_{ta} - \ell_{ta^*})]$. In the literature, in stochastic regimes, it is common in the literature to consider the reward, which corresponds to the inverse sign of the loss plus an appropriate constant. Still, to clarify the correspondence with the adversarial regime, we will proceed with the loss here.

Lower Bound Let $\mathsf{Reg}_T^\pi(\nu^*)$ be the regret when the underlying distribution of arms is ν^*. A bandit algorithm π is *consistent* if for any absolute constant $\alpha > 0$ and any loss distribution ν^* it holds that $\mathbb{E}[\mathsf{Reg}_T^\pi(\nu^*)] = o(T^\alpha)$. Then we have the following distribution-dependent lower bound:

Theorem 2.1 (Lai and Robbins 1985; Burnetas and Katehakis 1996)**.** *Consider a multi-armed bandit with Bernoulli distribution ν^* with mean $\mu \in [0,1]^k$.*[3] *Then for any consistent algorithm π,*

$$\liminf_{T \to \infty} \frac{\mathsf{Reg}_T^\pi(\nu^*)}{\log T} \geq \sum_{a \colon \Delta_a > 0} \frac{\Delta_a}{\mathrm{kl}\left(\mu_a, \mu_{a^*}\right)}, \tag{2.2}$$

where $\mathrm{kl}\left(x, y\right) = x \log(x/y) + (1-x)\log((1-x)/(1-y))$ is the KL divergence of the Bernoulli distributions with mean x from that with mean y.

This implies that in the stochastic regime, by Pinsker's inequality, $\mathrm{kl}\left(x,y\right) \geq 2(x-y)^2$, the optimal regret is approximately expressed as $\mathsf{Reg}_T = O(\sum_{a \colon \Delta_a > 0} \frac{\log T}{\Delta_a})$.

What is the favorable way to choose the arm at time t from the information obtained up to time $t-1$? Let $N_a(t) = |\{s \leq t-1 : A_s = a\}|$ be the number of times arm a is selected before the t-th round, and let $\widehat{\mu}_{t-1}(a) = \frac{1}{N_a(t)} \sum_{s=1}^{t-1} \ell_{sa} \mathbb{1}[A_s = a]$ be the empirical expected loss of arm a at round t. One of the most naive ideas is to choose an arm that minimizes the empirical expected losses so far, *i.e.*, $A_t = \arg\min_{a \in [k]} \widehat{\mu}(a)$. However, this only exploits and does not explore, and thus the learner suffers a linear regret. In the following, we present the two most representative algorithms that achieve the regret upper bound of $O(\log T)$.

2.3.2 UCB Algorithm

One of the most classical algorithms in the stochastic regime is the Upper Confidence Bound (UCB) algorithm (Auer et al., 2002a). The UCB algorithm optimistically estimates the expected value (μ_a) of the underlying loss distribution and selects arms based

[3] In fact, similar lower bounds can be shown for more general and unbounded distributions. See Lai and Robbins (1985); Burnetas and Katehakis (1996); Lattimore and Szepesvári (2020a) for details.

on the estimates. Specifically, the UCB algorithm selects the arm at time t by the following procedure:

$$A_t = \operatorname*{arg\,min}_{a \in [k]} \widehat{\nu}_{t-1}(a) - \sqrt{\frac{2 \log T}{N_a(t)}} \, .$$

The argument of $\arg\max$ in the above expression represents the empirical mean of the loss $\widehat{\nu}_{t-1}(a)$ minus the correction term $\sqrt{\frac{2\log T}{N_a(t)}}$, which increases as the number of times it is chosen decreases and corresponds to an upper confidence bound of the expected reward of each arm a. In other words, the procedure is to decide what arm to take *optimistically* based on the information obtained so far, and this idea called *optimism in the face of uncertainty* is the basic design principle of algorithms for the stochastic regime (Auer et al., 2002a; Abbasi-Yadkori et al., 2011). In fact, this enables us to deal with the exploration and exploitation tradeoff mentioned above: we exploit the knowledge so far by basically relying on empirical average losses, but the correction term allows us to moderately pull uninformed arms for exploration. It is known that this UCB algorithm can achieve a regret upper bound of $O(\sum_{a \neq a^*} \frac{\log T}{\Delta_a})$ (Auer et al., 2002a). An extension of this, the KL-UCB algorithm, is known to be able to achieve the theoretical limit of Theorem 2.1 (Cappé et al., 2013). The UCB algorithm is relatively simple to analyze and has been extended for many structured bandits such as combinatorial bandits (Kveton et al., 2015) and linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011).

2.3.3 Thompson Sampling

Thompson sampling (Thompson, 1933) is another one of the most fundamental algorithms for bandits and is known for its strong empirical performance in the stochastic regime (Chapelle and Li, 2011).

The idea of Thompson sampling is very simple and based on Bayes statistics. First, we determine an appropriate prior distribution as the distribution that each arm's expected loss $(\mu_a)_{a \in [k]}$ follows. We then sample from the current distribution of each arm and select the arm with the smallest sampled value. Finally, we update the distribution of each arm using the results observed at this round to obtain the posterior distribution.

Thompson sampling is one of the most promising algorithms for a variety of online decision-making problems such as the multi-armed bandits (Kaufmann et al., 2012b), linear bandits (Agrawal and Goyal, 2013b), and partial monitoring (Vanchinathan et al., 2014; Tsuchiya et al., 2020), and the effectiveness of Thompson sampling has been investigated both empirically (Chapelle and Li, 2011) and theoretically (Kaufmann et al., 2012a; Agrawal and Goyal, 2013a). It is known that Thompson sampling can achieve a theoretical lower bound of Theorem 2.1 as KL-UCB (Kaufmann et al., 2012a; Agrawal and Goyal, 2013a).

2.4 Adversarial Regime and its Algorithms

The stochastic regime we have seen so far assumes that the losses of each arm follow a certain distribution in an i.i.d. manner, which is a relatively strong assumption for a realistic problem. Can we formulate the bandit problem under weaker assumptions? In this section, we formally define the adversarial regime for multi-armed bandits and then discuss achievable regret bounds. We then introduce and analyze fundamental algorithms in the adversarial regime, which is heavily exploited in this dissertation.

2.4.1 Definitions of Adversarial Regime and its Lower Bound

In the *adversarial regime* (a.k.a. non-stochastic regime), we do not assume any stochastic structure for the losses, and they can be chosen in an arbitrary manner only with the assumption that they are bounded.. We only assume that the loss vector is bounded, and for simplicity in this chapter we assume that $\ell_t \in [0, 1]^k$.

In the adversarial regime, there are two types of loss-generating schemes: oblivious adversary and adaptive adversary. The oblivious adversary determines the loss vectors ℓ_t before the game starts, whereas the adaptive adversary can decide ℓ_t depending on the past history until the $(t-1)$-th round, $(A_s)_{s=1}^{t-1}$. In the analysis of the regret we consider (defined in *e.g.,* (2.1) for multi-armed bandits), which we consider in this dissertation, these differences do not make a difference in the analysis, as we only need to evaluate the loss difference from a fixed arm $a^* \in [k]$.

Lower Bound In the adversarial regime, the best possible upper bound is of $\Omega(\sqrt{kT})$ (Auer et al., 2002b). Formally, we have the following lower bound:

Theorem 2.2 (Auer et al. 2002b, Theorem 5.1)**.** *For any algorithm π for multi-armed bandits, there exists a sequence of losses such that*

$$\mathsf{Reg}_T^\pi \geq \frac{1}{20} \min\left\{\sqrt{kT},\, T\right\} .$$

It is known that the follow-the-regularized-leader framework (explained later) with (negative) Tsallis entropy regularization achieves $O(\sqrt{kT})$-regret bounds (Audibert and Bubeck, 2009; Abernethy et al., 2015), which we will prove in Section 2.4.6.

2.4.2 Loss Estimation

Since the value of losses is arbitrarily determined in the adversarial regime, the empirical expected losses $\widehat{\mu}_a(t)$ used in the stochastic regime are almost useless. Instead, the inverse weighted estimator $\widehat{\ell}_t \in \mathbb{R}^k$ defined in the following is commonly used:

$$\widehat{\ell}_{ta} = \frac{\ell_{ta} \mathbb{1}[A_t = a]}{p_{ta}} . \tag{2.3}$$

This estimator is also referred to as the inverse propensity score or inverse probability score (IPS) in the field of economics (Wooldridge, 2002). This estimator is common in the literature and useful for its unbiasedness, *i.e.,*

$$\mathbb{E}_{A_t \sim p_t}\left[\widehat{\ell}_t \,\middle|\, p_t\right] = \ell_t .$$

The downside of this inverse weighted estimator is that the worst-case variance can be very large. This can be a major cause of suffering large regret in some algorithms.

Hence, with appropriately chosen $m_t \in [0, 1]^k$, we sometimes use the following reduced variance estimator:

$$\widehat{\ell}_{ta} = \frac{(\ell_{ta} - m_{ta}) \mathbb{1}[A_t = a]}{p_{ta}} + m_{ta} \quad \text{for} \quad a \in [k] .$$

One can see that this estimator is also unbiased. In this chapter, we will only focus on the inverse weighted estimator in (2.3).

2.4.3 Exp3 Algorithm

In the adversarial regime, the *Exp3* (Exponential weight algorithm for Exploration and Exploitation, Auer et al., 2002b) algorithm is one of the most representative and important algorithms. In the literature, Exp3 is called by various names, such as Multiplicative Weight Update (MWU), Hedge, or, more simply, exponential weight. A variety of algorithms in the literature are based on the Exp3 algorithm; for example, the Exp4 algorithm for contextual bandits (Auer et al., 2002b), the Exp2 algorithm for adversarial linear bandits (Dani et al., 2008; Bubeck et al., 2012), ComBand for combinatorial bandits (Cesa-Bianchi and Lugosi, 2012), and exploration by optimization for partial monitoring (Lattimore and Szepesvári, 2020b; Tsuchiya et al., 2023a), to mention a few.

The Exp3 algorithm determines the arm selection probability $p_t \in \mathcal{P}_k$ at time $t \in [T]$ as follows:

$$p_{ta} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{sa}\right)}{\sum_{b\in[k]} \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{sb}\right)} \quad \text{for } a \in [k]\,, \tag{2.4}$$

where $\eta_t > 0$ is the learning rate. The smaller the cumulative estimated loss, the greater the probability of selection.

It is known that the Exp3 algorithm with $\eta_t = \Theta(\log(k)/(kT))$ can achieve the regret upper bound of $O(\sqrt{kT\log k})$ (proven in Theorem 2.3). The Exp3 algorithm is a special case of the follow-the-regularized-leader framework, which will be presented in a subsequent section, and its general analysis results will be used to prove this upper bound.

2.4.4 Follow-the-Regularized-Leader

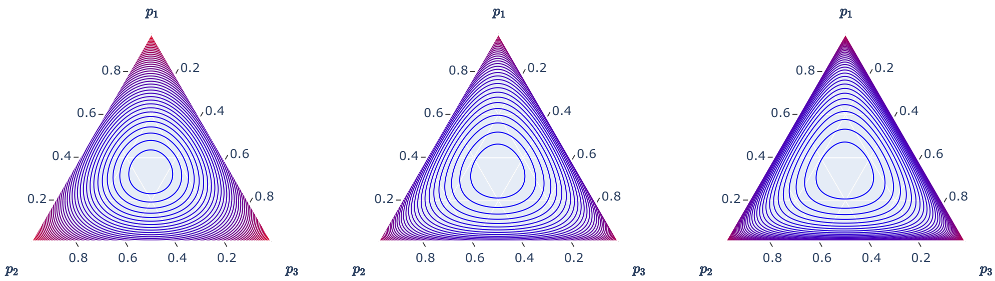
Follow-the-Regularized-Leader (FTRL) is a generalization of the Exp3 algorithm. In the FTRL framework, we choose arm selection probability at time t to minimize the expectation of cumulative estimated loss so far $\left\langle \sum_{s=1}^{t-1} \widehat{\ell}_s, p \right\rangle$ plus a convex regularizer $\psi_t(p)$. In other words, a probability vector $p_t \in \mathcal{P}_k$ over the action set $[k]$ is given as

$$p_t \in \operatorname*{arg\,min}_{p\in\mathcal{P}_k} \left\langle \sum_{s=1}^{t-1} \widehat{\ell}_s,\, p \right\rangle + \psi_t(p)\,,$$

where the vector $\widehat{\ell}_t \in \mathbb{R}^k$ is an unbiased estimator of ℓ_t and ψ_t is a convex regularizer. If there were not a convex regularizer, the cumulative expected loss would be completely trusted, and the arm selection probability would be a point on the border of the probability simplex. However, an appropriately chosen convex regularizer prevents this situation and allows us to tackle the exploitation and exploitation tradeoff.

Remark. Originally, FTRL and online mirror descent, which will be explained in the next section, were not designed exclusively for the bandit problem but are closely related to the history of online learning and online optimization. For further details, see *e.g.,* Orabona (2019); Lattimore and Szepesvári (2020a).

Remark. Here we only consider the case using p_t directly as the arm selection probability, but when applying FTRL to the bandit problem, the output of FTRL are sometimes transformed to compute the arm-selection probability. One of its roles is to reduce the variance of the loss estimator $\widehat{\ell}_t$, and in Chapters 4 and 5, we consider FTRL to perform such a transformation.



(a) negative Shannon entropy

(b) negative Tsallis entropy with an exponent of $1/2$

(c) log barrier

Figure 2.1: Contour maps of typical regularizers used in FTRL on two-dimensional probability simplex

Regularizes In the following, we introduce several common regularizers used in FTRL. The most common form of regularizer is the one that can be written as $\psi_t(p) = \frac{1}{\eta_t}\phi(p)$ with a certain learning rate η_t and $\phi \colon \mathcal{P}_k \to \mathbb{R}$. The following functions are typical examples of ϕ.

- *Negative Shannon entropy* (a.k.a. negentropy, entropic regularizer) is defined by

$$\phi(p) = -\sum_{a=1}^{k} p_a \log\left(\frac{1}{p_a}\right) = -H(p) =: \psi^{\mathsf{nS}}(p)\,. \tag{2.5}$$

  One can easily check that if we use negative Shannon entropy with learning rate η_t, FTRL becomes the Exp3 algorithm, *i.e.,* $p_t \in \mathcal{P}_k$ is expressed as (2.4).

- *Negative α-Tsallis entropy* for $\alpha \in (0,1)$ (Tsallis, 1988) is defined by[4]

$$\phi(p) = \frac{1 - \sum_{a=1}^{k} p_a^{\alpha}}{1-\alpha} =: \psi_{\alpha}^{\mathsf{nT}}(p)\,.$$

  It is known that FTRL with negative α-Tsallis entropy using absolute constant α can achieve the minimax regret of $O(\sqrt{kT})$ (Audibert and Bubeck, 2009; Abernethy et al., 2015), which matches the lower bound in Theorem 2.2 and will be proven in the following section (Theorem 2.4).

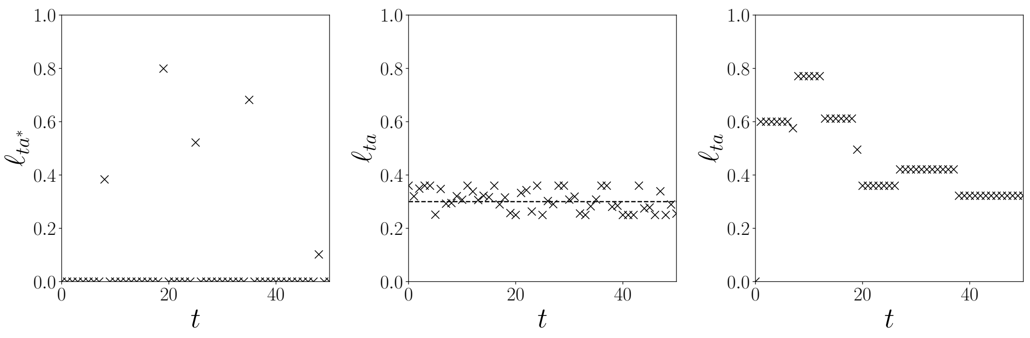
- *Log barrier regularizer* is defined by

$$\phi(p) = -\sum_{a=1}^{k} \log(p_a) =: \psi^{\mathsf{LB}}(p)\,.$$

In the following, we will often omit "negative" and just write Shannon entropy and Tsallis entropy.

What is the relationship between these three regularizers? One of the most important facts when considering their application to the bandit problem is that the Tsallis entropy interpolates the Shannon entropy and the log-barrier regularizer. To confirm this, consider the second derivative of the regularizer, an important concept that determines the behavior of FTRL. The Hessian of each of the Shannon entropy, α-Tsallis entropy, and log-barrier regularizer is $\mathrm{diag}(\Theta(1/p_a))$, $\mathrm{diag}(\Theta(1/p_a^{2-\alpha}))$, $\mathrm{diag}(\Theta(1/p_a^2))$. From this

[4] The definition of Tsallis entropy may be accompanied with a constant factor or linear term difference, but these are mostly for the sake of brevity of description.



(a) losses with small L^* **(b)** losses with small Q_p **(c)** losses with small V_p

Figure 2.2: Examples of "easy" loss data. Figures 2.2a, 2.2b, and 2.2c show that examples of ℓ_t such that its first-order quantity, second-order quantity, and path-length quantity are small.

fact, we can expect α-Tsallis entropy to behave like a Shannon entropy when $\alpha \to 1$ and a log-barrier regularizer when $\alpha \to 0$.[5] Figure 2.1 shows these regularizers on the two-dimensional probability simplex. We can see that as the regularizers change from the Shannon entropy to the log barrier, the contour lines of the regularizers become denser around the boundary.

We sometimes use a *hybrid regularizer* that is a linear combination of the above regularizers and their variants. The hybrid regularizer is introduced basically in order to stabilize the behavior of the arm selection probability in the aim of achieving various objectives (Bubeck et al. 2018; Luo et al. 2018; Zheng et al. 2019; Lee et al. 2020; Ito 2021a; Erez and Koren 2021; Ito et al. 2022b,a; Tsuchiya et al. 2023a,b, to name a few), and will be heavily exploited in this dissertation (Chapters 4, 5, and 6).

It is worth noting that another common regularizer in FTRL is squared 2-norm $\phi(p) = \|p\|^2$. However, while this regularizer is sometimes used in the field of online optimization, it is rarely used in bandit problems. In order for the bandit algorithm to be "stable", a regularizer must be Legendre, namely as it approaches the border of a feasible region of FTRL ($\mathcal{P}_k$ in multi-armed bandits), the norm of the gradient of the regularizer goes to ∞, and this is why the squared 2-norm is basically not used in the bandit problem.

2.4.5 Data-dependent Bounds

The adversarial regime is a very pessimistic setting since it allows arbitrary loss sequences. As shown in Theorem 2.2, the best achievable bound is $O(\sqrt{kT})$ in this regime. But, what if the underlying losses are benign to handle? Can we improve the regret bound for such loss sequences? Here, benign losses could be, for example, losses for an optimal arm is 0 in almost all rounds or losses that have basically the same loss values as the previous round. For these benign losses, the regret upper bound that depends on the quantity measuring "easiness" is called the *data-dependent bound*.

Typical examples of data-dependent bounds are first-order bounds dependent on the cumulative loss and second-order bounds dependent on sample variances in losses. Allenberg et al. (2006) provided an algorithm with a first-order regret bound of $O(\sqrt{kL^* \log k})$ for $L^* = \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \langle \ell_t, a \rangle$. Second-order regret bounds have been shown in some

[5] In the α-Tsallis entropy if we let $\alpha \to 1$ directly and use L'Hôpital's theorem, one can see that it coincides with the Shannon entropy except for a constant factor and a constant additive term. Also, in the case of $\alpha \to 0$, by modifying the definition of regularizer only by a constant factor as done by Zimmert and Seldin (2021), one can confirm that it agrees with log-barrier, up to a constant factor and a constant additive term.

studies, *e.g.,* by Hazan and Kale (2011); Wei and Luo (2018); Bubeck et al. (2018). In particular, Bubeck et al. (2018) provided the regret bound of $O(\sqrt{Q_2 \log k})$ for $Q_2 = \sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|^2$. Other examples of data-dependent bounds include path-length bounds in the form of $O(\sqrt{kV_1 \log T})$ for $V_1 = \sum_{t=1}^{T-1} \|\ell_t - \ell_{t+1}\|_1$. Figure 2.2 shows examples of the losses whose data-dependent quantities become smaller compared to the worst-case quantity of $\Theta(T)$. We will investigate the first-order, second-order, and path-length bounds for combinatorial semi-bandits in Chapter 6.

A sparsity-dependent bound is another important data-dependent bound (Kwon and Perchet, 2016; Bubeck et al., 2019b, 2018; Wei and Luo, 2018; Zheng et al., 2019). The study on a sparse-dependent bound was initiated by Kwon and Perchet (2016), who provided the algorithm achieving $\mathsf{Reg}_T = \tilde{O}(\sqrt{sT})$ and proved the matching (up to logarithmic factor) lower bound of $\mathsf{Reg}_T = \Omega(\sqrt{sT})$. We will further discuss and investigate a sparsity-dependent bound for multi-armed bandits in Chapter 5.

2.4.6 Analysis of Follow-the-Regularized-Leader

In this section, we prove the bound of $O(\sqrt{kT \log k})$ for the Exp3 algorithm presented above, the minimax optimal bound of $O(\sqrt{kT})$, and the first-order bound of $O(\sqrt{kL^* \log T})$ introduced in Section 2.4.5.

General Result for FTRL The regret analysis of FTRL boils down to the evaluation of $\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, p_t - p \right\rangle$, which is bounded using the following lemma:

Lemma 2.1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex body and $g_1, \dots, g_T \in \mathbb{R}^k$. Let $\psi_1, \dots, \psi_T, \psi_{T+1}$ be convex and differentiable functions and*

$$x_t = \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\langle \sum_{s=1}^{t-1} g_s, x \right\rangle + \psi_t(x) \,.$$

Then, for any $u \in \mathcal{X}$ it holds that

$$\begin{aligned}\sum_{t=1}^{T} \langle g_t, x_t - u \rangle \leq & \underbrace{\sum_{t=1}^{T} \big(\psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1})\big) + \psi_{T+1}(u) - \psi_1(x_1)}_{\text{penalty term}} \\ & + \underbrace{\sum_{t=1}^{T} \left(\langle g_t, x_t - x_{t+1} \rangle - D_{\psi_t}(x_{t+1}, x_t)\right)}_{\text{stability term}} \,. \end{aligned} \tag{2.6}$$

We refer to the terms in (2.6) as *penalty* and *stability* terms following, *e.g.,* Zimmert and Seldin (2021). The first term of the stability term increases if the variation of FTRL outputs in the adjacent rounds is large, whereas the penalty term comes from the strength of the regularization. The proof of this lemma is standard in the literature and similar results can be found in Lattimore and Szepesvári (2020a, Chapter 28) and Orabona (2019, Chapter 7). In the following, we give representative results and their proof for multi-armed bandits based on the above lemma. Before going into the main results, we provide an auxiliary lemma for bounding the stability term.

Lemma 2.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex body and $\psi \colon \mathcal{X} \to \mathbb{R}$ be a convex and twice differentiable function in the interior of $\mathcal{X}$. Then for any $g \in \mathbb{R}^d$ and $x, y \in \mathcal{X}$ there exists $\alpha \in [0, 1]$ such that $z = \alpha x + (1 - \alpha) y$ and*

$$\langle g, x - y \rangle - D_\psi(y, x) \leq \frac{1}{2} \|g\|^2_{(\nabla^2 \psi(z))^{-1}} \,.$$

Proof. By Taylor's theorem, there exists $\alpha \in [0,1]$ such that $z = \alpha x + (1-\alpha)y$ and $D_\psi(x,y) = \frac{1}{2}\|x-y\|^2_{\nabla^2\psi(z)}$. Using this, we have

$$\begin{aligned}\langle g, x-y\rangle - D_\psi(y,x) &\le \|g\|_{(\nabla^2\psi(z))^{-1}}\|x-y\|_{\nabla^2\psi(z)} - \frac{1}{2}\|x-y\|^2_{\nabla^2\psi(z)} \\ &\le \frac{1}{2}\|g\|^2_{(\nabla^2\psi(z))^{-1}} ,\end{aligned}$$

where the first inequality follows by Cauchy–Schwarz inequality and the last inequality takes the worst-case with respect to $\|x-y\|^2_{(\nabla^2\psi(z))}$. □

Remark. In this chapter, we will use Lemma 2.2 to bound the stability term in (2.6) in order to illustrate the effect of the Hessian of the regularizer on the regret upper bound, but we do not necessarily need to use this lemma. In fact, in our analysis in Chapters 4 to 6, we will simply take the worst-case for y (corresponding to p_{t+1}). This analysis has the advantage that a tight upper bound can sometimes be obtained, and a straightforward analysis can be performed, especially when we use a hybrid regularizer.

FTRL with negative Shannon entropy regularizer (Exp3) We first show that the Exp3 algorithm achieves nearly optimal regret bounds in the adversarial regime:

Theorem 2.3. *Consider the multi-armed bandit problem with $\ell_t \in [0,1]^k$. Then FTRL with the negative Shannon entropy $\psi_t(p) = -\frac{1}{\eta}H(p)$ and $\eta = \sqrt{2\log k/(kT)}$ achieves*

$$\mathsf{Reg}_T \le \sqrt{2kT\log k} .$$

Remark. In the analysis of the Exp3 algorithm, it was more common to use the potential-based proof (Auer et al., 2002b). Since the analysis based on FTRL is more straightforward, we here provide the proof based on it.

Proof. It holds that

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{tA_t} - \sum_{t=1}^T \ell_{ta^*}\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - e_{a^*}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \left\langle \widehat{\ell}_t, p_t - e_{a^*}\right\rangle\right] ,$$

where the last equality follows by $\mathbb{E}\left[\widehat{\ell}_t \,\middle|\, p_t\right] = \ell_t$. Then using Lemma 2.1,

$$\sum_{t=1}^T \left\langle \widehat{\ell}_t, p_t - e_{a^*}\right\rangle \le \frac{H(p_1)}{\eta} + \sum_{t=1}^T \left(\left\langle p_t - p_{t+1}, \widehat{\ell}_t\right\rangle - D_{\psi_t}(p_{t+1}, p_t)\right) .$$

The upper bound of the penalty term in the RHS of the last inequality is bounded by $\frac{\log k}{\eta}$ since the entropy is bounded by $\log k$.

Hence in the following, we consider the stability term. When $p_{tA_t} \le p_{t+1,A_t}$, it is trivial that the stability term is bounded by 0 since $\widehat{\ell}_t \ge 0$. Hence we consider the case of $p_{tA_t} > p_{t+1,A_t}$ in the following. Using Lemma 2.2, the stability term is bounded as

$$\left\langle p_t - p_{t+1}, \widehat{\ell}_t\right\rangle - D_{\psi_t}(p_{t+1}, p_t) \le \frac{1}{2}\|\widehat{\ell}_t\|^2_{\mathrm{diag}((\eta q_{ta})_a)} = \frac{\eta}{2}\sum_{a=1}^k q_{ta}\widehat{\ell}^2_{ta} \le \frac{\eta}{2} p_{tA_t}\widehat{\ell}^2_{tA_t} ,$$

where $q_t = \alpha p_t + (1-\alpha)p_{t+1}$ for some $\alpha \in [0,1]$ and the last inequality follows by $p_{tA_t} > p_{t+1,A_t}$ and the definition of $\widehat{\ell}_t$. Summing up the above arguments,

$$\mathsf{Reg}_T \le \mathbb{E}\left[\frac{\log k}{\eta} + \frac{\eta}{2}\sum_{t=1}^T p_{tA_t}\widehat{\ell}^2_{tA_t}\right] \le \frac{\log k}{\eta} + \frac{\eta k T}{2} = \sqrt{2kT\log k} ,$$

which completes the proof of Theorem 2.3. □

FTRL with negative Tsallis entropy regularizer The Exp3 algorithm is suboptimal compared to the minimax optimal bounds by only $\log k$ multipcalitive factor. Can we improve this? This can be done by balancing the penalty and stability terms in terms of their dependence on the arm selection probabilities, which can be made possible by the Tsallis entropy regularizer.

Theorem 2.4. *Consider the multi-armed bandit problem with $\ell_t \in [0,1]^k$. Then FTRL with the negative Tsallis entropy $\psi_t(p) = -\frac{4}{\eta}\sum_{a=1}^k \sqrt{p_a}$ and $\eta_t = \frac{8}{\sqrt{T}}$ achieves*

$$\mathsf{Reg}_T \le 2\sqrt{2kT}\,.$$

Remark. The same results can be obtained for the Tsallis entropy with an exponent of any absolute constant $\alpha \in (0,1)$, while here we only consider the case of $\alpha = 1/2$ for notational simplicity.

Proof. It holds that

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{tA_t} - \sum_{t=1}^T \ell_{ta^*}\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - e_{a^*}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \left\langle \widehat{\ell}_t, p_t - e_{a^*}\right\rangle\right],$$

where the last equality follows by $\mathbb{E}\left[\widehat{\ell}_t \,\middle|\, p_t\right] = \ell_t$. Then using Lemma 2.1,

$$\sum_{t=1}^T \left\langle \widehat{\ell}_t, p_t - e_{a^*}\right\rangle \le \frac{4\sum_{a=1}^k \sqrt{p_{1a}}}{\eta} + \sum_{t=1}^T \left(\left\langle p_t - p_{t+1}, \widehat{\ell}_t\right\rangle - D_{\psi_t}(p_{t+1}, p_t)\right)\,.$$

The penalty term is bounded by $\frac{4\sqrt{k}}{\eta}$ by Cauchy-Schwarz inequality.

In the following we consider the stability term. We rely on a similar argument as that of the proof of Corollary 2.3. When $p_{tA_t} \le p_{t+1,A_t}$, it is trivial that the stability term is bounded by 0 since $\widehat{\ell}_t \ge 0$. Hence we consider the case of $p_{tA_t} > p_{t+1,A_t}$ in the following. Using Lemma 2.2, the stability term is bounded as

$$\begin{aligned}&\left\langle p_t - p_{t+1}, \widehat{\ell}_t\right\rangle - D_{\psi_t}(p_{t+1}, p_t)\\ &\le \frac{1}{2}\|\widehat{\ell}_t\|^2_{\mathrm{diag}((\eta q_{ta}^{3/2})_a)} = \frac{\eta}{2}\sum_{a=1}^k q_{ta}^{3/2}\widehat{\ell}_{ta}^2 \le \frac{\eta}{2} p_{tA_t}^{3/2}\widehat{\ell}_{tA_t}^2\,,\end{aligned}$$

where $q_t = \alpha p_t + (1-\alpha)p_{t+1}$ for some $\alpha \in [0,1]$ and the last inequality follows by $p_{tA_t} > p_{t+1,A_t}$ and the definition of $\widehat{\ell}_t$.

Summing up the above arguments and using Cauchy-Schwarz inequality,

$$\mathsf{Reg}_T \le \mathbb{E}\left[\frac{4\sqrt{k}}{\eta} + \frac{\eta}{2}\sum_{t=1}^T p_{tA_t}^{3/2}\widehat{\ell}_{tA_t}^2\right] \le \frac{4\sqrt{k}}{\eta} + \frac{\eta T\sqrt{k}}{2} = 2\sqrt{2kT}\,,$$

which completes the proof of Theorem 2.4. □

FTRL with log-barrier regularizer

Theorem 2.5. *Consider the multi-armed bandit problem with $\ell_t \in [0,1]^k$. Then FTRL with the log-barrier regularizer $\psi_t(p) = -\frac{1}{\eta_t}\psi^{\mathsf{LB}}(p)$ and*

$$\eta_t = \frac{c}{\sqrt{1 + \sum_{s=1}^{t-1}\ell_{sA_s}^2}} \quad \textit{with} \quad c = \sqrt{k\log T}$$

achieves

$$\mathsf{Reg}_T \le 4\sqrt{k\log T\,(1+L^*)} + k\,,$$

where $L^* = \min_{a\in[k]}\sum_{t=1}^T \ell_{ta}$.

This upper bound is the first-order bound explained in Section 2.4.5, and when the cumulative loss of the optimal arm is of a constant order, the regret is bounded by $O(\sqrt{k \log T} + k)$, which is much smaller than the worst-case bound of $\sqrt{kT}$.

Proof. Define $p^* \in \mathcal{P}_k$ by

$$p^* = \left(1 - \frac{k}{T}\right) e_{a^*} + \frac{1}{T}\mathbf{1}\,.$$

Then, using the definition of the algorithm,

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - p^* \rangle\right] + \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p^* - e_{a^*} \rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^T \left\langle \widehat{\ell}_t, p_t - p^* \right\rangle\right] + k\,,$$

where the inequality follows from $\mathbb{E}\left[\widehat{\ell}_t \,\middle|\, p_t\right] = \ell_t$, the definition of p^*, and Cauchy-Schwarz inequality. Then using Lemma 2.1,

$$\sum_{t=1}^T \left\langle \widehat{\ell}_t, p_t - e_{a^*} \right\rangle \leq \frac{\psi^{\mathsf{LB}}(p^*)}{\eta_{T+1}} + \sum_{t=1}^T \left(\left\langle p_t - p_{t+1}, \widehat{\ell}_t \right\rangle - D_{\psi_t}(p_{t+1}, p_t) \right)\,.$$

Since $p_i^* \geq 1/T$, the penalty term is bounded as

$$\frac{\psi^{\mathsf{LB}}(p^*)}{\eta_{T+1}} = \frac{1}{\eta_{T+1}} \sum_{a=1}^k \log(1/p_a^*) \leq \frac{k \log T}{\eta_{T+1}}\,.$$

In the following we consider the stability term. We rely on a similar argument as that of the proof of Corollary 2.3. When $p_{tA_t} \leq p_{t+1,A_t}$, it is trivial that the stability term is bounded by 0 since $\widehat{\ell}_t \geq 0$. Hence we consider the case of $p_{tA_t} > p_{t+1,A_t}$ in the following. Using Lemma 2.2, the stability term is bounded as

$$\begin{aligned}&\left\langle p_t - p_{t+1}, \widehat{\ell}_t \right\rangle - D_{\psi_t}(p_{t+1}, p_t) \\ &\leq \frac{1}{2} \|\widehat{\ell}_t\|^2_{\mathrm{diag}((\eta_t q_{ta}^2)_a)} = \frac{\eta_t}{2} \sum_{a=1}^k q_{ta}^2 \widehat{\ell}_{ta}^2 \leq \frac{\eta_t}{2} p_{tA_t}^2 \widehat{\ell}_{tA_t}^2 = \frac{\eta_t \ell_{tA_t}^2}{2}\,,\end{aligned}$$

where $q_t = \alpha p_t + (1-\alpha) p_{t+1}$ for some $\alpha \in [0,1]$ and the last inequality follows by $p_{tA_t} > p_{t+1,A_t}$ and the definition of $\widehat{\ell}_t$.

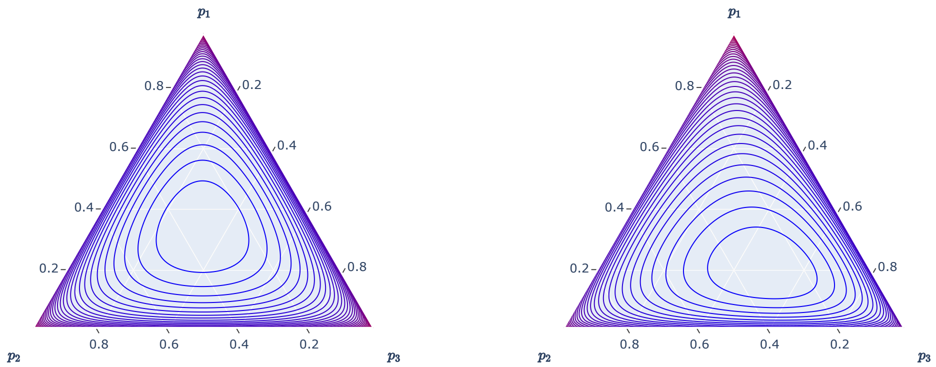
Taking the summation over $t \in [T]$ and using the definition of η_t , the stability term is further bounded as

$$\begin{aligned}\sum_{t=1}^T \eta_t \ell_{tA_t}^2 &= \sum_{t=1}^T \frac{c \ell_{tA_t}^2}{\sqrt{1 + \sum_{s=1}^{t-1} \ell_{sA_s}^2}} \leq 2 \sum_{t=1}^T \frac{c \ell_{tA_t}^2}{\sqrt{\sum_{s=1}^t \ell_{sA_s}^2} + \sqrt{\sum_{s=1}^{t-1} \ell_{sA_s}^2}} \\ &= 2c \sum_{t=1}^T \left(\sqrt{\sum_{s=1}^t \ell_{sA_s}^2} - \sqrt{\sum_{s=1}^{t-1} \ell_{sA_s}^2} \right) = 2c \sqrt{\sum_{t=1}^T \ell_{tA_t}^2}\,.\end{aligned}$$

(Note that the case of $\sum_{s=1}^t \ell_{tA_s}^2 = 0$ is trivially bounded, although the precise argument is omitted for simplicity.) Summing up the above arguments,

$$\mathsf{Reg}_T \leq \mathbb{E}\left[\frac{k \log T}{\eta_{T+1}} + c \sqrt{\sum_{t=1}^T \ell_{tA_t}^2}\right] + k \leq 2 \sqrt{k \log T \left(1 + \mathbb{E}\left[\sum_{t=1}^T \ell_{tA_t}^2\right]\right)} + k\,.$$

Combining this with $\mathbb{E}\left[\sum_{t=1}^T \ell_{tA_t}^2\right] \leq \mathsf{Reg}_T + \min_{a \in [k]} \sum_{t=1}^T \ell_{ta} = \mathsf{Reg}_T + L^*$ and the fact that $x \leq \sqrt{ax + b}$ for $a, b > 0$ implies $x \leq a + 2\sqrt{b}$ completes the proof of Theorem 2.5. □



(a) $q_t = (1/3, 1/3, 1/3)^\top$ **(b)** $q_t = (0.2, 0.3, 0.5)^\top$

Figure 2.3: Contour maps of the Bregman divergence $D_{\psi_t}(p, p_{t-1})$ induced by the log-barrier regularizer with different q_t on two-dimensional probability simplex

2.4.7 Online Mirror Descent

Online mirror descent (OMD) can also be regarded as a generalization of the Exp3 algorithm. In the OMD framework, a probability vector $p_t \in \mathcal{P}_k$ over the action set $[k]$ is computed so that it minimizes the expected loss $\left\langle \widehat{\ell}_{t-1}, p \right\rangle$ based on the estimated loss vector at the previous round plus Bregman divergence $D_{\psi_t}(p, p_{t-1})$, namely

$$p_t \in \operatorname*{arg\,min}_{p \in \mathcal{P}_k} \left\langle \widehat{\ell}_{t-1},\, p \right\rangle + D_{\psi_t}(p, p_{t-1}) \,,$$

where the vector $\widehat{\ell}_t \in \mathbb{R}^k$ is an unbiased estimator of ℓ_t and ψ_t is a convex regularizer. It is common to consider a form of $\psi_t = \frac{1}{\eta_t}\psi$ with a learning rate η_t. When $\eta_t = \eta$ for all $t \in [T]$, one can confirm that OMD with (negative) Shannon entropy regularizer in (2.5) coincides with the Exp3 algorithm. Figure 2.3 shows the contour maps of Bregman divergence induced by the log-barrier regularize on the two-dimensional probability simplex.

When $\eta_t = \eta$ for all $t \in [T]$, the regularizer is Legendre, and the feasible region satisfies the appropriate conditions, OMD coincides with FTRL. There are also several important differences. Since OMD only considers only the loss $\widehat{\ell}_t$ of the previous round, it is known that OMD can achieve good performance when the learner needs to track abrupt changes in the loss vectors. For example, it can achieve the path-length bounds introduced in Section 2.4.5 with an optimal order of $O(\sqrt{kV_\infty})$ (Bubeck et al., 2019a). In contrast, FTRL is known to have an advantage over OMD in that it is more robust in the stochastic regime with some adversarial noise (Amir et al., 2020). This dissertation mainly aims to achieve a good performance in such an environment and thus heavily relys on FTRL.

2.5 Best of Both Worlds: Simultaneously Achieving Optimality for Both Stochastic and Adversarial Regimes

Motivation As we have seen, the regret upper bound achievable in the stochastic and adversarial regimes are different: in the stochastic regime, $O(\log T)$ can be achieved by the UCB algorithm and Thompson sampling, and $O(\sqrt{T})$ in the adversarial regime by the Exp3 algorithm or more generally by FTRL. While algorithms for the stochastic regime is based on estimating the underlying expected loss based on the empirical average loss, algorithms for the adversarial regime basically does not involve such a thing.

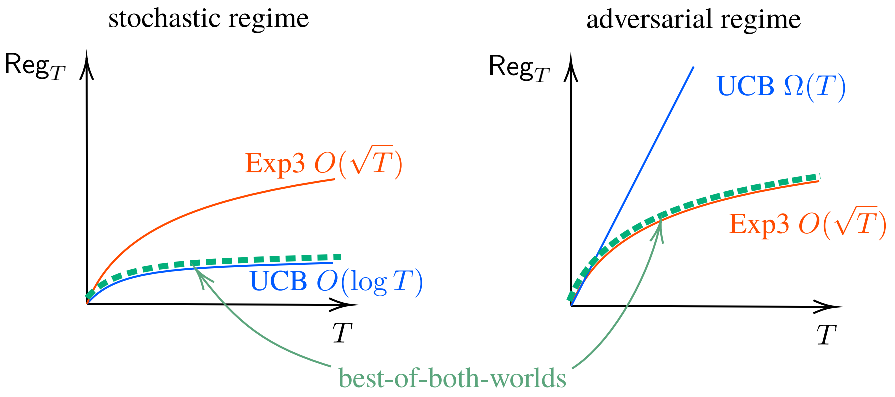


Figure 2.4: Idea of best-of-both-worlds

We now consider formulating a real-world problem as a bandit problem. Then we realize that it is unclear which regime's algorithms are better suited to practical applications. In fact, it is known that algorithms specialized for the stochastic regime suffer a linear regret even in the almost stochastic environment (Zimmert and Seldin, 2021). In contrast, classical algorithms for the adversarial regime work poorly in the stochastic regime since the algorithm needs to consider the worst-case scenarios. Since it is difficult to know the underlying regime in practical scenarios, it is desirable to obtain an algorithm that obtains an $O(\log T)$ regret for the stochastic regime and an $O(\sqrt{T})$ regret for the adversarial regime *without* knowing the underlying environment.

Definition To achieve this goal, particularly in the classical multi-armed bandits, the Best-of-Both-Worlds (BOBW) algorithms that perform well in both stochastic and adversarial regimes have been developed. The concept of a best-of-both-worlds algorithm in the stochastic and adversarial regimes is illustrated in Figure 2.4. We formally define the BOBW algorithm as follows:

Definition 2.1 (Best-of-both-worlds algorithm for multi-armed bandit problems)**.** Consider a multi-armed bandit problem and $\ell_t \in [0,1]^k$ for all $t \in [T]$. A bandit algorithm π for multi-armed bandits has the *best-of-both-worlds* property if the (pseudo-)regret in the adversarial regime satisfies $\mathsf{Reg}_T^\pi \leq \tilde{O}(\sqrt{T})$ and in the stochastic regime satisfies $\mathsf{Reg}_T^\pi \leq O((\log T)^\alpha)$ for some absolute constant $\alpha > 0$.[6]

For more general problem class P, we define the best-of-both-worlds propoerty as follows:

Definition 2.2 (Best-of-both-worlds algorithm for general bandit problems)**.** Consider a bandit problem P. Suppose that the optimal dependence on horizon T is $f(T)$ for the adversarial regime and $g(T)$ for the stochastic regime. Then a bandit algorithm π for P has the *best-of-both-worlds* property if the (pseudo-)regret in the adversarial regime is bounded as $\mathsf{Reg}_T^\pi = \tilde{O}(f(T))$ and in the stochastic regime is bounded as $\mathsf{Reg}_T^\pi \leq \tilde{O}(g(T))$, where $\tilde{O}(\cdot)$ ignores the logarithmic factors.

2.5.1 History

BOBW for multi-armed bandits The study of the BOBW algorithm started with a seminal paper by Bubeck and Slivkins (2012) and was followed by several other studies (Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017). The basic idea of the approaches by Bubeck and Slivkins (2012); Auer and Chiang (2016) is

[6] In the literature, the best-of-both-worlds algorithm has not been formally defined.

to first run the algorithm for stochastic regime assuming that the underlying environment is stochastic. Then, if a precisely designed hypothesis test shows that this assumption does not hold, it switches to the algorithm for the adversarial regime. On the other hand, the approaches by Seldin and Slivkins (2014); Seldin and Lugosi (2017) first assume the adversarial regime, and when the environment is found to be stochastic, it switches to an algorithm for the stochastic regime. However, their performance guarantees are suboptimal and not adaptive enough due to the lack of realistic performance caused by the explicit switching of algorithms. Therefore, it would be desirable if it is possible to obtain the BOBW guarantee without switching explicitly between the algorithms specialized for the stochastic and adversarial regimes. It would be further desirable if we could obtain the BOBW guarantee by a simple algorithm.

This desire was accomplished by an OMD-based algorithm by by Wei and Luo (2018). Following this work, the celebrated Tsallis-INF algorithm, which is FTRL with the Tsallis entropy regularization, was proposed by Zimmert and Seldin (2019, 2021). This Tsallis-INF algorithm achieves a $O(\sqrt{kT})$ regret bound for the adversarial regime and $O(\sum_{a \neq a^*} \frac{\log T}{\Delta_a})$ for the stochastic regime, and importantly empirically performs significantly well compared to the previous BOBW algorithms. The use of Tsallis entropy is not considered to have any information-theoretic meaning, and Tsallis entropy was just considered to achieve a minimax optimal regret upper bound $O(\sqrt{kT})$ in an adversarial regime, as discussed in 2.4.6. In fact, as described below, by allowing a certain loose of regret of the logarithmic factor in the adversarial regime, the BOBW guarantees are obtained with FTRL with log-barrier and that with Shannon entropy regularizer.

In the multi-armed bandit problem, the regret upper bound that can actually be achieved is not $O(\sum_{a \neq a^*} \frac{\log T}{\Delta_a})$ but (2.2). Hence, one might wonder if it is possible to achieve the truly optimal regret upper bound given by (2.2) in the stochastic regime while maintaining good performance in the adversarial regime. This is an important open problem as discussed in Chapter 7. However, as a first step toward this, there is an algorithm that takes into account the variance (Ito et al., 2022b). In particular, the algorithm therein achieves a regret bound of $O(\sum_{a \neq a^*} (\frac{\sigma_a^2}{\Delta_a} + 1) \log T)$ for loss variances σ_a^2 of arm a.

BOBW beyond bandits Very recently, BOBW algorithms have been extensively investigated in many online-decision making problems beyond the multi-armed bandits. For structured bandits, BOBW algorithms have been developed for the problem of prediction with expert advice (de Rooij et al., 2014; Gaillard et al., 2014; Luo and Schapire, 2015; Mourtada and Gaïffas, 2019), online learning with feedback graphs (Erez and Koren, 2021; Ito et al., 2022a; Rouyer et al., 2022), online linear optimization (Huang et al., 2017), online submodular minimization (Ito, 2022), dueling bandits (Saha and Gaillard, 2022), linear bandits (Lee et al., 2021), combinatorial semi-bandits (Wei and Luo, 2018; Zimmert et al., 2019; Ito, 2021a; Tsuchiya et al., 2023b), position-based model (Chen et al., 2022a), contextual bandits (Pacchiano et al., 2022), partial monitoring (Tsuchiya et al., 2023a,c), and episodic Markov decision processes (Jin and Luo, 2020; Jin et al., 2021). There are also a decoupled setting, in which a different arm can be selected for exploration and exploitation (Rouyer and Seldin, 2020), a setting with switching cost, where we incur certain losses when the learner changes the arm to choose in addition to the regret (Rouyer et al., 2021; Amir et al., 2022), delayed feedback setting, where there is a delay until the loss of the selected arm is observed (Masoudian et al., 2022), and a setting where the underlying loss distributions are heavy-tailed (Huang et al., 2022).

Very recently, there have been several numbers of studies that aim to achieve BOBW and data-dependent bounds simultaneously. Wei and Luo (2018) devised an algorithm to achieve BOBW and first-order bounds simultaneously, Ito (2021c); Ito et al. (2022b); Tsuchiya et al. (2023b) to achieve BOBW and first-, second-, and path-length bounds, Tsuchiya et al. (2023c) to achieve BOBW and sparsity-dependent bounds.

2.5.2 Self-Bounding Technique

It is becoming very common to use FTRL to realize BOBW algorithms, and a *self-bounding technique* is one of the key techinques to prove a BOBW guarantee (Gaillard et al., 2014; Wei and Luo, 2018; Zimmert and Seldin, 2021). In the self-bounding technique, we first derive upper and lower bounds of regret using a variable depending on the arm selection probability and then derive a regret bound by combining the upper and lower bounds.

In the following, we introduce one of the strategies to give an intuition. Suppose that we can derive the upper and lower bounds of regret using a random variable $P \in [0, T]$ satisfying $\mathsf{Reg}_T \leq O(\mathrm{polylog}(T)\sqrt{P})$. The bound for the adversarial regime of $\tilde{O}(\sqrt{T})$ can be directly obtained by taking the worst-case with respect to P in the last inequality. For the stochastic regime, suppose that we have a lower bound of $\mathsf{Reg}_T \geq O(P)$, which can often be readily obtained in stochastic regimes by the definition of regret. We then can derive the (poly-)logarithmic regret bound from these bounds as

$$\mathsf{Reg}_T = 2\mathsf{Reg}_T - \mathsf{Reg}_T \leq O(\mathrm{polylog}(T)\sqrt{P} - P) \leq O(\mathrm{polylog}(T)) .$$

We can prove a BOBW guarantee in a similar way as above by deriving different types of upper bounds depending on desirable properties we want in the regret upper bound.

2.5.3 Intermediate Regime between Stochastic and Adversarial Regimes

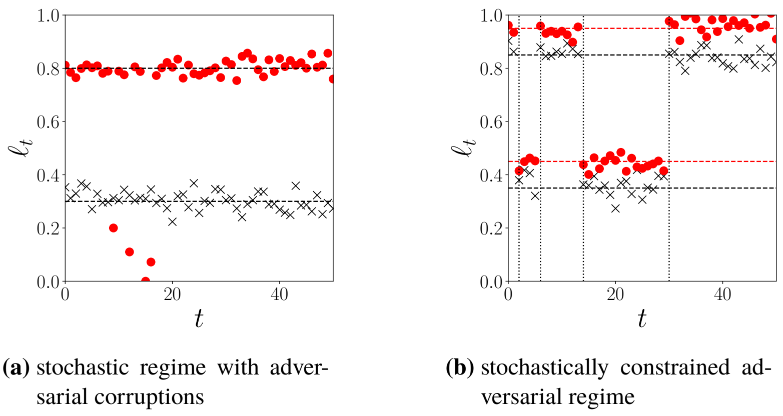
So far, we have discussed BOBW algorithms, which achieve optimality simultaneously in both stochastic and adversarial regimes. However, both regimes have extreme characteristics. The stochastic regime assumes that losses are obtained in an i.i.d. manner, which is a somewhat strong assumption in a practical scenario. In contrast, the adversarial regime assumes arbitrary bounded losses, which is a very pessimistic assumption. Then, a natural question arises. Is there an intermediate regime between these regimes, or is there a regime that interpolates between them? In this section, we first introduce the stochastic regime with adversarial corruptions and the stochastically constrained adversarial regime as intermediate regimes between the stochastic regime and the adversarial regime. We then present the adversarial regime with a self-bounding constraint, which includes all the regimes introduced so far as special cases.

Stochastic Regime with Adversarial Corruptions One of the most representative intermediate regimes is the *stochastic regime with adversarial corruptions* (Lykouris et al., 2018). In this regime, a temporary loss $\ell'_t \in [0, 1]^k$ is sampled from an unknown distribution ν^*, and then the adversary corrupts ℓ'_t to ℓ_t. We define the corruption level $C \in [0, T]$ by

$$C = \mathbb{E}\left[\sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty\right] .$$

If $C = 0$, this regime coincides with the stochastic regime, and if $C = T$, this regime corresponds to the adversarial regime. Figure 2.5a shows an example of the stochastic regime with adversarial corruptions.

Stochastically Constrained Adversarial Regime The stochastically constrained adversarial regime was first considered by Wei and Luo (2018) and also discussed in Zimmert and Seldin (2021) in the context of the multi-armed bandit problem. We say that a



(a) stochastic regime with adversarial corruptions

(b) stochastically constrained adversarial regime

Figure 2.5: Examples of intermediate regimes between stochastic and adversarial regimes with $k = 2$. Figures 2.5a and 2.5b are the examples of stochastic regimes with adversarial corruptions and stochastically constrained adversarial regimes, respectively. The black crosses and red circles denote $(\ell_{t1})_t$ and $(\ell_{t2})_t$, respectively.

regime is the stochastically constrained adversarial regime if for any $a \neq a^*$ there exists $\tilde{\Delta}_{a,a^*} > 0$ such that

$$\mathbb{E}_{\ell_t \sim \nu^*}[\ell_{ta} - \ell_{ta^*} | \ell_1, \dots, \ell_{t-1}] \geq \tilde{\Delta}_{a,a^*} \, .$$

From a practical standpoint, this setting is modeling a case where, for example, the differences of probabilities of purchasing products are always the same for all the time, but the willingness to buy varies uniformly with the weather. Figure 2.5a shows an example of the stochastically constrained adversarial regimes.

Adversarial Regime with a Self-Bounding Constraint The following *adversarial regime with a self-bounding constraint*, developed originally in the multi-armed bandits (Zimmert and Seldin, 2021), includes the regimes that appeared so far.

Definition 2.3. Let $\Delta \in [0, 1]^k$ and $C \geq 0$. The regime is in an *adversarial regime with a* (Δ, C, T) self-bounding constraint if it holds for any algorithm that

$$\mathsf{Reg}_T \geq \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t} - C\right] \, .$$

One can see that the regimes that have appeared so far are included in the adversarial regime with a self-bounding constraint. Note that these regimes can also be defined for structured bandits and formal definitions of them are given in the following chapters. For multi-armed bandits, the optimal robustness of this regime has been investigated by Ito (2021b). They showed that $\mathsf{Reg}_T = \Omega(\frac{k}{\Delta_{\min}} + \sqrt{\frac{Ck}{\Delta_{\min}}})$ under a certain condition, where $\Delta_{\min} = \min_{a \neq a^*} \Delta_a$ with Δ in Definition 5.1.

Remark. Importantly, most FTRL-based BOBW algorithms can achieve a near-optimal regret in an adversarial regime with a self-bounding constraint without knowing C, and in fact, most of the algorithms introduced in Section 2.5 can do so (Zimmert and Seldin 2021; Ito 2021c; Jin et al. 2021; Masoudian and Seldin 2021; Tsuchiya et al. 2023a, to mention a few). This can be easily proven by a slight modification of the self-bounding technique. BOBW algorithms that achieve good performance in the intermediate regime are sometimes called best-of-three-worlds algorithms or best-of-all-algorithms. In fact, however, algorithms with best-of-three-worlds or best-of-all-worlds properties are often simply called best-of-both-worlds, and we will follow this convention in this dissertation.

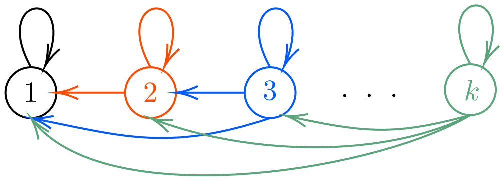


Figure 2.6: An example of feedback graphs G used in the problem of online learning with feedback graphs.

2.6 Structured Bandits

In this section, we introduce several structured bandits, which is a model of multi-armed bandits with structure. In particular, we discuss online learning with feedback graphs, partial monitoring, and combinatorial bandits. Not only do they have many interesting theoretical properties, but more importantly, they have applications to many real-world problems, which will also be discussed.

2.6.1 Full Information Setting

We start by introducing the full information setting before discussing structured bandits. In the full information setting, the entire loss vector $\ell_t = (\ell_1, \dots, \ell_k)^\top$ can be observed independently of which arm the learner takes, whereas in the multi-armed bandits setting, only the loss for the selected arm ℓ_{tA_t} is observed.

One of the most representative algorithms in the full information setting is the Hedge algorithm (Freund and Schapire, 1997). This corresponds to the Exp3 algorithm described above in which the estimated vector $\widehat{\ell_t}$ is replaced by the actually observed vector ℓ_t. (Note that historically, the Hedge algorithm was developed before the Exp3 algorithm.)

Using this Hedge algorithm, we can achieve a regret upper bound of $O(\sqrt{T \log k})$ in the adversarial regime, which matches the lower bound (Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006). One can confirm that this bound holds by checking that the stability term, which appeared in the analysis of Section 2.4.6, is reduced to a smaller value since the entire loss vector is observed. It is also known that we can achieve an optimal regret of $O(\frac{\log k}{\Delta})$ in the stochastic regime (Mourtada and Gaïffas, 2019).

2.6.2 Online Learning with Feedback Graphs

We next introduce online learning with feedback graphs (Mannor and Shamir, 2011) (a.k.a. bandits with feedback graphs), which interpolates between multi-armed bandits and online learning (in fact, even extrapolating!). In this problem, we are given a directed feedback graph $G = (V, E)$, where $V = [k]$ is the set of arms (or actions), and $E \subseteq V \times V$ is the structure of feedback for choosing arms. At each round $t = 1, 2, \dots, T$, the learner chooses an action $A_t \in V$ and the adversary simultaneously selects a loss function $\ell_t \colon V \to [0, 1]$. The learner then incurs the loss of $\ell_t(A_t)$ and observes feedback of $\ell_t(a)$ for all a such that the feedback graph G has an edge from A_t to a.

The model of online learning with feedback graphs includes a variety of sequential decision-making problems. If G consists of only self-loops, *i.e.,* if $E = \{(a, a) \colon a \in V\}$, the problem corresponds to the multi-armed bandit problem. If G is a complete directed graph with self-loops, *i.e.,* $E = V \times V$, the problem corresponds to the full information setting.

Algorithm 2.2: The procedure of partial monitoring game $\mathcal{G} = (\mathcal{L}, \Phi)$

```
input: horizon $T$, number of actions $k$, number of outcomes $d$, set of feedback symbols $\Sigma$
The learner observes $\mathcal{L}$ and $\Phi$.
for $t = 1, 2, \ldots, T$ do
    The learner chooses an action $A_t \in [k]$.
    The opponent simultaneously chooses an outcome $x_t \in [d]$.
    The learner suffers an unobserved loss $\mathcal{L}_{A_t x_t}$, and receives a feedback symbol $\sigma_t = \Phi_{A_t x_t}$.
```

Online learning with feedback graphs was initiated by Mannor and Shamir (2011) and was given a very important characterization by Alon et al. (2015). In particular, Alon et al. (2015) classified the difficulty of problems according to the structure of the feedback graph and revealed the regret achievable in each problem class (see Alon et al. 2015 for details). Their algorithm is based on the Exp3 algorithm, which again indicates the usefulness of the Exp3 algorithm.

Application An amusing application is the newsvender problem. In this problem, one can choose the amount of newspapers to print from $[k]$, and the goal is to maximize the profit by appropriately choosing the amount of newspapers to print on each day. An important characteristic of this problem is that it has a structure where the more copies one prints, the more information one gains. For example, if you print 100 newspapers, you do not know how much profit you get for printing 200 newspapers, whereas if you print 200 newspapers, you know how much profit you get for printing 100 newspapers. The feedback graph corresponding to this problem is shown in Figure 2.6.

2.6.3 Partial Monitoring

Partial monitoring (PM) is a general sequential decision-making problem with limited feedback, which can be seen as a generalization of the bandit problem and online learning with feedback graphs. A PM game $\mathcal{G} = (\mathcal{L}, \Phi)$ is defined by the pair of a loss matrix $\mathcal{L} \in [0,1]^{k \times d}$ and feedback matrix $\Phi \in \Sigma^{k \times d}$, where k is the number of actions, d is the number of outcomes, and Σ is a set of feedback symbols. The game is sequentially played by a learner and opponent for T rounds. At the beginning of the game, the learner observes $\mathcal{L}$ and Φ. At every round $t \in [T]$, the opponent chooses an outcome $x_t \in [d]$, and then the learner chooses an action $A_t \in [k]$, suffers an unobserved loss $\mathcal{L}_{A_t x_t}$, and receives a feedback symbol $\sigma_t = \Phi_{A_t x_t}$, where $\mathcal{L}_{ax}$ is the (a, x)-th element of $\mathcal{L}$. The procedure of a PM game is given in Algorithm 2.2. In general, the learner cannot directly observe the outcome and loss, and can only observe the feedback symbol. As in the case of online learning with feedback graphs, all PM games are classified into four disjoint classes based on how much information the feedback matrix gives about the loss matrix, and achievable regrets are different in each class. In particular, PM games can be classified into trivial, easy, hard, and hopeless games, for which their minimax regrets are 0, $\tilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$, and $\Theta(T)$, respectively (Bartók et al., 2011; Lattimore and Szepesvári, 2019b). Further detailed descriptions and many properties of PM are given in Chapters 3, 4, and 5.

Application Partial monitoring includes many online decision-making problems as a special case. Here, we introduce several important examples.

- Multi-armed bandits: PM includes the vanilla multi-armed bandits with finite loss support as a special case. For example, k-armed Bernoulli bandits is expressed as $k \times 2^k$ loss and feedback matrices, and in particular, when $k = 2$, they can be expressed as

  $$\mathcal{L} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} , \quad \Phi = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} .$$

  It is known that the multi-armed bandit belongs to the class of easy games.

- Apple tasting, matching pennies: Suppose that we sequentially need to determine whether a product (apple) flowing on a conveyor belt is P (positive) or N (negative). As commodity inspectors, we have three choices of actions: (i) determine that the commodity is P, (ii) determine that the commodity is N, or (iii) disassemble and inspect the commodity and determine whether it is P or N. Note that the commodity cannot be determined whether it is P or N without actually disassembling it. If P (resp. N) is wrongly determined to be N (resp. P), the learner suffers a loss of $c_{\mathsf{P}\to\mathsf{N}} > 0$ (resp. $c_{\mathsf{P}\to\mathsf{N}} > 0$). Suppose that decomposition and inspection costs $c_q \geq 0$. How should the learner choose action so that they can minimize the cost as much as possible? This problem, called apple tasting or matching pennies, can be defined using the following loss matrix $\mathcal{L}$ and feedback matrix Φ:

  $$\mathcal{L} = \begin{pmatrix} 0 & c_{\mathsf{N}\to\mathsf{P}} \\ c_{\mathsf{P}\to\mathsf{N}} & 0 \\ c_q & c_q \end{pmatrix} , \quad \text{and} \quad \Phi = \begin{pmatrix} \text{None} & \text{None} \\ \text{None} & \text{None} \\ \mathsf{P} & \mathsf{N} \end{pmatrix} .$$

  Many other problems can be formulated in the same manner. For example, in the problem of determining whether a large number of e-mails sequentially delivered to a mailbox are ham or spam, a certain cost is incurred when we misclassify them, and it is impossible to tell whether they are actually ham or spam without asking a human. It is known that this problem falls into easy or hard games depending on the parameters in $\mathcal{L}$. For example, when $c_{\mathsf{N}\to\mathsf{P}} = c_{\mathsf{P}\to\mathsf{N}} = 1$, the problem is easy if $c_q \in (0, 1/2)$, hard if $c_q > 1/2$, and trivial if $c_q = 0$ (Lattimore and Szepesvári, 2020b).

- Dynamic pricing: Dynamic pricing is another important example of PM games. In the dynamic pricing game, the learner corresponds to a seller, and the opponent corresponds to a buyer. At each round $t \in [T]$, the seller sells an item with a specific price of $A_t \in [k]$, and the buyer comes with an evaluation price $x_t \in [d]$ for the item, where the selling price and the evaluation price correspond to the action and outcome, respectively. The buyer buys the item if the selling price A_t is smaller than or equal to x_t and not otherwise. The seller can only know if the buyer bought the item (denoted as feedback 0) or did not buy the item (denoted as 1). The goal of the seller is to minimize the cumulative loss, and there are two types of definitions for the loss, where each induced game falls into the easy and hard games. We call them *dp-easy* and *dp-hard* games, respectively.

  In both cases, the seller incurs the constant loss $c > 0$ when the item is not bought due to the loss of opportunity to sell the item. When the item is not bought, the loss incurred to the seller is different between these settings. The seller in the dp-easy game *does not* take the buyer's evaluation price into account. In other words, the seller gains the selling price A_t as a reward (equivalently incurs $-A_t$ as a loss). In particular, the loss for selling price A_t and evaluation price x_t is given by

  $$\mathcal{L}_{A_t x_t} = -A_t \mathbb{1}[A_t \leq x_t] + c\mathbb{1}[A_t > x_t] .$$

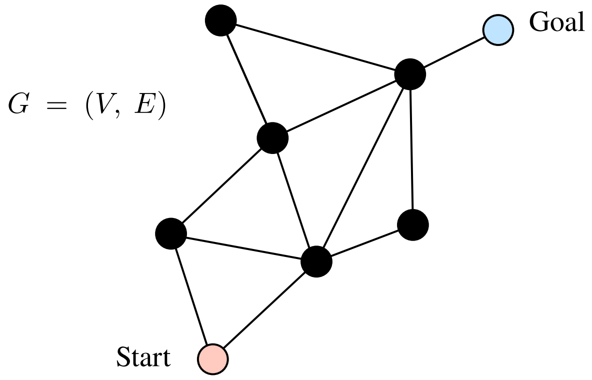


Figure 2.7: An example of network routing formalized as combinatorial bandits. In this example, $d = |E| = 12$.

In a matrix form, the loss matrix $\mathcal{L} \in \mathbb{R}^{k \times k}$ and feedback matrix $\Phi \in \mathbb{R}^{k \times k}$ of the dp-easy game are given by[7]

$$\mathcal{L} = \begin{pmatrix} -1 & -1 & \dots & -1 \\ c & -2 & \dots & -2 \\ \vdots & \ddots & \ddots & \vdots \\ c & \dots & c & -k \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \dots & 1 & 0 \end{pmatrix},$$

where the feedback 0 (resp. 1) denotes the events that the buyer do (resp. does not) buy products. This setting can be regarded as a generalized version of the online posted price mechanism, which was addressed in, *e.g.,* Blum et al. (2004) and Cesa-Bianchi et al. (2006), and an example of easy games.

On the other hand, the seller in the dp-hard game *does* take the buyer's evaluation price into account when the item is bought. In other words, the seller incurs the difference between the opponent evaluation and the selling price $x_t - A_t$ as a loss because the seller could have made more profit if the seller had sold at the price x_t. Namely, the loss incurred at time t is given by

$$\mathcal{L}_{A_t x_t} = (x_t - A_t)\mathbb{1}[A_t \leq x_t] + c\mathbb{1}[A_t > x_t] \,.$$

In the matrix form, the loss matrix and feedback matrix of the dp-hard setting is given by

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & \dots & k-1 \\ c & 0 & \dots & k-2 \\ \vdots & \ddots & \ddots & \vdots \\ c & \dots & c & 0 \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \dots & 1 & 0 \end{pmatrix},$$

This setting is also addressed in Cesa-Bianchi et al. (2006) and belongs to the class of hard games.

2.6.4 Combinatorial Bandits

In combinatorial bandits, the learner and environment play the game sequentially. The learner is given an action set $\mathcal{A} \subset \{0,1\}^d$, where $d \in \mathbb{N}$ is the dimension of the action set.[8] For every round $t \in [T]$, the environment chooses a loss $\ell(t) \in [0,1]^d$, and

[7] $\mathcal{L} \in [0,1]^k$ is not satisfied here, but we present a loss matrix following the convention. Dividing by k is enough to get $\mathcal{L} \in [0,1]^k$.

[8] Here, d corresponds to the number of arms k in multi-armed bandits. In this dissertation, d is used instead of k, following the convention.

the learner then chooses an action $a(t) \in \mathcal{A}$ (also called a *super-arm*), incurs a loss $\langle \ell(t), a(t) \rangle$. In the combinatorial bandit problem, there are two types of feedback for the learner: *semi-bandit feedback* and *full-bandit feedback*. In the semi-bandit setting, the learner observes $\ell_i(t)$ for all $i \in [d]$ such that $a_i(t) = 1$, whereas in the full-bandit setting, the learner observes only the sum of losses for the taken action, that is, $\sum_{i \in [d]: a_i(t)=1} \ell_i(t)$. We refer to each index $i \in [d]$ as *base-arm* i. Further detailed descriptions of combinatorial semi-bandits are given in Chapter 6.

Application One of the representative problems of combinatorial bandits is the advertisement placement problem. Consider the following setup: a company managing a website has d advertisements at hand that can be placed on the web, and selects $m < d$ of them to place on the web (Anantharam et al., 1987; Chen et al., 2013). Suppose that the company can observe whether the placed ads are actually clicked or not, and based on the results they determine a strategy for ad placement for the next time period. The goal of the company is to maximize its cumulative reward by properly selecting and placing the ads. This problem can be viewed as a combinatorial semi-bandits problem with the action set $\mathcal{A} = \{a \in \{0,1\}^d : \|a\|_1 = m\}$. This problem belongs to a class of m-set problems among combinatorial bandits.

Another representative example is the online shortest path problem used for network routing (Gai et al., 2012). Figure 2.7 illustrates this problem. In this problem, we are given a connected directed weighted graph $G = (V, E)$ with start $s \in V$ and goal $g \in V$. Here E corresponds to the set of base-arms, and thus $|E| = d$. Under these conditions, we consider how to minimize the cumulative transportation cost when sending packets from S to G sequentially. This problem can be viewed as a combinatorial bandits problem by setting the action set to $\mathcal{A} = \{a \in \{0,1\}^d : a \in E\}$. Here, $a \in E$ means that if $a_i = 1$ then edge i is included in E. This corresponds to a semi-bandits setting if the weights of each edge are observed individually and to a full-bandits setting if only the sum of the weights of each edge can be observed.

The combinatorial semi-bandit problem also includes many practical problems such as multi-task bandits (Cesa-Bianchi and Lugosi, 2012), crowdsourcing (ul Hassan and Curry, 2016), learning spectrum allocations (Gai et al., 2012), and recommender systems (Qin et al., 2014).

Chapter 3

Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring

In this chapter, we investigate finite stochastic partial monitoring, which is a general model for sequential learning with limited feedback. While Thompson sampling is one of the most promising algorithms for a variety of online decision-making problems, its properties for stochastic partial monitoring have not been theoretically investigated, and the existing algorithm relies on a heuristic approximation of the posterior distribution. To mitigate these problems, we present a novel Thompson-sampling-based algorithm, which enables us to exactly sample the target parameter from the posterior distribution. Besides, to theoretically justify the proposed algorithm, we consider the linearized variant of the partial monitoring problem with local observability. For this problem, we prove that a special case of the new algorithm achieves the logarithmic distribution-dependent expected pseudo-regret $O(\log T)$. This result is the first regret bound of Thompson sampling for partial monitoring, which also becomes the first logarithmic regret bound of Thompson sampling for linear bandits.

3.1 Introduction

Partial monitoring (PM) is a general sequential decision-making problem with limited feedback (Rustichini, 1999; Piccolboni and Schindelhauer, 2001). PM is attracting broad interest because it includes a wide range of problems such as the multi-armed bandit problem (Lai and Robbins, 1985), a linear optimization problem with full or bandit feedback (Zinkevich, 2003; Dani et al., 2008), dynamic pricing (Kleinberg and Leighton, 2003), and label efficient prediction (Cesa-Bianchi et al., 2005).

A PM game can be seen as a sequential game that is played by two players: a learner and an opponent. At every round, the learner chooses an action, while the opponent chooses an outcome. Then, the learner suffers an unobserved loss and receives a feedback symbol, both of which are determined from the selected action and outcome. The main characteristic of this game is that the learner cannot directly observe the outcome and loss. The goal of the learner is to minimize his/her cumulative loss over all rounds. The performance of the learner is evaluated by the regret, which is defined as the difference between the cumulative losses of the learner and the optimal action (*i.e.,* the action whose expected loss is the smallest).

There are mainly two types of PM games, which are the *stochastic* and *adversarial* settings (Piccolboni and Schindelhauer, 2001; Bartók et al., 2011). In the stochastic setting, the outcome at each round is determined from the *opponent's strategy*, which is a probability vector over the opponent's possible choices. On the other hand, in the adversarial setting, the outcomes are arbitrarily decided by the opponent. We refer to the PM game with finite actions and finite outcomes as a *finite* PM game. In this chapter, we focus on the stochastic finite game.

One of the first algorithms for PM was considered by Piccolboni and Schindelhauer (2001). They proposed the FeedExp3 algorithm, the key idea of which is to use an unbiased estimator of the losses. They showed that the FeedExp3 algorithm attains $\tilde{O}(T^{3/4})$ minimax regret for a certain class of PM games, and showed that any algorithm suffers linear minimax regret $\Omega(T)$ for the other class. Here T is the time horizon and the notation $\tilde{O}(\cdot)$ hides polylogarithmic factors. The upper bound $\tilde{O}(T^{3/4})$ is later improved by Cesa-Bianchi et al. (2006) to $O(T^{2/3})$, and they also provided a game with a matching lower bound.

In the seminal paper by Bartók et al. (2011), they classified PM games into four classes based on their minimax regrets. To be more specific, they classified games into trivial, easy, hard, and hopeless games, where their minimax regrets are 0, $\tilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$, and $\Theta(T)$, respectively. Note that the easy game is also called a *locally observable* game. After their work, several algorithms have been proposed for the finite PM problem (Bartók et al., 2012; Vanchinathan et al., 2014; Komiyama et al., 2015a). For the problem-dependent regret analysis, Komiyama et al. (2015a) proposed an algorithm that achieves $O(\log T)$ regret with the optimal constant factor. However, it requires solving a time-consuming optimization problem with infinitely many constraints at each round. In addition, this algorithm relies on the forced exploration to achieve the optimality, which makes the empirical performance near-optimal only after prohibitively many rounds, say, 10^5 or 10^6.

Thompson sampling (TS, Thompson, 1933) is one of the most promising algorithms on a variety of online decision-making problems such as the multi-armed bandit (Lai and Robbins, 1985) and the linear bandit (Agrawal and Goyal, 2013b), and the effectiveness of TS has been investigated both empirically (Chapelle and Li, 2011) and theoretically (Kaufmann et al., 2012a; Agrawal and Goyal, 2013a; Honda and Takemura, 2014). In the literature on PM, Vanchinathan et al. (2014) proposed a TS-based algorithm called BPM-TS (Bayes-update for PM based on TS) for stochastic PM, which empirically achieved state-of-the-art performance. Their algorithm uses Gaussian approximation to handle the complicated posterior distribution of the opponent's strategy. However, this approximation is somewhat heuristic and can degrade the empirical performance due to the discrepancy from the exact posterior distribution. Furthermore, no theoretical guarantee is provided for BPM-TS.

Our goals are to establish a new TS-based algorithm for stochastic PM, which allows us to sample the opponent's strategy parameter from the exact posterior distribution, and investigate whether the TS-based algorithm can achieve sub-linear regret in stochastic PM. Using the accept-reject sampling, we propose a new TS-based algorithm for PM (TSPM), which is equipped with a numerical scheme to obtain a posterior sample from the complicated posterior distribution. We derive a logarithmic regret upper bound $O(\log T)$ for a variant of the proposed algorithm, TSPM-Gaussian, on the locally observable game under a linearized variant of the problem. This is the first regret bound for TS on the locally observable game. Moreover, our setting includes the linear bandit problem and our result is also the first logarithmic expected regret bound of TS for the linear bandit, whereas a high-probability bound was provided, for example, in Agrawal and Goyal (2013b). Finally, we compare the performance of TSPM with existing algorithms in numerical experiments, and show that TSPM and TSPM-Gaussian outperform existing algorithms.

3.2 Preliminaries

A PM game with k actions and d outcomes is defined by a pair of a loss matrix $\mathcal{L} \in \mathbb{R}^{k \times d}$ and a feedback matrix $\Phi \in \Sigma^{k \times d}$, where Σ is a set of feedback symbols. In this chapter,

Table 3.1: List of symbols used in this chapter.

Symbol	Meaning
$k, d \in \mathbb{N}$	number of actions and outcomes
Σ	set of feedback symbols
m	number of feedback symbols
$\mathcal{L} \in \mathbb{R}^{k \times d}$	loss matrix
$\Phi \in \Sigma^{k \times d}$	feedback matrix
$p^* \in \mathcal{P}_d$	opponent's strategy
$S_a \in \{0,1\}^{m \times d}$	signal matrix of action a
$A_t \in [k]$	action taken at round t
$x_t \in [d]$	outcome chosen by opponent at round t
$\sigma_t \in \Sigma$	feedback symbol observed at round t in discrete setting
$y(t) \in \{0,1\}^m$	feedback symbol vector observed at round t in linear setting
$N_a(t)$	number of times action a is taken before time $t \in [T]$
$\mathcal{C}_a \subset \mathcal{P}_d$	cell of action a
$F_t(p)$	unnormalized posterior distribution in (3.1)
$f_t(p)$	probability density function corresponding to $F_t(p)$
$G_t(p)$	unnormalized proposal distribution for $F_t(p)$ in (3.2)
$g_t(p)$	probability density function corresponding to $G_t(p)$
$q_a^{(t)} \in \mathcal{P}_d$	empirical feedback distribution of action a by time t
$q_{a,n} \in \mathcal{P}_d$	empirical feedback distribution of action a after the action is taken n times

we let $\Sigma = [m] = \{1, \dots, m\}$ without loss of generality.

A PM game can be seen as a sequential game that is played by two players: the learner and the opponent. At each round $t = 1, 2, \dots, T$, the learner selects action $A_t \in [k]$, and at the same time the opponent selects an outcome based on the opponent's strategy $p^* \in \mathcal{P}_d$, where $\mathcal{P}_n = \{p \in \mathbb{R}^n : p_k \geq 0, \sum_{k=1}^n p_k = 1\}$ is the $(n-1)$-dimensional probability simplex. The outcome x_t of each round is an independent and identically distributed sample from p^*, and then, the learner suffers loss $\mathcal{L}_{A_t x_t}$ at round t. The learner cannot directly observe the value of this loss, but instead observes the *feedback symbol* $\sigma_t = \Phi_{A_t x_t} \in [A]$. The setting explained above has been widely studied in the literature of stochastic PM (Bartók et al., 2011; Komiyama et al., 2015a), and we call this the *discrete* setting. In Section 3.4, we also introduce a *linear* setting for theoretical analysis, which is a slightly different setting from the discrete one.

The learner aims to minimize the cumulative loss over T rounds. The expected loss of action a is given by $L_a^\top p^*$, where L_a is the a-th column of $\mathcal{L}^\top$. We say action a is *optimal* under strategy p^* if $(L_a - L_b)^\top p^* \leq 0$ for any $b \neq a$. We assume that the optimal action is unique, and without loss of generality that the optimal action is action 1. Let $\Delta_a = (L_a - L_1)^\top p^* \geq 0$ for $a \in [k]$ and $N_a(t)$ be the number of times action a is selected before the t-th round. When the time step is clear from the context, we use n_a instead of $N_a(t)$. We use the (pseudo-)regret to measure the performance:

$$\mathsf{Reg}_T = \mathbb{E}\left[\widehat{\mathsf{Reg}}_T\right] \quad \text{for} \quad \widehat{\mathsf{Reg}}_T = \sum_{t=1}^{T} \Delta_{A_t} = \sum_{a=1}^{k} \Delta_a N_a(T+1)\,.$$

This is the relative performance of the algorithm against the *oracle*, which knows the optimal action 1 before the game starts.

We introduce the following definitions to clarify the class of PM games, for which we develop an algorithm and derive a regret upper bound. The following cell decomposition is the concept to divide the simplex $\mathcal{P}_d$ based on the loss matrix to identify the optimal action, which depends on the opponent's strategy p^*.

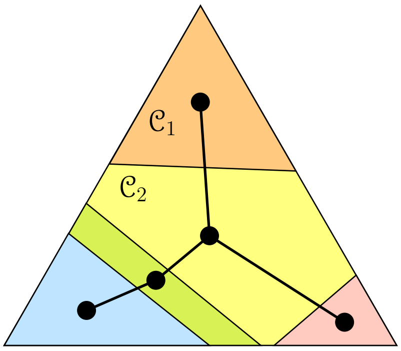


Figure 3.1: An example of cell decomposition. The points • correspond to Pareto-optimal actions. Cells $\mathcal{C}_1$ and $\mathcal{C}_2$ are neighbors.

Definition 3.1 (Cell decomposition and Pareto-optimality)**.** For action $a \in [k]$, *cell* a

$$\mathcal{C}_a = \left\{ p \in \mathcal{P}_d \colon (L_a - L_b)^\top p \le 0, \ \forall b \ne a \right\}$$

is the set of opponent's strategies for which action a is optimal. Action a is *Pareto-optimal* if there exists an opponent's strategy p^* under which action a is optimal.

Each cell is a convex closed polytope. Next, we define *neighbors* between two Pareto-optimal actions, which intuitively means that the two actions "touch" each other on their surfaces. An example of cell decomposition is given in Figure 3.1.

Definition 3.2 (Neighbors and neighborhood action)**.** Two Pareto-optimal actions a and b are *neighbors* if $\mathcal{C}_a \cap \mathcal{C}_b$ is an $(d-2)$-dimensional polytope. For two neighboring actions $a, b \in [k]$, the *neighborhood action set* is defined as

$$N^+_{a,b} = \{ c \in [k] \colon \mathcal{C}_a \cap \mathcal{C}_b \subseteq \mathcal{C}_c \} \ .$$

Note that the neighborhood action set $N^+_{a,b}$ includes actions a and b from its definition. Next, we define the *signal matrix*, which encodes the information of the feedback matrix Φ so that we can utilize the feedback information.

Definition 3.3 (Signal matrix)**.** The signal matrix $S_a \in \{0,1\}^{m \times d}$ of action a is defined as

$$(S_a)_{\sigma,x} = \mathbb{1}[\Phi_{ax} = \sigma] \, .$$

Note that if we define the signal matrix as above, $S_a p^* \in \mathbb{R}^m$ is a probability vector over feedback symbols of action a. The following *local observability* condition separates easy and hard games, This condition intuitively means that the information obtained by taking actions in the neighborhood action set $N^+_{a,b}$ is sufficient to distinguish the loss difference between actions a and b.

Definition 3.4 (Local observability)**.** A partial monitoring game is said to be *locally observable* if for all pairs a, b of neighboring actions, $L_a - L_b \in \oplus_{c \in N^+_{a,b}} \operatorname{Im} S_c^\top$,

$$L_a - L_b \in \oplus_{c \in N^+_{a,b}} \operatorname{Im} S_c^\top \, ,$$

where $\operatorname{Im} V$ is the image of the linear map V, and $V \oplus W$ is the direct sum between the vector spaces V and W.

We also consider the concept of the *strong local observability* condition, which implies the above local observability condition.

Definition 3.5 (Strong local observability)**.** A partial monitoring game is said to be *strongly locally observable* if for all pairs $a, b \in [k]$,

$$L_a - L_b \in \operatorname{Im} S_a^\top \oplus \operatorname{Im} S_b^\top \,.$$

This condition was assumed in the theoretical analysis in Vanchinathan et al. (2014), and we also assume this condition in theoretical analysis in Section 3.4. Note that the strong local observability means that, for any $a \neq b$, there exists $z_{a,b} \neq 0 \in \mathbb{R}^{2m}$ such that $L_a - L_b = (S_a^\top, S_b^\top)\, z_{a,b}$.

Notation Let $\|\cdot\|$ and $\|\cdot\|_p$ be the Euclidean norm and p-norm, and let $\|x\|_A = \sqrt{x^\top A x}$ be the norm induced by the positive semidefinite matrix $A \succeq 0$. The vector $e_i \in \mathbb{R}^d$ is the i-th orthonormal basis of $\mathbb{R}^d$, and $\mathbf{1}_n = [1, \dots, 1]^\top$ is the n-dimensional all-one vector. Let $q_a^{(t)}$ be the empirical feedback distribution of action a at round t, *i.e.,* $q_a^{(t)} = [n_{a1}/n_a, \dots, n_{am}/n_a]^\top \in \mathcal{P}_m$, where $n_{ay} = \sum_{s=1}^{t} \mathbb{1}[A_s = a, \sigma_s = y]$ and $n_a = \sum_{y=1}^{m} n_{ay}$. The notation used in this chapter is summarized in Table 3.1.

Methods for Sampling from Posterior Distribution We briefly review the methods to draw a sample from the posterior distribution. While TS is one of the most promising algorithms, the posterior distribution can be in a quite complicated form, which makes obtaining a sample from it computationally hard. To overcome this issue, a variety of approximate posterior sampling methods have been considered, such as Gibbs sampling, Langevin Monte Carlo, Laplace approximation, and the bootstrap (Russo et al., 2018, Section 5). Recent work (Lu and Van Roy, 2017) proposed a flexible approximation method, which can even efficiently be applied to quite complex models such as neural networks. However, more recent work revealed that algorithms based on such an approximation procedure *can* suffer a linear regret (Phan et al., 2019), even if the approximation error in terms of the α-divergence is small enough.

Although BPM-TS is one of the best methods for stochastic PM, it approximates the posterior by a Gaussian distribution in a heuristic way, which can degrade the empirical performance due to the distributional discrepancy from the exact posterior distribution. Furthermore, no theoretical guarantee is provided for BPM-TS. To address this issue, we mitigate these problems by providing a new algorithm for stochastic PM, which allows us to exactly draw samples from the posterior distribution. We also give theoretical analysis for the proposed algorithm.

3.3 Thompson-sampling-based Algorithm for Partial Monitoring

In this section, we present a new algorithm for stochastic PM games, where we name the algorithm TSPM (TS-based algorithm for PM). The algorithm is given in Algorithm 3.1, and we will explain the subroutines in the following.

3.3.1 Accept-Reject Sampling

We adopt the accept-reject sampling (Casella et al., 2004) to *exactly* draw samples from the posterior distribution. The accept-reject sampling is a technique to draw samples from a specific distribution f, and a key feature is to use a *proposal distribution* g, from which we can easily draw a sample and whose ratio to f, that is f/g, is bounded by a constant value R. To obtain samples from f, (i) we generate samples $X \sim g$; (ii) accept

Algorithm 3.1: TSPM Algorithm

```
Input: prior parameter $\lambda > 0$
Set $B_0 \leftarrow \lambda I_d, b_0 \leftarrow 0$.
Take each action for $n \geq 1$ times.
for $t = 1, 2, \ldots, T$ do
    Sample $\tilde{p}_t \sim \pi(p \mid \{A_s, \sigma_s\}_{s=1}^{t-1})$ based on the accept-reject sampling (Algorithm 3.2).
    Take action $A_t = \arg\min_{a \in [k]} L_a^\top \tilde{p}_t$ and observe feedback $\sigma_t$.
    Update $B_t \leftarrow B_{t-1} + S_{A_t}^\top S_{A_t},\ b_t \leftarrow b_{t-1} + S_{A_t}^\top e_{\sigma_t}$.
```

Algorithm 3.2: Accept-Reject Sampling

```
Input: constant $R \in [0, 1]$
while true do
    Sample $\tilde{p}_t \sim g_t(p)$ (Algorithm 3.3).
    Sample $\tilde{u} \sim \mathcal{U}([0, 1])$.
    if $R\tilde{u} < F_t(\tilde{p}_t)/G_t(\tilde{p}_t)$ then
        return $\tilde{p}_t$.
```

Algorithm 3.3: Sampling from $g_t(p)$

```
Compute $\tilde{B}_t, \tilde{b}_t$ from $B_t, b_t$.
repeat
    Sample $p^{(\alpha)} \sim \mathcal{N}(\tilde{B}_t^{-1}\tilde{b}_t, \tilde{B}_t^{-1})$.
until $p^{(\alpha)} \in \mathcal{P}_{d-1}$ ;
return $\tilde{p} = [p^{(\alpha)\top}, 1 - \sum_{i=1}^{d-1}(p^{(\alpha)})_i]^\top$.
```

X with probability $f(X)/Rg(X)$. Note that f and g do not have to be normalized when the acceptance probability is calculated.

Let $\pi(p)$ be a prior distribution for p. Then an unnormalized density of the posterior distribution for p can be expressed as

$$F_t(p) = \pi(p) \prod_{a=1}^{k} \exp\left(-n_a D\left(q_a^{(t)} \| S_a p\right)\right) , \tag{3.1}$$

the detailed derivation of which is given in Section 3.6.1. We use the proposal distribution with unnormalized density

$$G_t(p) = \pi(p) \prod_{a=1}^{k} \exp\left(-\frac{1}{2} n_a \|q_a^{(t)} - S_a p\|^2\right) . \tag{3.2}$$

Based on these distributions, we use Algorithm 3.2 for exact sampling from the posterior distribution, where $\mathcal{U}([0, 1])$ is the uniform distribution over $[0, 1]$ and $g_t(p)$ is the distribution corresponding to the unnormalized density $G_t(p)$ in (3.2). The following proposition shows that setting $R = 1$ realizes the exact sampling.

Proposition 3.1. *Let $f_t(p)$ be the distribution corresponding to the unnormalized density $F_t(p)$ in* (3.1). *Then, the output of Algorithm 3.2 with $R = 1$ follows $f_t(p)$.*

This proposition can easily be proved by Pinsker's inequality, which is detailed in Section 3.6.1.

In practice, $R \in [0, 1]$ is a parameter to balance the amount of over-exploration and the computational efficiency. As R decreases from 1, the algorithm tends to accept a point p far from the mode. The case $R = 0$ corresponds to the TSPM algorithm where the proposal distribution is used without the accept-reject sampling, which we call *TSPM-Gaussian*. As we will see in Section 3.4, TSPM-Gaussian corresponds to exact sampling of the posterior distribution when the feedback follows a Gaussian distribution rather than a multinomial distribution.

TSPM-Gaussian can be related to BPM-TS (Vanchinathan et al., 2014) in the sense that both of them use samples from Gaussian distributions. Nevertheless, they use different Gaussians and TSPM-Gaussian performs much better than BPM-TS as we will see in the experiments. Details on the relation between TSPM-Gaussian and BPM-TS are described in Section 3.6.3.

In general, we can realize efficient sampling with a small number of rejections if the proposal distribution and the target distribution are close to each other. On the other hand, in our problem, the densities in (3.1) and (3.2) for each fixed point p exponentially decay with the number of samples n_a if the empirical feedback distribution $q_a^{(t)}$ converges. This means that $F_t(p)$ and $G_t(p)$ have an exponentially large relative gap in most rounds. Nevertheless, the number of rejections does not increase with t as we will see in the experiments, which suggests that the proposal distribution approximates the target distribution well.

3.3.2 Sampling from Proposal Distribution

When we consider Gaussian density $\mathcal{N}(0, \lambda I_d)$ truncated over $\mathcal{P}_d$ as a prior, the proposal distribution also has the Gaussian density $\mathcal{N}(B_t^{-1} b_t,\ B_t^{-1})$ over $\mathcal{P}_d$, where

$$B_t = \lambda I_d + \sum_{a=1}^{k} n_a S_a^\top S_a = B_{t-1} + S_{A_t}^\top S_{A_t}\,,\ \ b_t = \sum_{a=1}^{k} n_a S_a^\top q_a^{(t)} = b_{t-1} + S_{A_t}^\top e_{\sigma_t}\,. \quad (3.3)$$

Here note that the probability simplex $\mathcal{P}_d$ is in an $(d-1)$-dimensional space and a sample from $\mathcal{N}(0, \lambda I_d)$ is not contained in $\mathcal{P}_d$ with probability one. In the literature, *e.g.*, Altmann et al. (2014), sampling methods for Gaussian distributions truncated on a simplex have been discussed. We use one of these procedures summarized in Algorithm 3.3, where we first sample $d-1$ elements of p from another Gaussian distribution and determine the remaining element by the constraint $\sum_{i=1}^{d} p_i = 1$.

Proposition 3.2. *Sampling from $g_t(p)$ is equivalent to Algorithm 3.3 with*

$$\tilde{B}_t = C_t - 2D_t + f_t \mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top\,, \quad \tilde{b}_t = f_t \mathbf{1}_{d-1} - g_t + b_t^{(\alpha)} - b^{(d)}\mathbf{1}_{d-1}\,,$$

where $B_t = \begin{bmatrix} C_t & g_t \\ g_t^\top & f_t \end{bmatrix}$ for $C_t \in \mathbb{R}^{d-1\times d-1}$, $g_t \in \mathbb{R}^{d-1}$, $f_t \in \mathbb{R}$, $b_t = [{b_t^{(\alpha)}}^\top, b_t^{(d)}]^\top \in \mathbb{R}^{d-1} \times \mathbb{R}$, and $D_t = \frac{1}{2}(g_t \mathbf{1}_{d-1}^\top + \mathbf{1}_{d-1} g_t^\top)$.

We give the proof of this proposition for self-containedness in Section 3.6.2.

3.4 Theoretical Analysis

This section considers a regret upper bound of the TSPM-Gaussian algorithm. In the theoretical analysis, we consider a *linear* setting of PM called linear partial monitoring. In the linear PM, the learner suffers the expected loss $L_{A_t}^\top p^*$ as in the discrete setting, and receives feedback vector

$$y(t) = S_{A_t} p^* + \epsilon_t \quad \text{for} \quad \epsilon_t \sim \mathcal{N}(0, I_d)\,, \qquad (3.4)$$

instead of σ_t whereas the one-hot representation of $y(t)$ is distributed by the probability vector $S_a p^*$ in the discrete setting. In general, the linear PM setting considered in (3.4) does not include discrete PM. However, if we allow ϵ_t to be an action-dependent sub-Gaussian distribution instead of a Gaussian distribution, then linear PM includes discrete PM as a special case, and such a formulation is recently investigated (Kirschner et al., 2023). It is worth noting the linear PM also includes the linear bandit problem, where the feedback vector is expressed as $L_a^\top p^* + \epsilon_t$.

In the linear PM, $G_t(p)$ in (3.2) becomes the exact posterior distribution rather than a proposal distribution. The definition of the cell decomposition for this setting is largely the same as that of discrete setting and detailed in Section 3.6.5. Therefore, TS with exact posterior sampling in the linear PM corresponds to TSPM-Gaussian. In the linear PM, the unknown parameter p^* is in $\mathbb{R}^d$ rather than in $\mathcal{P}_d$, and therefore we consider the prior $\pi(p) = \mathcal{N}(0, \lambda I_d)$ over $\mathbb{R}^d$, where the posterior distribution becomes $\mathcal{N}(B_t^{-1} b_t,\ B_t^{-1})$.

There are a few works that analyze TS for the PM because of its difficulty. For example in Vanchinathan et al. (2014), an analysis of the TS-based algorithm (BPM-TS) is not given despite the fact that its performance is better than the algorithm based on a confidence ellipsoid (BPM-LEAST). Zimmert and Lattimore (2019) considered the theoretical aspect of a variant of TS for the linear PM in view of the Bayes regret, but this algorithm is based on the knowledge on the time horizon and different from the family of TS used in practice. More specifically, their algorithm considers the posterior distribution for *regret* (not pseudo-regret), and an action is chosen according to the posterior probability that each arm minimizes the *cumulative* regret. Thus, the time horizon also needs to be known.

Types of Regret Bounds We focus on the *(a) problem-dependent (b) expected pseudo-regret.* (a) In the literature, a *minimax* (or *problem-independent*) regret bound has mainly been considered, for example, to classify difficulties of the PM problem (Bartók et al., 2010; Bartók et al., 2011). On the other hand, a *problem-dependent* regret bound often reflects the empirical performance more clearly than the minimax regret (Bartók et al., 2012; Vanchinathan et al., 2014; Komiyama et al., 2015a). For this reason, we consider this problem-dependent regret bound. (b) In complicated settings of bandit problems, a *high-probability regret bound* has mainly been considered (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013b), which bounds the pseudo-regret with high probability $1-\delta$. Though such a bound can be transformed to an expected regret bound, this type of analysis often sacrifices the tightness since a linear regret might be suffered with small probability δ. This is why the analysis in Vanchinathan et al. (2014) for BPM-LEAST finally yielded an $\tilde{O}(\sqrt{T})$ expected regret bound whereas their high-probability bound is $O(\log T)$.

3.4.1 Regret Upper Bound

In the following theorem, we show that logarithmic problem-dependent expected regret is achievable by the TSPM-Gaussian algorithm.

Theorem 3.1 (Regret upper bound)**.** *Consider any finite stochastic linear partial monitoring game with Gaussian noise. Assume that the game is strongly locally observable and* $\Delta_a = (L_a - L_1)^\top p^* > 0$ *for any* $a \neq 1$*. Then, the regret of TSPM-Gaussian satisfies for sufficiently large* T *that*

$$\mathsf{Reg}_T = O\left(\frac{mk^2 d \max_{a\in[k]} \Delta_a \log T}{\Lambda^2}\right),$$

where $\Lambda := \min_{a\neq 1} \Lambda_a$ *for* $\Lambda_a = \Delta_a / \|z_{1,a}\|$ *with* $z_{1,a}$ *defined after Definition 3.5.*

Remark. In the proof of Theorem 3.1, it is sufficient to assume that $L_1 - L_a \in \operatorname{Im} S_1^\top \oplus \operatorname{Im} S_a^\top$ for $a \in [k]$, which is weaker than the strong local observability, though it is still sometimes stronger than the local observability condition.

The proof of Theorem 3.1 is given in Section 3.6.5. This result is the first problem-dependent bound of TS for PM, which also becomes the first logarithmic regret bound of TS for linear bandits.

The norm of $z_{a,b}$ in Λ intuitively indicates the difficulty of the problem. Whereas we can estimate $(S_a p, S_b p)$ with noise through taking actions a and b, the actual interest is the gap of the losses $p^\top(L_a - L_b) = (S_a p, S_b p)^\top z_{a,b}$. Thus, if $\|z_{a,b}\|$ is large, the gap estimation becomes difficult since the noise is enhanced through $z_{a,b}$.

Unfortunately, the derived bound in Theorem 3.1 has quadratic dependence on k, which does not seem tight. This quadratic dependence comes from the difficulty of the *expected* regret analysis. In general, we evaluate the regret before and after the convergence of the statistics separately. Whereas the latter one usually becomes dominant, the main difficulty comes from the analysis of the former one, which might become large with small probability (Agrawal and Goyal, 2012; Kaufmann et al., 2012a; Agrawal and Goyal, 2013a).

In our analysis, we were not able to bound the former one within a non-dominant order, though it is still logarithmic in T. In fact, our analysis shows that the regret after convergence is $O(\sum_{a\neq 1} \Delta_a \frac{m}{\Lambda^2} \log T)$ as shown in Lemma 3.10 in Section 3.6.5, which will become the regret with high probability. In particular, if we consider the classic bandit problem as a PM game, we can confirm that the derived bound after convergence becomes the best possible bound

$$O\left(\sum_{a\neq 1} \frac{\log T}{\Delta_a}\right)$$

by considering Λ_a depending on each suboptimal arm a as the difficulty measure instead of Λ. Still, deriving a regret bound for the term before convergence within an non-dominant order is an important future work.

3.4.2 Technical Difficulties of the Analysis

The main difficulty of this regret analysis is that PM requires consideration of the statistics of *all* actions when the number of selections $N_a(t)$ of some action a is evaluated. This is in stark contrast to the analysis of the classic bandit problems, where it becomes sufficient to evaluate statistics of action a and optimal action 1. This makes the analysis remarkably complicated in PM, where we need to separately consider the randomness caused by the feedback and TS.

To overcome this difficulty, we handle the effect of actions of no interest in two different novel ways depending on each decomposed regret. The first one is to evaluate the worst-case effect of these actions based on an argument (Lemma 3.4) related to the law of the iterated logarithm (LIL), which is sometimes used in the best-arm identification literature to improve the performance (Jamieson et al., 2014). The second one is to bound the action-selection probability of TS using an argument of (super-)martingale (Theorem 3.4), which is of independent interest. Whereas such a technique is often used for the construction of confidence bounds (Abbasi-Yadkori et al., 2011), we reveal that it is also useful for evaluation of the regret of TS.

We only focused on the Gaussian noise $\epsilon_t \sim \mathcal{N}(0, I_d)$, rather than the more general sub-Gaussian noise. This restriction to the Gaussian noise comes from the essential difficulty of the problem-dependent analysis of TS, where lower bounds for some probabilities are needed whereas the sub-Gaussian assumption is suited for obtaining upper bounds. To the best of our knowledge, the problem-dependent regret analysis for TS on the sub-Gaussian case has never been investigated even for the multi-armed bandit setting, which is quite simple compared to that of PM. In the literature of the problem-dependent regret analysis, the noise distribution is restricted to distributions with explicitly given forms, *e.g.,* Bernoulli, Gaussian, or more generally a one-dimensional canonical exponential family (Kaufmann et al., 2012a; Agrawal and Goyal, 2013a; Korda et al., 2013). Their analysis relies on the specific characteristic of the distribution to bound the problem-dependent regret.

3.5 Experiments

In this section, we numerically compare the performance of TSPM and TSPM-Gaussian against existing methods, which are RandomPM (the algorithm which selects action randomly), FeedExp3 (Piccolboni and Schindelhauer, 2001), and BPM-TS (Vanchinathan et al., 2014). Recently, Lattimore and Szepesvári (2019a) considered the sampling-based algorithm called Mario sampling for easy games. Mario sampling coincides with TS (except for the difference between pseudo-regret and regret with known time horizon) mentioned in the last section when any pair of actions is a neighbor. As shown in Section 3.6.6, this property is indeed satisfied for dp-easy games defined in the following. Therefore, the performance is essentially the same between TSPM with $R = 1$ and Mario sampling. To compare the performance, we consider a dynamic pricing problem introduced in Chapter 2, which is a typical example of PM games. We conducted experiments on the discrete setting because the experiments for PM has mainly focused on the discrete setting.

In the dynamic pricing game, the player corresponds to a seller, and the opponent corresponds to a buyer. At each round, the seller sells an item for a specific price A_t, and the buyer comes with an evaluation price x_t for the item, where the selling price and the evaluation price correspond to the action and outcome, respectively. The buyer buys the item if the selling price A_t is smaller than or equal to x_t and not otherwise. The seller can only know if the buyer bought the item (denoted as feedback 0) or did not buy the item (denoted as 1). The seller aims to minimize the cumulative "loss", and there are two types of definitions for the loss, where each induced game falls into the easy and hard games. We call them *dp-easy* and *dp-hard* games, respectively.

In both cases, the seller incurs the constant loss $c > 0$ when the item is not bought due to the loss of opportunity to sell the item. In contrast, when the item is not bought, the loss incurred to the seller is different between these settings. The seller in the dp-easy game *does not* take the buyer's evaluation price into account. In other words, the seller gains the selling price A_t as a reward (equivalently incurs $-A_t$ as a loss). Therefore, the loss for the selling price A_t and the evaluation x_t is

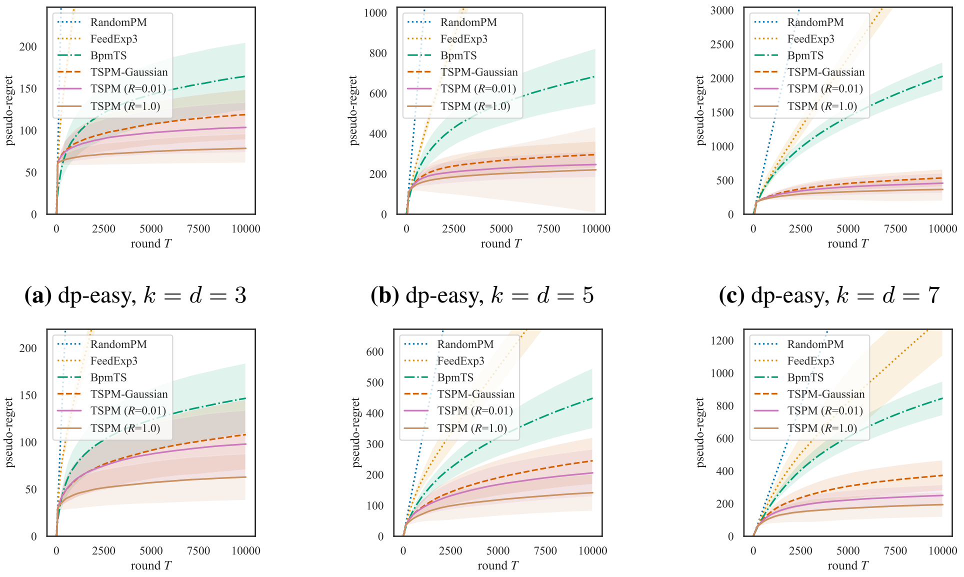
$$\mathcal{L}_{A_t x_t} = -A_t \mathbb{1}[A_t \leq x_t] + c\mathbb{1}[A_t > x_t] \,.$$

This setting can be regarded as a generalized version of the online posted price mechanism, which was addressed in, *e.g.,* Blum et al. (2004) and Cesa-Bianchi et al. (2006), and an example of strongly locally observable games.

On the other hand, the seller in dp-hard game *does* take the buyer's evaluation price into account when the item is bought. In other words, the seller incurs the difference between the opponent evaluation and the selling price $x_t - A_t$ as a loss because the seller could have made more profit if the seller had sold the item at the price x_t. Therefore, the loss incurred at time t is

$$\mathcal{L}_{A_t x_t} = (x_t - A_t)\mathbb{1}[A_t \leq x_t] + c\mathbb{1}[A_t > x_t] \,.$$

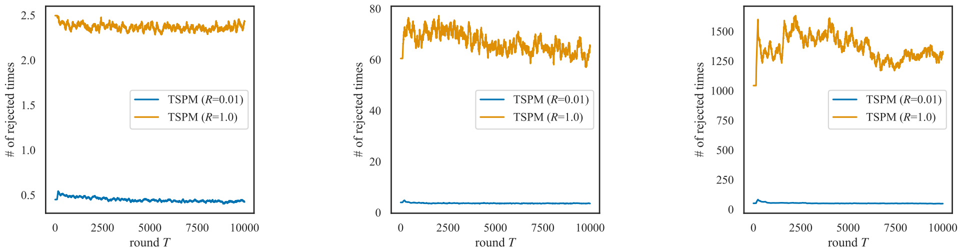
This setting is also addressed in Cesa-Bianchi et al. (2006), and belongs to the class of hard games. Note that our algorithm can also be applied to a hard game, though there is no theoretical guarantee.



(**a**) dp-easy, $k = d = 3$ (**b**) dp-easy, $k = d = 5$ (**c**) dp-easy, $k = d = 7$

(**d**) dp-hard, $k = d = 3$ (**e**) dp-hard, $k = d = 5$ (**f**) dp-hard, $k = d = 7$

Figure 3.2: Regret-round plots of algorithms. The solid lines indicate the average over 100 independent trials. The thin fillings are the standard error.



(**a**) dp-easy, $k = d = 3$ (**b**) dp-easy, $k = d = 5$ (**c**) dp-easy, $k = d = 7$

Figure 3.3: The number of rejected times by the accept-reject sampling. The solid lines indicate the average over 100 independent trials after taking moving average with window size 100.

Setup In both dp-easy and dp-hard games, we fixed $k = d \in \{3, 5, 7\}$ and $c = 2$. We fixed the time horizon T to 10000 and simulated 100 times. For FeedExp3 and BPM-TS, the setup of hyperparameters follows their original papers. For TSPM, we set $\lambda = 0.001$, and R was selected from $\{0.01, 1.0\}$. Here, recall that TSPM with $R = 1$ and $R = 0$ correspond to the exact sampling and TSPM-Gaussian, respectively, and a smaller value of R gives the higher acceptance probability in the accept-reject sampling. Therefore, using small R makes the algorithm time-efficient, although it can worsen the performance since it over-explores the tail of the posterior distributions. To stabilize sampling from the proposal distribution in Algorithm 3.3, we used an initialization that takes each action $n = 10m$ times. The detailed settings of the experiments with more results are given in Section 3.6.7.

Results Figure 3.2 is the empirical comparison of the proposed algorithms against the benchmark methods. This result shows that, in all cases, the TSPM with exact sampling gives the best performance. TSPM-Gaussian also outperforms BPM-TS even though both of them use Gaussian distributions as posteriors. Besides, the experimental results suggest that our algorithm performs reasonably well even for a hard game. It can be observed that the proposed methods outperform BPM-TS more significantly for a larger number of outcomes. Further discussion for this observation is given in Section 3.6.3.

Figure 3.3 shows the number of rejections at each time step in the accept-reject sampling. We counted the number of times that either Line 4 in Algorithm 3.2 or Line 4 in Algorithm 3.3 was not satisfied. In the accept-reject sampling, it is desirable that the frequency of rejection does not increase as the time-step t and does not increase rapidly with the number of outcomes. We can see that the former one is indeed satisfied. For the latter property, the frequency of rejection becomes unfortunately large when exact sampling ($R = 1$) is conducted. Still, we can substantially improve this frequency by setting R to be a small value or zero, which still keeps regret tremendously better than that of BPM with almost the same time-efficiency as BPM-TS.

3.6 Deferred Discussion and Proofs

3.6.1 Posterior Distribution and Proposal Distribution in Section 3.3

In this section, we discuss the representation of the posterior distribution and its relation with the proposal distribution.

Proposition 3.3. *$F_t(p)$ in (3.1) is proportional to the posterior distribution of the opponent's strategy, and $F_t(p) \le G_t(p)$ for all $p \in \mathcal{P}_d$.*

Proof. The posterior distribution of the opponent's strategy parameter $\pi\big(p \mid \{A_s, \sigma_s\}_{s=1}^t\big)$ is rewritten as

$$\begin{aligned}\pi\big(p \mid \{A_s, \sigma_s\}_{s=1}^t\big) &\propto \pi\big(p, \{A_s, \sigma_s\}_{s=1}^t\big)\\ &\propto \pi(p) \prod_{s=1}^{t} \mathbb{P}\{\sigma_s \mid A_s, p\}\\ &= \pi(p) \prod_{a=1}^{k} \prod_{y=1}^{A} (S_{a,y} p)^{n_{ay}}\\ &\propto \pi(p) \prod_{a=1}^{k} \exp\Big(-n_a D\left(q_a^{(t)} \| S_a p\right)\Big),\end{aligned}$$

where $S_{a,y}$ is the y-th row of the signal matrix $S_a \in \{0,1\}^{m \times d}$, and note that $q_a^{(t)}$ is the empirical feedback distribution of action a at time t, that is, $q_a^{(t)} = [n_{a1}/n_a, \dots, n_{am}/n_a]^\top \in \mathcal{P}_m$ for $n_{ay} = \sum_{s=1}^t \mathbb{1}[A_s = a, \sigma_s = y]$ and $n_a = \sum_{y=1}^m n_{ay}$.

Next, we show that $F_t(p) \le G_t(p)$ holds for all $p \in \mathcal{P}_d$. Using Pinsker's inequality, the unnormalized posterior distribution $F_t(p)$ can be bounded from above as

$$\begin{aligned}F_t(p) &= \pi(p) \prod_{a=1}^{k} \exp\Big(-n_a D\left(q_a^{(t)} \| S_a p\right)\Big)\\ &\le \pi(p) \prod_{a=1}^{k} \exp\Big(-\frac{1}{2} n_a \|q_a^{(t)} - S_a p\|_1^2\Big) \quad \text{(by Pinsker's inequality)}\\ &= \pi(p) \exp\Big(-\frac{1}{2} \sum_{a=1}^{k} n_a \|q_a^{(t)} - S_a p\|_1^2\Big)\\ &\le \pi(p) \exp\Big(-\frac{1}{2} \sum_{a=1}^{k} n_a \|q_a^{(t)} - S_a p\|^2\Big) \quad \Big(\text{by } \|q_a^{(t)} - S_a p\|_1 \ge \|q_a^{(t)} - S_a p\|\Big)\\ &= G_t(p)\,.\end{aligned}$$

□

Remark. The unnormalized density $G_t(p)$ is indeed Gaussian. Recalling that B_t and b_t are defined in (3.3) as

$$B_t = \sum_{a=1}^{k} n_a S_a^\top S_a = \sum_{s=1}^{t} S_{A_s}^\top S_{A_s} = B_{t-1} + S_{A_t}^\top S_{A_t},\; b_t = \sum_{a=1}^{k} n_a S_a^\top q_a^{(t)} = b_{t-1} + S_{A_t}^\top e_{\sigma_t}\,,$$

we have

$$\begin{aligned}
\sum_{a=1}^{k} n_a \|q_a^{(t)} - S_a p\|^2 &= \sum_{a=1}^{k} n_a (q_a^{(t)} - S_a p)^\top (q_a^{(t)} - S_a p) \\
&= p^\top \underbrace{\Big(\sum_{a=1}^{k} n_a S_a^\top S_a\Big)}_{B_t} p, -2 \underbrace{\Big(\sum_{a=1}^{k} n_a S_a^\top q_a^{(t)}\Big)^\top}_{b_t} p + \underbrace{\sum_{a=1}^{k} n_a \|q_a^{(t)}\|^2}_{c_t} \\
&= p^\top B_t p - 2 b_t^\top p + c_t \\
&= (p - B_t^{-1} b_t)^\top B_t (p - B_t^{-1} b_t) + c_t - b_t^\top B_t^{-1} b_t\,.
\end{aligned}$$

Therefore, we have

$$\exp\Big(-\frac{1}{2}\sum_{a=1}^{k} n_a \|q_a^{(t)} - S_a p\|^2\Big) \propto \exp\Big(-\frac{1}{2}(p - B_t^{-1} b_t)^\top B_t (p - B_t^{-1} b_t)\Big)\,.$$

3.6.2 Proof of Proposition 3.2

We will see that the procedure of sampling $\tilde{p}_t$ from $g_t(p)$ and Algorithm 3.3 are equivalent. First, we derive the Gaussian density of $g_t(p)$ projected onto $\{p \in \mathbb{R}^d : \sum_{i=1}^d p_i = 1\}$.

For simplicity, we omit the subscript t and write, *e.g.,* B instead of B_t. We define $p = [{p^{(\alpha)}}^\top, p_d]^\top \in \mathbb{R}^{d-1} \times \mathbb{R}$. Let $h = B^{-1} b$, and define $h = [{h^{(\alpha)}}^\top, h_d]^\top \in \mathbb{R}^{d-1} \times \mathbb{R}$. Let $B = \begin{bmatrix} C & g \\ g^\top & f \end{bmatrix}$, where $C \in \mathbb{R}^{d-1 \times d-1}, d \in \mathbb{R}^{d-1}$, and $f \in \mathbb{R}$. Also, let $b = [{b^{(\alpha)}}^\top, b^{(d)}]^\top \in \mathbb{R}^{d-1} \times \mathbb{R}$.

Using the decomposition

$$(p - B^{-1} b)^\top B (p - B^{-1} b) = \underbrace{p^\top B p}_{\text{(a)}} - 2 \underbrace{h^\top B p}_{\text{(b)}} + h^\top B h\,.$$

We then rewrite each term by restricting the domain of p so that it satisfies the condition $\sum_{i=1}^d p_i = 1$. Now the first term (a) is rewritten as

$$\begin{aligned}
\text{(a)} &= {p^{(\alpha)}}^\top C p^{(\alpha)} + 2 {p^{(\alpha)}}^\top g p_d + f p_d^2 \\
&= {p^{(\alpha)}}^\top C p^{(\alpha)} + 2 \underbrace{{p^{(\alpha)}}^\top g \Big(1 - \sum_{i=1}^{d-1} p_i\Big)}_{\text{(a1)}} + f \underbrace{\Big(1 - \sum_{i=1}^{d-1} p_i\Big)^2}_{\text{(a2)}}\,.
\end{aligned}$$

The term (a1) is rewritten as

$$\begin{aligned}
\text{(a1)} &= {p^{(\alpha)}}^\top g - {p^{(\alpha)}}^\top g \sum_{i=1}^{d-1} p_i \\
&= {p^{(\alpha)}}^\top g - {p^{(\alpha)}}^\top g \mathbf{1}_{d-1}^\top p^{(\alpha)} \\
&= {p^{(\alpha)}}^\top g - {p^{(\alpha)}}^\top D p^{(\alpha)} \quad \Big(D = \frac{1}{2}\left(g \mathbf{1}_{d-1}^\top + \mathbf{1}_{d-1} g^\top\right)\Big)\,,
\end{aligned}$$

and the term (a2) is rewritten as

$$\begin{aligned}
\text{(a2)} &= \Big(1 - \sum_{i=1}^{d-1} p_i\Big)^2 \\
&= 1 - 2\sum_{i=1}^{d-1} p_i + \Big(\sum_{i=1}^{d-1} p_i\Big)^2 \\
&= 1 - 2\mathbf{1}_{d-1}^\top p^{(\alpha)} + {p^{(\alpha)}}^\top \mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top p^{(\alpha)}\,.
\end{aligned}$$

Therefore,

$$\text{(a)} = {p^{(\alpha)}}^\top \underbrace{(C - 2D + f\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top)}_{\tilde{B}} p^{(\alpha)} - 2(f\mathbf{1}_{d-1} - g)^\top p^{(\alpha)} + f\,.$$

With regard to the term (b), we have

$$\begin{aligned}
\text{(b)} &= b^\top p \\
&= {b^{(\alpha)}}^\top {p^{(\alpha)}}^\top + b^{(d)} p_d \\
&= (b^{(\alpha)} - b^{(d)}\mathbf{1}_{d-1})^\top p^{(\alpha)} + b^{(d)}\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&(p - B^{-1}b)^\top B(p - B^{-1}b) \\
&= {p^{(\alpha)}}^\top \tilde{B} p^{(\alpha)} - 2(\underbrace{f\mathbf{1}_{d-1} - g + b^{(\alpha)} - b^{(d)}\mathbf{1}_{d-1}}_{\tilde{b}})^\top p^{(\alpha)} + f - 2b^{(d)} + h^\top B h \\
&= (p^{(\alpha)} - \tilde{B}^{-1}\tilde{b})^\top \tilde{B}(p^{(\alpha)} - \tilde{B}^{-1}\tilde{b}) + f - 2b^{(d)} - \tilde{b}^\top \tilde{B}^{-1}\tilde{b} + b^\top B^{-1} b\,,
\end{aligned}$$

where the last equality follows by $h^\top B h = b^\top B^{-1} b$. From the above argument, the density $\mathcal{N}(\tilde{B}^{-1}b,\ \tilde{B}^{-1})$ is the Gaussian distribution of $g_t(p)$ on $\{p \in \mathbb{R}^d\colon\ \sum_{i=1}^d p_i = 1\}$. Therefore, the $p = [{p^{(\alpha)}}^\top,\ 1 - \sum_{i=1}^{d-1}(p^{(\alpha)})_i]^\top$ for $p^{(\alpha)} \sim \mathcal{N}(\tilde{B}^{-1}b,\ \tilde{B}^{-1})$ is supported over $\{p \in \mathbb{R}^d : \sum_{i=1}^d p_i = 1\}$.

If the sample $p^{(\alpha)}$ from $\mathcal{N}(\tilde{B}^{-1}b,\ \tilde{B}^{-1})$ is in $\mathcal{P}_{d-1}$, then we can obtain the last element $p^{(d)}$ by $p^{(d)} = 1 - \sum_{i=1}^{d-1}(p^{(\alpha)})_i$. Otherwise, the probability that $p^{(\alpha)}$ is the first $(d-1)$ elements of the sample from $g_t(p)$ is zero, and hence, $[{p^{(\alpha)}}^\top, p^{(d)}]^\top$ cannot be a sample from $g_t(p)$. Therefore, sampling $\tilde{p}_t$ from $g_t(p)$ and Algorithm 3.3 are equivalent.

3.6.3 Relation between TSPM-Gaussian and BPM-TS

In this section, we discuss the relation between TSPM-Gaussian and BPM-TS (Vanchinathan et al., 2014).

Underlying Feedback Structure Here, we discuss the underlying feedback structure behind TSPM-Gaussian and BPM-TS.

We first consider the underlying feedback structure behind BPM-TS. In the following, we see that the feedback structure

$$y(t) = S_{A_t} p + S_{A_t}\epsilon\,,\ \epsilon \sim \mathcal{N}(0, I_d)$$

induces the posterior distribution in BPM-TS. Under this feedback structure, we have $y(t) \sim \mathcal{N}(S_{A_t}p, S_{A_t}S_{A_t}^\top)$.

When we take the prior distribution $\pi(p)$ as $\mathcal{N}(0, \sigma_0^2 I_d)$, the posterior distribution for the opponent's strategy parameter can be written as

$$
\begin{aligned}
&\pi\big(p \mid \{A_s, y(s)\}_{s=1}^t\big) \\
&\propto \pi(p) \prod_{s=1}^t \pi(y(s) \mid A_s,\, p) \\
&= \pi(p) \prod_{s=1}^t \mathbb{P}_{y \sim \mathcal{N}(S_{A_s} p, S_{A_s} S_{A_s}^\top)} \{y = y(s)\} \\
&= \exp\left(-\frac{p^\top p}{2\sigma_0^2}\right) \prod_{s=1}^t \exp\left(-\frac{1}{2}(y(s) - S_{A_s} p)^\top (S_{A_s} S_{A_s}^\top)^{-1} (y(s) - S_{A_s} p)\right) \\
&= \exp\Bigg(-\frac{1}{2}\left(p^\top \left(\frac{1}{\sigma_0^2} I_d + \sum_{s=1}^t S_{A_s}^\top (S_{A_s} S_{A_s}^\top)^{-1} S_{A_s}\right) p\right) \\
&\qquad\qquad - 2\left(\sum_{s=1}^t y(s)^\top (S_{A_s} S_{A_s}^\top)^{-1} S_{A_s} p\right) + (\text{a term independent of } p)\Bigg) \\
&\propto \exp\left(-\frac{1}{2}(p^\top B_t^{\mathrm{BPM}} p - 2 b_t^{\mathrm{BPM}^\top} p)\right) \\
&\propto \exp\left(-\frac{1}{2}(p - B_t^{\mathrm{BPM}^{-1}} b_t^{\mathrm{BPM}})^\top B_t^{\mathrm{BPM}} (p - B_t^{\mathrm{BPM}^{-1}} b_t^{\mathrm{BPM}})\right),
\end{aligned}
$$

where

$$
\begin{aligned}
B_t^{\mathrm{BPM}} &= \frac{1}{\sigma_0^2} I_d + \sum_{s=1}^t S_{A_s}^\top (S_{A_s} S_{A_s}^\top)^{-1} S_{A_s} = B_{t-1}^{\mathrm{BPM}} + S_{A_t}^\top (S_{A_t} S_{A_t}^\top)^{-1} S_{A_t}\,, \\
b_t^{\mathrm{BPM}} &= \sum_{s=1}^t S_{A_s}^\top (S_{A_s} S_{A_s}^\top)^{-1} y(s) = b_{t-1}^{\mathrm{BPM}} + S_{A_t}^\top (S_{A_t} S_{A_t}^\top)^{-1} y(t)\,.
\end{aligned}
$$

Therefore, the posterior distribution $\pi\big(p \mid \{A_s, y(s)\}_{s=1}^t\big)$ is

$$
\frac{1}{\sqrt{(2\pi)^d |B_t^{\mathrm{BPM}^{-1}}|}} \exp\left(-\frac{1}{2}(p - B_t^{\mathrm{BPM}^{-1}} b_t^{\mathrm{BPM}})^\top B_t^{\mathrm{BPM}} (p - B_t^{\mathrm{BPM}^{-1}} b_t^{\mathrm{BPM}})\right).
$$

and this distribution indeed corresponds to the posterior distribution in BPM-TS (Vanchinathan et al., 2014) with $B_t^{\mathrm{BPM}} = \Sigma_t^{-1}$.

Using the same argument, we can confirm that the feedback structure

$$
y_t = S_a p + \epsilon\,,\ \epsilon \sim \mathcal{N}(0, I_d)\,.
$$

induces

$$
\bar{g}_t(p) \coloneqq \frac{1}{\sqrt{(2\pi)^d |B_t^{-1}|}} \exp\left(-\frac{1}{2}\left\|p - B_t^{-1} b_t\right\|_{B_t}^2\right),
$$

which corresponds to the posterior distribution for TSPM in linear partial monitoring.

Covariances in TSPM-Gaussian and BPM-TS In the linear partial monitoring, TSPM assumes noise with covariance I_d, which is compatible with the fact that the discrete setting can be regarded as linear PM with I_d-sub-Gaussian noise. On the other hand, BPM-TS assumes covariance $S_a S_a^\top$, and in general $I_d \preceq S_a S_a^\top$ holds. Therefore, BPM-TS assumes unnecessarily larger covariance, which makes learning slow down.

3.6.4 Preliminaries for Regret Analysis

In this section, we give some technical lemmas, which are used for the derivation of the regret bound in Section 3.6.5. Here, we write $X \succeq Y$ to denote $X - Y \succeq 0$. For $a, b \in \mathbb{R}$, let $a \wedge b$ be a if $a \le b$ otherwise b, and $a \vee b$ be b if $a \le b$ otherwise a. We use $h(a) := \mathbb{P}_{X \sim \chi^2_d} \{X \ge a\}$ to evaluate the behavior of the posterior samples, where χ^2_d is the chi-squared distribution with d degree of freedom.

3.6.4.1 Basic Lemmas

Fact 3.2 (Moment generating function of squared-Gaussian distribution)**.** Let X be the random variable following the standard normal distribution. Then, the moment generating function of X^2 is $\mathbb{E}\big[\exp(\xi X^2)\big] = (1-2\xi)^{-1/2}$ for $\xi < 1/2$.

Lemma 3.1 (Chernoff bound for chi-squared random variable)**.** *Let X be the random variable following the chi-squared distribution with k degree of freedom. Then, for any $x \ge 0$ and $0 \le \xi < 1/2$,*

$$\mathbb{P}\{X \ge a\} \le \mathrm{e}^{-\xi a}(1-2\xi)^{-\frac{k}{2}}.$$

Proof. By Markov's inequality, the LHS can be bounded as

$$\begin{aligned}
\mathbb{P}\{X \ge x\} &= \mathbb{P}\left\{\sum_{i=1}^{k} X_i^2 \ge x\right\} \; (X_1, \ldots, X_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)) \\
&= \mathbb{P}\left\{\exp\Big(\xi \sum_{i=1}^{k} X_i^2\Big) \ge \exp(\xi a)\right\} \\
&\le \mathrm{e}^{-\xi a}\Big(\mathbb{E}\Big[\mathrm{e}^{\xi X_1^2}\Big]\Big)^k \quad \text{(by Markov's inequality)} \\
&= \mathrm{e}^{-\xi a}(1-2\xi)^{-\frac{k}{2}} \quad \text{(by Fact 3.2)} \; ,
\end{aligned}$$

which completes the proof. □

3.6.4.2 Property of Strong Local Observability

Recall that $\Delta_a = (L_a - L_1)^\top p^* > 0$ for $a \in [k]$, which is the difference of the expected loss of actions a and 1. Using this define

$$\epsilon := \left(\frac{1}{2\sqrt{m}} \min_{a \ne 1} \frac{\Delta_a}{\|z_{1,a}\|}\right) \wedge \left(\min_{p \in \mathcal{C}_1^c} \frac{4}{3}\|p - p^*\|\right) \; , \tag{3.5}$$

which is used throughout the proof of this section and Section 3.6.5. The following lemma provides the key property of the strong local observability condition.

Lemma 3.2. *For any partial monitoring game with strong local observability and $p \in \mathbb{R}^d$, any of the conditions 1–3 in the following is not satisfied:*

1. $L_1^\top p > L_a^\top p$ *(Worse action a looks better under p.)*
2. $\|S_1 p - S_1 p^*\| \le \epsilon$
3. $\|S_a p - S_a p^*\| \le \epsilon$.

Proof. We prove this by contradiction. Assume that there exists $p \in \mathbb{R}^d$ such that conditions 1–3 are simultaneously satisfied.

Now, by conditions 2 and 3, we have

$$
\begin{aligned}
|S_1 p - S_1 p^*| &\preceq \epsilon \mathbf{1}_m \,, \\
|S_a p - S_a p^*| &\preceq \epsilon \mathbf{1}_m \,.
\end{aligned}
$$

Here, $|\cdot|$ is the element-wise absolute value and $\preceq$ means that the inequality $\leq$ holds for each element. Therefore,

$$
\left| \begin{pmatrix} S_1 \\ S_a \end{pmatrix} (p - p^*) \right| \preceq \epsilon \mathbf{1}_{2m} \,. \tag{3.6}
$$

On the other hand, by the strong local observability condition, for any $k \neq 1$, there exists $z_{1,a} \neq 0 \in \mathbb{R}^{2m}$ such that

$$
(L_1 - L_a)^\top = z_{1,a}^\top \begin{pmatrix} S_1 \\ S_a \end{pmatrix} . \tag{3.7}
$$

Now, we have

$$
\begin{aligned}
& z_{1,a}^\top \begin{pmatrix} S_1 \\ S_a \end{pmatrix} (p - p^*) \\
& \leq \|z_{1,a}\| \left\| \begin{pmatrix} S_1 \\ S_a \end{pmatrix} (p - p^*) \right\| \quad \text{(by Cauchy-Schwarz inequality)} \\
& \leq \sqrt{2m} \epsilon \|z_{1,a}\| \quad \text{(by (3.6))} \,,
\end{aligned} \tag{3.8}
$$

and

$$
\begin{aligned}
& z_{1,a}^\top \begin{pmatrix} S_1 \\ S_a \end{pmatrix} (p - p^*) \\
& = (L_1 - L_a)^\top (p - p^*) \quad \text{(by (3.7))} \\
& = (L_1 - L_a)^\top p + (L_a - L_1)^\top p^* \\
& \geq \Delta_k \quad \text{(by Condition 1 \& def. of } \Delta_a) \,.
\end{aligned} \tag{3.9}
$$

Therefore, from (3.8) and (3.9), we have

$$
\Delta_a \leq \sqrt{2m} \epsilon \|z_{1,a}\| \,.
$$

This inequality does not hold for all $a \neq 1$ for the predefined value of ϵ, since we have

$$
\epsilon \leq \frac{1}{2\sqrt{m}} \min_{a \neq 1} \frac{\Delta_a}{\|z_{1,a}\|} \,.
$$

Therefore, the proof is completed by contradiction. □

Remark. The similar result holds when the optimal action 1 is replaced with action $a \neq b$ such that $\Delta_{a,b} := (L_a - L_b)^\top p^* > 0$ by taking ϵ satisfying

$$
\epsilon \leq \frac{1}{2\sqrt{m}} \min_{a \neq b : \Delta_{a,b} > 0} \frac{\Delta_{a,b}}{\|z_{a,b}\|} \,.
$$

From Lemma 3.2, we have the following corollary.

Corollary 3.1. *For any $p \in \mathbb{R}^d$ satisfying $p \in \mathcal{C}_a$ and $\|S_1 p - S_1 p^*\| \leq \epsilon$, we have*

$$
\|S_a p - S_a p^*\| > \epsilon \,.
$$

Proof. Note that $p \in \mathcal{C}_a$ is equivalent to $(L_1 - L_a)^\top p^* > 0$ for any $a \neq 1$. Therefore, the desired result directly follows from Lemma 3.2. □

The next lemma is the property of Mahalanobis distance corresponding to $\bar{g}_t(p)$.

Lemma 3.3. *Define* $\mathcal{T}_a = \{p \in \mathbb{R}^d \colon \|S_a p - S_a p^*\| > \epsilon\}$. *Assume that* $N_a(t) \geq n_a$, $\|S_a \widehat{p}_t - S_a p^*\| \leq \epsilon/4$. *Then, for any* $0 \leq \xi < 1/2$

$$h\Big(\inf_{p \in \mathcal{T}_a} \|B_t^{1/2}(p - \widehat{p}_t)\|^2\Big) \leq \exp\Big(-\frac{9}{16}\xi\epsilon^2 n_a\Big)(1-2\xi)^{-d/2}\,.$$

Proof. To bound the LHS of the above inequality, we bound $\|B_t^{1/2}(p - \widehat{p}_t)\|^2$ from below for $p \in \mathcal{T}_a$. Using the triangle inequality and the assumptions, we have

$$\begin{aligned}\|S_a(p - \widehat{p}_t)\| &\geq \|S_a p - S_a p^*\| - \|S_a \widehat{p}_t - S_a p^*\| \\ &> \epsilon - \epsilon/4 > 0\,. \end{aligned} \tag{3.10}$$

Therefore, we have

$$\begin{aligned}\|B_t^{1/2}(p - \widehat{p}_t)\|^2 &\geq \sum_{b \in [k]} N_b(t) \|S_b(p - \widehat{p}_t)\|^2 \quad \text{(by def. of } B_t) \\ &\geq n_a \|S_a(p - \widehat{p}_t)\|^2 \quad (N_a(t) \geq n_a) \\ &> \frac{9}{16}\epsilon^2 n_a \quad \text{(by (3.10))}\,.\end{aligned}$$

By the Chernoff bound for a chi-squared random variable in Lemma 3.1, we now have

$$h(x) \leq \mathrm{e}^{-\xi x}(1-2\xi)^{-d/2}\,,$$

for any $a \geq 0$ and $0 \leq \xi < 1/2$. Hence, using the fact that $\|B_t^{1/2}(p - \widehat{p}_t)\|^2$ follows the chi-squared distribution with d degree of freedom, we have

$$\begin{aligned}h\Big(\inf_{p \in \mathcal{T}_a} \|B_t^{1/2}(p - \widehat{p}_t)\|^2\Big) &\leq h\Big(\frac{9}{16}\epsilon^2 n_a\Big) \\ &\leq \exp\Big(-\frac{9}{16}\xi\epsilon^2 n_a\Big)(1-2\xi)^{-d/2}\,,\end{aligned}$$

which completes the proof. □

3.6.4.3 Statistics of Uninterested Actions

For any $a \neq i$ and $n_a \in [T]$, define

$$\begin{aligned}Z_{n_a} &:= n_a \|q_{a,n_a} - S_a p^*\|^2\,, \\ Z_{\backslash i} &:= \sum_{a \neq i} \max_{n_a \in [T]} Z_{n_a}\,.\end{aligned}$$

In this section, we bound $\mathbb{E}\big[Z_{\backslash i}\big]$ from above. Note that $Z_{\backslash i}$ is independent of the randomness of Thompson sampling.

Lemma 3.4 (Upper bound for the expectation of $Z_{\backslash i}$).

$$\mathbb{E}\big[Z_{\backslash i}\big] \leq 4k\Big(\log T + \frac{m}{2}\log 2 + 1\Big)\,.$$

Proof. Recall that in linear partial monitoring, the feedback $y(t) \in \mathbb{R}^m$ for action a is given by

$$y_t = S_a p^* + \epsilon\,,\ \epsilon \sim \mathcal{N}(0, I_m)$$

at round $t \in [T]$, Therefore, $y(t) - S_k p^* \sim \mathcal{N}(0, I_m)$. Since $q_{a,n_a} = \frac{1}{n_a}\sum_{s\in[T]:A_s=a} y(s)$ for any $n_a \in [T]$, we have

$$q_{a,n_a} - S_a p^* = \frac{1}{n_a}\sum_{s\in[T]:A_s=a}(y(s) - S_a p^*) \sim \mathcal{N}(0, I_m/n_a)\,.$$

Therefore,

$$\sqrt{n_a}(q_{a,n_a} - S_a p^*) \sim \mathcal{N}(0, I_m)\,,$$

and thus

$$n_k\|q_{a,n_a} - S_a p^*\|^2 = \|\sqrt{n_a}(q_{a,n_a} - S_a p^*)\|^2 \sim \chi^2_m\,.$$

Therefore, for any $0 \le \xi < 1/2$,

$$\begin{aligned}
\mathbb{E}\left[\max_{n_a\in[T]} Z_{n_a}\right] &= \int_0^\infty \mathbb{P}\left\{\max_{n_a\in[T]} Z_{n_a} \ge x\right\}\mathrm{d}x\\
&\le \int_0^\infty [1 \wedge T\cdot\mathbb{P}\{Z_1 \ge x\}]\,\mathrm{d}x \quad \text{(by the union bound)}\\
&\le \int_0^\infty \left[1 \wedge T\cdot \mathrm{e}^{-\xi x}(1-2\xi)^{-\frac{m}{2}}\right]\mathrm{d}x \quad \big(\text{by } Z_1 \sim \chi^2_m \text{ and Lemma 3.1}\big)\\
&= \int_0^{x^*}\mathrm{d}x + \int_{x^*}^\infty T\cdot \mathrm{e}^{-\xi x}(1-2\xi)^{-\frac{m}{2}}\,\mathrm{d}x\\
&\le x^* + T\cdot\int_{x^*}^\infty \mathrm{e}^{-\xi x}(1-2\xi)^{-\frac{m}{2}}\,\mathrm{d}x\\
&= x^* + T(1-2\xi)^{-\frac{m}{2}}\left[-\frac{1}{\xi}\mathrm{e}^{-\xi x}\right]_{x^*}^\infty\\
&= \frac{1}{\xi}\left\{\log T - \frac{m}{2}\log(1-2\xi) + 1\right\},
\end{aligned}$$

where $x^* \coloneqq \frac{1}{\xi}\big\{\log T - \frac{m}{2}\log(1-2\xi)\big\}$. Therefore, taking $\xi = 1/4$, we have

$$\begin{aligned}
\mathbb{E}[Z_{\setminus i}] &= \mathbb{E}\left[\sum_{a\ne i}\max_{n_k\in[T]} Z_{n_a}\right]\\
&\le \sum_{a\ne i}\mathbb{E}\left[\max_{n_a\in[T]} Z_{n_a}\right]\\
&\le (k-1)\frac{1}{\xi}\left\{\log T - \frac{m}{2}\log(1-2\xi) + 1\right\}\\
&\le 4k\left(\log T + \frac{m}{2}\log 2 + 1\right),
\end{aligned}$$

which completes the proof. □

3.6.4.4 Mahalanobis Distance Process

Discussions in this section are essentially very similar to Abbasi-Yadkori et al. (2011, Lemma 11), but their results are not directly applicable and we give the full derivation for self-containedness. To maximize the applicability here we only assume sub-Gaussian noise rather than a Gaussian one.

Let ϵ_t be zero-mean 1-sub-Gaussian random variable, which satisfies

$$\mathbb{E}\left[\mathrm{e}^{\lambda^\top \epsilon_t}\right] \le \mathrm{e}^{-\frac{\|\lambda\|^2}{2}}$$

for any $\lambda \in \mathbb{R}^d$.

Lemma 3.5. *For any vector $v \in \mathbb{R}^d$ and positive definite matrix $V \in \mathbb{R}^{d\times d}$ such that $V \succ I$,*

$$\mathbb{E}_{\epsilon_t}\left[\mathrm{e}^{\frac{\|\epsilon_t+v\|^2_{V^{-1}}}{2}}\right] \le \frac{\sqrt{|V|}}{\sqrt{|V-I|}}\mathrm{e}^{\frac{1}{2}v^\top (V-I)^{-1}v}\,.$$

Proof. For any $x \in \mathbb{R}^d$

$$\mathbb{E}_{\lambda\sim\mathcal{N}(0,V^{-1})}\left[\mathrm{e}^{\lambda^\top x}\right] = \mathrm{e}^{\frac{\|x\|^2_{V^{-1}}}{2}}\,.$$

Therefore, by letting $x = \epsilon_t + v$ we see that

$$\mathrm{e}^{\frac{\|\epsilon_t+v\|^2_{V^{-1}}}{2}} = \mathbb{E}_{\lambda\sim\mathcal{N}(0,V^{-1})}\left[\mathrm{e}^{\lambda^\top(\epsilon_t+v)}\right].$$

As a result, by the definition of sub-Gaussian random variables, we have

$$\begin{aligned}
&\mathbb{E}_{\epsilon_t}\left[\mathrm{e}^{\frac{\|\epsilon_t+v\|^2_{V^{-1}}}{2}}\right] = \mathbb{E}_{\lambda\sim\mathcal{N}(0,V^{-1})}\left[\mathbb{E}_{\epsilon_t}\left[\mathrm{e}^{\lambda^\top(\epsilon_t+v)}\right]\right]\\
&= \mathbb{E}_{\lambda\sim\mathcal{N}(0,V^{-1})}\left[\mathrm{e}^{\lambda^\top v}\mathbb{E}_{\epsilon_t}\left[\mathrm{e}^{\lambda^\top \epsilon_t}\right]\right]\\
&\le \mathbb{E}_{\lambda\sim\mathcal{N}(0,V^{-1})}\left[\mathrm{e}^{\lambda^\top v}\mathrm{e}^{\|\lambda\|^2/2}\right]\\
&= \frac{1}{(2\pi)^{d/2}\sqrt{|V^{-1}|}}\int \mathrm{e}^{\lambda^\top v}\mathrm{e}^{\|\lambda\|^2/2}\mathrm{e}^{-\|\lambda\|^2_V/2}\mathrm{d}\lambda\\
&= \frac{1}{(2\pi)^{d/2}\sqrt{|V^{-1}|}}\int \mathrm{e}^{-\frac{1}{2}\left(\lambda^\top(V-I)\lambda-2v^\top\lambda\right)}\mathrm{d}\lambda\\
&= \frac{\sqrt{|V-I|}}{(2\pi)^{d/2}\sqrt{|V^{-1}||V-I|}}\int \mathrm{e}^{-\frac{1}{2}\left((\lambda-(V-I)^{-1}v)^\top(V-I)(\lambda-(V-I)^{-1}v)-v^\top(V-I)^{-1}v\right)}\mathrm{d}\lambda\\
&= \frac{\sqrt{|V|}}{\sqrt{|V-I|}}\mathrm{e}^{\frac{1}{2}v^\top(V-I)^{-1}v}\,.
\end{aligned}$$

□

Lemma 3.6.

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\left(\|\widehat{p}_t - p^*\|^2_{B_t} - \|\widehat{p}_{t-1} - p^*\|^2_{B_{t-1}}\right)\right)\,\middle|\, \widehat{p}_{t-1},\, B_{t-1},\, S_{i(t-1)}\right] \le \sqrt{\frac{|B_t|}{|B_{t-1}|}}\,.$$

Proof. Let $Z_t := -\lambda p^* + \sum_{s=1}^t S_{A_s}^\top \epsilon_s$, and we have

- $B_t = \lambda I + \sum_{s=1}^{t} S_{A_s}^\top S_{A_s}$,
- $b_t = \sum_{s=1}^{t} S_{A_s}^\top y(s) = B_t p^* + Z_t$,
- $\widehat{p}_t = B_t^{-1} b_t = p^* + B_t^{-1} Z_t$.

In the following, we omit the conditioning on $(\widehat{p}_{t-1}, B_{t-1}, S_{A_{t-1}})$ for notational simplicity. We also define $S_t \coloneqq S_{A_t}$ for notational simplicity only in the proof of this lemma.

Let us define $C_t \coloneqq S_t B_{t-1} S_t^\top$ and $d_t \coloneqq S_t B_{t-1}^{-1} Z_{t-1} = S_t(\widehat{p}_t - p^*)$. Then, using the Sherman-Morrison-Woodbury formula we have

$$\begin{aligned}
&\|\widehat{p}_t - p^*\|_{B_t}^2 - \|\widehat{p}_{t-1} - p^*\|_{B_{t-1}}^2 \\
&= Z_t^\top B_t^{-1} Z_t - Z_{t-1}^\top B_{t-1}^{-1} Z_{t-1} \\
&= (Z_{t-1}^\top + \epsilon_t^\top S_t)(B_{t-1}^{-1} - B_{t-1}^{-1} S_t^\top (I + S_t B_{t-1}^{-1} S_t^\top)^{-1} S_t B_{t-1}^{-1})(Z_{t-1} + S_t^\top \epsilon_t) - Z_{t-1}^\top B_{t-1}^{-1} Z_{t-1} \\
&= (Z_{t-1}^\top + \epsilon_t^\top S_t) B_{t-1}^{-1} (Z_{t-1} + S_t^\top \epsilon_t) - Z_{t-1}^\top B_t^{-1} Z_{t-1} \\
&\quad - (Z_{t-1}^\top + \epsilon_t^\top S_t) B_{t-1}^{-1} S_t^\top (I + S_t B_{t-1}^{-1} S_t^\top)^{-1} S_t B_{t-1}^{-1} (Z_{t-1} + S_t^\top \epsilon_t) \\
&= \epsilon_t^\top S_t B_{t-1}^{-1} S_t^\top \epsilon_t + 2 Z_{t-1}^\top B_{t-1}^{-1} S_t^\top \epsilon_t \\
&\quad - (Z_{t-1}^\top + \epsilon_t^\top S_t) B_{t-1}^{-1} S_t^\top (I + S_t B_{t-1}^{-1} S_t^\top)^{-1} S_t B_{t-1}^{-1} (Z_{t-1} + S_t^\top \epsilon_t) \\
&= \epsilon_t^\top C_t \epsilon_t + 2 d_t^\top \epsilon_t - (d_t^\top + \epsilon_t^\top C_t)(I + C_t)^{-1}(d_t + C_t \epsilon_t) \\
&= \epsilon_t^\top C_t (I - (I + C_t)^{-1} C_t) \epsilon_t + 2 d_t^\top (I - (I + C_t)^{-1} C_t) \epsilon_t - d_t^\top (I + C_t)^{-1} d_t \\
&= \epsilon_t^\top C_t (I + C_t)^{-1} \epsilon_t + 2 d_t^\top (I + C_t)^{-1} \epsilon_t - d_t^\top (I + C_t)^{-1} d_t \\
&= \left\|\epsilon_t + C_t^{-1} d_t\right\|_{C_t (I + C_t)^{-1}}^2 - d_t^\top (I + C_t)^{-1} C_t^{-1} d_t - d_t^\top (I + C_t)^{-1} d_t \\
&= \left\|\epsilon_t + C_t^{-1} d_t\right\|_{C_t (I + C_t)^{-1}}^2 - d_t^\top (I + C_t)^{-1} (I + C_t^{-1}) d_t \,.
\end{aligned}$$

Therefore, Lemma 3.5 with $V \coloneqq \left(C_t (I + C_t)^{-1}\right)^{-1} = (I + C_t) C_t^{-1}$, $v \coloneqq C_t^{-1} d_t$ yields

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1}{2}\left(\|\widehat{p}_t - p^*\|_{B_t}^2 - \|\widehat{p}_{t-1} - p^*\|_{B_{t-1}}^2\right)\right)\right] \\
&\le \frac{\sqrt{|(I + C_t) C_t^{-1}|}}{\sqrt{|(I + C_t) C_t^{-1} - I|}} \mathrm{e}^{\frac{1}{2} d_t^\top C_t^{-1} ((I + C_t) C_t^{-1} - I)^{-1} C_t^{-1} d_t} \mathrm{e}^{-\frac{1}{2} d_t^\top (I + C_t)^{-1} (I + C_t^{-1}) d_t} \\
&\le \frac{\sqrt{|(I + C_t) C_t^{-1}|}}{\sqrt{|C_t^{-1}|}} \mathrm{e}^{\frac{1}{2} d_t^\top C_t^{-1} (C_t^{-1})^{-1} C_t^{-1} d_t} \mathrm{e}^{-\frac{1}{2} d_t^\top (I + C_t)^{-1} (I + C_t^{-1}) d_t} \\
&= \sqrt{|(I + C_t)|} \\
&= \sqrt{\frac{|B_t|}{|B_{t-1}|}},
\end{aligned}$$

where see, *e.g.,* Abbasi-Yadkori et al. (2011, Lemma 11) for the last equality. □

3.6.4.5 Norms under Perturbations

In the following two lemmas, we give some analysis of norms under perturbations.

Lemma 3.7. *Let A be a positive definite matrix. Let $a \in \mathbb{R}^d$ and $\epsilon > 0$ be such that $\epsilon < \|a\|/3$. Then*

$$\min_{x:\|x\|\le 2\epsilon} \max_{x':\|x'\|\le\epsilon} \left\{(a+x+x')^\top A(a+x+x')\right\} = \min_{x'':\|x''\|\le\epsilon} \left\{(a+x'')^\top A(a+x'')\right\} .$$

Proof. By considering the Lagrangian multiplier we see that any stationary point of the function $(a+x'')^\top A(a+x'')$ over $\{(x,x') : \|x\| \le 2\epsilon,\ \|x'\| \le \epsilon\}$ satisfies

$$\begin{aligned} &A(a+x+x') - \lambda_1 x = 0\,,\\ &A(a+x+x') - \lambda_2 x' = 0\,,\\ &x^\top x = 4\epsilon^2\,,\\ &x'^\top x' = \epsilon^2\,, \end{aligned} \tag{3.11}$$

and therefore $\lambda_1 x = \lambda_2 x'$. Considering the last two conditions of (3.11) we have $\lambda_2 = \pm 2\lambda_1$, implying that

$$x' = -(3A - 2\lambda_1 I)Aa \tag{3.12}$$

or

$$x' = (A - 2\lambda_1 I)Aa \tag{3.13}$$

for λ_1 satisfying $x'^\top x' = \epsilon^2$.

Note that it holds for any positive definite matrix B that

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} a(B+\lambda I)^{-2}a = a(B+\lambda I)^{-4}a = \left\|(B+\lambda I)^{-2}a\right\|^2 ,$$

which is positive almost everywhere, meaning that $a(B+\lambda I)^{-2}a$ is strictly convex with respect to $\lambda \in \mathbb{R}$. Therefore, there exists at most two λ_1''s satisfying (3.12) and $x'^\top x' = \epsilon^2$, and there exists at most two λ_1''s satisfying (3.13) and $x'^\top x' = \epsilon^2$. In summary, there at most four stationary points of $(a+x'')^\top A(a+x'')$ over $\{(x,x') : \|x\| \le 2\epsilon,\ \|x'\| \le \epsilon\}$.

On the other hand, two optimization problems

$$\min_{x:\|x\|\le 2\epsilon} \min_{x':\|x'\|\le\epsilon} \left\{(a+x+x')^\top A(a+x+x')\right\} = \min_{x'':\|x''\|\le 3\epsilon} \left\{(a+x'')^\top A(a+x'')\right\}$$

and

$$\max_{x:\|x\|\le 2\epsilon} \max_{x':\|x'\|\le\epsilon} \left\{(a+x+x')^\top A(a+x+x')\right\} = \max_{x'':\|x''\|\le 3\epsilon} \left\{(a+x'')^\top A(a+x'')\right\}$$

can be easily solved by an elementary calculation and the optimal values are equal to those corresponding to (3.12).

Therefore, the optimal solutions of the two minimax problems

$$\max_{x:\|x\|\le 2\epsilon} \min_{x':\|x'\|\le\epsilon} \left\{(a+x+x')^\top A(a+x+x')\right\} \tag{3.14}$$

and

$$\min_{x:\|x\|\le 2\epsilon} \max_{x':\|x'\|\le\epsilon} \left\{(a+x+x')^\top A(a+x+x')\right\} \tag{3.15}$$

correspond to two points corresponding to (3.13).

We can see again from an elementary calculation that the optimal solutions for two optimization problems

$$\min_{x'':\|x''\|\le\epsilon}\left\{(a+x'')^\top A(a+x'')\right\}$$
$$\max_{x'':\|x''\|\le\epsilon}\left\{(a+x'')^\top A(a+x'')\right\}$$

have the same necessary and sufficient conditions as (3.13) and we complete the proof by noticing that (3.14) is less than (3.15). □

Lemma 3.8. *Let $A \succeq nS_1^\top S_1$ be a positive-definite matrix with minimum eigenvalue at least $\lambda > 0$. Then, for any $\widehat{p} \in \mathbb{R}^d$ and $\epsilon > 0$ satisfying $\epsilon < \|\widehat{p}-p^*\|/3$,*

$$\|\widehat{p}-p^*\|_m^2 - \inf_{p:\|p-p^*\|\le 2\epsilon}\sup_{p':\|p'-p\|\le\epsilon}\left\|p'-\widehat{p}\right\|_A^2 \ge \epsilon\sqrt{n\lambda}\,\|S_1(\widehat{p}-p^*)\| \ .$$

Proof. Let $a = \widehat{p}-p^*$. By Lemma 3.7, we have

$$\begin{aligned}&\inf_{p:\|p-p^*\|\le 2\epsilon}\sup_{p':\|p'-p\|\le\epsilon}\left\|p'-\widehat{p}\right\|_A^2\\ &=\inf_{x:\|x\|\le 2\epsilon}\sup_{x':\|x'\|\le\epsilon}\left\|a+x+x'\right\|_A^2\\ &=\inf_{x:\|x\|\le\epsilon}\|a+x\|_A^2 \ .\end{aligned}$$

Now define $\mathcal{S}_{\epsilon',A} = \{x : \|x\|_A \le \epsilon'\}$. Then, we see that $\mathcal{S}_{\epsilon\sqrt{\lambda},A} \subset \{x : \|x\| \le \epsilon\}$. Therefore, an elementary calculation using the Lagrange multiplier technique shows

$$\begin{aligned}\inf_{x:\|x\|\le\epsilon}\left\|p'-\widehat{p}\right\|_A^2 &\le \inf_{x\in\mathcal{S}_{\epsilon\sqrt{\lambda},A}}\|p-\widehat{p}\|_A^2\\ &=\left(\|a\|_A-\epsilon\sqrt{\lambda}\right)^2 \ .\end{aligned}$$

As a result, we see that

$$\begin{aligned}\|p^*-\widehat{p}\|_m^2 - \inf_{p:\|p-p^*\|\le 2\epsilon}\sup_{p':\|p'-p\|\le\epsilon}\left\|p'-\widehat{p}\right\|_A^2 &\ge \|a\|_A^2-\left(\|a\|_A-\epsilon\sqrt{\lambda}\right)^2\\ &=\epsilon\sqrt{\lambda}\left(\|a\|_A+\|a\|_A-\epsilon\sqrt{\lambda}\right)\\ &\ge\epsilon\sqrt{\lambda}\left(\|a\|_A+\|a\|\sqrt{\lambda}-\epsilon\sqrt{\lambda}\right)\\ &=\epsilon\sqrt{\lambda}\left(\|a\|_A+\sqrt{\lambda}(\|a\|-\epsilon)\right)\\ &\ge\epsilon\sqrt{\lambda}\,\|a\|_A\\ &\ge\epsilon\sqrt{n\lambda}\,\|S_1a\| \ .\end{aligned}$$

□

For the subsets of $\mathbb{R}^n$, $\mathcal{X}$ and $\mathcal{Y}$, let $\mathcal{X}+\mathcal{Y} := \{x+y : x\in\mathcal{X},\ y\in\mathcal{Y}\}$ be the Minkowski sum, and let $B_r^n(p)$ be the n-dimensional Euclidian ball of radius r at point $p\in\mathbb{R}^n$ (the superscript n can be omitted when it is clear from context). We also let ϵ' be

$$\epsilon' := \frac{\epsilon}{\left(16\max_{i\in[k]}\|S_i\|\right)\vee\left(\frac{1}{\sqrt{m}}\max_{i\in[k]}\frac{\|L_i-L_1\|}{\|z_{1,i}\|}\right)}\ , \tag{3.16}$$

which is also used throughout the proof of this section and Section 3.6.5 as ϵ in (3.5).

Theorem 3.3. *Let $\epsilon'' \in (0, \epsilon)$ be a constant for ϵ defined in* (3.5)*. Let $\widehat{p} \in \mathcal{C}_k + B^d_{\epsilon'}(0)$ be satisfying $\|S_k(\widehat{p} - p^*)\| \leq \epsilon''$. Then, there exists $\delta > 0$ satisfying for any $n \geq 0$ and $A \succeq nS_1^\top S_1 + \lambda I$ that*

$$\|p^* - \widehat{p}\|_m^2 - \inf_{p:\|p-p^*\|\leq 2\epsilon} \sup_{p':\|p'-p\|\leq\epsilon} \|p' - \widehat{p}\|_A^2 \geq \epsilon\delta\sqrt{\lambda n}\,.$$

Proof. Recall that $\epsilon'' < \epsilon \leq \min_{p\in\mathcal{C}_1^c} \|p - p^*\|/3$. It is enough from Lemma 3.8 to prove that

$$\delta := \min_{\widehat{p}\in\{p\in\mathcal{C}_k + B^d_{\epsilon'}(0):\|S_k(p-p^*)\|\leq\epsilon''\}} \|S_1(\widehat{p} - p^*)\|$$

is positive.

We prove by contradiction and the proof is basically same as that of Lemma 3.2 but more general in the sense that the condition on $\widehat{p}$ is not $\widehat{p} \in \mathcal{C}_k$ but $\widehat{p} \in \mathcal{C}_k + B^d_{\epsilon'}(0)$. Assume that $\delta = 0$, that is, there exists $\widehat{p} \in \mathcal{C}_k + B^d_{\epsilon'}(0)$ satisfying $\|S_k(p - p^*)\| \leq \epsilon''\}$ and $\|S_1(\widehat{p} - p^*)\| = 0$. Note that $\|S_1(\widehat{p} - p^*)\| = 0$ implies $\|S_1(\widehat{p} - p^*)\| \leq \epsilon''$. Therefore, we now have following conditions on $\widehat{p}$:

- $\widehat{p} \in \mathcal{C}_k + B^d_{\epsilon'}(0)$
- $\|S_1(\widehat{p} - p^*)\| \leq \epsilon''$
- $\|S_k(\widehat{p} - p^*)\| \leq \epsilon''$.

Following the same argument as the proof of Lemma 3.2, we have

$$z_{1,k}^\top \begin{pmatrix} S_1 \\ S_k \end{pmatrix} (\widehat{p} - p^*) \leq \sqrt{2m}\epsilon''\|z_{1,k}\|\,. \tag{3.17}$$

On the other hand, since $\widehat{p} \in \mathcal{C}_k + B^d_{\epsilon'}(0)$ we can take $\bar{p} \in \mathcal{C}_k$ such that $\|\widehat{p} - \bar{p}\| \leq \epsilon'$. Hence,

$$\begin{aligned} z_{1,k}^\top \begin{pmatrix} S_1 \\ S_k \end{pmatrix} (\widehat{p} - p^*) &= (L_1 - L_k)^\top(\widehat{p} - p^*) \\ &= -(L_k - L_1)^\top(\widehat{p} - \bar{p}) + (L_1 - L_k)^\top\bar{p} + (L_k - L_1)^\top p^* \\ &\geq -(L_k - L_1)^\top(\widehat{p} - p^*) + \Delta_k\,. \quad \text{(by } \bar{p} \in \mathcal{C}_k \text{ and def. of } \Delta_k\text{)} \end{aligned} \tag{3.18}$$

From (3.17) and (3.18), we have

$$\Delta_k - (L_k - L_1)^\top(\widehat{p} - p^*) \leq \sqrt{2m}\epsilon''\|z_{1,k}\|\,. \tag{3.19}$$

Now, the left hand side of (3.19) is bounded from below as

$$\begin{aligned} \Delta_k - (L_k - L_1)^\top(\widehat{p} - \bar{p}) &\geq \Delta_k - \|L_k - L_1\|\|\widehat{p} - \bar{p}\| \\ &\geq \Delta_k - \|L_k - L_1\|\epsilon' \\ &= \Delta_k - \|L_k - L_1\|\frac{\epsilon}{\frac{1}{\sqrt{m}}\max_i \frac{\|L_1 - L_i\|}{\|z_{1,i}\|}} \\ &= \Delta_k - \|L_k - L_1\|\frac{\frac{1}{2\sqrt{m}}\min_i \frac{\Delta_i}{\|z_{1,i}\|}}{\frac{1}{\sqrt{m}}\max_i \frac{\|L_1 - L_i\|}{\|z_{1,i}\|}} \\ &\geq \Delta_k - \Delta_k/2\,. \end{aligned}$$

On the other hand, using the definition of ϵ'', the right hand side of (3.19) is bounded from above as

$$\sqrt{2m}\epsilon''\|z_{1,k}\| < \Delta_k/2\,.$$

Therefore, the proof is completed by contradiction. □

3.6.4.6 Exit Time Analysis

We next consider the exit time. Let $\mathcal{A}_t$ be an event deterministic given $\mathcal{F}_t$, and $\mathcal{B}_t$ be a random event such that if $\mathcal{B}_t$ occurred then $\mathcal{A}_{t'}$ never occurs for $t' = t+1, t+2, \ldots$. Let P_t, $t = 1, 2, \ldots, T$, be a stochastic process satisfying $P_t \le \mathbb{P}\{\mathcal{B}_t | \mathcal{F}_t\}$ a.s. and P_t^{-1} is a supermartingale with respect to the filtration induced by $\mathcal{F}_t$.

Theorem 3.4. *Let τ be the stopping time defined as*

$$\tau = \begin{cases} \min\{t \in [T] : \mathcal{A}_t\} & \text{if } \mathcal{A}_t \text{ occurs for some } t \in [T]. \\ T+1 & \text{otherwise.} \end{cases} \tag{3.20}$$

Then we almost surely have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\mathcal{A}_t] \,\middle|\, \mathcal{F}_\tau \right] \le \begin{cases} P_\tau^{-1} & \tau \le T, \\ 0 & \tau = T+1. \end{cases}$$

We prove this theorem based on the following lemma.

Lemma 3.9. *Let $(Q_i)_{i=1}^{\infty} \subset [0,1]$ be an arbitrary stochastic process such that $(Q_i^{-1})_{i=1}^{\infty}$ is a supermartingale with respect to a filtration $(\mathcal{G}_i)_{i=1}^{\infty}$. Then, for any $\mathcal{G}_0 \subset \mathcal{G}_1$,*

$$\mathbb{E}\left[\sum_{i=1}^{T} \prod_{j=1}^{i} (1 - Q_j) \,\middle|\, \mathcal{G}_0 \right] \le \mathbb{E}\left[Q_1^{-1} | \mathcal{G}_0\right] - 1 \quad \text{a.s.}$$

Proof. Let

$$N_k((Q_i, \mathcal{G}_i)_{i=1}^{\infty}, \mathcal{G}_0) = \mathbb{E}\left[\sum_{i=1}^{k} \prod_{j=1}^{i} (1 - Q_j) \,\middle|\, \mathcal{G}_0 \right]$$

$$\overline{N}_k((Q_i, \mathcal{G}_i)_{i=1}^{\infty}, \mathcal{G}_0) = \mathbb{E}\left[\sum_{i=1}^{\infty} \prod_{j=1}^{i} (1 - Q_j) \,\middle|\, \mathcal{G}_0 \right] \quad \text{where } Q_j = Q_k \text{ for } j > k.$$

We show $\overline{N}_k((Q_i, \mathcal{G}_i)_{i=1}^{\infty}, \mathcal{G}_0) \le \mathbb{E}[Q_1^{-1} | \mathcal{G}_0] - 1$ a.s. for any $(Q_i,\, \mathcal{G}_i)_{i=1}^{\infty}$, $\mathcal{G}_0 \subset \mathcal{G}_1$ and $k \in \mathbb{N}$ by induction. First, for $k = 1$ the statement holds since

$$\begin{aligned} \overline{N}_1((Q_i, \mathcal{G}_i)_{i=1}^{\infty}, \mathcal{G}_0) &= \mathbb{E}\left[\sum_{i=1}^{\infty} \prod_{j=1}^{i} (1 - Q_1) \,\middle|\, \mathcal{G}_0 \right] \\ &= \mathbb{E}\left[Q_1^{-1} - 1 \mid \mathcal{G}_0\right] \\ &= \mathbb{E}\left[Q_1^{-1} \mid \mathcal{G}_0\right] - 1 \end{aligned}$$

Next, assume that the statement holds for all $(Q_i,\, \mathcal{G}_i)_{i=1}^{k}$, $\mathcal{G}_0 \subset \mathcal{G}_1$ and $k \le k_0$. Then, we almost surely have

$$\begin{aligned} \overline{N}_{k_0+1}((Q_i, \mathcal{G}_i)_{i=1}^{\infty}, \mathcal{G}_0) &= \mathbb{E}\left[(1 - Q_1)\mathbb{E}\left[1 + \sum_{i=2}^{\infty} \prod_{j=2}^{i} (1 - Q_j) \,\middle|\, \mathcal{G}_1 \right] \,\middle|\, \mathcal{G}_0 \right] \\ &= \mathbb{E}\left[(1 - Q_1)(1 + \overline{N}_{k_0}((Q_i, \mathcal{G}_i)_{i=2}^{\infty}, \mathcal{G}_1)) \mid \mathcal{G}_0\right] \\ &\le \mathbb{E}\left[(1 - Q_1)\mathbb{E}[Q_2^{-1} \mid \mathcal{G}_1] \mid \mathcal{G}_0\right] \quad \text{(assumption of the induction)} \\ &\le \mathbb{E}\left[Q_1^{-1} \mid \mathcal{G}_0\right] - 1 \quad (Q_i^{-1} \text{ is a supermartingale.}) \end{aligned}$$

We obtain the lemma from

$$\mathbb{E}\left[\sum_{i=1}^{k}\prod_{j=1}^{i}(1-Q_j)\,\middle|\,\mathcal{G}_0\right] = N_k((Q_i,\mathcal{G}_i)_{i=1}^{\infty},\mathcal{G}_0) \le \overline{N}_k((Q_i,\mathcal{G}_i)_{i=1}^{\infty},\mathcal{G}_0) \quad \text{a.s.}$$

□

Proof of Theorem 3.4. The statement is obvious for the case $\tau = T+1$ and we consider the other case in the following.

Let τ_i be the time of the i-th occurrence of $\mathcal{A}_t$. More formally, we define τ_i as the stopping time $\tau_1 = \tau$ and

$$\tau_{i+1} = \begin{cases} \min\left\{t\in[T] : \sum_{t'=1}^{T}\mathbb{1}[\mathcal{A}_{t'}] = i+1\right\} & \sum_{t'=1}^{T}\mathbb{1}[\mathcal{A}_{t'}] \ge i+1, \\ \tau_i + 1 & \text{otherwise.} \end{cases}$$

Then $(P'_i) = (P_{\tau_i})$ is a stochastic process measurable by the filtration induced by $(\mathcal{F}'_i) = (\mathcal{F}_{\tau_i})$. By Lemma 3.9 we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}[\mathcal{A}_t]\,\middle|\,\mathcal{F}_\tau\right] &= \mathbb{E}\left[\sum_{n=1}^{T}\mathbb{1}\left[\sum_{t=1}^{T}\mathbb{1}[\mathcal{A}_t]\ge n\,\middle|\,\mathcal{F}_\tau\right]\right] \\
&\le 1+\mathbb{E}\left[\sum_{n=2}^{T}\mathbb{1}\left[\sum_{t=1}^{T}\mathbb{1}[\mathcal{A}_t]\ge n\,\middle|\,\mathcal{F}_\tau\right]\right] \\
&\le 1+\mathbb{E}\left[\sum_{i=1}^{T}\prod_{j=1}^{i}(1-P'_j)\,\middle|\,\mathcal{F}'_1\right] \\
&\le 1+\mathbb{E}\left[(P'_1)^{-1}|\mathcal{F}'_1\right]-1 \\
&= P_\tau^{-1}\,.
\end{aligned}$$

□

3.6.5 Regret Analysis of TSPM Algorithm

In this section, we give the proof of Theorem 3.1. Note that the cells are defined for the decomposition of $\mathbb{R}^d$, not $\mathcal{P}_d$. In other words, cell $\mathcal{C}_a$ is here defined as $\mathcal{C}_a = \left\{p\in\mathbb{R}^d : \text{action } a \text{ is optimal}\right\}$. For the linear setting, the empirical feedback distribution $q_a^{(t)}$ and $q_{a,n}$ are defined as

$$\begin{aligned}
q_a^{(t)} &= \frac{1}{N_a(t)}\sum_{s\in[t-1]:A_s=a} y(s)\,, \\
q_{a,n} &= \text{the value of } q_a^{(t)} \text{ after taking action } a \text{ for } n \text{ times.}
\end{aligned}$$

Recall that $\widehat{p}_t = B_t^{-1}b_t$, which is the mode of $\bar{g}_t(p)$.

3.6.5.1 Regret Decomposition

Here, we break the regret into several terms. For any $i\in[k]$, we define events

$$\begin{aligned}
\mathcal{A}_i(t) &= \left\{\|S_i\widehat{p}_t - S_ip^*\| \le \frac{\epsilon}{4}\right\}, \\
\tilde{\mathcal{A}}_i(t) &= \left\{\|S_i\tilde{p}_t - S_ip^*\| \le \epsilon\right\}.
\end{aligned}$$

We first decompose the regret as

$$
\begin{aligned}
\widehat{\mathsf{Reg}}_T &= \sum_{t=1}^{T} \Delta_{A_t} \\
&\le \sum_{t=1}^{T} \Big(\Delta_{A_t} \mathbb{1}\Big[\tilde{\mathcal{A}}_1(t)\Big] + \max_{j \in [k]} \Delta_j \mathbb{1}\Big[\tilde{\mathcal{A}}_1^c(t)\Big] \Big) \\
&= \sum_{i \neq 1} \sum_{t=1}^{T} \Delta_i \mathbb{1}\Big[A_t = i,\, \tilde{\mathcal{A}}_1(t)\Big] + \max_{j \in [k]} \Delta_j \sum_{t=1}^{T} \mathbb{1}\Big[\tilde{\mathcal{A}}_1^c(t)\Big] \\
&\le \sum_{i \neq 1} \Delta_i \sum_{t=1}^{T} \Bigg(\underbrace{\mathbb{1}\Big[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t)\Big]}_{\text{(A)}} + \underbrace{\mathbb{1}[A_t = i,\, \mathcal{A}_i^c(t)]}_{\text{(B)}} \Bigg) + \max_{j \in [k]} \Delta_j \sum_{t=1}^{T} \mathbb{1}\Big[\tilde{\mathcal{A}}_1^c(t)\Big] .
\end{aligned}
\tag{3.21}
$$

To decompose the last term, we define the following notation. We define for any $i \in [k]$

$$P_i(t) \coloneqq \mathbb{P}\{\tilde{p}_t \in \mathcal{C}_i \mid \mathcal{F}_t\} .$$

We also define

$$\mathcal{C}_{i,t} \coloneqq \mathcal{C}_i \cap B_{\epsilon'}(\widehat{p}_t) ,$$

where ϵ' is defined in (3.16), and

$$\bar{i}_t \coloneqq \operatorname*{arg\,max}_{i \in [k]} \mathbb{P}\{\tilde{p}_t \in \mathcal{C}_{i,t} \mid \mathcal{F}_t\} .$$

We define $\bar{p}_t$ as an arbitrary point in $\mathcal{C}_{\bar{i}_t,t}$. Then, we define

$$\bar{\mathcal{A}}_i(t) \coloneqq \left\{ \|S_i \bar{p}_t - S_i p^*\| \le \frac{\epsilon}{8} \right\} .$$

Using these notations, the last term in (3.21) can be decomposed as

$$
\begin{aligned}
\mathbb{1}\Big[\tilde{\mathcal{A}}_1^c(t)\Big] &\le \sum_{i=1}^{k} \mathbb{1}\Big[\bar{p}_t \in \mathcal{C}_i,\, \tilde{\mathcal{A}}_1^c(t)\Big] \\
&= \sum_{i=1}^{k} \mathbb{1}\Big[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, \tilde{\mathcal{A}}_1^c(t)\Big] + \sum_{i=1}^{k} \mathbb{1}\Big[\bar{p}_t \in \mathcal{C}_k,\, \bar{\mathcal{A}}_i(t),\, \tilde{\mathcal{A}}_1^c(t)\Big] \\
&\le \underbrace{\sum_{i=1}^{k} \mathbb{1}\big[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t)\big]}_{\text{(C)}} + \underbrace{\mathbb{1}\Big[\bar{p}_t \in \mathcal{C}_1,\, \bar{\mathcal{A}}_1(t),\, \tilde{\mathcal{A}}_1^c(t)\Big]}_{\text{(D)}} + \underbrace{\sum_{i=2}^{k} \mathbb{1}\big[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_k(t)\big]}_{\text{(E)}} .
\end{aligned}
$$

We will bound the expectation of each term in the following and complete the proof of

Theorem 3.1 as

$$\begin{aligned}
\mathsf{Reg}_T &= \sum_{i\neq 1} \Delta_i \left(O\left(\frac{1}{\epsilon^2}\log T\right) + O\left(\frac{k}{\epsilon^2}\log T\right)\right) \\
&\quad + \max_{j\in[k]} \Delta_j \left(\sum_{i=1}^{k} O\left(\frac{kd}{\epsilon^2}\log T\right) + O(1) + \sum_{i=2}^{k} O(1)\right) \\
&= O\left(\max\left\{ \frac{k\sum_{i\in[k]}\Delta_i}{\epsilon^2}, \frac{k^2 d \max_{i\in[k]}\Delta_i}{\epsilon^2} \right\} \log T \right) \\
&= O\left(\frac{mk^2 d \max_{i\in[k]}\Delta_i \log T}{\Lambda^2} \right) ,
\end{aligned}$$

where the last transformation follows from the definition of ϵ in (3.5).

3.6.5.2 Analysis for Case (A)

Lemma 3.10. *For any* $i \neq 1$*,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t) \right] \right] \leq \frac{64}{9\epsilon^2}\log T + 2^{d/2} .$$

To prove Lemma 3.10, we prove the following lemma using Corollary 3.1 and Lemma 3.3.

Lemma 3.11. *For any* $0 \leq \xi < 1/2$*,*

$$\mathbb{P}\{\tilde{p}_t \in \mathcal{V}_i \mid \mathcal{A}_i(t),\, N_i(t) > n_i\} \leq \exp\Bigl(-\frac{9}{16}\xi\epsilon^2 n_i\Bigr)(1-2\xi)^{-d/2} ,$$

where $\mathcal{V}_i := \{p \in \mathcal{C}_i : \|S_1 p - S_1 p^*\| \leq \epsilon\}$*.*

Proof. Since $\tilde{p}_t \sim \mathcal{N}(\widehat{p}_t, B_t^{-1})$ for $\widehat{p}_t = B_t^{-1} b_t$, the squared Mahalanobis distance $\|B_t^{1/2}(\tilde{p}_t - \widehat{p}_t)\|^2$ follows the chi-squared distribution with d degree of freedom. Therefore, we have

$$\mathbb{P}\{\tilde{p}_t \in \mathcal{V}_i \mid \mathcal{A}_i(t),\, N_i(t) > n_i\} \leq h\Bigl(\inf_{p\in\mathcal{V}_i} \|B_t^{1/2}(p - \widehat{p}_t)\|^2 \Bigr) ,$$

where $h(x) = \mathbb{P}_{X\sim\chi^2_d}\{X \geq x\}$. To use Lemma 3.3, we check the condition of Lemma 3.3 is indeed satisfied. First, it is obvious that the assumptions $N_i(t) \geq n_i$ and $\|S_i\widehat{p}_t - S_i p^*\| < \epsilon/4$ are satisfied. Besides, $p \in \mathcal{V}_i$ implies $p \in \mathcal{T}_i = \{p \in \mathbb{R}^d : \|S_i p - S_i p^*\| \geq \epsilon\}$ from Corollary 3.1. Thus, applying Lemma 3.3 concludes the proof. □

Proof of Lemma 3.10. For any $n_i > 0$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t) \right] \\
&= \sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t),\, N_i(t) \leq n_i \right] + \sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t),\, N_i(t) > n_i \right] \\
&\leq n_i + \sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t),\, N_i(t) > n_i \right] .
\end{aligned}$$

The second term is bounded from above as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t),\, N_i(t) > n_i\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{P}\left\{A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t),\, N_i(t) > n_i\right\} \\
&\le \sum_{t=1}^{T}\mathbb{P}\left\{A_t = i,\, \tilde{\mathcal{A}}_1(t) \,\middle|\, \mathcal{A}_i(t),\, N_i(t) > n_i\right\} \\
&= \sum_{t=1}^{T}\mathbb{P}\left\{A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \tilde{p}_t \in \mathcal{C}_i \,\middle|\, \mathcal{A}_i(t),\, N_i(t) > n_i\right\} \quad (A_t = i \text{ implies } \tilde{p}_t \in \mathcal{C}_i) \\
&\le \sum_{t=1}^{T}\mathbb{P}\{\tilde{p}_t \in \mathcal{V}_i \mid \mathcal{A}_i(t),\, N_i(t) > n_i\}\,.
\end{aligned}$$

To obtain an upper bound for $\mathbb{P}\{\tilde{p}_t \in \mathcal{V}_i \mid \mathcal{A}_i(t),\, N_i(t) > n_i\}$, we use Lemma 3.11. By taking $n_i = \frac{16}{9}\frac{1}{\xi\epsilon^2}\log T$ with $\xi = 1/4$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left[A_t = i,\, \tilde{\mathcal{A}}_1(t),\, \mathcal{A}_i(t)\right]\right] &\le n_i + \sum_{t=1}^{T}\mathbb{P}\{\tilde{p}_t \in \mathcal{V}_i \mid \mathcal{A}_i(t),\, N_i(t) > n_i\} \\
&\le n_i + \sum_{t=1}^{T}\exp\Big(-\frac{9}{16}\xi\epsilon^2 n_i\Big)(1-2\xi)^{-d/2} \\
&= \frac{16}{9}\frac{1}{\xi\epsilon^2}\log T + (1-2\xi)^{-d/2} \\
&= \frac{64}{9\epsilon^2}\log T + 2^{d/2}\,,
\end{aligned}$$

where the second inequality follows by Lemma 3.11. This completes the proof. □

3.6.5.3 Analysis for Case (B)

Lemma 3.12. *For any $i \neq 1$,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}[A_t = i,\, \mathcal{A}_i^c(t)]\right] \le \frac{256k\left(\log T + \frac{m}{2}\log 2 + 1\right)}{\epsilon^2} + \frac{16A^2}{\epsilon^2}$$

The regret in this case can intuitively be bounded because as the round proceeds the event $A_t = i$ makes $S_i\widehat{p}_t$ close to $S_i p^*$, which implies that the expected number of times the event $\mathcal{A}_i^c(t)$ occurs is not large.

Before going to the analysis of Lemma 3.12, we prove useful inequalities between $\|q_i^{(t)} - S_i p^*\|$, $\|q_i^{(t)} - S_i\widehat{p}_t\|$, and $\|S_i\widehat{p}_t - S_i p^*\|$.

Lemma 3.13. *Assume $N_i(t) > 0$. Then,*

$$\|q_i^{(t)} - S_i\widehat{p}_t\|^2 \le \frac{Z_{\setminus i}}{N_i(t)} + \|q_i^{(t)} - S_i p^*\|^2\,.$$

Proof. Recall that $\widehat{p}_t$ is the maximizer of $\bar{g}_t(p)$, and we have

$$\begin{aligned}
\widehat{p}_t &= \operatorname*{arg\,max}_{p\in\mathbb{R}^d} \bar{g}_t(p) \\
&= \operatorname*{arg\,max}_{p\in\mathbb{R}^d} \prod_{i=1}^{k}\exp\Big\{-\frac{1}{2}N_i(t)\|q_i^{(t)} - S_i p\|^2\Big\} = \operatorname*{arg\,min}_{p\in\mathbb{R}^d}\sum_{i=1}^{k}N_i(t)\|q_i^{(t)} - S_i p\|^2\,.
\end{aligned}$$

Using this and the definition of $Z_{\setminus i}$, we have

$$\begin{aligned}N_i(t)\|q_i^{(t)} - S_i\widehat{p}_t\|^2 &\le \sum_{a\in[k]} N_a(t)\|q_a^{(t)} - S_a\widehat{p}_t\|^2 \\ &\le \sum_{a\in[k]} N_a(t)\|q_a^{(t)} - S_a p^*\|^2 \\ &\le Z_{\setminus i} + N_i(t)\|q_i^{(t)} - S_i p^*\|^2 \,.\end{aligned}$$

Dividing by $N_i(t)$ on the both sides completes the proof. □

Lemma 3.14. *Assume that $\mathcal{A}_i^c(t)$ and $N_i(t) > 0$ hold. Then,*

$$\|q_i^{(t)} - S_i p^*\| > \frac{1}{2}\left(\frac{\epsilon}{4} - \sqrt{\frac{Z_{\setminus i}}{N_i(t)}}\right) . \tag{3.22}$$

Proof. By the triangle inequality,

$$\begin{aligned}\|q_i^{(t)} - S_i p^*\| &\ge \|S_i\widehat{p}_t - S_i p^*\| - \|q_i^{(t)} - S_i\widehat{p}_t\| \\ &> \frac{\epsilon}{4} - \sqrt{\frac{Z_{\setminus i}}{N_i(t)} + \|q_i^{(t)} - S_i p^*\|^2} \quad \text{(by } \mathcal{A}_i^c(t) \text{ and Lemma 3.13)} \\ &\ge \frac{\epsilon}{4} - \sqrt{\frac{Z_{\setminus i}}{N_i(t)}} - \|q_i^{(t)} - S_i p^*\| \quad \left(\text{by } \sqrt{x+y} \le \sqrt{x} + \sqrt{y} \text{ for } x, y \ge 0\right) ,\end{aligned}$$

which is equivalent to (3.22). □

Proof of Lemma 3.12. We first bound the expectation conditioned on $Z_{\setminus i}$, and then take the expectation for $Z_{\setminus i}$. Now,

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[A_t = i,\, \mathcal{A}_i^c(t)] \,\middle|\, Z_{\setminus i}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \mathcal{A}_i^c(t),\, N_i(t) \le \frac{64 Z_{\setminus i}}{\epsilon^2}\right] \,\middle|\, Z_{\setminus i}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \mathcal{A}_i^c(t),\, N_i(t) > \frac{64 Z_{\setminus i}}{\epsilon^2}\right] \,\middle|\, Z_{\setminus i}\right] \\ &\le \frac{64 Z_{\setminus i}}{\epsilon^2} + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \mathcal{A}_i^c(t),\, N_i(t) > \frac{64 Z_{\setminus i}}{\epsilon^2}\right] \,\middle|\, Z_{\setminus i}\right] \quad (A_t = i \text{ for all } t \in [T]) \,.\end{aligned}$$

The first term becomes $256k\left(\log T + \frac{m}{2}\log 2 + 1\right)/\epsilon^2$ by taking expectation over $Z_{\setminus i}$ using Lemma 3.4. Then, we bound the second term. From Lemma 3.14, $\mathcal{A}_i^c(t)$ and

$N_i(t) > \frac{64 Z_{\setminus i}}{\epsilon^2}$ imply $\|q_i^{(t)} - S_i p^*\| > \epsilon/16$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \mathcal{A}_i^c(t),\, N_i(t) > \frac{64 Z_{\setminus i}}{\epsilon^2}\right] \middle| Z_{\setminus i}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \|q_i^{(t)} - S_i p^*\| > \frac{\epsilon}{16}\right]\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, \bigcup_{y \in [A]} |(q_i^{(t)})_y - (S_i)_y p^*| > \frac{\epsilon}{16\sqrt{m}}\right]\right] \\
&\le \mathbb{E}\left[\sum_{y=1}^{m} \sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, |(q_i^{(t)})_y - (S_i)_y p^*| > \frac{\epsilon}{16\sqrt{m}}\right]\right] \\
&\le \mathbb{E}\left[\sum_{y=1}^{m} \sum_{n_i=1}^{T} \sum_{t=1}^{T} \mathbb{1}\left[A_t = i,\, N_i(t) = n_i,\, |(q_i^{(t)})_y - (S_i)_y p^*| > \frac{\epsilon}{16\sqrt{m}}\right]\right] \\
&= \mathbb{E}\left[\sum_{y=1}^{m} \sum_{n_i=1}^{T} \mathbb{1}\left[\bigcup_{t=1}^{T} \left\{A_t = i,\, N_i(t) = n_i,\, |(q_i^{(t)})_y - (S_i)_y p^*| > \frac{\epsilon}{16\sqrt{m}}\right\}\right]\right] \\
&\qquad \text{(The event } \{A_t = i,\, N_i(t) = n_i\} \text{ occurs at most once for fixed } n_i.) \\
&\le \sum_{y=1}^{m} \sum_{n_i=1}^{T} \mathbb{P}\left\{|(q_{i,n_i})_y - (S_i)_y p^*| > \frac{\epsilon}{4\sqrt{m}}\right\} \\
&\le \sum_{y=1}^{m} \sum_{n_i=1}^{T} 2\exp\left(-2n_i \left(\frac{\epsilon}{4\sqrt{m}}\right)^2\right) \quad \text{(by Hoeffding's inequality)} \\
&\le 2m \sum_{n_i=1}^{\infty} \exp\left(-\frac{n_i \epsilon^2}{8m}\right) \\
&= 2m \frac{1}{\exp\left(\frac{\epsilon^2}{8m}\right) - 1} \\
&\le 2m \frac{1}{\frac{\epsilon^2}{8m}} \quad (\text{by } \mathrm{e}^x \ge 1 + x) \\
&= \frac{16 m^2}{\epsilon^2}\,.
\end{aligned}
$$

By summing up the above argument, the proof is completed. □

3.6.5.4 Analysis for Case (C)

Before going to the analysis of cases (C), (D), and (E), we recall some notations. Recall that

$$P_i(t) = \mathbb{P}\{\tilde{p}_t \in \mathcal{C}_i \mid \mathcal{F}_t\}\,,$$

$\mathcal{C}_{i,t} = \mathcal{C}_i \cap B_{\epsilon'}(\widehat{p}_t)$, $\bar{i}_t = \arg\max_{i \in [k]} \mathbb{P}\{\tilde{p}_t \in \mathcal{C}_{i,t} | \mathcal{F}_t\}$, and $\bar{p}_t$ is an arbitrary point in $\mathcal{C}_{\bar{i}_t,t}$. Also recall that

$$\bar{\mathcal{A}}_i(t) = \left\{\|S_i \bar{p}_t - S_i p^*\| \le \frac{\epsilon}{8}\right\}.$$

Lemma 3.15. *For any $i \in [k]$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t)\right]\right] \le \frac{k}{p_0}\left(\frac{2^5 d \log T}{\epsilon^2} + \mathrm{e}^{\lambda \|p^*\|^2/2}\left(\frac{1}{\lambda T} + \frac{L}{d\lambda}\right)^{d/2} \frac{1}{1 - \mathrm{e}^{-\epsilon^2/2^5}}\right),$$

where $L = \max_i \sqrt{\mathrm{trace}(S_i^\top S_i)} = \max_i \|S_i\|_{\mathrm{F}}$.

Before proving the above lemma, we give two lemmas.

Lemma 3.16.

$$\mathbb{P}\{\tilde{p}_t \in \mathcal{C}_{\bar{i}_t} \mid \mathcal{F}_t\} \ge \mathbb{P}\{\tilde{p}_t \in \mathcal{C}_{\bar{i}_t,t} \mid \mathcal{F}_t\} \ge p_0/k\,,$$

where $p_0 \coloneqq 1 - h((\lambda\epsilon')^2)$.

Proof. First, we prove

$$\mathbb{P}\left\{\tilde{p}_t \in \bigcup_{i\in[k]} \mathcal{C}_{i,t} \,\middle|\, \mathcal{F}_t\right\} \ge 1 - h((\lambda\epsilon')^2)\,.$$

This follows from

$$\begin{aligned}
\mathbb{P}\left\{\tilde{p}_t \notin \bigcup_{i\in[k]} \mathcal{C}_{i,t} \,\middle|\, \mathcal{F}_t\right\} &= \mathbb{P}\{\tilde{p}_t \in B_{\epsilon'}(\widehat{p}_t) \mid \mathcal{F}_t\} \\
&\le h\left(\inf_{p \in \{p : \|p - \widehat{p}_t\| > \epsilon'\}} \|B_t^{1/2}(p - \widehat{p}_t)\|^2\right) \\
&\le h\Big(\lambda \|p - \widehat{p}_t\|^2\Big) \\
&\le h((\lambda\epsilon')^2)\,.
\end{aligned}$$

Using the definition of $\bar{i}_t$ completes the proof. □

Lemma 3.17. *For any $i \in [k]$, the event $\bar{\mathcal{A}}_i^c(t)$ implies $\|S_i\widehat{p}_t - S_i p^*\| \ge \epsilon/16$.*

Proof. Using the triangle inequality, we have

$$\begin{aligned}
\|S_i\widehat{p}_t - S_i p^*\| &\ge \|S_i\bar{p}_t - S_i p^*\| - \|S_i\bar{p}_t - S_i\widehat{p}_t\| \\
&\ge \epsilon/8 - \|S_i\|\|\bar{p}_t - \widehat{p}_t\| \\
&\ge \epsilon/8 - \|S_i\| \frac{\epsilon}{16 \max_i \|S_i\|} \\
&\ge \epsilon/8 - \epsilon/16 = \epsilon/16\,.
\end{aligned}$$

□

Proof of Lemma 3.15. For any n_0, which is specified later, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) < n_0\right]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) \ge n_0\right]\right]
\end{aligned}$$

The first term can be bounded by $(p_0/k)^{-1} \cdot n_0$ from Lemma 3.16. The rigorous proof can be obtained by the almost same argument as the following analysis of the second term using Theorem 3.4.

Then, we will bound the second term. Specifically, we will prove that for $n_0 = \frac{d \log T}{(\epsilon/16)^2}$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) \geq n_0\right]\right] = O(1)\,.$$

First we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) \geq n_0\right]\right] \\
&\leq \sum_{m=n_0}^{\infty} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) = m\right]\right].
\end{aligned}$$

Let

$$\tau = \min\left\{t : \bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) = m\right\} \wedge (T+1)$$

be the first time such that $\bar{p}_t \in \mathcal{C}_i$, $\bar{\mathcal{A}}_i^c(t)$ and $N_i(t) = m$ occur. Letting $\mathcal{A}_t := \left\{\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) = m\right\}$, $\mathcal{B}_t := \{A_t = i\}$ and $P_t := p_0/k$ in Theorem 3.4, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) = m\right]\right] \leq \frac{k}{p_0}\mathbb{P}\{\tau \leq T\}\,. \tag{3.23}$$

Here $\tau \leq T$ implies that

$$\begin{aligned}
\|\widehat{p}_\tau - p^*\|_{B_\tau} &= (\widehat{p}_\tau - p^*)^\top \left(\lambda I + \sum_{j \in [k]} N_j(\tau) S_j^\top S_j\right)(\widehat{p}_\tau - p^*) \\
&\geq m(\widehat{p}_\tau - p^*)^\top \left(S_i^\top S_i\right)(\widehat{p}_\tau - p^*) \\
&= m\left\|S_i(\widehat{p}_\tau - p^*)\right\|^2 \geq m(\epsilon/16)^2\,,
\end{aligned}$$

where the last inequality follows from Lemma 3.17. Therefore we have

$$\begin{aligned}
\mathbb{E}\bigl[\exp(\|\widehat{p}_\tau - p^*\|_{B_\tau}^2/2)\bigr] &\geq \mathbb{E}\bigl[\mathbb{1}[\tau \leq T]\exp(\|\widehat{p}_\tau - p^*\|_{B_\tau}^2/2)\bigr] \\
&\geq \exp(m(\epsilon/16)^2/2)\mathbb{P}\{\tau \leq T\}\,. \qquad (3.24)
\end{aligned}$$

Note that $|B_\tau| \leq |B_T| \leq (1 + TL/d)^d$ for $L = \max_i \sqrt{\mathrm{trace}(S_i^\top S_i)} = \max_i \|S_i\|_{\mathrm{F}}$ by Lemma 10 of Abbasi-Yadkori et al. (2011), where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Therefore we have

$$\begin{aligned}
\mathbb{E}[\exp(\|\widehat{p}_\tau - p^*\|_{B_\tau}/2)] &\leq \mathbb{E}\left[\sqrt{|B_\tau|} \cdot \frac{\exp(\|\widehat{p}_\tau - p^*\|_{B_\tau}^2/2)}{\sqrt{|B_\tau|}}\right] \\
&\leq (1 + TL/d)^{d/2}\mathbb{E}\left[\frac{\exp(\|\widehat{p}_\tau - p^*\|_{B_\tau}^2/2)}{\sqrt{|B_\tau|}}\right] \\
&\leq (1 + TL/d)^{d/2}\mathbb{E}\left[\frac{\exp(\|\widehat{p}_0 - p^*\|_{B_0}^2/2)}{\sqrt{|B_0|}}\right] \qquad (3.25) \\
&= \left(1 + \frac{TL}{d\lambda}\right)^{d/2} \mathrm{e}^{\lambda\|p^*\|^2/2}\,, \qquad (3.26)
\end{aligned}$$

where (3.25) holds since $\frac{\exp(\|\widehat{p}_t - p^*\|^2_{B_t}/2)}{\sqrt{|B_t|}}$ is a supermartingale from Lemma 3.6. Combining (3.23), (3.24), and (3.26), we obtain

$$\begin{aligned}\sum_{m=n_0}^{\infty} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i^c(t),\, N_i(t) = m\right]\right] &\le \frac{k}{p_0}\left(\frac{1}{\lambda} + \frac{TL}{d\lambda}\right)^{d/2} \mathrm{e}^{\lambda\|p^*\|^2/2} \sum_{m=n_0}^{\infty} \mathrm{e}^{-m(\epsilon/16)^2/2} \\ &\le \frac{k}{p_0}\left(\frac{1}{\lambda} + \frac{TL}{d\lambda}\right)^{d/2} \mathrm{e}^{\lambda\|p^*\|^2/2} \frac{\mathrm{e}^{-n_0\epsilon^2/2}}{1 - \mathrm{e}^{-(\epsilon/16)^2/2}}\,.\end{aligned}$$

By choosing $n_0 = \frac{d \log T}{(\epsilon/16)^2}$ we obtain the lemma. □

3.6.5.5 Analysis for Case (D)

Lemma 3.18. *For any $i \in [k]$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i(t),\, \tilde{\mathcal{A}}_i^c(t)\Big]\right] \le \frac{48}{9}\frac{d+2}{\epsilon^2}\frac{k}{p_0}\,.$$

Remark. To prove the regret upper bound, it is enough to prove Lemma 3.18 only for $i = 1$. However, for the sake of generality, we prove the lemma for any $i \in [k]$.

Before proving Lemma 3.18, we give two following lemmas.

Lemma 3.19. *For any $i \in [k]$, the event $\bar{\mathcal{A}}_i(t)$ implies $\mathcal{A}_i(t)$.*

Proof. Using the triangle inequality, we have

$$\begin{aligned}\|S_i p^* - S_i \widehat{p}_t\| &\le \|S_i p^* - S_i \bar{p}_t\| + \|S_i \bar{p}_t - S_i \widehat{p}_t\| \\ &\le \epsilon/8 + \|S_i\| \cdot \frac{\epsilon}{16 \max_i \|S_i\|} < \epsilon/4\,,\end{aligned}$$

which completes the proof. □

Now, Lemma 3.18 can be intuitively proven because from Lemma 3.19, $\bar{\mathcal{A}}_i(t)$ implies $\mathcal{A}_i(t)$, and the events $\mathcal{A}_i(t)$ and $\tilde{\mathcal{A}}_i^c(t)$ does not simultaneously occur many times.

Let $t = \sigma_1, \dots, \sigma_m$ be the time of the first m times that the event $\{\bar{p}_t \in \mathcal{C}_i,\, \mathcal{A}_i(t),\, N_i(t) = n_i\}$ occurred (not $\{\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i(t),\, N_i(t) = n_i\}$). In other words, we define

- σ_1 : the first time that $\bar{p}_t \in \mathcal{C}_i$, $\mathcal{A}_i(t)$ and $N_i(t) = n_i$ occurred
- σ_2 : the second time that $\bar{p}_t \in \mathcal{C}_i$, $\mathcal{A}_i(t)$ and $N_i(t) = n_i$ occurred
- $\dots$.

Now we prove the following lemma using Lemma 3.3.

Lemma 3.20. *For any $0 \le \xi < 1/2$,*

$$\mathbb{P}\Big\{\tilde{\mathcal{A}}_i^c(t) \,\Big|\, \mathcal{A}_i(t),\, \sigma_k = t\Big\} \le \exp\left(-\frac{9}{16}\xi\epsilon^2 n_i\right)(1 - 2\xi)^{-d/2}\,. \qquad (3.27)$$

Proof. Recall that $\mathcal{T}_i = \left\{p \in \mathbb{R}^d : \|S_i p - S_i p^*\| > \epsilon\right\}$. We follow a similar argument as the analysis for Lemma 3.11. Since $\tilde{p}_t \sim \mathcal{N}(B_t^{-1} b_t, B_t^{-1})$, the squared Mahalanobis distance $\|B_t^{1/2}(p - \widehat{p}_t)\|^2$ follows the chi-squared distribution with d degree of freedom. Hence, for $h(x) = \mathbb{P}_{X \sim \chi_d^2}\{X \ge x\}$, we have

$$\mathbb{P}\Big\{\tilde{\mathcal{A}}_i^c(t) \,\Big|\, \mathcal{A}_{i,n_i},\, \sigma_k = t\Big\} \le h\left(\inf_{p \in \mathcal{T}_i} \|B_t^{1/2}(p - \widehat{p}_t)\|^2\right)\,.$$

Then, (3.27) directly follows from Lemma 3.3. □

Proof of Lemma 3.18. From Lemma 3.19, the event $\bar{\mathcal{A}}_i(t)$ implies $\mathcal{A}_i(t)$. Hence, it is enough to derive the upper bound for

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \mathcal{A}_i(t),\, \tilde{\mathcal{A}}_i^c(t)\right]\right]$$

instead of the bound for

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i(t),\, \tilde{\mathcal{A}}_i^c(t)\right]\right].$$

Using Lemma 3.20, we can bound the term for case (D) from above as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \mathcal{A}_i(t),\, \tilde{\mathcal{A}}_i^c(t)\right]\right] \\
&= \mathbb{E}\left[\sum_{n_i=1}^{T}\sum_{t=1}^{T} \mathbb{1}\left[\mathcal{A}_i(t),\, \tilde{\mathcal{A}}_i^c(t),\, N_i(t) = n_i\right]\right] \\
&= \sum_{n_i=1}^{T}\sum_{t=1}^{T} \mathbb{P}\left\{\mathcal{A}_i(t),\, \tilde{\mathcal{A}}_i^c(t),\, N_i(t) = n_i\right\} \\
&= \sum_{n_i=1}^{T}\sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\left\{\mathcal{A}_i(t),\, \tilde{\mathcal{A}}_i^c(t),\, \sigma_k = t\right\} \quad \text{(the event } \{\sigma_k = t\} \text{ is exclusive for fixed } n_i) \\
&= \sum_{n_i=1}^{T}\sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{\mathcal{A}_i(t),\, \sigma_k = t\}\mathbb{P}\left\{\tilde{\mathcal{A}}_i^c(t) \,\middle|\, \mathcal{A}_i(t),\, \sigma_k = t\right\} \\
&\le \sum_{n_i=1}^{T}\sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{\mathcal{A}_i(t),\, \sigma_k = t\} C\mathrm{e}^{-n_i \iota} \quad \text{(by Lemma 3.20)} \\
&= \sum_{n_i=1}^{T} C\mathrm{e}^{-n_i \iota} \sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{\mathcal{A}_i(t),\, \sigma_k = t\} \\
&\le \sum_{n_i=1}^{T} C\mathrm{e}^{-n_i \iota} \sum_{t=1}^{T}\sum_{k=1}^{T} \mathbb{P}\{\sigma_k = t\} \\
&\le \sum_{n_i=1}^{T} C\mathrm{e}^{-n_i \iota} \sum_{k=1}^{T} \mathbb{P}\{\sigma_k \text{ exists}\} \\
&\le \sum_{n_i=1}^{T} C\mathrm{e}^{-n_i \iota} \sum_{k=1}^{T} \left(1 - \frac{p_0}{k}\right)^{k-1} \quad \text{(by } \tilde{p}_{\sigma_s} \notin \mathcal{C}_i \text{ for } s = 1, \ldots, k-1) \\
&\le 3C \frac{1}{\mathrm{e}^{\iota} - 1} \frac{k}{p_0} \\
&\le \frac{48}{9} \frac{d+2}{\epsilon^2} \frac{k}{p_0},
\end{aligned}$$

where $\iota = \frac{9\xi\epsilon^2}{16}, C = (1 - 2\xi)^{-\frac{d}{2}}$, and in the last inequality we select the optimal ξ and use $1 + x \le \mathrm{e}^x$. □

3.6.5.6 Analysis for Case (E)

Lemma 3.21. *For any $i \neq 1$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{p}_t \in \mathcal{C}_i,\, \bar{\mathcal{A}}_i(t)\right]\right] \leq \frac{2^{5d/2+7}\Gamma(d/2+1)\mathrm{e}^{\lambda^2\|p^*\|^2/2}}{\delta^2\epsilon^{d+2}\lambda^{d/2+1}}\,, \tag{3.28}$$

where ϵ is defined in (3.5) *and satisfies $\epsilon \leq \min_{p\in\mathcal{C}_1^c}\|p-p^*\|/3$, and*

$$\delta := \min_{\widehat{p}:(L_1-L_i)^\top\widehat{p}\geq 0,\, \|S_i(\widehat{p}-p^*)\|\leq\epsilon/8} \|S_1(\widehat{p}-p^*)\|\;.$$

We prove Lemma 3.21 using Lemma 3.6 and Theorem 3.4.

Remark. The upper bound in (3.28) goes to infinite when we set $\lambda = 0$, that is, a flat prior is used. However, this is not the essential effect of the prior but just comes from the minimum eigenvalue of B_1. In fact, we can see from the proof that a similar bound can be obtained for $\lambda = 0$ if we run some deterministic initialization until B_t becomes positive definite.

Proof. We evaluate each term in the summation using Theorem 3.4 with

$$\begin{aligned}
\mathcal{A}_t &= \{\bar{p}_t \in \mathcal{C}_i,\, \|S_i(\bar{p}_t - p^*)\| \leq \epsilon/8,\, N_1(t) = n\}\,,\\
\mathcal{B}_t &= \{\tilde{p}_t \in \mathcal{C}_1\}\,.
\end{aligned}$$

for $n \in [T]$. Recall that

$$\bar{g}_t(p) = \frac{1}{\sqrt{(2\pi)^d|B_t^{-1}|}}\exp\left(-\frac{1}{2}\|p-\widehat{p}_t\|_{B_t}^2\right)$$

is the probability density function of $\widehat{p}_t$ given $\mathcal{F}_t = \{B_t, b_t\}$. Using τ defined in (3.20), it holds for any $\tau \in [T]$ that

$$\begin{aligned}
\mathbb{P}\{\mathcal{B}_\tau|\mathcal{F}_\tau\} &= \mathbb{P}\{\tilde{p}_\tau \in \mathcal{C}_1 \mid \mathcal{F}_\tau\}\\
&= \int_{p\in\mathcal{C}_1}\bar{g}_\tau(p)\mathrm{d}p\\
&\geq \int_{p:\|p-p^*\|\leq 3\epsilon}\bar{g}_\tau(p)\mathrm{d}p\\
&\geq \sup_{p:\|p-p^*\|\leq 2\epsilon}\int_{p':\|p'-p\|\leq\epsilon}\bar{g}_\tau(p')\mathrm{d}p' \qquad (3.29)\\
&\geq \sup_{p:\|p-p^*\|\leq 2\epsilon}\inf_{p':\|p'-p\|\leq\epsilon}\bar{g}_\tau(p')\mathrm{Vol}(\{p'' : \|p''-p\| \leq \epsilon\})\\
&= \frac{(\sqrt{\pi}\epsilon)^d}{\Gamma(d/2+1)}\sup_{p:\|p-p^*\|\leq 2\epsilon}\inf_{p':\|p'-p\|\leq\epsilon}\bar{g}_\tau(p')\\
&= \frac{(\epsilon/\sqrt{2})^d\sqrt{|B_\tau|}}{\Gamma(d/2+1)}\exp\left\{-\frac{1}{2}\left(\inf_{p:\|p-p^*\|\leq 2\epsilon}\sup_{p':\|p'-p\|\leq\epsilon}\|p'-\widehat{p}_\tau\|_{B_\tau}^2\right)\right\}\\
&\geq \frac{(\epsilon/\sqrt{2})^d\sqrt{|B_\tau|}}{\Gamma(d/2+1)}\exp\left\{-\frac{\|\widehat{p}_\tau - p^*\|_{B_\tau}^2 - \epsilon\delta\sqrt{\lambda n}}{2}\right\}\,, \qquad (3.30)
\end{aligned}$$

where (3.29) follows since $\{p : \|p-p^*\| \leq 3\epsilon\} \supset \{p' : \|p'-p_0\| \leq \epsilon\}$ for any p_0 such that $\|p_0 - p^*\| \leq 2\epsilon$, and the last inequality follows from Theorem 3.3. To apply Theorem 3.3, we used Lemma 3.19.

Now we define a stochastic process corresponds to (3.30) as

$$P_t = \frac{(\epsilon/\sqrt{2})^d \sqrt{|B_t|}}{\Gamma(d/2+1)} \exp\left\{ -\frac{\|\widehat{p}_t - p^*\|_{B_t}^2 - \epsilon\delta\sqrt{\lambda n}}{2} \right\} .$$

Then, by Lemma 3.6,

$$\begin{aligned}
\mathbb{E}[P_{t+1}^{-1}|\mathcal{F}_t] &\le \frac{\Gamma(d/2+1)}{(\epsilon/\sqrt{2})^d} \mathrm{e}^{-\epsilon\delta\sqrt{\lambda n}/2} \mathbb{E}\left[\frac{1}{\sqrt{|B_{t+1}|}} \mathbb{E}\left[\exp\left(\frac{\|\widehat{p}_t - p^*\|_{B_{t+1}}^2}{2} \right) \middle| \mathcal{F}_t,\, S_{i(t)} \right] \middle| \mathcal{F}_t \right] \\
&\le \frac{\Gamma(d/2+1)}{(\epsilon/\sqrt{2})^d} \mathrm{e}^{-\epsilon\delta\sqrt{\lambda n}/2} \mathbb{E}\left[\frac{1}{\sqrt{|B_t|}} \exp\left(\frac{\|\widehat{p}_t - p^*\|_{B_t}^2}{2} \right) \middle| \mathcal{F}_t \right] \\
&= P_t^{-1} ,
\end{aligned}$$

which means that P_t^{-1} is a supermartingale. Therefore we can apply Theorem 3.4 and obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\widehat{p}_t \in \mathcal{C}_i,\, \|S_i(\bar{p}_t - p^*)\| \le \epsilon/8,\, N_1(t) = n] \right] &\le \mathbb{E}\left[\mathbb{1}[\tau \le T] P_\tau^{-1} \right] \\
&\le \mathbb{E}\left[P_\tau^{-1} \right] \\
&\le \mathbb{E}\left[P_1^{-1} \right] \\
&= \frac{\Gamma(d/2+1)\mathrm{e}^{\lambda^2\|p^*\|^2/2}}{(\epsilon\sqrt{\lambda/2})^d} \mathrm{e}^{-\epsilon\delta\sqrt{\lambda n}/2} .
\end{aligned}$$

Finally we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\bar{p}_t \in \mathcal{C}_i,\, \|S_i(\bar{p}_t - p^*)\| \le \epsilon/8] \right] \\
&= \sum_{n=1}^{T} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\bar{p}_t \in \mathcal{C}_i,\, \|S_i(\bar{p}_t - p^*)\| \le \epsilon/8,\, N_1(t) = n] \right] \\
&\le \frac{\Gamma(d/2+1)\mathrm{e}^{\lambda^2\|p^*\|^2/2}}{(\epsilon\sqrt{\lambda/2})^d} \sum_{n=1}^{\infty} \mathrm{e}^{-\epsilon\delta\sqrt{\lambda n}/2} \\
&\le \frac{\Gamma(d/2+1)\mathrm{e}^{\lambda^2\|p^*\|^2/2}}{(\epsilon\sqrt{\lambda/2})^d} \int_0^\infty \mathrm{e}^{-\epsilon\delta\sqrt{\lambda x}/2} \mathrm{d}x \\
&= \frac{\Gamma(d/2+1)\mathrm{e}^{\lambda^2\|p^*\|^2/2}}{(\epsilon\sqrt{\lambda/2})^d} \frac{2}{(\epsilon\delta\sqrt{\lambda}/2)^2} \Gamma(2) \\
&= \frac{2^{d/2+3}\Gamma(d/2+1)\mathrm{e}^{\lambda^2\|p^*\|^2/2}}{\delta^2\epsilon^{d+2}\lambda^{d/2+1}} ,
\end{aligned}$$

which completes the proof. □

3.6.6 Property of Dynamic Pricing Games

In this section, we investigate a property of dp-easy games.

Proposition 3.4. *Consider any dp-easy games with $c > -1$. Then, any two actions in the game are neighbors.*

Remark. In section 6.6, we considered dp-easy games with $c > 0$, but this can be relaxed to $c > -1$ to prove Proposition 3.4.

Proof. Take any two different actions $a, b \in [k]$ such that $a < b$. From the definition of the loss matrix in dp-easy games, we have $e_a \in \mathcal{C}_a$ and $e_b \in \mathcal{C}_b$.

First, we will find $\alpha \in [0, 1]$ such that

$$\alpha e_a + (1 - \alpha) e_b \in \mathcal{C}_a \cap \mathcal{C}_b \,. \tag{3.31}$$

From the definition of the loss matrix, the i-th element of $L(\alpha e_a + (1 - \alpha) e_b) \in \mathcal{P}_d$ is

$$\begin{cases} -i & (1 \le i \le a) \\ \alpha c + (1 - \alpha) \cdot (-i) & (a + 1 \le i \le b) \\ c & (b < i \le k) \end{cases} . \tag{3.32}$$

It is easy to see that the indices which give the minimum value in (3.32) is a or b. Thus, to achieve the condition (3.31), the following should be satisfied,

$$-a = \alpha c + (1 - \alpha) \cdot (-b) \,,$$

which is equivalent to

$$\alpha = \frac{b - a}{c + b} (=: \alpha^*) \,.$$

Note that we have $0 \le \alpha \le 1$ for any $c > -1$.

Next, we introduce the following definitions.

$$\begin{aligned} p^{(a,b)} &:= \alpha^* e_a + (1 - \alpha^*) e_b \in \mathcal{C}_a \cap \mathcal{C}_b \,, \\ \mathrm{Ball}_\epsilon^{(a,b)} &:= \left\{ p \in \mathcal{P}_d \colon \|p - p^{(a,b)}\| \le \epsilon \right\} , \\ L^{(x)} &:= L(p^{(a,b)} + x) \in \mathbb{R}^k \,. \end{aligned}$$

To prove the proposition, it is enough to prove the following: there exists $\epsilon > 0$, $\mathrm{Ball}_\epsilon^{(a,b)} \subset \mathcal{C}_a \cup \mathcal{C}_b$.

To prove this, it is enough to prove that, there exists $\epsilon > 0$,

$$\min_{x \in \mathbb{R}^d \colon \|x\| \le \epsilon} \min_{i \in [k] \setminus \{a,b\}} \left((L^{(x)})_i - (L^{(x)})_a \right) \vee \left((L^{(x)})_i - (L^{(x)})_b \right) > 0 \,. \tag{3.33}$$

We will prove (3.33) in the following. Take any $i \in [k] \setminus \{a, b\}$ and

$$\epsilon := \min_{i : 1 \le i < a} \frac{1}{2} \frac{a - i}{\|L_a - L_i\|} \wedge \min_{i : a < i < b} \frac{1}{2} \frac{(1 - \alpha^*)(b - i)}{\|L_i - L_b\|} \wedge \min_{i : b < i \le k} \frac{1}{2} \frac{c + a}{\|L_a - L_i\|} \,.$$

Note that the ϵ used here is different from the one used in the proof of the regret upper bounds.

Case (A): When $1 \le i < a$, using Cauchy–Schwarz inequality, we have

$$\begin{aligned} \left((L^{(x)})_i - (L^{(x)})_a \right) \vee \left((L^{(x)})_i - (L^{(x)})_b \right) &\ge (L^{(x)})_i - (L^{(x)})_a \\ &= (-i + L_i^\top x) - (-a + L_a^\top x) \\ &= (a - i) - (L_a - L_i)^\top x \\ &\ge (a - i) - \|L_a - L_b\| \|x\| \\ &\ge (a - i) - \epsilon \|L_a - L_i\| \\ &\ge \frac{1}{2}(a - i) \\ &> 0 \,. \end{aligned}$$

The arguments for cases (B) and (C) follow in the similar manner as case (A).
Case (B): When $a < i < b$, we have

$$\begin{aligned}
&\Big((L^{(x)})_i - (L^{(x)})_a\Big) \vee \Big((L^{(x)})_i - (L^{(x)})_b\Big) \\
&\geq (L^{(x)})_i - (L^{(x)})_b \\
&= \Big\{\alpha^* c + (1-\alpha^*)\cdot(-i) + L_i^\top x\Big\} - \Big\{\alpha^* c + (1-\alpha^*)\cdot(-b) + L_k^\top x\Big\} \\
&= (1-\alpha^*)(b-i) - (L_i - L_b)^\top x \\
&\geq (1-\alpha^*)(b-i) - \epsilon(L_i - L_b)^\top \\
&\geq \frac{1}{2}(1-\alpha^*)(b-i) > 0\,.
\end{aligned}$$

Case (C): When $b < i \leq k$, we have

$$\begin{aligned}
\Big((L^{(x)})_i - (L^{(x)})_a\Big) \vee \Big((L^{(x)})_i - (L^{(x)})_b\Big) &\geq (L^{(x)})_i - (L^{(x)})_a \\
&= (c + L_i^\top x) - (-a + L_a^\top x) \\
&\geq c + a - \|L_a - L_i\|\|x\| \\
&\geq c + a - \epsilon\|L_a - L_i\| \\
&\geq \frac{1}{2}(c+a) \\
&> 0\,.
\end{aligned}$$

Summing up the argument for cases (A) to (C), the proof is completed. □

3.6.7 Details and Additional Results of Experiments

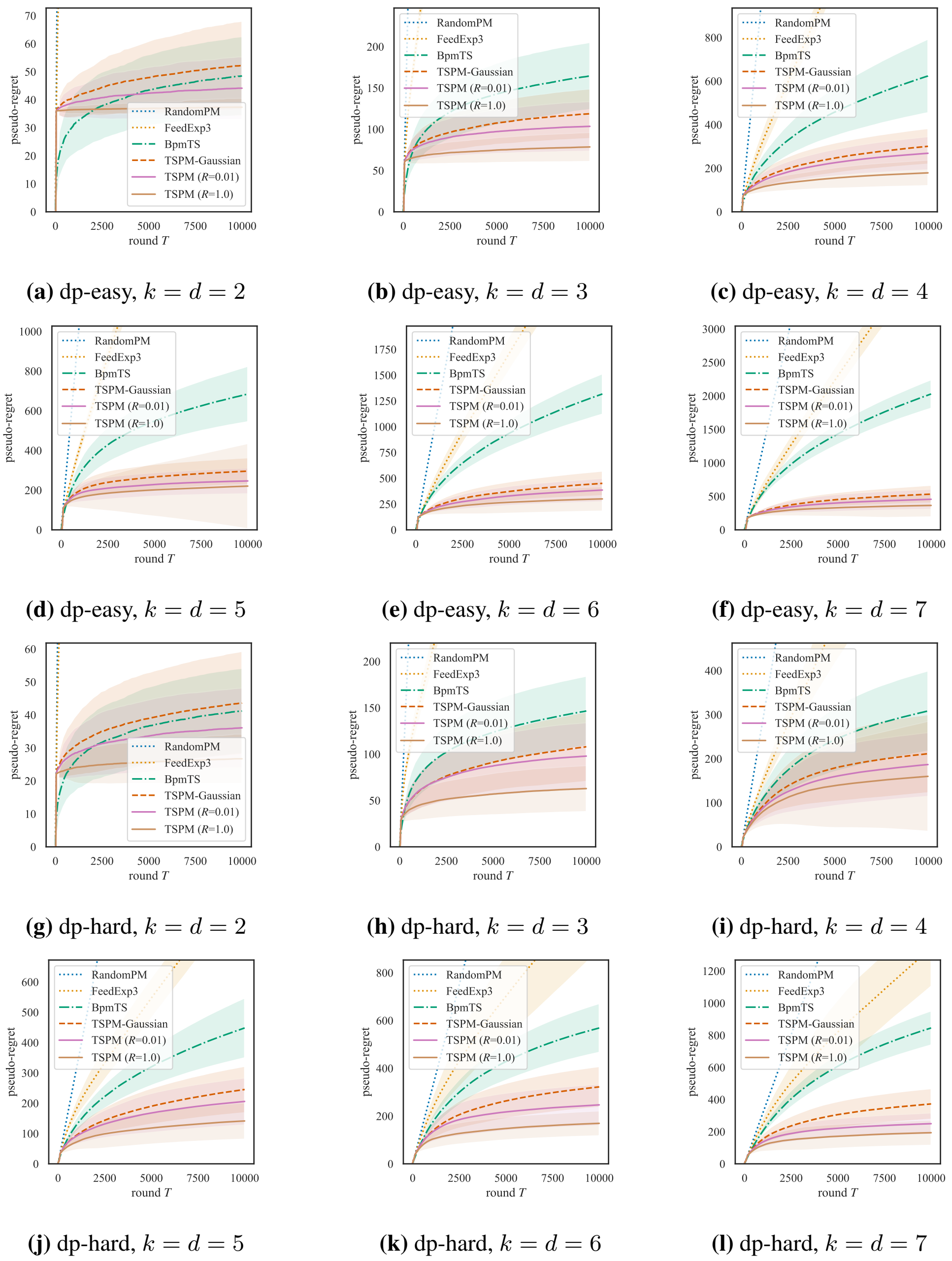
Here we give the specific values of the opponent's strategy used in Section 6.6 and show the extended experimental results for performance comparison. Table 3.2 summarizes the values of opponent's strategy used in this section and Section 6.6. Figure 3.4 shows the empirical comparison of the proposed algorithms against the benchmark methods, and Figure 3.5 shows the number of the rejected times. We can see the same tendency as Section 6.6, that is, TSPM performs the best and the number of rejections does not increase with the time step t.

Table 3.2: The values of the opponent's strategy.

# of outcomes d	opponent's strategy p^*
2	$[0.7, 0.3]$
3	$[0.5, 0.3, 0.2]$
4	$[0.3, 0.3, 0.3, 0.1]$
5	$[0.2, 0.3, 0.3, 0.1, 0.1]$
6	$[0.2, 0.2, 0.3, 0.1, 0.1, 0.1]$
7	$[0.2, 0.2, 0.3, 0.1, 0.1, 0.05, 0.05]$

3.7 Conclusion

This chapter investigated Thompson sampling (TS) for stochastic partial monitoring from the algorithmic and theoretical viewpoints. We provided a new algorithm that enables exact sampling from the posterior distribution, and numerically showed that the proposed algorithm outperforms existing methods. Besides, we provided an upper bound



(a) dp-easy, $k = d = 2$ **(b)** dp-easy, $k = d = 3$ **(c)** dp-easy, $k = d = 4$

(d) dp-easy, $k = d = 5$ **(e)** dp-easy, $k = d = 6$ **(f)** dp-easy, $k = d = 7$

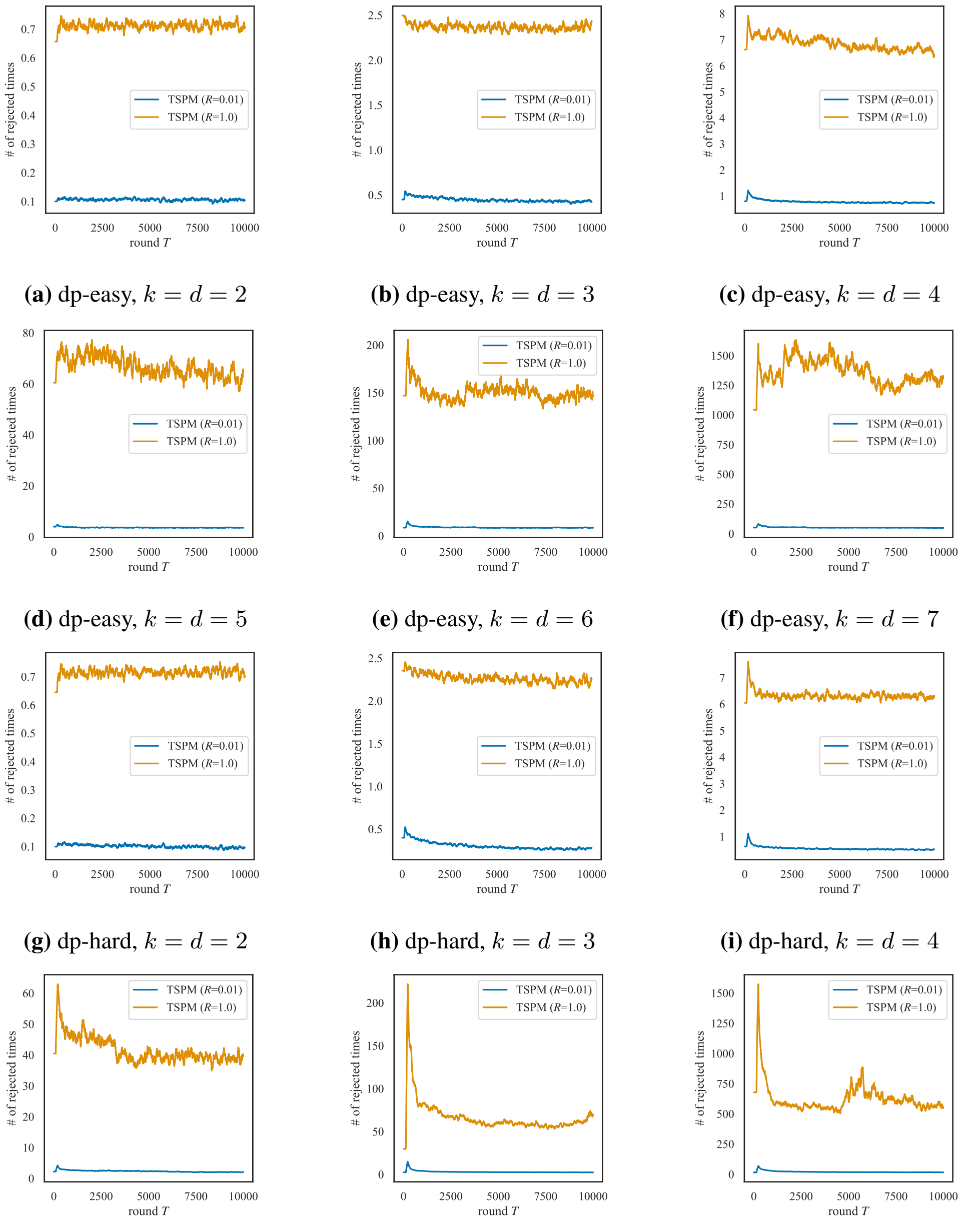
(g) dp-hard, $k = d = 2$ **(h)** dp-hard, $k = d = 3$ **(i)** dp-hard, $k = d = 4$

(j) dp-hard, $k = d = 5$ **(k)** dp-hard, $k = d = 6$ **(l)** dp-hard, $k = d = 7$

Figure 3.4: Regret-round plots of the algorithms. The solid lines indicate the average over 100 independent trials. The thin fillings are the standard error.

for the problem-dependent logarithmic expected pseudo-regret for the linearized version of the partial monitoring. To our knowledge, this bound is the first logarithmic problem-dependent expected pseudo-regret bound of a TS-based algorithm for linear bandit problems and strongly locally observable partial monitoring games.

There are several remaining questions. As mentioned in Section 3.4, Kirschner et al. (2020) considered linear partial monitoring with the feedback structure $y(t) = S_{A_t} p^* + \epsilon_t$, where $(\epsilon_t)_{t=1}^T$ is a sequence of independent sub-Gaussian noise vector in $\mathbb{R}^d$. This setting is the generalization of our linear setting, where $(\epsilon_t)_{t=1}^T$ are i.i.d. Gaussian vectors. Therefore, a natural question that arises is whether we can extend our analysis on TSPM-Gaussian to the sub-Gaussian case, although we believe it would be not straightforward as discussed in Section 3.4. It is also an important open problem to derive a regret bound on TSPM using the exact posterior sampling for the discrete partial mon-



(a) dp-easy, $k = d = 2$ **(b)** dp-easy, $k = d = 3$ **(c)** dp-easy, $k = d = 4$

(d) dp-easy, $k = d = 5$ **(e)** dp-easy, $k = d = 6$ **(f)** dp-easy, $k = d = 7$

(g) dp-hard, $k = d = 2$ **(h)** dp-hard, $k = d = 3$ **(i)** dp-hard, $k = d = 4$

(j) dp-hard, $k = d = 5$ **(k)** dp-hard, $k = d = 6$ **(l)** dp-hard, $k = d = 7$

Figure 3.5: The number of rejected times by the accept-reject sampling. The solid lines indicate the average over 100 independent trials after taking moving average with window size 100.

itoring. Although we conjecture that the algorithm also achieves logarithmic regret for the setting, there still remain some difficulties in the analysis. In particular, we have to handle the KL divergence in $f_t(p)$ and consider the restriction of the support of the opponent's strategy to $\mathcal{P}_d$, which make the analysis much more complicated. Besides, it is worth noting that the theoretical analysis of TS for hard games has never been theoretically investigated. We believe that in general TS suffers linear regret in the minimax sense due to its greediness. However, we conjecture that TS can achieve the sub-linear regret for some specific instances of hard games in the sense of the problem-dependent regret, as empirically observed in the experiments. Finally, it is an important open problem to derive the minimax regret for anytime TS-based algorithms. This needs more detailed analysis on $o(\log T)$ terms in the regret bound, which were dropped in our main result.

Chapter 4

Best of Both Worlds Algorithms for Partial Monitoring

This chapter continues to consider the partial monitoring problem with k-actions and d-outcomes. In the previous chapter, we constructed an algorithm that achieves good performance in the stochastic regime, but the underlying environment is not necessarily stochastic. To address this issue, in this chapter, we aim to construct the first best-of-both-worlds partial monitoring algorithms that performs well not only in the stochastic regime but also in the adversarial regime. In particular, we show that for non-degenerate locally observable games, the regret is $O(m^2 k^4 \log(T) \log(k_\Pi T)/\Delta_{\min})$ in the stochastic regime and $O(mk^{3/2}\sqrt{T \log(T) \log k_\Pi})$ in the adversarial regime, where T is the number of rounds, m is the maximum number of distinct observations per action, $\Delta_{\min}$ is the minimum suboptimality gap, and k_Π is the number of Pareto optimal actions. Moreover, we show that for globally observable games, the regret is $O(c_\mathcal{G}^2 \log(T) \log(k_\Pi T)/\Delta_{\min}^2)$ in the stochastic regime and $O((c_\mathcal{G}^2 \log(T) \log(k_\Pi T))^{1/3} T^{2/3})$ in the adversarial regime, where $c_\mathcal{G}$ is a game-dependent constant. We also provide regret bounds for a stochastic regime with adversarial corruptions. Our algorithms are based on the follow-the-regularized-leader framework and are inspired by the approach of exploration by optimization and the adaptive learning rate in the field of online learning with feedback graphs.

4.1 Introduction

Partial monitoring (PM) is a general sequential decision-making problem with limited feedback, which can be seen as a generalization of the bandit problem. A PM game $\mathcal{G} = (\mathcal{L}, \Phi)$ is defined by the pair of a loss matrix $\mathcal{L} \in [0,1]^{k \times d}$ and feedback matrix $\Phi \in \Sigma^{k \times d}$, where k is the number of actions, d is the number of outcomes, and Σ is a set of feedback symbols. The game is sequentially played by a learner and opponent for $T \geq 3$ rounds. At the beginning of the game, the learner observes $\mathcal{L}$ and Φ. At every round $t \in [T]$, the opponent chooses an outcome $x_t \in [d]$, and then the learner chooses an action $A_t \in [k]$, suffers an unobserved loss $\mathcal{L}_{A_t x_t}$, and receives a feedback symbol $\sigma_t = \Phi_{A_t x_t}$, where $\mathcal{L}_{ax}$ is the (a, x)-th element of $\mathcal{L}$. In general, the learner cannot directly observe the outcome or loss, and can only observe the feedback symbol. The learner's goal is to minimize their cumulative loss over all rounds. The performance of the learner is evaluated by the regret Reg_T, which is defined as the difference between the cumulative loss of the learner and the single optimal action a^* fixed in hindsight, that is, $a^* = \arg\min_{a \in [k]} \mathbb{E}\big[\sum_{t=1}^T \mathcal{L}_{a x_t}\big]$ and $\mathsf{Reg}_T = \mathbb{E}\big[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\big] = \mathbb{E}\big[\sum_{t=1}^T \langle \ell_{A_t} - \ell_{a^*}, e_{x_t} \rangle\big]$, where $\ell_a \in \mathbb{R}^d$ is the a-th row of $\mathcal{L}$, and $e_x \in \{0,1\}^d$ is the x-th standard basis of $\mathbb{R}^d$.

PM has been investigated in two regimes: the *stochastic* and *adversarial* regimes.

In the stochastic regime, outcomes $(x_t)_{t=1}^T$ are sampled from a fixed distribution ν^* in an i.i.d. manner, whereas in the adversarial regime, the outcomes are arbitrarily decided from the set of outcomes $[d]$ possibly depending on the history of the actions $(A_s)_{s=1}^{t-1}$.

Some of the first investigations on PM originate from the work by Rustichini (1999); Piccolboni and Schindelhauer (2001). The seminal work was conducted by Cesa-Bianchi et al. (2006); Bartók et al. (2011), the latter of which showed that all PM games can be classified into four classes based on their minimax regrets. They classified PM games into trivial, easy, hard, and hopeless games, for which their minimax regrets are 0, $\tilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$, and $\Theta(T)$, respectively. The easy and hard games are also called *locally observable* and *globally observable* games, respectively.

PM algorithms have been established for both the stochastic and adversarial regimes. In the adversarial regime, the most common form of algorithms is an *Exp3-type* one (Freund and Schapire, 1997; Auer et al., 2002b). Recently, Lattimore and Szepesvári (2020b) showed that an Exp3-type algorithm with the approach of *exploration by optimization* obtains the aforementioned minimax bounds. Notably, they proved the regret bounds of $O(mk^{3/2}\sqrt{T \log k})$ for non-degenerate locally observable games, and $O((c_{\mathcal{G}}T)^{2/3}(\log k)^{1/3})$ for globally observable games, where $m \leq \min\{|\Sigma|, d\}$ is the maximum number of distinct observations per action and $c_{\mathcal{G}}$ is a game-dependent constant defined in Section 4.5. PM has also been investigated in the stochastic regime and some algorithms exploiting the stochastic structure of the problem can achieve $O(\log T)$ regret bounds (Vanchinathan et al., 2014; Komiyama et al., 2015a; Tsuchiya et al., 2020).

Algorithms assuming the stochastic model for losses can suffer linear regret in the adversarial regime, whereas algorithms for the adversarial regime tend to perform poorly in the stochastic regime. Since knowing the underlying regime is difficult in practice, obtaining favorable performance for both the stochastic and adversarial regimes *without* knowing the underlying regime is desirable.

To achieve this goal, particularly in the classical multi-armed bandits, the Best-of-Both-Worlds (BOBW) algorithms that perform well in both stochastic and adversarial regimes have been developed. The first BOBW algorithm was developed in a seminal paper by Bubeck and Slivkins (2012), and the celebrated Tsallis-INF algorithm was recently proposed by Zimmert and Seldin (2021). BOBW algorithms have also been developed beyond the multi-armed bandits (*e.g.,* Gaillard et al. 2014; Luo and Schapire 2015; Erez and Koren 2021; Zimmert et al. 2019; Lee et al. 2021; Jin and Luo 2020; Huang et al. 2022; Saha and Gaillard 2022), whereas they have never been investigated in PM.

Some BOBW algorithms are known to perform well also in the *stochastic regime with adversarial corruptions* (Lykouris et al., 2018), which is an intermediate regime between the stochastic and adversarial regimes. This regime is advantageous in practice, since the stochastic assumption on outcomes is too strong whereas the adversarial assumption is too pessimistic. Therefore it is also practically important to develop BOBW algorithms that cover this intermediate regime.

4.1.1 Contributions of this Chapter

This study establishes new BOBW algorithms for PM based on the Follow-the-Regularized-Leader (FTRL) framework (McMahan, 2011). We rely on two recent theoretical advances: (i) the Exp3-type algorithm for PM developed with the approach of exploration by optimization (Lattimore and Szepesvári, 2020b) and (ii) the adaptive learning rate for online learning with feedback graphs (Ito et al., 2022b), for which BOBW algorithms have been developed (Erez and Koren, 2021; Ito et al., 2022a; Rouyer et al., 2022; Kong et al., 2022). Note that it is known that the FTRL with the (negative) Shannon entropy regularizer corresponds to the Exp3 algorithm.

The regret bounds of the proposed algorithms are as follows. We define the number

Table 4.1: Regret upper bounds for PM. The constant $C \geq 0$ is the corruption level, and $\mathcal{R}^{\mathrm{loc}}$ and $\mathcal{R}^{\mathrm{glo}}$ are the regret upper bounds of the proposed algorithm in the stochastic regime for locally and globally games, respectively. "observ." means observability. TSPM is the bound by Tsuchiya et al. (2020); refer to the paper for the definition of Λ'. ExpPM is by Lattimore and Szepesvári (2020b). PM-DEMD is by Komiyama et al. (2015a), and $D(\nu^*)$ is a distribution-dependent constant.

observ.	algorithm	stochastic (stoc.)	adversarial	stoc. w/ corruptions
locally	TSPM	$O\left(\frac{mk^2 d \log(T)}{\Lambda'^2}\right)$	–	–
obs.	ExpPM	–	$O(mk^{3/2}\sqrt{T \log k})$	–
	Proposed	$O\left(\frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}}\right)$	$O(mk^{3/2}\sqrt{T \log(T) \log k_\Pi})$	$\mathcal{R}^{\mathrm{loc}} + \sqrt{C\mathcal{R}^{\mathrm{loc}}}$
globally	PM-DMED	$O(D(\nu^*) \log T)$	–	–
obs.	ExpPM	–	$O((c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3})$	–
	Proposed	$O\left(\frac{c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T)}{\Delta_{\min}^2}\right)$	$O((c_{\mathcal{G}} T)^{2/3} (\log(T) \log(k_\Pi T))^{1/3})$	$\mathcal{R}^{\mathrm{glo}} + (C^2 \mathcal{R}^{\mathrm{glo}})^{1/3}$

of Pareto optimal actions by $k_\Pi \leq k$, and the minimum suboptimality gap by $\Delta_{\min} = \min_{a \in [k] \setminus \{a^*\}} \Delta_a$, where $\Delta_a = (\ell_a - \ell_{a^*})^\top \nu^* \geq 0$ for $a \in [k]$ is the loss gap between action a and optimal action a^*. We show that for non-degenerate locally observable games, the regret is $O(m^2 k^4 \log(T) \log(k_\Pi T)/\Delta_{\min})$ in the stochastic regime and $O(mk^{3/2}\sqrt{T \log(T) \log k_\Pi})$ in the adversarial regime. We also show that for globally observable games, the regret is $O(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T)/\Delta_{\min}^2)$ in the stochastic regime and $O((c_{\mathcal{G}} T)^{2/3} (\log(T) \log(k_\Pi T))^{1/3})$ in the adversarial regime. In addition, we also consider some intermediate regimes, such as the stochastic regime with adversarial corruptions (Lykouris et al., 2018), which we define in PM based on the corruptions on outcomes. To our knowledge, the proposed algorithms are the first BOBW algorithms for PM. Table 5.2 lists the regret bounds provided in this study and summarizes comparisons with existing work. Our algorithm is not the best in the strict sense. For example in the stochastic regime, compared to Komiyama et al. (2015a), the dependence on T of their bound is $\log T$, whereas that of ours is $(\log T)^2$. Nevertheless, this kind of looseness often appears in the BOBW literature (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Seldin and Lugosi, 2017; Ito et al., 2022a) and it is an important future work to close this gap as was done by Zimmert and Seldin (2021) in the case of multi-armed bandits.

4.1.2 Technical Summary

For locally observable games, we develop the algorithm based on the approach of *exploration by optimization* (Lattimore and Szepesvári, 2020b) with the Shannon entropy regularizer. This approach is promising especially in locally observable games for bounding a component of regret, in which we consider a certain optimization problem with respect to the action selection probability. To obtain BOBW guarantees, we consider using a self-bounding technique (Zimmert and Seldin, 2021). In the self-bounding technique, we first derive upper and lower bounds of regret using a random variable depending on the action selection probability, and then derive a regret bound by combining the upper and lower bounds. However, using the exploration by optimization may make some action selection probabilities extremely small, preventing derivation of a meaningful lower bound. To handle this problem, we consider an optimization over a *restricted* feasible set. This restriction enables us to lower bound the regret such that the self-bounding technique is applicable, and we show that even with the optimization over the restricted feasible set, the component of regret is favorably bounded. In addition, we consider the upper truncation of the learning rate developed by Ito et al. (2022a) to collaborate with the theory of exploration by optimization.

For globally observable games, we develop the algorithm using the Shannon entropy regularizer as for locally observable games. To derive BOBW guarantees, we use the technique of adaptive learning rate developed in online learning with feedback graphs by Ito et al. (2022a), but in a modified way. Their work uses a regularization called hybrid regularizers, which combines a Shannon entropy of the compensation of the action selection probability with typical regularizers (Zimmert et al., 2019; Ito et al., 2022b,a). We think that naively applying this regularization also yields BOBW guarantees, but it loses the closed form of the action selection probability in FTRL updates and requires solving an optimization problem each round. This study shows that we can obtain the BOBW guarantee even only with the standard Shannon entropy regularization, and consequently, the proposed algorithm does not need to solve the optimization problem every round and can be implemented efficiently.

4.1.3 Related Work

In the adversarial regime, FeedExp3 is a first Exp3-type algorithm, which has a first non-asymptotic regret bound (Piccolboni and Schindelhauer, 2001) and is known to achieve a minimax regret of $O(T^{2/3})$ (Cesa-Bianchi et al., 2006). Since then, Exp3-type algorithms have been used in many contexts. Bartók (2013) relied on an Exp3-type algorithm as a subroutine of their algorithm. Lattimore and Szepesvári (2019b) showed that for a variant of the locally observable game (point-locally observable games), an Exp3-type algorithm achieves an $O(\sqrt{T})$ regret. Recently, Lattimore and Szepesvári (2020b) showed that an Exp3-type algorithm using exploration by optimization can obtain bounds with good leading constants for both easy and hard games. There are also a few algorithms that are not Exp3-type (Bartók et al., 2011; Foster and Rakhlin, 2012).

PM has also been investigated in the stochastic regime, although less extensively than the adversarial regime (Bartók et al., 2012). One study (Komiyama et al., 2015a) is based on DMED (Honda and Takemura, 2011), in which the algorithm heavily exploits the stochastic structure, and the algorithm was shown to achieve an $O(\log T)$ regret with a distribution-optimal constant factor for globally observable games. Two other approaches (Vanchinathan et al., 2014; Tsuchiya et al., 2020) are based on Thompson sampling (Thompson, 1933). They focus on another variant of locally observable games (strongly locally observable games), and the algorithms presented a strong empirical performance in the stochastic regime with an $O(\log T)$ regret bound (Tsuchiya et al., 2020).

It is worth noting that PM has been studied in a variety of contexts with somewhat different settings, *e.g.,* with feedback graphs (Alon et al., 2015) or with linear feedback (Lin et al., 2014). While our focus in this chapter is the locally and globally observable games, there has been some literature for hopeless games; we basically cannot do anything with the current definition of the regret, but some research has been done by modifying the definition of the regret (Rustichini, 1999; Mannor and Shimkin, 2003; Perchet, 2011; Mannor et al., 2014).

4.2 Background

Notation Let $\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ be the Euclidian, ℓ_1-, and ℓ_∞-norms for a vector x respectively, and $\|A\|_\infty = \max_{i,j}|A_{ij}|$ be the maximum norm for a matrix A. Let $\mathcal{P}_k = \{p \in [0,1]^k : \|p\|_1 = 1\}$ be the $(k-1)$-dimensional probability simplex. A vector $e_a \in \{0,1\}^k$ is the a-th standard basis of $\mathbb{R}^k$, and $\mathbf{1}$ is the all-one vector.

Partial Monitoring Consider any PM game $\mathcal{G} = (\mathcal{L}, \Phi)$. Let $m \leq |\Sigma|$ be the maximum number of distinct symbols in a single row of $\Phi \in \Sigma^{k \times d}$ over all rows. In the following, we introduce several concepts in PM. Different actions a and b are *duplicate* if $\ell_a = \ell_b$. We can decompose possible distributions of d outcomes in $\mathcal{P}_d$ based on the loss matrix: for every action $a \in [k]$, *cell* $\mathcal{C}_a = \{u \in \mathcal{P}_d : \max_{b \in [k]} (\ell_a - \ell_b)^\top u \leq 0\}$ is the set of probability vectors in $\mathcal{P}_d$ for which action a is optimal. Each cell is a convex closed polytope. Let $\dim(\mathcal{C}_a)$ be the dimension of the affine hull of $\mathcal{C}_a$. If $\mathcal{C}_a = \emptyset$, action a is *dominated*. For non-dominated actions, if $\dim(\mathcal{C}_a) = d - 1$ then action a is *Pareto optimal*, and if $\dim(\mathcal{C}_a) < d - 1$ then action a is *degenerate*. We denote the set of Pareto optimal actions by Π, and the number of Pareto optimal actions by $k_\Pi = |\Pi|$. Two Pareto optimal actions $a, b \in \Pi$ are *neighbors* if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d - 2$, and this notion is used to define the difficulty of PM games. It is known that the undirected graph induced by the above neighborhood relations is connected (see *e.g.,* Bartók et al. 2012, Lattimore and Szepesvári 2020a, Lemma 37.7), and this is useful for loss difference estimations between distinct Pareto optimal actions. A PM game is called non-degenerate if it has no degenerate actions. An example of cell decomposition is given in Figure 3.1 in Chapter 3. From hereon, we assume that PM game $\mathcal{G}$ is non-degenerate and contains no duplicate actions. The following *observability* conditions characterize the difficulty of PM games.

Definition 4.1. Neighbouring actions a and b are *globally observable* if there exists function $w_e : [k] \times \Sigma \to \mathbb{R}$ such that

$$\sum_{c=1}^{k} w_e(c, \Phi_{cx}) = \mathcal{L}_{ax} - \mathcal{L}_{bx} \text{ for all } x \in [d] \,. \tag{4.1}$$

Neighbouring actions a and b are *locally observable* if there exists $w_e = w_{ab}$ satisfying (4.1) and $w_e(c, \sigma) = 0$ for $c \notin \{a, b\}$. A PM game is called globally (resp. locally) observable if all neighboring actions are globally (resp. locally) observable.

It is easy to see from the above definition that any locally observable games are globally observable, and this chapter assumes that $\mathcal{G}$ is globally observable.

Loss Difference Estimation Next, we introduce a method of loss difference estimations used in PM. Let $\mathcal{H}$ be the set of all functions from $[k] \times \Sigma$ to $\mathbb{R}^d$. In the following, we show that for globally observable games we can estimate loss differences between *any* Pareto optimal actions using some $G \in \mathcal{H}$ based on (4.1).

Lemma 4.1 (Lattimore and Szepesvári 2020b, Lemma 4)**.** *Consider any globally observable game. Then there exists a function $G \in \mathcal{H}$ such that for all $b, c \in \Pi$, we have*

$$\sum_{a=1}^{k} (G(a, \Phi_{ax})_b - G(a, \Phi_{ax})_c) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \textit{ for all } x \in [d] \,. \tag{4.2}$$

This result straightforwardly follows from the fact that the graph induced by the set of Pareto optimal actions is connected. Let $\mathscr{T}$ be a tree over Π induced by the neighborhood relations. Lattimore and Szepesvári (2020b) provides the following example of G:

$$G(a, \sigma)_b = \sum_{e \in \mathrm{path}_{\mathscr{T}}(b)} w_e(a, \sigma) \ \text{ for } \ a \in \Pi \,, \tag{4.3}$$

where $\mathrm{path}_{\mathscr{T}}(b)$ is the set of edges from $b \in \Pi$ to an arbitrarily chosen root $c \in \Pi$ on $\mathscr{T}$.

Intermediate Regimes between Stochastic and Adversarial Regimes Here, we discuss intermediate regimes between the stochastic and adversarial regimes: the stochastic regime with adversarial corruptions and an adversarial regime with a self-bounding constraint.

The stochastic regime with adversarial corruptions was originally considered by Lykouris et al. (2018) in the classical multi-armed bandits. We define this regime in PM by considering the corruptions on the sequence of outcomes $(x_t)_{t=1}^T$. In this regime, a temporary outcome $x'_t \in [d]$ is sampled from an unknown distribution ν^*, and the adversary then corrupts x'_t to x_t without knowing A_t. We define the corruption level by $C = \mathbb{E}\big[\sum_{t=1}^T \|\mathcal{L}e_{x_t} - \mathcal{L}e_{x'_t}\|_\infty\big] \geq 0$. If $C = 0$, this regime corresponds to the stochastic regime, and if $C \geq T$, this regime corresponds to the adversarial regime. As we will see, the proposed algorithms work without knowing the corruption level C. We also define another intermediate regime, a *stochastically constrained adversarial regime*, in Section 4.6.1.

In this work, we consider an *adversarial regime with a self-bounding constraint*, developed in the multi-armed bandits (Zimmert and Seldin, 2021) and includes the regimes that appeared so far.

Definition 4.2. Let $\Delta \in [0,1]^k$ and $C \geq 0$. The environment is in an *adversarial regime with a* (Δ, C, T) self-bounding constraint if it holds for any algorithm that $\mathsf{Reg}_T \geq \mathbb{E}\big[\sum_{t=1}^T \Delta_{A_t} - C\big]$.

We can show that the regimes that have appeared so far are included in the adversarial regime with a self-bounding constraint; the details are discussed in Section 4.6.1.

In this study, we assume that there exists a unique optimal action. This assumption has been employed by many studies aiming to develop BOBW algorithms (Gaillard et al., 2014; Luo and Schapire, 2015; Wei and Luo, 2018; Ito, 2021a; Zimmert and Seldin, 2021).

4.3 Follow-the-Regularized-Leader

This section recalls the FTRL framework, which was introduced in Chapter 2, and provides some fundamental bounds used in the analysis. We recall that Π is the set of Pareto optimal actions. In the FTRL framework, a probability vector $p_t \in \mathcal{P}_k$ over the action set $[k]$ is given as

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{P}(\Pi)} \left\langle \sum_{s=1}^{t-1} \widehat{y}_s, q \right\rangle + \psi_t(q), \quad p_t = \mathcal{T}_t(q_t), \tag{4.4}$$

where the set $\mathcal{P}(\mathcal{B}) \coloneqq \{p \in \mathcal{P}_k : p_a = 0 \text{ for } a \notin \mathcal{B}\}$ for $\mathcal{B} \subset [k]$ is a convex closed polytope on the probability simplex with nonzero elements at indices in $\mathcal{B}$, $\widehat{y}_s \in \mathbb{R}^k$ is an estimator of the loss at round t, $\psi_t : \mathcal{P}_k \to \mathbb{R}$ is a convex regularizer, and $\mathcal{T}_t : \mathcal{P}(\Pi) \to \mathcal{P}_k$ is a map from q_t to an action selection probability vector p_t. We use the Shannon entropy for ψ_t, which is defined as

$$\psi_t(p) = \frac{1}{\eta_t} \sum_{a=1}^k p_a \log(p_a) = -\frac{1}{\eta_t} H(p). \tag{4.5}$$

We can easily check that if we use the Shannon entropy with learning rate η_t, $q_t \in \mathcal{P}(\Pi)$ is expressed as

$$q_{t,a} = \frac{\mathbb{1}[a \in \Pi] \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{y}_{sa}\right)}{\sum_{b \in \Pi} \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{y}_{sb}\right)} \quad \text{for } a \in [k]. \tag{4.6}$$

We set an estimator to $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{tA_t}$ (Lattimore and Szepesvári, 2020b), where for locally observable games, G_t is obtained by minimizing a certain optimization problem, whereas for globally observable games G_t is set to (4.3). The regret analysis of FTRL boils down to the evaluation of $\sum_{t=1}^T \sum_{a=1}^k p_{ta}(\widehat{y}_{ta} - \widehat{y}_{ta^*})$. We can decompose this quantity into

$$\sum_{t=1}^{T}\sum_{a=1}^{k} p_{ta}(\widehat{y}_{ta} - \widehat{y}_{ta^*}) \le \underline{\sum_{t=1}^{T}\Big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\Big) + \psi_{T+1}(e_{a^*}) - \psi_1(q_1)}$$

$$+ \underline{\sum_{t=1}^{T}\Big(\langle q_t - q_{t+1}, \widehat{y}_t\rangle - D_t(q_{t+1}, q_t)\Big)} + \underline{\sum_{t=1}^{T}\sum_{a=1}^{k}(q_{t,a} - p_{ta})(\widehat{y}_{ta} - \widehat{y}_{ta^*})} \tag{4.7}$$

where the inequality follows from the standard analysis of the FTRL framework as given in Lemma 2.1 in Chapter 2, and $D_t : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$ is the *Bregman divergence* induced by ψ_t, *i.e.,* $D_t(p, q) = \psi_t(p) - \psi_t(q) - \langle \nabla\psi_t(q), p - q\rangle$. We refer to the terms with dashed, wavy, and straight underlines in (5.2) as the *penalty*, *stability*, and *transformation* terms, respectively.

We use a self-bounding technique to bound the regret in the stochastic regime, which requires a lower bound of the regret. To this end, we introduce parameters $Q(a^*)$ and $\bar{Q}(a^*)$ given by

$$Q(a^*) = \sum_{t=1}^{T}(1 - q_{t,a^*}) \quad \text{and} \quad \bar{Q}(a^*) = \mathbb{E}\left[Q(a^*)\right] . \tag{4.8}$$

Note that $0 \le \bar{Q}(a^*) \le T$ for any $a^* \in [k]$. Based on the quantity $\bar{Q}(a^*)$, the regret in the adversarial regime with a self-bounding constraint can be bounded from below as follows.

Lemma 4.2. *In the adversarial regime with a self-bounding constraint, if there exists* $c \in (0, 1]$ *such that* $p_{ta} \ge c\, q_{t,a}$ *for* $t \in [T]$ *and* $a \in [k]$*, the regret is bounded as* $\mathsf{Reg}_T \ge c\,\Delta_{\min}\bar{Q}(a^*) - C$ *.*

All omitted proofs are given in Section 4.6. This lemma is used to derive polylogarithmic regret bounds in the adversarial regime with a self-bounding constraint.

4.4 Locally Observable Case

This section provides a BOBW algorithm for locally observable games and derives its regret bounds.

4.4.1 Exploration by Optimization in PM

We first briefly explain the approach of exploration by optimization by Lattimore and Szepesvári (2020b), based on which our algorithm for locally observable games is developed. In locally observable games, the achievable regret is generally smaller than in globally observable games. Hence, we need to exploit this easiness to achieve small regret, for which we rely on exploration-by-optimization. Intuitively, in locally observable games, a loss estimator may suffer a large variance because an informative action might not be selected due to its large losses. To overcome this issue, Lattimore and Szepesvári (2020b) proposed exploration-by-optimization, which improves regret bound by optimizing the stability that corresponds to the variance.

The key idea behind the approach is to minimize a part of a regret upper bound of an Exp3-type algorithm (equivalently, FTRL with the Shannon entropy). In particular, they consider the optimization on variables $G : [k] \times \Sigma \to \mathbb{R}^k$ and $p \in \mathcal{P}_k$. Their algorithm computes every round the function G and the action selection probability vector p by optimizing a part of the regret upper bound of FTRL, expressed as

$$\underset{G \in \mathcal{H},\, p \in \mathcal{P}_k}{\text{minimize}} \quad \max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x}{\eta} + \frac{\text{bias}_q(G; x)}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^k p_a \left\langle q, \xi\left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right], \tag{4.9}$$

where $\xi(x) = \mathrm{e}^{-x} + x - 1$ (we abuse the notation by applying ξ in an element-wise manner), and

$$\text{bias}_q(G; x) = \left\langle q,\, \mathcal{L} e_x - \sum_{a=1}^k G(a, \Phi_{ax}) \right\rangle + \max_{c \in \Pi} \left(\sum_{a=1}^k G(a, \Phi_{ax})_c - \mathcal{L}_{cx} \right) \tag{4.10}$$

is the bias function. In the optimization problem (4.9), the first term corresponds to the transformation term, the second term corresponds to the regret for using a biased estimator, and the third term comes from a part of the stability term. Note that the bias term does not appear when G satisfies (4.2). Note also that the optimization problem in (4.9) is convex and can be solved numerically by using standard solvers as discussed in Lattimore and Szepesvári (2020b).

4.4.2 Proposed Algorithm

This section describes the proposed algorithm for locally observable games. Although exploration-by-optimization significantly improves the regret bound for locally observable games, they only consider the adversarial regimes, and some modification is required for making it valid also for the stochastic regime. To obtain BOBW guarantees, we often rely on a self-bounding technique, which requires a certain lower bound on the action selection probability p (Gaillard et al., 2014; Wei and Luo, 2018; Zimmert and Seldin, 2021). However, solving the optimization problem (4.9) may result in $p_a = 0$ for a certain $a \in [k]$, which precludes the use of the technique. The proposed algorithm considers the minimization problem over a restricted feasible set for p instead of over $\mathcal{P}_k$. Let $\mathcal{P}'_k(q)$ for $q \in \mathcal{P}(\Pi)$ be $\mathcal{P}'_k(q) = \{p \in \mathcal{P}_k : p_a \geq q_a/(2k) \text{ for all } a \in [k]\} \subset \mathcal{P}_k$. We then consider the following optimization problem:

$$\underset{G \in \mathcal{H},\, p \in \mathcal{P}'_k(q)}{\text{minimize}} \quad \max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x}{\eta} + \frac{\text{bias}_q(G; x)}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^k p_a \left\langle q, \xi\left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right], \tag{4.11}$$

which implies that the solution p of the optimization problem (4.11) satisfies $p \geq q/(2k)$. This property is useful when applying the self-bounding technique to bound the regret in the stochastic regime (possibly with adversarial corruptions). We define the optimal value of the optimization problem (4.11) by $\text{opt}'_q(\eta)$ and its truncation at round t by $V'_t = \max\{0, \text{opt}'_{q_t}(\eta_t)\}$.

Regularizer and Learning Rate We use the Shannon entropy with learning rate η_t in (4.5) as a regularizer. The learning rate η_t is defined as follows. Let $\beta'_1 = c_1 \geq 1$ and

$$\beta'_{t+1} = \beta'_t + \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^t H(q_s)}}, \quad \beta_t = \max\left\{B, \beta'_t\right\}, \quad \text{and} \quad \eta_t = \frac{1}{\beta_t} \tag{4.12}$$

Algorithm 4.1: BOBW algorithm for locally observable games

```
input: $B$
for $t = 1, 2, \dots$ do
    Compute $\eta_t$ using (4.12) and $q_t$ using (4.6)
    Solve (4.11) with $\eta \leftarrow \eta_t$ and $q \leftarrow q_t$ to determine $V_t' = \max\{0, \mathrm{opt}_{q_t}'(\eta_t)\}$
      and the corresponding solution $p_t$ and $G_t$.
    Sample $A_t \sim p_t$, observe $\sigma_t \in \Sigma$, compute $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{tA_t}$, update $\beta_t'$
      using (4.12).
```

for $c_1 > 0$ (determined in Theorem 4.1). The fundamental idea of this learning rate was developed by Ito et al. (2022a), and we use its variant by the upper truncation of β_t'. The truncation is required when applying the following lemma to bound $\mathrm{opt}_q'(\eta)$.

Lemma 4.3. *For non-degenerate locally observable games and $\eta \leq 1/(2mk^2)$, we have*

$$\mathrm{opt}_*'(\eta) := \sup_{q \in \mathcal{P}_k} \mathrm{opt}_q'(\eta) \leq 3m^2k^3 \,.$$

This lemma is a slightly stronger version of Lattimore and Szepesvári (2020b, Proposition 8), in which the same upper bound is derived for the minimum value over larger feasible set $\mathcal{P}_k \supset \mathcal{P}_k'(q)$ in (4.9) instead of (4.11). Since the objective function of (4.9) and (4.11) originally comes from a component of the regret, this lemma means that the restriction of the feasible set does not harm the regret bound. Algorithm 4.1 provides the proposed algorithm for locally observable games.

4.4.3 Regret Analysis for Locally Observable Games

With the above algorithm, we can prove the following regret bound for locally observable games.

Theorem 4.1. *Consider any locally observable non-degenerate partial monitoring game. If we run Algorithm 4.1 with $B \geq 2mk^2$ and $c_1 = \Theta\big(mk^{3/2}\sqrt{(\log T)/(\log k_\Pi)}\big)$, we have the following bounds. For the adversarial regime with a (Δ, C, T) self-bounding constraint, we have*

$$\mathsf{Reg}_T = O\left(\frac{m^2k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + \sqrt{\frac{Cm^2k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}}}\right), \quad (4.13)$$

and for the adversarial regime, we have

$$\mathsf{Reg}_T = O\big(mk^{3/2}\sqrt{T \log(T) \log k_\Pi}\big) + B \log k_\Pi \,.$$

Note that (4.13) with $C = 0$ yields the bound in the stochastic regime. The bound for the adversarial regime is a factor of $\sqrt{\log(T)\log(k_\Pi)/\log k}$ worse for large enough T than the algorithm by Lattimore and Szepesvári (2020b). This comes from the difficulty of obtaining the BOBW guarantee, where we need to aggressively change the learning rate when the environment looks not so much adversarial. Note that exactly solving the optimization problem (4.11) is not necessary, and we discuss regret bounds for this case in Section 4.6.11. In the rest of this section, we provide a sketch of the analysis.

We start by decomposing the regret as follows.

Lemma 4.4. $\mathsf{Reg}_T \leq \mathbb{E}\Big[\sum_{t=1}^T \big(\eta_{t+1}^{-1} - \eta_t^{-1}\big) H(q_{t+1}) + H(q_1)/\eta_1 + \sum_{t=1}^T \eta_t V_t'\Big] \,.$

This can be proven by refining the analysis of the penalty term of Theorem 6 in Lattimore and Szepesvári (2020b), in which we rely on the standard analysis in (5.2), and the first and remaining terms correspond to the penalty term and the sum of the transformation and stability terms, respectively. As will be shown in the proof of Theorem 4.1, the RHS of Lemma 4.4 can be bounded in terms of $\sum_{t=1}^{T} H(q_t)$, for which we have the following bound.

Lemma 4.5. *For any $a^* \in [k]$, we have $\sum_{t=1}^{T} H(q_t) \le Q(a^*) \log(\mathrm{e} k_\Pi T / Q(a^*))$.*

We can show this lemma similarly to Ito et al. (2022a, Lemma 4) by noting that $q_{t,a} = 0$ for $a \notin \Pi$. Finally, we are ready to prove Theorem 4.1. Here, we only sketch the proof and provide the complete proof can be found in Section 4.6.6.

Proof sketch of Theorem 4.1. We prove this theorem by bounding the RHS of Lemma 4.4.

(Bounding the penalty term) Since β'_{t+1} is non-decreasing and $\beta'_t \le \beta_t$ from the definition of learning rate in (4.12), it holds that

$$\sum_{t=1}^{T} \left({\eta_{t+1}}^{-1} - {\eta_t}^{-1}\right) H(q_{t+1}) \le \sum_{t=1}^{T} (\beta'_{t+1} - \beta'_t) H(q_{t+1}) = \sum_{t=1}^{T} \frac{c_1 \sqrt{\log k_\Pi}\, H(q_{t+1})}{\sqrt{\log k_\Pi + \sum_{s=1}^{t} H(q_s)}}$$

$$\le c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \frac{2 H(q_{t+1})}{\sqrt{\sum_{s=1}^{t+1} H(q_s)} + \sqrt{\sum_{s=1}^{t} H(q_s)}} \le 2 c_1 \sqrt{\log k_\Pi} \sqrt{\sum_{t=1}^{T} H(q_t)}, \tag{4.14}$$

where the second inequality follows from $0 \le H(q_{t+1}) \le \log k_\Pi$, and the last inequality follows by sequentially applying $b/(\sqrt{a+b} + \sqrt{a}) = \sqrt{a+b} - \sqrt{a}$ for $a, b > 0$, the telescoping argument, $\sqrt{a+b} - \sqrt{b} \le \sqrt{a}$ for $a, b \ge 0$, and $H(q_{T+1}) \le H(q_1)$.

(Bounding the sum of the transformation and part of stability terms) It holds that

$$\sum_{t=1}^{T} \eta_t V'_t \le \max_{s \in [T]} V'_s \sum_{t=1}^{T} \eta_t \le 3 m^2 k^3 \sum_{t=1}^{T} \eta_t \le \frac{3 m^2 k^3 (1 + \log T)}{c_1} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{t=1}^{T} H(q_t)}\,, \tag{4.15}$$

where the second inequality follows from Lemma 4.3 and the last inequality follows since the lower bound $\beta'_t = c_1 + \sum_{u=1}^{t-1} \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{u} H(q_s)}} \ge \frac{c_1 t}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}}$ implies that

$$\sum_{t=1}^{T} \eta_t \le \sum_{t=1}^{T} \frac{1}{\beta'_t} \le \sum_{t=1}^{T} \frac{1}{c_1 t} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{s=1}^{t} H(q_s)} \le \frac{1 + \log T}{c_1} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{t=1}^{T} H(q_t)}\,.$$

(Summing up arguments and applying a self-bounding technique) By bounding the RHS of Lemma 4.4 by (4.14) and (4.15) with $c_1 = \Theta\left(m k^{3/2} \sqrt{\log(T)/\log k_\Pi}\right)$, we have $\mathsf{Reg}_T = O\left(m k^{3/2} \sqrt{\log(T) \sum_{t=1}^{T} H(q_t)} + m k^{3/2} \sqrt{\log(T) \log k_\Pi}\right) + 2 m k^2 \log k_\Pi\,.$ Since $\sum_{t=1}^{T} H(q_t) \le T \log k_\Pi$, the desired bound for the adversarial regime is obtained. We consider the adversarial regime with a self-bounding constraint in the following. Here, we only consider the case of $Q(a^*) \ge \mathrm{e}$, since otherwise we easily obtain the desired bound. Note that Lemma 5.2 with $Q(a^*) \ge \mathrm{e}$ implies $\sum_{t=1}^{T} H(q_t) \le$

Algorithm 4.2: BOBW algorithm for globally observable games

```
for $t = 1, 2, \dots$ do
    Compute $q_t$ using (4.6).
    Compute $a_t, b_t$ in (4.17), $\gamma'_t, \gamma_t$ in (4.16), and $p_t$ from $q_t$ by (4.18).
    Sample $A_t \sim p_t$, observe $\sigma_t \in \Sigma$, compute $\widehat{y}_t = G(A_t, \sigma_t)/p_{tA_t}$, and update
      $\beta_t$ using (4.16).
```

$Q(a^*) \log(k_\Pi T)$. Hence, for any $\lambda > 0$

$$\begin{aligned}
\mathsf{Reg}_T &= (1+\lambda)\mathsf{Reg}_T - \lambda \mathsf{Reg}_T \\
&\le \mathbb{E}\Big[(1+\lambda) O\left(m k^{3/2} \sqrt{\log(T)\log(k_\Pi T) Q(a^*)}\right) - \frac{\lambda \Delta_{\min} Q(a^*)}{2k}\Big] + \lambda C \\
&\le O\big(\mathcal{R}^{\mathrm{loc}} + \lambda(\mathcal{R}^{\mathrm{loc}} + C) + \mathcal{R}^{\mathrm{loc}}/\lambda\big) ,
\end{aligned}$$

where the first inequality follows by Lemma 4.2 with $c = 1/(2k)$, and the second inequality follows from $a\sqrt{x} - bx/2 \le a^2/(2b)$ for $a, b, x \ge 0$ and $\mathcal{R}^{\mathrm{loc}} = m^2 k^4 \log(T) \log(k_\Pi T)/\Delta_{\min}$. Appropriately choosing λ gives the desired bound. □

4.5 Globally Observable Case

This section proposes an algorithm for globally observable games and derives its BOBW regret bound. We use G defined in (4.3) and let $c_{\mathcal{G}} = \max\{1, k\|G\|_\infty\}$ be the game-dependent constant.

4.5.1 Proposed Algorithm

In the proposed algorithm for globally observable games, we use the regularizer ψ_t in (4.5) as used in the locally observable case, but with different parameters. We define $\beta_t, \gamma_t \in \mathbb{R}$ by $\beta_1 = \max\{c_2, 2c_{\mathcal{G}}\}$ and

$$\gamma'_t = \frac{1}{4} \frac{c_1 b_t}{c_1 + \left(\sum_{s=1}^t b_s\right)^{1/3}} , \quad \beta_{t+1} = \beta_t + \frac{c_2 b_t}{\gamma'_t \left(c_1 + \sum_{s=1}^{t-1} \frac{b_s a_{s+1}}{\gamma'_s}\right)^{1/2}} , \quad \gamma_t = \gamma'_t + \frac{c_{\mathcal{G}}}{2\beta_t} , \tag{4.16}$$

where c_1 and c_2 are parameters satisfying $c_1 \ge \max\{1, \log k_\Pi\}$, and a_t and b_t are defined by

$$a_t = H(q_t) = -\sum_{a \in \Pi} q_{t,a} \log(q_{t,a}) \quad \text{and} \quad b_t = 1 - \max_{a \in \Pi} q_{t,a} . \tag{4.17}$$

Note that we have $\psi_t(0) = 0$, and using $\beta_t \ge \beta_1 \ge 2c_{\mathcal{G}}$ and $b_t \le \sum_{a=1}^k q_{t,a} \le 1$ we have $\gamma_t \le c_1 b_t/(4c_1) + c_{\mathcal{G}}/(2c_{\mathcal{G}}) \le 1/2$. We use the following transform from q_t to p_t:

$$p_t = \mathcal{T}_t(q_t) = (1 - \gamma_t) q_t + \frac{\gamma_t}{k} \mathbf{1} . \tag{4.18}$$

Algorithm 4.2 presents the proposed algorithm for globally observable games.

4.5.2 Regret Analysis for Globally Observable Games

With the above algorithm, we can prove the following regret bound for globally observable games.

Theorem 4.2. *Consider any globally observable partial monitoring game. If we run Algorithm 4.2 with* $c_1 = \Theta\big(\big(c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)\big)^{1/3}\big)$ *and* $c_2 = \Theta\big(\sqrt{c_{\mathcal{G}}^2 \log T}\big)$ *, we have the following bounds. For the adversarial regime with a* (Δ, C, T) *self-bounding constraint, we have*

$$\mathsf{Reg}_T = O\left(\frac{c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)}{\Delta_{\min}^2} + \left(\frac{C^2 c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)}{\Delta_{\min}^2} \right)^{1/3} \right), \quad (4.19)$$

and for the adversarial regime, we have

$$\mathsf{Reg}_T = O\big(\big(c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)\big)^{1/3} T^{2/3}\big),$$

where in the last big-O notation, the terms of $o(\mathrm{poly}(k, c_G)(T\log T)^{2/3})$ *are ignored.*

Note that (4.19) with $C = 0$ yields the bound in the stochastic regime. The bound for the adversarial regime is a factor of $(\log(T)\log(k_\Pi T)/\log k)^{1/3}$ worse than the algorithm by Lattimore and Szepesvári (2020b). This comes from the difficulty of obtaining the BOBW guarantee, where we need to aggressively change the learning rate when the environment looks not so much adversarial.

We begin the analysis by decomposing the regret as follows.

Lemma 4.6. *The regret of Algorithm 4.2 is bounded as*

$$\begin{aligned}\mathsf{Reg}_T \le \mathbb{E}\Bigg[&\sum_{t=1}^T \gamma_t + \sum_{t=1}^T \big(\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)\big) \\ &+ \sum_{t=1}^T \big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\big) + \psi_{T+1}(e_{a^*}) - \psi_1(q_1) \Bigg].\end{aligned}$$

This lemma can be proven based on the fact that we can estimate loss differences between Pareto optimal actions, and boundedness of $\mathcal{L}$, combined with the standard analysis of FTRL given in (5.2). Note that the first, second, and last terms correspond to the transformation, stability, and penalty terms, respectively. We can bound the stability term on the RHS of Lemma 4.6 as follows.

Lemma 4.7. *If* ψ_t *is given by* (4.5) *and* b_t *is defined by* (4.17)*, then we have*

$$\mathbb{E}[\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)] \le \mathbb{E}\big[2c_{\mathcal{G}}^2 b_t/(\beta_t\gamma_t)\big]. \quad (4.20)$$

Remark. Globally observable PM is a generalization of the weakly observable setting in online learning with feedback graphs (Alon et al., 2015). For this online learning problem, the regularizer in the form of $-H(p) - H(\mathbf{1} - p)$ rather than (4.6) is introduced in Ito et al. (2022a) to make the LHS of (4.20) easy to bound. However, FTRL with this regularizer requires solving a convex optimization every round. This study shows that the LHS of (4.20) can be favorably bounded without the regularization of $-H(\mathbf{1} - p)$. The key to the proof of this lemma is that for any $a' \in [k]$ it holds that $\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) = \langle \widehat{y}_t - \widehat{y}_{ta'}\mathbf{1}, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \le \beta_t \sum_{a=1}^k q_{t,a}\xi\left((\widehat{y}_{ta} - \widehat{y}_{ta'})/\beta_t\right)$, which enables us to bound the stability term with b_t in (4.17), leading to the regret upper bound depending on $Q(a^*)$ in Proposition 4.1.

Using the definition of β_t and γ_t in (4.16) with Lemmas 4.6 and 4.7, we can bound the regret as follows.

Proposition 4.1. *Assume β_t and γ_t are given by* (4.16)*. Then, the regret is bounded as* $\mathsf{Reg}_T = O\Big(\mathbb{E}\Big[c_1 B_T^{2/3} + \tilde{c}\sqrt{c_1^2 + (\log k_\Pi + A_T)\left(c_1 + B_T^{1/3}\right)}\Big] + \beta_1 \log k_\Pi\Big)$, *where* $A_T = \sum_{t=1}^T a_t$, $B_T = \sum_{t=1}^T b_t$, *and* $\tilde{c} = O\left(\frac{1}{\sqrt{c_1}}\left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2\right)\right) = O\Big(\frac{c_1}{\sqrt{\log(k_\Pi T)}}\Big)$.

The proof of this lemma is similar to Proposition 2 of Ito et al. (2022a). Now we are ready to prove Theorem 4.2, whose proof is sketched below and completed in Appendix 4.6.10.

Proof sketch of Theorem 4.2. We first consider the adversarial regime. In the adversarial regime, Proposition 4.1 with $A_T \le T \log k_\Pi$ and $B_T \le T$ immediately leads to

$$\begin{aligned}\mathsf{Reg}_T &= O\Big(c_1 T^{2/3} + \tilde{c}\sqrt{c_1^2 + (\log k_\Pi + T\log k_\Pi)(c_1 + T^{1/3})}\Big) \\ &= O\big((c_1 + \tilde{c}\sqrt{\log k_\Pi})T^{2/3}\big)\,. \end{aligned} \tag{4.21}$$

We next consider the adversarial regime with a self-bounding constraint. Here, we only consider the case of $Q(a^*) > \max\{\mathrm{e}, c_1^3\}$, since otherwise we can easily obtain the desired bound. Note that $A_T \le Q(a^*)\log(k_\Pi T)$ by Lemma 5.2 with $Q(a^*) \ge \mathrm{e}$ and $B_T = \sum_{t=1}^T (1 - \max_{a\in\Pi} q_{t,a}) \le \sum_{t=1}^T (1 - q_{t,a^*}) = Q(a^*)$. Then, Proposition 4.1 with these inequalities and $Q(a^*) > c_1^3$ gives

$$\begin{aligned}\mathsf{Reg}_T &\le O\Big(\mathbb{E}\Big[c_1 Q(a^*)^{2/3} + \tilde{c}\sqrt{\log(k_\Pi T) Q(a^*)^{4/3}}\Big]\Big) \\ &\le O\big((c_1 + \tilde{c}\sqrt{\log(k_\Pi T)})\bar{Q}(a^*)^{2/3}\big)\,. \end{aligned} \tag{4.22}$$

By (4.21) and (4.22), there exists $\widehat{c} = O\big(c_1 + \tilde{c}\sqrt{\log(k_\Pi T)}\big)$ satisfying $\mathsf{Reg}_T \le \widehat{c}T^{2/3}$ for the adversarial regime and $\mathsf{Reg}_T \le \widehat{c}\bar{Q}(a^*)^{2/3}$ for the adversarial regime with a self-bounding constraint. Recalling the definitions of c_1 and c_2, we have $\widehat{c} = O\left((c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T))^{1/3}\right)$, which gives the desired bounds for the adversarial regime. For the adversarial regime with a self-bounding constraint, using $\mathsf{Reg}_T \le \widehat{c}\bar{Q}(a^*)^{2/3}$ and Lemma 4.2 with $c = 1/2$ for any $\lambda \in (0,1]$ it holds that

$$\mathsf{Reg}_T = (1+\lambda)\mathsf{Reg}_T - \lambda\mathsf{Reg}_T \le (1+\lambda)\widehat{c}\cdot\bar{Q}(a^*)^{2/3} - \lambda\Delta_{\min}\bar{Q}(a^*)/2 + \lambda C\,.$$

Taking the worst case of this with respect to $\bar{Q}(a^*)$ and taking $\lambda \in (0,1]$ appropriately gives the desired bound for the adversarial regime with a self-bounding constraint. □

4.6 Deferred Discussion and Proofs

4.6.1 Intermediate Regimes between Stochastic and Adversarial Regimes in Partial Monotoring

This section details the discussion on intermediate regimes between stochastic and adversarial regimes given in Section 4.2. This section first defines the *stochastically constrained adversarial regime* in PM, and then shows that the stochastic regime, adversarial regime, stochastically constrained adversarial regime, and stochastic regime with adversarial corruptions are indeed adversarial regimes with a self-bounding constraint defined in Definition 5.1.

The stochastically constrained adversarial regime was initially considered by Wei and Luo (2018) and also discussed in Zimmert and Seldin (2021) in the context of the multi-armed bandit problem. We say that the environment is the stochastically constrained adversarial regime if for any $a \neq a^*$ there exists $\tilde{\Delta}_{a,a^*} > 0$ such that $\mathbb{E}_{x_t \sim \nu^*}[\mathcal{L}_{ax_t} - \mathcal{L}_{a^* x_t} | x_1, \dots, x_{t-1}] \geq \tilde{\Delta}_{a,a^*}$.

Next, we show that the stochastic regime, adversarial regime, stochastically constrained adversarial regime, and stochastic regime with adversarial corruptions are indeed included in the adversarial regime with a self-bounding constraint. We first consider the stochastic regime. Indeed, if outcomes $(x_t)_t$ follow a distribution ν^* independently for $t = 1, 2, \dots, T$, we have $\mathsf{Reg}_T = \max_{a^* \in [k]} \mathbb{E}[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})] = \mathbb{E}[\sum_{t=1}^T \Delta_{A_t}]$, where we define $\Delta \in [0,1]^k$ by $\Delta_a = \mathbb{E}_{x \sim \nu^*}[\mathcal{L}_{ax} - \mathcal{L}_{a^* x}]$. This implies that the stochastic regime is in the adversarial regime with a $(\Delta, 0, T)$ self-bounding constraint. We next consider the stochastic regime with adversarial corruptions. In fact, using the definition of the corruption level C, we have

$$\begin{aligned}
&\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{A_t x_t'} - \mathcal{L}_{a^* x_t'}\right)\right] + \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{A_t x_t} - \mathcal{L}_{A_t x_t'}\right)\right] + \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{a^* x_t'} - \mathcal{L}_{a^* x_t}\right)\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^T \Delta_{A_t}\right] - 2C \,,
\end{aligned}$$

which implies that the stochastic regime with adversarial corruption with corruption levels C is an adversarial regime with a $(\Delta, 2C, T)$ self-bounding constraint. It is also easy to see that adversarial regimes are the adversarial regime with a $(\Delta, 2T, T)$ self-bounding constraint, and the stochastically constrained adversarial regime are the adversarial regime with a $(\Delta, 0, T)$ self-bounding constraint by defining $\Delta \in [0,1]^k$ by $\Delta_a = \tilde{\Delta}_{a,a^*}$.

4.6.2 Proof of Lemma 4.2

Proof. Note that the environment is the adversarial regime with a self-bounding constraint with $\Delta \in [0,1]^k$ such that $\Delta_a \geq \Delta_{\min}$ for all $a \in [k] \setminus \{a^*\}$. Hence, the regret is then bounded as

$$\begin{aligned}
\mathsf{Reg}_T &\geq \mathbb{E}\left[\sum_{t=1}^T \Delta_{A_t}\right] - C = \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k p_{ta} \Delta_a\right] - C \\
&\geq \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k c\, q_{t,a} \Delta_a\right] - C \geq c\, \Delta_{\min} \bar{Q}(a^*) - C \,,
\end{aligned}$$

where the first inequality follows from Definition 5.1, the equality follows from $A_t \sim p_t$, the second inequality follows from the definition of p_t given in (5.1), and the last inequality follows from the assumption $p_{ta} \geq c\, q_{t,a}$ for all $t \in [T], a \in [k]$ and the definition of $\bar{Q}(a^*)$ given in (4.8). This completes the proof of Lemma 4.2. □

4.6.3 Proof of Lemma 4.3

Before proving Lemma 4.3, we review the definition and property of the water transfer operator W_ν introduced by Lattimore and Szepesvári (2019c). We refer to $\mathscr{T} \subset [k] \times [k]$ representing the edges of a directed tree with vertices $[k]$ as *in-tree* with vertex set $[k]$ and define $\mathcal{E} = \{(a,b) \in [k] \times [k] : a \text{ and } b \text{ are neighbors}\}$.

Lemma 4.8 (Lattimore and Szepesvári, 2019c). *Assume that partial monitoring game $\mathcal{G}$ is non-degenerate and locally observable and let $\nu \in \mathcal{P}_d$. Then there exists a function $W_\nu : \mathcal{P}_k \to \mathcal{P}_k$ such that the following hold for all $q \in \mathcal{P}_k$: (a) $(W_\nu(q) - q)^\top \mathcal{L}\nu \le 0$; (b) $W_\nu(q)_a \ge q_a/k$ for all $a \in [k]$; and (c) there exists an in-tree $\mathscr{T} \subset \mathcal{E}$ over $[k]$ such that $W_\nu(q)_a \le W_\nu(q)_b$ for all $(a,b) \in \mathscr{T}$.*

Using this, we prove the generalized version of Lattimore and Szepesvári (2020b, Proposition 8), where the proof follows a quite similar argument as their proof therein.

Proof of Lemma 4.3. We define the set of functions that satisfy (4.2) by

$$\mathcal{H}_\circ = \left\{ G : (e_b - e_c)^\top \sum_{a=1}^{k} G(a, \Phi_{ax}) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \text{ for all } b, c \in \Pi \text{ and } x \in [d] \right\} .$$

Take any $q \in \mathcal{P}_k$. By Sion's minimax theorem, we have

$$\begin{aligned} \mathrm{opt}'_q(\eta) &\le \min_{G \in \mathcal{H}_\circ,\, p \in \mathcal{P}'_k(q)} \max_{\nu \in \mathcal{P}_d} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu + \frac{1}{\eta^2} \sum_{x=1}^{d} \nu_x \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a,\Phi_{ax})}{p_a} \right) \right\rangle \right] \\ &= \max_{\nu \in \mathcal{P}_d} \min_{G \in \mathcal{H}_\circ,\, p \in \mathcal{P}'_k(q)} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu + \frac{1}{\eta^2} \sum_{x=1}^{d} \nu_x \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a,\Phi_{ax})}{p_a} \right) \right\rangle \right] , \end{aligned}$$

where the first inequality follows since we added the constraint that $G \in \mathcal{H}_\circ$, which makes the bias term zero. Take any $\nu \in \mathcal{P}_d$ and let $\mathscr{T}$ be the in-tree over $[k]$ decided based on Lemma 4.8. Using these variables, we define the action selection probability vector $p \in \mathcal{P}'_k(q)$ by

$$p = (1-\gamma)u + \frac{\gamma}{k}\mathbf{1} , \quad \text{where} \quad u = W_\nu(q) , \quad \text{and} \quad \gamma = \frac{\eta m k^2}{2} .$$

Here, $W_\nu : \mathcal{P}_k \to \mathcal{P}_k$ is the water operator. It is worth noting that from the assumption that $\eta \le 1/(mk^2)$, we have $\gamma \le 1/2$ and $p_a \ge u_a/2 = W_\nu(q)_a/2 \ge q_a/(2k)$, where the last inequality follows from Part (b) of Lemma 4.8, and this indeed implies $p \in \mathcal{P}'_k(q)$.

We take $G \in \mathcal{H}_\circ$ defined in (4.3), where we recall that $G(a,\sigma)_b = \sum_{e \in \mathrm{path}_{\mathcal{T}}(b)} w_e(a,\sigma)$. By Lattimore and Szepesvári (2020b, Lemma 20) and the assumption that $\mathcal{G}$ is non-degenerate, w_e can be chosen so that $\|w_e\|_\infty \le m/2$. Since paths in $\mathcal{T}$ have length at most k, we have $\|G\|_\infty \le km/2$. From the above definitions, for any $x \in [d]$ we have

$$\frac{\eta G(a,\Phi_{ax})}{p_a} \ge -\frac{\eta m k^2}{2\gamma} = -1 .$$

Hence, using Parts (b) and (c) of Lemma 4.8, we have

$$\begin{aligned} \frac{1}{\eta^2} \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a,\Phi_{ax})}{p_a} \right) \right\rangle &\le \sum_{a=1}^{k} \frac{1}{p_a} \sum_{b=1}^{k} q_b \left(G(a,\Phi_{ax})_b \right)^2 \\ &\le 2 \sum_{a=1}^{k} \frac{1}{u_a} \sum_{b=1}^{k} q_b \left(G(a,\Phi_{ax})_b \right)^2 \\ &= 2 \sum_{b=1}^{k} \sum_{a=1}^{k} \frac{q_b}{u_a} \left(\sum_{e \in \mathrm{path}_{\mathcal{T}}(b)} w_e(a,\Phi_{ax}) \right)^2 \\ &\le \frac{m^2}{2} \sum_{b=1}^{k} \sum_{a=1}^{k} \frac{q_b}{u_a} \left(\sum_{e \in \mathrm{path}_{\mathcal{T}}(b)} \mathbb{1}[a \in e] \right)^2 \\ &\le 2m^2k^3 , \end{aligned}$$

where the first inequality follows from

$$\xi(x) = \exp(-x) + x - 1 \leq x^2 \text{ for } x \geq -1\,, \tag{4.23}$$

the second inequality follows since $p_a \geq u_a/2$, the third inequality follows since $\|w_e\|_\infty \leq m/2$, and the last inequality follows from Part (b) of Lemma 4.8 to implying that $q_b \leq k u_b$ and Part (c) implying that $u_a \geq u_b$ for $a \in \mathrm{path}_{\mathcal{T}}(b)$. Finally,

$$\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu = \frac{1}{\eta}(u-q)^\top \mathcal{L}\nu + \frac{\gamma}{\eta}\left(\frac{1}{k}\mathbf{1} - u\right)^\top \mathcal{L}\nu \leq \frac{\gamma}{\eta}\left(\frac{1}{k}\mathbf{1} - u\right)^\top \mathcal{L}\nu \leq mk^2\,,$$

where the first inequality follows from Part (a) of Lemma 4.8. Summing up the above arguments, we have $\mathrm{opt}'_q(\eta) \leq 3m^2k^3$, which completes the proof of Lemma 4.3. □

4.6.4 Proof of Lemma 4.4

We first analyze the stability term in (5.2) for ψ_t defined in (4.5).

Lemma 4.9. *If ψ_t is given by* (4.5)*, it holds for any $\ell \in \mathbb{R}^k$ and $p, q \in \mathcal{P}_k$ that*

$$\langle \ell, p - q\rangle - D_t(q,p) \leq \beta_t \sum_{a=1}^{k} p_a \xi\left(\frac{\ell_a}{\beta_t}\right)\,,$$

where we recall that $\xi(x) = \exp(-x) + x - 1$.

Proof. For any $x, y \in (0,1)$, we let $d(y,x) \geq 0$ be the Bregman divergence over $(0,1)$ induced by $\psi(x) = x\log x$, *i.e.,*

$$d(y,x) = y\log y - x\log x - (\log x + 1)(y-x) = y\log\frac{y}{x} + x - y\,.$$

Using this, the Bregman divergence induced by $\psi_t(p) = (1/\eta_t)\sum_{a=1}^k p_a \log(p_a) = \beta_t \sum_{a=1}^k p_a \log(p_a)$ in (4.5) can be written as

$$D_t(q,p) = \psi_t(p) - \psi_t(q) - \langle \nabla\psi_t(q), p - q\rangle = \beta_t \sum_{a=1}^{k} d(q_a, p_a)\,.$$

From this, we have

$$\langle \ell, p - q\rangle - D_t(q,p) \leq \sum_{a=1}^{k} \left(\ell_a(p_a - q_a) - \beta_t d(q_a, p_a)\right)\,. \tag{4.24}$$

We show

$$\ell_a(p_a - q_a) - \beta_t d(q_a, p_a) \leq \beta_t p_a \xi\left(\frac{\ell_a}{\beta_t}\right)\,. \tag{4.25}$$

As $\ell_a(p_a - q_a) - \beta_t d(q_a, p_a)$ is concave in q, its maximum subject to $q \in \mathbb{R}$ is attained when the derivative of it is equal to zero, *i.e.,*

$$\frac{\partial}{\partial q_a}\left(\ell_a(p_a - q_a) - \beta_t d(q_a, p_a)\right) = -\ell_a - \beta_t\left(\log q_a - \log p_a\right) = 0\,.$$

This implies that the maximum is attained when $q_a = q_a^* := p_a \exp(-\ell_a/\beta_t)$. Hence, we obtain (4.25) by

$$\begin{aligned}
&\ell_a(p_a - q_a) - \beta_t d(q_a, p_a) \le \ell_a(p_a - q_a^*) - \beta_t d(q_a^*, p_a) \\
&= \ell_a(p_a - q_a^*) - \beta_t\left(q_a^* \log q_a^* - p_a \log p_a - (\log p_a + 1)(q_a^* - p_a)\right) \\
&= \ell_a p_a - \beta_t\left(q_a^* \log p_a - p_a \log p_a - (\log p_a + 1)(q_a^* - p_a)\right) \\
&= \ell_a p_a + \beta_t(q_a^* - p_a) = \beta_t p_a \left(\exp\left(-\frac{\ell_a}{\beta_t}\right) + \frac{\ell_a}{\beta_t} - 1\right) \\
&= \beta_t p_a \xi\left(\frac{\ell_a}{\beta_t}\right),
\end{aligned}$$

where the second equality follows from $\log q_a^* = \log p_a - \ell_a/\beta_t$, and the fourth equality follows from $q_a^* = p_a \exp(-\ell_a/\beta_t)$. Combining (4.24) and (4.25) completes the proof. □

Proof of Lemma 4.4. Let $a^* = \arg\min_{a\in[k]} \mathbb{E}\big[\sum_{t=1}^T \mathcal{L}_{ax_t}\big] \in \Pi$ be the optimal action in hindsight. We have

$$\begin{aligned}
\mathsf{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{b=1}^k p_{tb}(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{b=1}^k (p_{tb} - q_{t,b})(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t}) + \sum_{t=1}^T \sum_{b=1}^k q_{t,b}(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t})\right] \quad (4.26)
\end{aligned}$$

The first term in (4.26) is equal to $\mathbb{E}\big[\sum_{t=1}^T (p_t - q_t)^\top \mathcal{L} e_{x_t}\big]$. The second term in (4.26) can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{b=1}^k q_{t,b}(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t})\right] = \mathbb{E}\left[\sum_{b=1}^k q_t^\top \mathcal{L} e_{x_t} - \mathcal{L}_{a^* x_t}\right] \\
&= \mathbb{E}\left[\sum_{b=1}^k q_t^\top \mathcal{L} e_{x_t} - q_t^\top \sum_{a=1}^k G_t(a, \Phi_{ax_t}) + \sum_{a=1}^k G_t(a, \Phi_{ax_t})_{a^*} - \mathcal{L}_{a^* x_t}\right] \\
&\quad + \mathbb{E}\left[q_t^\top \sum_{a=1}^k G_t(a, \Phi_{ax_t}) - \sum_{a=1}^k G_t(a, \Phi_{ax_t})_{a^*}\right] \\
&\le \mathbb{E}[\mathrm{bias}_{q_t}(G; x_t)] + \mathbb{E}\Big[q_t^\top \widehat{y}_t - \widehat{y}_{ta^*}\Big], \qquad (4.27)
\end{aligned}$$

where in the last inequality we used the definition in (4.10) and Lemma 4.1 with $a^* \in \Pi$ and $q_{t,a} = 0$ for $a \notin \Pi$. The sum over $t \in [T]$ of the last term in (4.27) can be bounded using (5.2) and the definition of the regularizer (4.5) as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \sum_{b=1}^k q_{t,b}(\widehat{y}_{tb} - \widehat{y}_{ta^*})\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \Big(\langle q_t - q_{t+1}, \widehat{y}_t\rangle - D_t(q_{t+1}, q_t)\Big)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \frac{\langle q_t, \xi(\eta_t \widehat{y}_t)\rangle}{\eta_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \frac{1}{\eta_t} \sum_{a=1}^k p_{ta} \left\langle q_t, \xi\left(\frac{\eta_t G(a, \sigma_t)}{p_{ta}}\right)\right\rangle\right], \\
&\qquad (4.28)
\end{aligned}$$

where in the second inequality we used the following inequality obtained by Lemma 4.9:

$$\langle \widehat{y}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \le \beta_t \sum_{a=1}^{k} q_{t,a} \xi \left(\frac{\widehat{y}_{ta}}{\beta_t} \right) = \frac{\langle q_t, \xi(\eta_t \widehat{y}_t) \rangle}{\eta_t} .$$

Using the definition of the optimization problem (4.11) and $V_t' = \max\{0, \mathrm{opt}'_{q_t}(\eta_t)\}$, we have

$$(p_t - q_t)^\top \mathcal{L} e_{x_t} + \mathrm{bias}_{q_t}(G; x_t) + \frac{1}{\eta_t} \sum_{a=1}^{k} p_{ta} \left\langle q_t, \xi \left(\frac{\eta_t G(a, \sigma_t)}{p_{ta}} \right) \right\rangle \le \eta_t V_t' \tag{4.29}$$

Summing up the arguments in (4.26), (4.27), (4.28), and (4.29), we have

$$\mathsf{Reg}_T \le \mathbb{E} \left[\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^{T} \eta_t V_t' \right] ,$$

which completes the proof. □

4.6.5 Proof of Lemma 5.2

Proof. For any $q \in \mathcal{P}(\Pi)$ and $a^* \in \Pi$, we have

$$\begin{aligned}
H(p) &= \sum_{a \in \Pi} q_a \log \frac{1}{q_a} = \sum_{a \in \Pi \setminus \{a^*\}} q_a \log \frac{1}{q_a} + q_{a^*} \log \left(1 + \frac{1 - q_{a^*}}{q_{a^*}} \right) \\
&\le (k_\Pi - 1) \sum_{a \in \Pi \setminus \{a^*\}} \frac{1}{k_\Pi - 1} q_a \log \frac{1}{q_a} + q_{a^*} \frac{1 - q_{a^*}}{q_{a^*}} \\
&\le (k_\Pi - 1) \cdot \frac{\sum_{a \in \Pi \setminus \{a^*\}} q_a}{k_\Pi - 1} \log \frac{k_\Pi - 1}{\sum_{a \in \Pi \setminus \{a^*\}} q_a} + q_{a^*} \frac{1 - q_{a^*}}{q_{a^*}} \\
&= (1 - q_{a^*}) \left(\log \frac{k_\Pi - 1}{1 - q_{a^*}} + 1 \right) \le (1 - q_{a^*}) \log \frac{\mathrm{e} k_\Pi}{1 - q_{a^*}} ,
\end{aligned} \tag{4.30}$$

where the first inequality follows from $\log(1 + x) \le x$ for $x \ge 0$, the last inequality follows from Jensen's inequality, and the last equality follows from $\sum_{a \in \Pi} q_a = 1$. Using (4.30), for any $a^* \in [k]$ we have

$$\begin{aligned}
\sum_{t=1}^{T} a_t = \sum_{t=1}^{T} H(q_t) &\le \sum_{t=1}^{T} (1 - q_{ta^*}) \log \frac{\mathrm{e} k_\Pi}{1 - q_{ta^*}} \\
&= T \sum_{t=1}^{T} \frac{1}{T} (1 - q_{t,a^*}) \log \frac{\mathrm{e} k_\Pi}{1 - q_{t,a^*}} \\
&\le T \left(\sum_{t=1}^{T} \frac{1}{T} (1 - q_{t,a^*}) \right) \log \frac{\mathrm{e} k_\Pi}{\sum_{t=1}^{T} \frac{1}{T} (1 - q_{t,a^*})} \\
&= T \frac{Q(a^*)}{T} \log \frac{\mathrm{e} k_\Pi T}{Q(a^*)} = Q(a^*) \left(\log \frac{\mathrm{e} k_\Pi T}{Q(a^*)} \right) ,
\end{aligned}$$

where in the second inequality we used Jensen's inequality since $f(x) = x \log(1/x)$ is concave, and in the third inequality we used the definition of $Q(a^*)$ in (4.8). □

4.6.6 Proof of Theorem 4.1

Proof. We prove this theorem by bounding the RHS of Lemma 4.4.

(Bounding the penalty term) Let $t_0 = \min\{t \in [T] : \beta'_t \geq B\}$. Then, the definition of the learning rate (4.12) gives that

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(q_{t+1}) = \sum_{t=1}^{T} \left(\beta_{t+1} - \beta_t \right) H(q_{t+1}) \\
&= \sum_{t=1}^{t_0-2} \left(\beta_{t+1} - \beta_t \right) H(q_{t+1}) + \left(\beta_{t_0} - \beta_{t_0-1} \right) H(q_{t+1}) + \sum_{t=t_0}^{T} \left(\beta_{t+1} - \beta_t \right) H(q_{t+1}) \\
&\leq 0 + \left(\beta'_{t_0} - \beta'_{t_0-1} \right) H(q_{t+1}) + \sum_{t=t_0}^{T} \left(\beta'_{t+1} - \beta'_t \right) H(q_{t+1}) \\
&\leq \sum_{t=1}^{T} \left(\beta'_{t+1} - \beta'_t \right) H(q_{t+1}) \,,
\end{aligned}
$$

where in the first inequality we used the fact that β'_{t+1} is non-decreasing, $\beta_{t+1} = \beta_t$ for $t \leq t_0 - 1$, $\beta'_t \leq \beta_t$, and $\beta'_t = \beta_t$ for $t \geq t_0$. Using this inequality, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(q_{t+1}) \leq \sum_{t=1}^{T} \left(\beta'_{t+1} - \beta'_t \right) H(q_{t+1}) \\
&= \sum_{t=1}^{T} \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}} \cdot H(q_{t+1}) \\
&= 2c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \frac{H(q_{t+1})}{\sqrt{\log k_\Pi + \sum_{s=1}^{t} H(q_s)} + \sqrt{\log k_\Pi + \sum_{s=1}^{t} H(q_s)}} \\
&\leq 2c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \frac{H(q_{t+1})}{\sqrt{\sum_{s=1}^{t+1} H(q_s)} + \sqrt{\sum_{s=1}^{t} H(q_s)}} \\
&= 2c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \left(\sqrt{\sum_{s=1}^{t+1} H(q_s)} - \sqrt{\sum_{s=1}^{t} H(q_s)} \right) \\
&= 2c_1 \sqrt{\log k_\Pi} \left(\sqrt{\sum_{s=1}^{T+1} H(q_s)} - \sqrt{H(q_1)} \right) \\
&\leq 2c_1 \sqrt{\log k_\Pi} \left(\sqrt{\sum_{s=2}^{T+1} H(q_s)} \right) \leq 2c_1 \sqrt{\log k_\Pi} \sqrt{\sum_{t=1}^{T} H(q_t)} \,,
\end{aligned}
\tag{4.31}
$$

where the second inequality follows from $0 \leq H(q_{t+1}) \leq \log k_\Pi$, the third inequality follows from the inequality $\sqrt{a+b} - \sqrt{b} \leq \sqrt{a}$ that holds for $a, b \geq 0$, and the last inequality follows since $H(q_{T+1}) \leq H(q_1)$.

(Bounding the sum of the transformation and part of stability term) Using the definition of β'_t in (4.12), we can bound β'_t as

$$
\beta'_t = c_1 + \sum_{u=1}^{t-1} \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{u} H(q_s)}} \geq \frac{c_1 t}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}} \,.
$$

Using this inequality, we have

$$\sum_{t=1}^{T} \eta_t \leq \sum_{t=1}^{T} \frac{1}{\beta'_t} \leq \sum_{t=1}^{T} \frac{1}{c_1 t} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{s=1}^{t} H(q_s)} \leq \frac{1 + \log T}{c_1} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{t=1}^{T} H(q_t)} \,. \tag{4.32}$$

Further, we have

$$\sum_{t=1}^{T} \eta_t V'_t \leq \max_{s \in [T]} V'_s \sum_{t=1}^{T} \eta_t = \left(\max_{s \in [T]} \max \left\{ 0,\, \mathrm{opt}'_*(\eta_s) \right\} \right) \sum_{t=1}^{T} \eta_t \leq 3 m^2 k^3 \sum_{t=1}^{T} \eta_t \,, \tag{4.33}$$

where in the last inequality we used Lemma 4.3 with $\eta_t \leq 1/(2mk^2)$.

(Summing up the above arguments with a self-bounding technique) By bounding the RHS of Lemma 4.4 using (4.31), (4.32), and (4.33), we have

$$\begin{aligned} \mathsf{Reg}_T &\leq 3m^2k^3 \mathbb{E}\left[\frac{1 + \log T}{c_1} \sqrt{1 + (\log k_\Pi)^{-1} \sum_{t=1}^{T} H(q_t)} \right] + 2 c_1 \sqrt{\log k_\Pi}\, \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} H(q_t)} \right] + \frac{\log k_\Pi}{\eta_1} \\ &= O\left(m k^{3/2} \sqrt{\log(T) \sum_{t=1}^{T} H(q_t)} + m k^{3/2} \sqrt{\log(T) \log k_\Pi} \right) + 2mk^2 \log k_\Pi \,, \end{aligned} \tag{4.34}$$

where we set $c_1 = \Theta\left(m k^{3/2} \sqrt{\frac{\log T}{\log k_\Pi}} \right)$.

The desired bound is obtained for the adversarial regime, since $\sum_{t=1}^{T} H(q_t) \leq T \log k_\Pi$. We consider the stochastic regime in the following. If $Q(a^*) \leq \mathrm{e}$, Lemma 5.2 implies $\sum_{t=1}^{T} H(q_t) \leq \mathrm{e} \log(k_\Pi T)$ since $k_\Pi T \geq \mathrm{e}$, and otherwise we have $\sum_{t=1}^{T} H(q_t) \leq Q(a^*) \log(k_\Pi T)$. In the former case, we can trivially obtain the desired bound immediately from (4.34). For the latter case, using the inequality $\sum_{t=1}^{T} H(q_t) \leq Q(a^*) \log(k_\Pi T)$, (4.33), and Lemma 4.2 with $c = 1/(2k)$, we have for any $\lambda > 0$ that

$$\begin{aligned} &\mathsf{Reg}_T = (1 + \lambda) \mathsf{Reg}_T - \lambda \mathsf{Reg}_T \\ &\leq \mathbb{E}\left[(1 + \lambda) O\left(m k^{3/2} \sqrt{\log(T) \log(k_\Pi T) Q(a^*)} \right) - \frac{\lambda \Delta_{\min}}{2k} Q(a^*) \right] + \lambda C \\ &\leq O\left(\frac{(1 + \lambda)^2 m^2 k^4 \log(T) \log(k_\Pi T)}{\lambda \Delta_{\min}} \right) + \lambda C \\ &= O\left(\frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + \lambda \left(\frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + C \right) + \frac{1}{\lambda} \frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} \right) \,, \end{aligned} \tag{4.35}$$

where the second inequality follows from $a\sqrt{x} - bx/2 \leq a^2/(2b)$, which holds for any $a, b, x \geq 0$. Taking

$$\lambda = O\left(\sqrt{ m^2 k^4 \log(T) \log(k_\Pi T) \Big/ \left(\frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + C \right) } \right)$$

completes the proof. □

4.6.7 Proof of Lemma 4.6

Proof. Let $a^* = \arg\min_{a\in[k]} \mathbb{E}\bigl[\sum_{t=1}^T \mathcal{L}_{ax_t}\bigr]$ be the optimal action in hindsight, where ties are broken so that $a^* \in \Pi$. Note that since action a with $\dim(\mathcal{C}_a) < d-1$ cannot be uniquely optimal, one can see that we can take action $b \in \Pi$ instead of such a with the same loss. We have

$$\begin{aligned}
\mathsf{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{A_t,x_t} - \mathcal{L}_{a^*,x_t}\right)\right] = \mathbb{E}\left[\sum_{t=1}^T \langle p_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \langle q_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle + \sum_{t=1}^T \gamma_t \left\langle \frac{1}{k}\mathbf{1} - q_t, \mathcal{L}e_{x_t}\right\rangle\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \langle q_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle + \sum_{t=1}^T \gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k q_{t,a}\left(\mathcal{L}_{ax_t} - \mathcal{L}_{a^*x_t}\right) + \sum_{t=1}^T \gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k q_{t,a}\left(\widehat{y}_{ta} - \widehat{y}_{ta^*}\right) + \sum_{t=1}^T \gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^T \langle q_t - e_{a^*}, \widehat{y}_t\rangle + \sum_{t=1}^T \gamma_t\right],
\end{aligned}$$

where the inequality follows from the boundedness of $\mathcal{L}$, the fourth equality follows since $a^* \in \Pi$, $q_{t,a} = 0$ for $a \notin \Pi$, and Lemma 4.1, and the fifth equality follows from the definitions of $\widehat{y}$ and $q_{t,a} = 0$ for $a \notin \Pi$. Combining the above inequality and (5.2) completes the proof. □

4.6.8 Proof of Lemma 4.7

Proof. We first bound the stability term. Using Lemma 4.9, for any $a' \in \mathcal{A}$ it holds that

$$\begin{aligned}
\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) &= \langle \widehat{y}_t - \widehat{y}_{ta'}\mathbf{1}, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \\
&\le \beta_t \sum_{a=1}^k q_{t,a}\xi\left(\frac{\widehat{y}_{ta} - \widehat{y}_{ta'}}{\beta_t}\right).
\end{aligned}$$

We evaluate the RHS of this inequality. As we define p_t by (4.18), we have $p_{ta} \ge \gamma_t/k$ for any $a \in [k]$. We first show that $|(\widehat{y}_{ta} - \widehat{y}_{ta'})/\beta_t| \le 1$ for all $a, a' \in [k]$. Let $\tau = \|G\|_\infty$. Recall that $c_{\mathcal{G}} = \max\{1, k\tau\}$. Then we have

$$\frac{\widehat{y}_t}{\beta_t} = \frac{G(a, \Phi_{ax})}{\beta_t\, p_{tA_t}} \ge -\frac{\tau}{\beta_t\, p_{tA_t}}\mathbf{1} \ge -\frac{1}{2}\mathbf{1},$$

where the inequalities here are element-wise, the first inequality follows from the definition of τ, and in the last inequality we used $p_{ta} \ge \gamma_t/k \ge c_{\mathcal{G}}/(2\beta_t k) \ge \tau/(2\beta_t)$ for all $a \in [k]$. In a similar manner we have

$$\frac{\widehat{y}_t}{\beta_t} = \frac{G(a, \Phi_{ax})}{\beta_t\, p_{tA_t}} \le \frac{\tau}{\beta_t\, p_{tA_t}}\mathbf{1} \le \frac{1}{2}\mathbf{1}.$$

These arguments conclude that $|(\widehat{y}_{ta} - \widehat{y}_{ta'})/\beta_t| \le |\widehat{y}_{ta}/\beta_t| + |\widehat{y}_{ta'}/\beta_t| \le 1$ for all $a, a' \in [k]$. Hence, we have

$$\begin{aligned}
\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) &\le \min_{a'\in[k]} \beta_t \sum_{a=1}^k q_{t,a}\left(\frac{\widehat{y}_{ta} - \widehat{y}_{ta'}}{\beta_t}\right)^2 \\
&= \frac{1}{\beta_t}\min_{a'\in[k]} \sum_{a=1}^k q_{t,a}\left(\widehat{y}_{ta} - \widehat{y}_{ta'}\right)^2 \\
&= \frac{1}{\beta_t}\min_{a'\in[k]} \sum_{a\ne a'} q_{t,a}\left(\widehat{y}_{ta} - \widehat{y}_{ta'}\right)^2, \qquad (4.36)
\end{aligned}$$

where the inequality follows from (4.23). Now, for any $a \in \mathcal{A}$ we have

$$\mathbb{E}[\widehat{y}_{ta}^2] = \mathbb{E}\left[\left(\frac{G(A_t, \Phi_{A_t x_t})}{p_{tA_t}}\right)^2\right] \le \mathbb{E}\left[\sum_{a=1}^{k} p_{ta} \frac{\|G\|_\infty^2}{p_{ta}^2}\right] \le \sum_{a=1}^{k} \frac{k\|G\|_\infty^2}{\gamma_t} = \frac{c_{\mathcal{G}}^2}{\gamma_t}, \tag{4.37}$$

where the last inequality follows from $p_{ta} \ge \gamma_t/k$. Hence, using (4.37) it holds that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\beta_t} \min_{a' \in [k]} \sum_{a \neq a'} q_{t,a} \left(\widehat{y}_{ta} - \widehat{y}_{ta'}\right)^2\right] &\le \mathbb{E}\left[\frac{2}{\beta_t} \min_{a' \in [k]} \sum_{a \neq a'} q_{t,a} \frac{c_{\mathcal{G}}^2}{\gamma_t}\right] \\ &= \mathbb{E}\left[\frac{2 \min_{a' \in [k]}(1 - q_{ta'})c_{\mathcal{G}}^2}{\beta_t \gamma_t}\right] = \mathbb{E}\left[\frac{2c_{\mathcal{G}}^2 b_t}{\beta_t \gamma_t}\right]. \end{aligned} \tag{4.38}$$

Combining (4.36) and (4.38) yields

$$\mathbb{E}[\langle \widehat{y}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t)] \le \mathbb{E}\left[\frac{2c_{\mathcal{G}}^2 b_t}{\beta_t \gamma_t}\right],$$

which completes the proof. □

4.6.9 Proof of Proposition 4.1

Proof. Note that the penalty term can be rewritten as

$$\begin{aligned}&\sum_{t=1}^{T} (\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})) + \psi_{T+1}(e_{a^*}) - \psi_1(q_1) \\ &= \sum_{t=1}^{T} (\beta_t - \beta_{t+1}) (-H(q_{t+1})) + \beta_1 H(q_1) = \sum_{t=1}^{T} (\beta_{t+1} - \beta_t)\, a_{t+1} + \beta_1 a_1 ,\end{aligned}$$

where we recall that the definition of a_t in (4.17). Combining this with Lemmas 4.6 and 4.7, we have

$$\mathsf{Reg}_T \le \underbrace{\sum_{t=1}^{T} \left(\gamma_t' + \frac{c_{\mathcal{G}}}{2\beta_t}\right)}_{\text{transformation term}} + \underbrace{\sum_{t=1}^{T} \frac{2c_{\mathcal{G}}^2 b_t}{\beta_t \gamma_t}}_{\text{stability term}} + \underbrace{\sum_{t=1}^{T} ((\beta_{t+1} - \beta_t) a_{t+1}) + \beta_1 a_1}_{\text{penalty term}} , \tag{4.39}$$

where the first, second, and remaining terms correspond to the transformation, stability, and penalty terms, respectively. We bound each term of the RHS in (4.39) in the following.

Note that $b_t \le 1$ and

$$b_t = 1 - \max_{a \in [k]} q_{t,a} \le - \max_{a \in [k]} q_{t,a} \log\left(\max_{a' \in [k]} q_{t,a'}\right) \le - \sum_{a \in [k]} q_{t,a} \log q_{t,a} = a_t \le \log k_\Pi , \tag{4.40}$$

where the first inequality follows from the inequality $1 - x \le -x \log x$ for $x > 0$. We define $z_t = \frac{a_{t+1} b_t}{\gamma_t'}$ and $Z_t = \sum_{s=1}^{t} z_s$.

(Bounding the penalty term) From the definition of γ'_t, we can bound z_t from below as

$$z_t = \frac{a_{t+1} b_t}{\gamma'_t} = \frac{4a_{t+1}}{c_1}\left(c_1 + B_t^{1/3}\right) \geq 4a_{t+1} \geq 4b_{t+1} \,, \tag{4.41}$$

where the second inequality follows from $b_t \leq a_t$ in (4.40). Further, we can bound z_t from above as

$$z_t = \frac{4a_{t+1}}{c_1}\left(c_1 + B_t^{1/3}\right) \leq 4\left(c_1 + B_t^{1/3}\right) \leq 4\left\{c_1 + \left(b_1 + \sum_{s=1}^{t-1} z_s\right)^{1/3}\right\} \leq 8\left(c_1 + Z_{t-1}\right) \,, \tag{4.42}$$

where the first inequality follows from $a_{t+1} \leq \log k_\Pi$ and $c_1 \geq \log k_\Pi$, and the second inequality follows from $B_t = b_1 + \sum_{s=1}^{t-1} b_{s+1} \leq b_1 + \sum_{s=1}^{t-1} z_s$, and the last inequality follows from $b_1 \leq 1 \leq c_1$. From this, since β_t satisfies $\beta_{t+1} - \beta_t = \frac{z_t}{a_{t+1}} \frac{c_2}{(c_1 + Z_{t-1})^{1/2}}$, we can bound the penalty term in (4.39) as

$$\begin{aligned} \sum_{t=1}^T (\beta_{t+1} - \beta_t) a_{t+1} &= c_2 \sum_{t=1}^T \frac{z_t}{\sqrt{c_1 + Z_{t-1}}} = 5c_2 \sum_{t=1}^T \frac{Z_t - Z_{t-1}}{4\sqrt{c_1 + Z_{t-1}} + \sqrt{c_1 + Z_{t-1}}} \\ &< 5c_2 \sum_{t=1}^T \frac{Z_t - Z_{t-1}}{\sqrt{c_1 + Z_t} + \sqrt{c_1 + Z_{t-1}}} = 5c_2 \sum_{t=1}^T \left(\sqrt{c_1 + Z_t} - \sqrt{c_1 + Z_{t-1}}\right) \leq 5c_2 \sqrt{Z_T} \,, \end{aligned} \tag{4.43}$$

where the first equality follows from the definitions of β_t and z_t, and the first inequality follows since

$$\sqrt{c_1 + Z_t} \leq \sqrt{c_1 + Z_{t-1}} + \sqrt{z_t} < 4\sqrt{c_1 + Z_{t-1}} \,,$$

where the last inequality follows from (4.42).

(Bounding the stability term and transformation terms) We define $w_t = \frac{b_t}{\gamma'_t}$ and $W_t = \sum_{s=1}^t w_s$. From the definition of γ'_t, we have

$$w_t = \frac{b_t}{\gamma'_t} = 4\left(1 + \frac{1}{c_1} B_t^{1/3}\right) \geq 4 \,. \tag{4.44}$$

Using $b_t \leq 1$, we can confirm that w_t satisfies

$$w_1 \leq 8 \,,\; w_{t+1} = 4\left(1 + \frac{1}{c_1} B_{t+1}^{1/3}\right) \leq \left(1 + \frac{1}{c_1}(B_t + 1)^{1/3}\right) \leq 2w_t \,,\; w_t \leq 4(1 + t^{1/3}) \,. \tag{4.45}$$

Then β_t can be bounded as

$$\begin{aligned} \beta_t &\geq c_2 + c_2 \sum_{s=1}^{t-1} \frac{w_s}{\sqrt{c_1 + Z_{s-1}}} \geq \frac{c_2}{\sqrt{c_1 + Z_t}}\left(1 + \sum_{s=1}^{t-1} w_s\right) \\ &= \frac{c_2}{\sqrt{c_1 + Z_t}}(1 + W_{t-1}) \qquad (4.46) \\ &\geq \frac{c_2 t}{\sqrt{c_1 + Z_t}} \,, \qquad (4.47) \end{aligned}$$

where the second inequality follows from (4.44).

Using the above inequalities, we can bound the stability term in (4.39) as

$$\sum_{t=1}^{T} \frac{b_t}{\gamma_t \beta_t} \leq \sum_{t=1}^{T} \frac{b_t}{\gamma_t' \beta_t} \leq \sum_{t=1}^{T} \frac{\sqrt{c_1 + Z_t}}{c_2} \frac{w_t}{1 + W_{t-1}} \leq \frac{\sqrt{c_1 + Z_T}}{c_2} \sum_{t=1}^{T} \frac{w_t}{1 + W_{t-1}}$$
$$\leq O\left(\frac{\sqrt{c_1 + Z_T}}{c_2} \log\left(1 + W_T\right) \right) \leq O\left(\frac{\sqrt{c_1 + Z_T}}{c_2} \log T \right) , \qquad (4.48)$$

where the first inequality follows from (4.46), the last inequality follows from (4.45), and the fourth inequality can be shown by taking the sum of the following inequality:

$$\log(1 + W_t) - \log(1 + W_{t-1}) = \log \frac{1 + W_t}{1 + W_{t-1}} = \log\left(1 + \frac{w_t}{1 + W_{t-1}}\right) \geq \frac{1}{2} \cdot \frac{w_t}{1 + W_{t-1}} ,$$

where the inequality follows from the fact that $\log(1 + x) \geq \frac{1}{2}x$ holds for any $x \in [0, 2]$ and that (4.45) implies $\frac{w_t}{1+W_{t-1}} \leq \frac{w_t}{1+w_t/2} \leq 2$ for all $t \in [T]$.

Using (4.47), we can bound the second part of the transformation term in (4.39) as

$$\sum_{t=1}^{T} \frac{1}{\beta_t} \leq \sum_{t=1}^{T} \frac{\sqrt{c_1 + Z_t}}{c_2 t} \leq \frac{\sqrt{c_1 + Z_T}}{c_2} \sum_{t=1}^{T} \frac{1}{t} = O\left(\frac{\sqrt{c_1 + Z_T}}{c_2} \log T \right) . \qquad (4.49)$$

In addition, from the definition of γ_t', we can bound the remaining part of the transformation term in (4.39) as

$$\sum_{t=1}^{T} \gamma_t' = \frac{c_1}{4} \sum_{t=1}^{T} \frac{b_t}{c_1 + B_t^{1/3}} \leq \frac{3c_1}{8} \sum_{t=1}^{T} \left(B_t^{2/3} - B_{t-1}^{2/3} \right) \leq \frac{3c_1}{8} B_T^{2/3} , \qquad (4.50)$$

where the first inequality follows from $y^{2/3} - x^{2/3} \geq \frac{2}{3}(y - x)y^{-1/3}$, which holds for any $y \geq x > 0$. Combining (4.43), (4.48), (4.49), and (4.50), we can bound the right-hand side of (4.39) as

$$\sum_{t=1}^{T} \left(\gamma_t + \frac{2c_{\mathcal{G}}^2 b_t}{\gamma_t \beta_t} + (\beta_{t+1} - \beta_t) a_{t+1} \right) + \beta_1 a_1$$
$$= \sum_{t=1}^{T} \left(\gamma_t' + \frac{c_{\mathcal{G}}}{2\beta_t} + \frac{2c_{\mathcal{G}}^2 b_t}{\gamma_t \beta_t} + (\beta_{t+1} - \beta_t) a_{t+1} \right) + \beta_1 a_1$$
$$= O\left(c_1 B_T^{2/3} + \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{c_1 + Z_T} + \beta_1 a_1 \right)$$
$$= O\left(c_1 B_T^{2/3} + \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{c_1 + \sum_{t=1}^{T} \frac{a_{t+1}}{c_1} \left(c_1 + B_t^{1/3} \right)} + \beta_1 a_1 \right)$$
$$= O\left(c_1 B_T^{2/3} + \frac{1}{\sqrt{c_1}} \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{c_1^2 + (\log k_\Pi + A_T) \left(c_1 + B_T^{1/3} \right)} + \beta_1 \log k_\Pi \right) ,$$

where in the third inequality we used (4.41) and in the last equality we used $a_{T+1} = O(\log k_\Pi)$. □

4.6.10 Proof of Theorem 4.2

Proof. We define c_1 and c_2 by

$$c_1 = \Theta\left(\left(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T) \right)^{1/3} \right) \quad \text{and} \quad c_2 = \Theta\left(\sqrt{c_{\mathcal{G}}^2 \log T} \right) , \qquad (4.51)$$

which implies that $\tilde{c} = c_1 / \sqrt{\log(k_\Pi T)}$. We have

$$B_T = \sum_{t=1}^{T} \left(1 - \max_{a \in \Pi} q_{t,a} \right) \leq \sum_{t=1}^{T} (1 - q_{t,a^*}) = Q(a^*) \,. \tag{4.52}$$

We first consider the adversarial regime. Since $A_T \leq T \log k_\Pi$ and $B_T \leq T$, using Proposition 4.1 we have

$$\begin{aligned}
\mathsf{Reg}_T &= O\left(c_1 T^{2/3} + \tilde{c}\sqrt{c_1^2 + (\log k_\Pi + T \log k_\Pi)(c_1 + T^{1/3})} + \beta_1 \log k_\Pi \right) \\
&= O\left(\left(c_1 + \tilde{c}\sqrt{\log k_\Pi} \right) T^{2/3} + \sqrt{\frac{\log k_\Pi}{\log(k_\Pi T)}} c_1^{3/2} T^{1/2} + \frac{c_1^2}{\sqrt{\log(k_\Pi T)}} + \beta_1 \log k_\Pi \right) .
\end{aligned} \tag{4.53}$$

We next consider the adversarial regime with a self-bounding constraint. When $Q(a^*) \leq c_1^3$ we can show that the obtained bound is smaller than the desired bound as follows. When $Q(a^*) \leq \mathrm{e} \leq c_1^3$, using Lemma 5.2 and (4.52), we have $A_T \leq \mathrm{e} \log(k_\Pi T)$ and $B_T \leq \mathrm{e}$. Hence, from Proposition 4.1, we have

$$\begin{aligned}
\mathsf{Reg}_T &= O\left(c_1 + \tilde{c}\sqrt{c_1^2 + \log(k_\Pi T) c_1} + \beta_1 \log k_\Pi \right) \\
&= O\left(\frac{c_1^2}{\sqrt{\log(k_\Pi T)}} + \beta_1 \log k_\Pi \right) = O\left(c_1^3 \right) \,.
\end{aligned}$$

When $\mathrm{e} < Q(a^*) \leq c_1^3$, using Lemma 5.2 and (4.52) we have $A_T \leq c_1^3 \log(k_\Pi T)$ and $B_T \leq c_1^3$. Hence, from Proposition 4.1, we have

$$\begin{aligned}
\mathsf{Reg}_T &= O\left(c_1^3 + \tilde{c}\sqrt{c_1^2 + \left(\log k_\Pi + c_1^3 \log(k_\Pi T) \right) c_1} + \beta_1 \log k_\Pi \right) \\
&= O\left(c_{\mathsf{g}}^2 \log(T) \log(k_\Pi T) \right) = O\left(c_1^3 \right) \,.
\end{aligned}$$

Hence, we only need to consider the case of $Q(a^*) > c_1^3$ in the following. Since $Q(a^*) \geq \mathrm{e}$ we have $A_T \leq Q(a^*) \log(k_\Pi T)$. Using Proposition 4.1 with this inequality, Lemma 5.2, and (4.52), we have

$$\begin{aligned}
\mathsf{Reg}_T &= O\left(\mathbb{E}\left[c_1 Q(a^*)^{2/3} + \tilde{c}\sqrt{c_1^2 + \left(\log k_\Pi + Q(a^*) \log(k_\Pi T) \right) \left(c_1 + Q(a^*)^{1/3} \right)} \right] + \beta_1 \log k_\Pi \right) \\
&\leq O\left(\mathbb{E}\left[c_1 Q(a^*)^{2/3} + \tilde{c}\sqrt{Q(a^*) \log(k_\Pi T) Q(a^*)^{1/3}} \right] \right) \\
&\leq O\left(\left(c_1 + \tilde{c}\sqrt{\log(k_\Pi T)} \right) \bar{Q}(a^*)^{2/3} \right) \,,
\end{aligned} \tag{4.54}$$

where the first inequality follows from $Q(a^*) > c_1^3$, and the second inequality follows from Jensen's inequality. Hence, by (4.53) and (4.54), there exists $\widehat{c} = O\left(c_1 + \tilde{c}\sqrt{\log(k_\Pi T)} \right)$ satisfying and $\mathsf{Reg}_T \leq \widehat{c} \bar{Q}(a^*)^{2/3}$ for the adversarial regime with a self-bounding constraint and $\mathsf{Reg}_T \leq \widehat{c} T^{2/3}$ for the adversarial regime.

Now, by recalling the definitions of c_1 and c_2 in (4.51), we have

$$\begin{aligned}
\widehat{c} &= O\left(\left(c_{\mathsf{g}}^2 \log(T) \log(k_\Pi T) \right)^{1/3} + \frac{1}{\sqrt{c_1}} \left(\frac{c_{\mathsf{g}}^2 \log T}{c_2} + c_2 \right) \sqrt{\log(k_\Pi T)} \right) \\
&= O\left(\left(c_{\mathsf{g}}^2 \log(T) \log(k_\Pi T) \right)^{1/3} \right) \,,
\end{aligned} \tag{4.55}$$

which gives the desired bounds for the adversarial regime.

For the adversarial regime with a self-bounding constraint, from the above inequality $\mathsf{Reg}_T \le \widehat{c}\bar{Q}(a^*)^{2/3}$ and Lemma 4.2 with $c = 1/2 \le 1 - \gamma_t$, we have for any $\lambda \in (0,1]$ that

$$\begin{aligned}\mathsf{Reg}_T &= (1+\lambda)\mathsf{Reg}_T - \lambda\mathsf{Reg}_T \le (1+\lambda)\widehat{c}\cdot\bar{Q}(a^*)^{2/3} - \frac{\lambda}{2}\Delta_{\min}\bar{Q}(a^*) + \lambda C \\ &\le O\left(\frac{(1+\lambda)^3\widehat{c}^3}{\lambda^2\Delta_{\min}^2}\right) + \lambda C = O\left(\left(1+\frac{1}{\lambda^2}\right)\frac{\widehat{c}^3}{\Delta_{\min}^2}\right) + \lambda C\,, \qquad (4.56)\end{aligned}$$

where the first inequality follows from the inequality $ax^{2/3} - b(x/2) \le 16a^3/(27b^2)$ for $a, b > 0$, and the last equality follows since $\lambda \in (0,1]$. Combining (4.55) and (4.56), and taking $\lambda = O\left(\frac{c_{\mathcal{G}}^2\log(T)\log(k_\Pi T)}{C\Delta_{\min}^2}\right)$, we have the desired result for the adversarial regime with a self-bounding constraint. □

4.6.11 Regret Bounds when the Optimization Problem is Not Exactly Solved

This section discusses the regret bound when the optimization problem (4.11) is not exactly solved, on which a similar discussion is given in Lattimore and Szepesvári (2020a, Chapter 37). We say that the optimization problem (4.11) can be solved with precision $\epsilon \ge 0$, if we can obtain $G \in \mathcal{H}$ and $p \in \mathcal{P}_k'(q)$ such that

$$\max_{x\in[d]}\left[\frac{(p-q)^\top\mathcal{L}e_x + \mathrm{bias}_q(G;x)}{\eta} + \frac{1}{\eta^2}\sum_{a=1}^k p_a\left\langle q, \xi\left(\frac{\eta G(a,\Phi_{ax})}{p_a}\right)\right\rangle\right] \le \mathrm{opt}_q'(\eta) + \epsilon\,.$$

Then if we run Algorithm 4.1 solving (4.11) with precision ϵ, one can see that we can obtain the following regret bounds. For the adversarial regime with a (Δ, C, T) self-bounding constraint, we have

$$\mathsf{Reg}_T = O\left(\frac{\left(mk^2 + \epsilon^2/(mk)\right)^2\log(T)\log(k_\Pi T)}{\Delta_{\min}} + \sqrt{\frac{C\left(mk^2+\epsilon^2/(mk)\right)^2\log(T)\log(k_\Pi T)}{\Delta_{\min}}}\right)\,,$$

and for the adversarial regime, we have

$$\mathsf{Reg}_T = O\left(mk^{3/2}\sqrt{T\log(T)\log k_\Pi} + \epsilon\,\frac{\sqrt{T\log(k_\Pi)\log(T)}}{mk^{3/2}}\right)\,.$$

Here, we give an overview of the analysis. Considering that the optimization problem in (4.11) can be solved with precision $\epsilon \ge 0$, the RHS of (4.29) can be replaced with $3m^2k^3 + \epsilon$. Then a similar analysis as the proof of Theorem 4.1 leads to

$$\mathsf{Reg}_T \le O\left(\left(mk^{3/2} + \frac{\epsilon}{mk^{3/2}}\right)\sqrt{\log(T)\sum_{t=1}^T H(q_t)}\right)\,.$$

Using $\sum_{t=1}^T H(q_t) \le T\log k_\Pi$ gives the bound for the adversarial regime. Replacing m^2k^4 with $\left(mk^2 + \frac{\epsilon}{mk}\right)^2$ in (4.35) and appropriately choose λ (note that we can take λ depending on ϵ), we obtain the desired bound for the adversarial regime with a self-bounding constraint.

4.7 Conclusion

In this chapter, we considered PM and provided the first best-of-both-worlds algorithms for both locally and globally observable games, both of which are based on the FTRL framework. For non-degenerate locally observable games, we showed that the regret is $O(m^2 k^4 \log(T) \log(k_\Pi T)/\Delta_{\min})$ in the stochastic regime and $O(mk^{3/2}\sqrt{T \log(T) \log k_\Pi})$ in the adversarial regime. To obtain this bound, we advanced the technique of exploration by optimization, which is a technique for bounding the stability component of regret, by considering the optimization problem over the restricted feasible set. This enabled us to use the self-bounding technique to prove the BOBW guarantee. We also show for globally observable games, the regret is $O(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T)/\Delta_{\min}^2)$ in the stochastic regime and $O((c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T))^{1/3} T^{2/3})$ in the adversarial regime. To obtain this goal, we modified the technique of adaptive learning rate developed in online learning with feedback graphs. Moreover, we developed a novel analysis to circumvent the optimization problem in computing the FTRL, which enabled us to implement the proposed algorithm efficiently.

Chapter 5

Stability-penalty-adaptive Follow the Regularized Leader: Sparsity, Game dependency, and Best of Both Worlds

Adaptivity to the difficulties of a problem is a key property in sequential decision-making problems to broaden the applicability of algorithms. Follow-the-regularized-leader has recently emerged as one of the most promising approaches for obtaining various types of adaptivity in bandit problems. In fact in the previous chapter, we observed that the follow-the-regularized-leader is a quite strong framework to achieve the best-of-both-worlds guarantee, which aims to achieve a near-optimal regret in both the stochastic and adversarial regimes. Aiming to further generalize this adaptivity, we develop a generic adaptive learning rate, called *stability-penalty-adaptive learning rate* for follow-the-regularized-leader. This learning rate yields a regret bound jointly depending on stability and penalty of the algorithm, into which the regret of follow-the-regularized-leader is typically decomposed. With this result, we establish several algorithms with three types of adaptivity: *sparsity*, *game-dependency*, and *best-of-both-worlds*. Sparsity frequently appears in real-world problems. However, existing sparse multi-armed bandit algorithms with k-arms assume that the sparsity level $s \leq k$ is known in advance, which is often not the case in real-world scenarios. To address this problem, with the help of the new learning rate framework, we establish *s-agnostic* algorithms with regret bounds of $\tilde{O}(\sqrt{sT})$ in the adversarial regime for T rounds, which matches the existing lower bound up to a logarithmic factor. Furthermore, leveraging the new adaptive learning rate framework and a novel analysis to bound the variation in follow-the-regularized-leader output in response to changes in a regularizer, we establish the first best-of-both-worlds algorithm with a sparsity-dependent bound. Additionally, we explore partial monitoring and demonstrate that the proposed learning rate framework allows us to achieve the best-of-both-worlds and game-dependent bounds simultaneously.

5.1 Introduction

This study considers the Multi-Armed Bandits (MAB) and Partial Monitoring (PM). In the MAB problem, the learner selects one of k arms, and the adversary simultaneously determines the loss of each arm, $\ell_t = (\ell_{t1}, \ldots, \ell_{tk})^\top$ in $[0,1]^k$ or $[-1,1]^k$. After that, the learner observes only the loss for the chosen arm. The learner's goal is to minimize the regret, which is the difference between the learner's total loss and the total loss of an optimal arm fixed in hindsight. PM is a generalization of MAB, and the learner observes feedback symbols instead of the losses.

One of the most promising frameworks for MABs and PM is Follow-the-Regularized-Leader (FTRL) (Auer et al., 2002b; Cesa-Bianchi et al., 2006), which determines the arm selection probability at each round by minimizing the sum of the cumulative (estimated) losses so far plus a convex regularizer. Note that the well-known Exp3 algorithm developed in Auer et al. (2002b) is equivalent to FTRL with the (negative) Shannon entropy regularizer. FTRL is also known to perform well for the classic expert problem (Freund and Schapire, 1997) and reinforcement learning (Zimin and Neu, 2013). Furthermore, when the problem is "benign", it is known that FTRL can exploit the underlying structure to adaptively improve its performance. Typical examples of such adaptive improvements are (i) *data-dependent bounds* and (ii) *Best-of-Both-Worlds (BOBW)*.

Data-dependent bounds have been investigated to enhance the adaptivity of algorithms to a given structure of losses in the *adversarial regime*, where feedback (*e.g.,* losses in MABs) is decided in an arbitrary manner. There are various examples of data-dependent bounds, and this study considers *sparsity-dependent bounds* and *game-dependent bounds*.

A sparsity-dependent bound is an important example of data-dependent bounds, as sparsity frequently appears in real-world problems. For example, in online advertisement allocation, it is often the case that only a fraction of the ads is clicked. Although there are some studies for sparse MABs (Kwon and Perchet, 2016; Bubeck et al., 2018; Zheng et al., 2019), all of them assume that (an upper bound of) sparsity level $s \geq \|\ell_t\|_0 = |\{i \in [k] : \ell_{ti} \neq 0\}|$ is known beforehand, which in many practical scenarios does not hold.

The concept of a game-dependent bound was recently introduced by Lattimore and Szepesvári (2020b) to derive a regret upper bound that depends on the game the learner is facing. As the authors suggest, one of the motivations for the game-dependent bound is that previous PM algorithms are "quite conservative and not practical for normal problems". For example, whereas the Bernoulli MAB is expressed as a PM, algorithms for PM do not always achieve the minimax regret of MAB (Auer et al., 2002b). The game-dependent bound enables the learner to automatically adapt to the essential difficulty of the game the algorithm is actually facing.

The BOBW algorithm aims to achieve near-optimal regret bounds in stochastic and adversarial regimes, where the feedback is stochastically generated in the stochastic regime. Since we often do not know the underlying regime, it is desirable for an algorithm to *simultaneously* obtain a near-optimal performance both for the stochastic and adversarial regimes. Bubeck and Slivkins (2012) developed the first BOBW algorithm, and Zimmert and Seldin (2021) proposed the well-known Tsallis-INF algorithm, which achieves the optimal regret for both regimes. The Tsallis-INF algorithm also achieves favorable regret guarantees in the *adversarial regime with a self-bounding constraint*, which interpolates between the stochastic and adversarial regimes.

To realize the aforementioned adaptivity in FTRL, the *adaptive learning rate* (a.k.a. time-varying learning rate) is one of the most representative approaches. This approach adjusts the learning rate based on previous observations. In the literature, adaptive learning rates have been designed to depend on *stability* or *penalty*, which are components of a regret upper bound of FTRL. The stability term increases if the variation of FTRL outputs in the adjacent rounds is large, and stability-dependent learning rates have been used in a considerable number of algorithms available in the literature, *e.g.,* McMahan (2011); Lattimore and Szepesvári (2020b); Orabona (2019) and references therein. In contrast, the penalty term comes from the strength of the regularization, and recently penalty-dependent learning rates were considered to achieve BOBW guarantees (Ito et al., 2022a; Tsuchiya et al., 2023a). However, existing stability-dependent (resp. penalty-dependent) learning rates are designed with the worst-case penalty (resp. stability), which could potentially limit the adaptivity and performance of FTRL. (There are numerous studies related to this chapter and we include additional related work in Section 5.7.2.)

5.1.1 Contributions of this Chapter

In this chapter, to further broaden the applicability of FTRL, we establish a generic framework for designing an adaptive learning rate that depends on both the stability and penalty components simultaneously, which we call a *Stability-Penalty-Adaptive (SPA) learning rate* (Definition 5.2). This enables us to bound a regret approximately by $\tilde{O}\left(\sqrt{\sum_{t=1}^{T} z_t h_{t+1}}\right)$ for a stability component (z_t) and a penalty component (h_t), which we call a *SPA regret bound* (Theorem 5.1). With appropriate selections of z_t and h_t, this result yields the three important adaptive bounds mentioned earlier, namely *sparsity, game-dependency, and BOBW*. In particular, our contributions are as follows (see also Tables 5.1 and 5.2):

- (Section 5.5.1) We initially provide new algorithms for sparse MABs as preliminaries for establishing a BOBW algorithm with a sparsity-dependent bound. In Section 5.5.1.1, we propose a novel estimator of the sparsity level, which is linked to a stability component and induces $L_2 = \sum_{t=1}^{T} \|\ell_t\|_2^2 \leq sT$. We demonstrate that a learning rate using this estimator with the Shannon entropy regularizer and $\tilde{\Theta}((kT)^{-2/3})$ uniform exploration immediately results in an $O(\sqrt{L_2 \log k})$ regret bound for $\ell_t \in [0,1]^k$. In Section 5.5.1.2, we investigate possibly negative losses $\ell_t \in [-1,1]^k$. We employ the time-invariant log-barrier proposed in Bubeck et al. (2018) and control the stability term. This allows us to achieve an $O(\sqrt{L_2 \log k})$ regret bound for losses in $[-1,1]^k$ even *without* the $\tilde{\Theta}((kT)^{-2/3})$ uniform exploration. This is a key factor for the BOBW property that we discuss next. Note that Section 5.5.1 serves as preliminary findings for the subsequent section.
- (Section 5.5.2) We establish a BOBW algorithm with a sparsity-dependent bound. In order to achieve this goal, we make another major technical development: we analyze the variation in the FTRL output when the regularizer changes (Lemma 5.8), which holds thanks to the time-invariant log-barrier and may be of independent interest. This analysis is necessary since we use a time-varying learning rate, whereas Bubeck et al. (2018) uses a constant learning rate. This technical development successfully allows us to achieve the goal (Theorem 5.2) in combination with the SPA learning rate developed in Section 5.4 and a technique for exploiting sparsity in Section 5.5.1.2.
- (Section 5.6) We show that the SPA learning rate established in Section 5.4 can also be used to achieve a game-dependent bound and a BOBW guarantee simultaneously, which further highlights the usefulness of the SPA learning rate.

5.2 Setup

This section introduces the preliminaries of this study. Sections 5.2.1 and 5.2.2 formulate the MAB and PM problems, respectively, and Section 5.2.3 defines regimes considered in this chapter.

Notation Let $\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ be the Euclidian, ℓ_1-, and ℓ_∞-norms for a vector x, respectively. Let $\|x\|_0$ be the number of non-zero elements for a vector x. Let $\mathcal{P}_k = \{p \in [0,1]^k : \|p\|_1 = 1\}$ be the $(k-1)$-dimensional probability simplex. A vector $e_i \in \{0,1\}^k$ is the i-th orthonormal basis of $\mathbb{R}^k$, and $\mathbf{1}$ is the all-one vector. Let $D_\psi : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$ be the *Bregman divergence* induced by $\psi \colon \mathbb{R}^k \to \mathbb{R}$, *i.e.,* $D_\psi(p,q) = \psi(p) - \psi(q) - \langle \nabla\psi(q), p - q\rangle$.

Table 5.1: Regret upper bounds with sparsity-dependent bounds. T is the time horizon. $s \leq k$ is the level of sparsity in losses. We define $L_2 = \sum_{t=1}^{T} \|\ell_t\|_2^2$ and $\|\ell_t\|_0 \leq s$ implies $L_2 = \sum_{t=1}^{T} \|\ell_t\|_2^2 \leq sT$ since $\|\ell_t\|_\infty \leq 1$. $\Delta_{\min}$ is the minimum suboptimality gap. Adv. and Stoc. are the abbreviations of the adversarial and stochastic regime, respectively.

Reference	s-agnostic?	Range of ℓ_{ti}	Regime	Regret bound
Kwon and Perchet (2016)	–	$[0,1]$	Adv.	$\Omega(\sqrt{sT})$
Kwon and Perchet (2016)	No	$[0,1]$	Adv.	$2\sqrt{e}\sqrt{sT\log(k/s)}$
Ours (Sec. 5.5.1.1, Cor. 5.1)	Yes	$[0,1]$	Adv.	$2\sqrt{2}\sqrt{L_2 \log k} + O((kT\log k)^{1/3})$
Bubeck et al. (2018)	No	$[-1,1]$	Adv.	$10\sqrt{L_2 \log k} + 20k\log T$
Ours (Sec. 5.5.1.2, Cor. 5.2)	Yes	$[-1,1]$	Adv.	$4\sqrt{2}\sqrt{L_2 \log k} + 2k\log T$
Ours (Sec. 5.5.2, Thm. 5.2)	Yes	$[-1,1]$	Adv. Stoc.	$4\sqrt{L_2 \log k \log T} + O(k\log T)$ $O(k\log(T)\log(kT)/\Delta_{\min})$

Table 5.2: Regret bounds for non-degenerate local PM games. V_t, V_t', and $\bar{V}'$ are game-dependent quantities satisfying $V_t \leq V_t' \leq \bar{V}$ (see Section 5.6 for definitions). $H(q_t)$ is the Shannon entropy for FTRL output q_t.

Reference	Game-dependent?	BOBW?	Order of regret bound
Many existing studies on PM	No	No	–
Lattimore and Szepesvári (2020b)	Yes	No	$\sqrt{\sum_{t=1}^{T} V_t \log k}$
Tsuchiya et al. (2023a)	No (only game-class-dependent)	Yes	$\sqrt{\bar{V}\sum_{t=1}^{T} H(q_{t+1})}$
Ours (Sec. 5.6, Cor. 5.3)	Yes	Yes	$\sqrt{\sum_{t=1}^{T} V_t' H(q_{t+1}) \log T}$

5.2.1 Multi-armed Bandits

In MAB with k-arms, at each round $t \in [T] := \{1, 2, \dots, T\}$, the environment determines the loss vector $\ell_t = (\ell_{t1}, \ell_{t2}, \dots, \ell_{tk})^\top$ in $[0,1]^k$ or $[-1,1]^k$, and the learner simultaneously chooses an arm $A_t \in [k]$ without knowing ℓ_t. After that, the learner observes only the loss ℓ_{tA_t} for the chosen arm. The performance of the learner is evaluated by the regret Reg_T, which is the difference between the cumulative loss of the learner and of the single optimal arm, that is, $a^* = \arg\min_{a\in[k]} \mathbb{E}\big[\sum_{t=1}^{T} \ell_{ta}\big]$ and $\mathsf{Reg}_T = \mathbb{E}\big[\sum_{t=1}^{T}(\ell_{tA_t} - \ell_{ta^*})\big]$, where the expectation is taken with respect to the internal randomness of the algorithm and the randomness of the loss vectors $(\ell_t)_{t=1}^{T}$.

5.2.2 Partial Monitoring

Formulation A PM game $\mathcal{G} = (\mathcal{L}, \Phi)$ with k-actions and d-outcomes is defined by a pair of a loss matrix $\mathcal{L} \in [0,1]^{k\times d}$ and feedback matrix $\Phi \in \Sigma^{k\times d}$, where Σ is a set of feedback symbols. The game is played in a sequential manner by a learner and an opponent across T rounds. The learner begins the game with knowledge of $\mathcal{L}$ and Φ. For every round $t \in [T]$, the opponent selects an outcome $x_t \in [d]$, and the learner simultaneously chooses an action $A_t \in [k]$. Then the learner suffers an unobserved loss $\mathcal{L}_{A_t x_t}$ and receives only a feedback symbol $\sigma_t = \Phi_{A_t x_t}$, where $\mathcal{L}_{ax}$ is the (a,x)-th element of $\mathcal{L}$. The learner's performance in the game is evaluated by the regret Reg_T as in the MAB case: $a^* = \arg\min_{a\in[k]} \mathbb{E}\big[\sum_{t=1}^{T} \mathcal{L}_{ax_t}\big]$ and $\mathsf{Reg}_T = \mathbb{E}\big[\sum_{t=1}^{T}(\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\big] = \mathbb{E}\big[\sum_{t=1}^{T} \langle \ell_{A_t} - \ell_{a^*}, e_{x_t}\rangle\big]$, where $\ell_a \in \mathbb{R}^d$ is the a-th row of $\mathcal{L}$.

Several concepts in PM Let $m \leq |\Sigma|$ be the maximum number of distinct symbols in a single row of $\Phi \in \Sigma^{k\times d}$. Different actions a and b are duplicate if $\ell_a = \ell_b$. We can decompose possible distributions of d outcomes in $\mathcal{P}_d$ based on the loss matrix. For every action $a \in [k]$, cell $\mathcal{C}_a = \{u \in \mathcal{P}_d : \max_{b \in [k]} (\ell_a - \ell_b)^\top u \leq 0\}$ is the set of probability vectors in $\mathcal{P}_d$ for which action a is optimal. Each cell is a closed convex polytope.

Define $\dim(\mathcal{C}_a)$ as the dimension of the affine hull of $\mathcal{C}_a$. Action a is said to be dominated if $\mathcal{C}_a = \emptyset$. For non-dominated actions, action a is said to be Pareto optimal if $\dim(\mathcal{C}_a) = d-1$, and degenerate if $\dim(\mathcal{C}_a) < d-1$. Let Π be the set of Pareto optimal actions. Two Pareto optimal actions $a, b \in \Pi$ are called neighbors if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d-2$, which is used to define the difficulty of PM games. A PM game is said to be non-degenerate if it has no degenerate actions. We assume that PM game $\mathcal{G}$ is non-degenerate and contains no duplicate actions.

The difficulty of PM games are characterized by the following observability conditions. Neighbouring actions a and b are locally observable if there exists $w_{ab} : [k] \times \Sigma \to \mathbb{R}$ such that $w_{ab}(c, \sigma) = 0$ for $c \notin \{a, b\}$ and $\sum_{c=1}^{k} w_{ab}(c, \Phi_{cx}) = \mathcal{L}_{ax} - \mathcal{L}_{bx}$ for all $x \in [d]$. A PM game is locally observable if all neighboring actions are locally observable, and this study considers locally observable games.

Loss difference estimation Let $\mathcal{H}$ be the set of all functions from $[k] \times \Sigma$ to $\mathbb{R}^d$. For any locally observable games, there exists $G \in \mathcal{H}$ such that for any $b, c \in \Pi$, $\sum_{a=1}^{k} (G(a, \Phi_{ax})_b - G(a, \Phi_{ax})_c) = \mathcal{L}_{bx} - \mathcal{L}_{cx}$ for all $x \in [d]$ Lattimore and Szepesvári (2020b). For example, we can take $G = G_0$ defined by $G_0(a, \sigma)_b = \sum_{e \in \mathrm{path}_{\mathscr{T}}(b)} w_e(a, \sigma)$ for $a \in \Pi$, where $\mathscr{T}$ is a tree over Π induced by neighborhood relations and $\mathrm{path}_{\mathscr{T}}(b)$ is the set of edges from $b \in \Pi$ to an arbitrarily chosen root $c \in \Pi$ on $\mathscr{T}$ (Lattimore and Szepesvári, 2020b). See Section 5.7.2 and (Lattimore and Szepesvári, 2020a, Chapter 37) for a more detailed explanation and background of PM.

5.2.3 Considered Regimes

We consider three regimes on the assumptions for losses in MABs and outcomes in PM. In the *stochastic regime*, a sequence of loss vector (ℓ_t) in MAB and that of outcome vector (x_t) in PM follow an unknown distribution ν^* in an i.i.d. manner. Define the minimum suboptimality gap in $\Delta_{\min} = \min_{a \neq a^*} \Delta_a$ for $\Delta_a = \mathbb{E}_{\ell_t \sim \nu^*}\left[(\ell_{ta} - \ell_{ta^*})\right]$ in MAB and $\Delta_a = \mathbb{E}_{x_t \sim \nu^*}\left[(\ell_a - \ell_{a^*})^\top e_{x_t}\right]$ in PM. Note that the definitions of ℓ in MAB and PM are different.

In contrast, the *adversarial regime* does not assume any stochastic structure for the losses or outcomes, and they can be chosen in an arbitrarily manner. In this regime, the environment can choose ℓ_t for MAB and x_t for PM depending on the past history until the $(t-1)$-th round, $(A_s)_{s=1}^{t-1}$.

We also consider a general regime, *adversarial regime with a self-bounding constraint* (Zimmert and Seldin, 2021).

Definition 5.1. Let $\Delta \in [0, 2]^k$ and $C \geq 0$. The environment is in an *adversarial regime with a* (Δ, C, T) self-bounding constraint if it holds for any algorithm that $\mathsf{Reg}_T \geq \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t} - C\right]$.

One can see that the stochastic and adversarial regimes are indeed instances of this regime, and that well-known *stochastic regimes with adversarial corruptions* (Lykouris et al., 2018) are also in this regime (see Zimmert and Seldin (2021) and Tsuchiya et al. (2023a) for definitions in MAB and PM, respectively).

We assume that there exists a unique optimal arm (or action) a^*, which was employed by many studies aiming at developing BOBW algorithms (Gaillard et al., 2014; Luo and Schapire, 2015; Wei and Luo, 2018; Zimmert and Seldin, 2021).

5.3 Preliminaries

This section provides preliminaries for developing and analyzing algorithms. We first introduce FTRL, upon which we develop our algorithms, and then describe the self-bounding technique, which is a common technique for proving a BOBW guarantee.

Follow-the-regularized-leader In the FTRL framework, an arm selection probability $p_t \in \mathcal{P}_k$ at round t is given by

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{P}_k} \left\langle \sum_{s=1}^{t-1} \widehat{y}_s,\, q \right\rangle + \psi_t(q) \quad \text{and} \quad p_t = \mathcal{T}_t(q_t)\,, \tag{5.1}$$

where $\widehat{y}_s \in \mathbb{R}^k$ is an estimator of loss ℓ_t at round t, $\psi_t : \mathcal{P}_k \to \mathbb{R}$ is a convex regularizer, and $\mathcal{T}_t : \mathcal{P}_k \to \mathcal{P}_k$ is a map from the output of FTRL q_t to an arm selection probability vector p_t.

In the analysis of FTRL, it is common to evaluate $\sum_{t=1}^T \langle \widehat{y}_t, p_t - p \rangle = \sum_{t=1}^T \langle \widehat{y}_t, q_t - p \rangle + \sum_{t=1}^T \langle \widehat{y}_t, p_t - q_t \rangle$ for some $p \in \mathcal{P}_k$. As introduce in Lemma 2.1 in Chapter 2, it is known that quantity $\sum_{t=1}^T \langle \widehat{y}_t, q_t - p \rangle$ is bounded from above by

$$\sum_{t=1}^T \underbrace{(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1}))}_{\text{penalty term}} + \psi_{T+1}(p) - \psi_1(q_1) + \sum_{t=1}^T \underbrace{(\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t))}_{\text{stability term}}\,. \tag{5.2}$$

We refer to the terms in (5.2) as a *penalty* and *stability* terms, and to the quantity $\langle \widehat{y}_t, p_t - q_t \rangle$ as a *transformation* term. Note that, though this study focuses on example in which $\Phi_{T+1}(p)$ is not dominant, this term may be dominant dependent on the choice of regularizers.

Self-bounding technique A self-bounding technique is a common method for proving a BOBW guarantee (Gaillard et al., 2014; Wei and Luo, 2018; Zimmert and Seldin, 2021). In the self-bounding technique, we first derive regret upper and lower bounds in terms of a variable dependent on the arm selection probability, and then derive a regret bound by combining the upper and lower bounds. We use a version proposed in Ito et al. (2022a). We consider $Q(i)$, $\bar{Q}(i)$, $P(i)$, and $\bar{P}(i)$ for $i \in [k]$ defined by $Q(i) = \sum_{t=1}^T (1 - q_{ti})$, $\bar{Q}(i) = \mathbb{E}\left[Q(i)\right]$, $P(i) = \sum_{t=1}^T (1 - p_{ti})$, and $\bar{P}(i) = \mathbb{E}\left[P(i)\right]$. Note that $\bar{Q}(i), \bar{P}(i) \in [0, T]$ for any $i \in [k]$. In terms of $\bar{Q}(i)$ or $\bar{P}(i)$, we can obtain the lower bound of the regret for the adversarial regime with a self-bounding constraint as follows:

Lemma 5.1 (Lemma 4.2 in Chapter 4). *In the adversarial regime with a self-bounding constraint (Definition 5.1), if there exists $c' \in (0, 1]$ such that $p_{ti} \geq c'\, q_{ti}$ for all $t \in [T]$ and $i \in [k]$, then $\mathsf{Reg}_T \geq \Delta_{\min} \bar{P}(a^*) - C \geq c'\, \Delta_{\min} \bar{Q}(a^*) - C$.*

It is known that the sums of the entropy $H(\cdot)$ of (p_t) is bounded by $P(i)$ as follows:

Lemma 5.2 (Ito et al. 2022a, Lemma 4). *Let $(q_t)_{t=1}^T$ be any sequence of probability vectors and define $Q(i) = \sum_{t=1}^T (1 - q_{ti})$. Then for any $i \in [k]$, $\sum_{t=1}^T H(q_t) \leq Q(i) \log(\mathrm{e}kT/Q(i))$.*

Based on Lemmas 5.1 and 5.2, it suffices to show $\mathsf{Reg}_T \lesssim \mathbb{E}\left[\sqrt{\sum_{t=1}^T H(q_t) \operatorname{polylog}(T)}\right]$ to prove a BOBW gurantee in MAB. This is because, for the adversarial regime, using $H(q_t) \leq \log k$ implies a $\tilde{O}(\sqrt{T})$ bound, and for the stochastic regime, using Lemmas 5.1 and 5.2 roughly bounds the regret as $\mathsf{Reg}_T = 2\mathsf{Reg}_T - \mathsf{Reg}_T \lesssim \sqrt{\bar{Q}(a^*) \operatorname{polylog}(T)} - \Delta_{\min} \bar{Q}(a^*) \lesssim {}^{\operatorname{polylog}(T)}/_{\Delta_{\min}}$.

5.4 Stability-Penalty-Adaptive Learning Rate and Regret Bound

This section proposes a new adaptive learning rate, which yields a regret upper bound dependent on both the stability component z_t and penalty component h_t for various choices of z_t and h_t. When we use a learning rate η_t, the analysis of FTRL boils down to the evaluation of

$$\widehat{\mathsf{Reg}}_T^{\mathsf{SP}} = \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) h_{t+1} + \lambda \sum_{t=1}^{T} \eta_t z_t \quad \text{for some} \quad \lambda > 0\,. \tag{5.3}$$

In particular, when we use the Exp3 algorithm, h_t is the Shannon entropy of the FTRL output at round t. This can be confirmed by checking the existing studies (*e.g.,* Ito et al. 2022a; Tsuchiya et al. 2023a) or the proofs in Sections 5.7.5, 5.7.6, 5.7.7.2, and 5.7.8. To favorably bound $\widehat{\mathsf{Reg}}_T^{\mathsf{SP}}$, we develop a new learning rate framework, which we call the jointly stability- and penalty-adaptive learning rate, or the *Stability-Penalty-Adaptive (SPA) learning rate* for short:

Definition 5.2 (Stability-penalty-adaptive learning rate)**.** Let $((h_t, z_t, \bar{z}_t))_{t=1}^T$ be non-negative reals such that $h_1 \geq h_t$ for all $t \in [T]$, $(\bar{z}_t h_1 + \sum_{s=1}^t z_s h_{s+1})_{t=1}^T$ is non-decreasing, and $\bar{z}_t h_1 \geq z_t h_{t+1}$ for all $t \in [T]$. Let $c_1, c_2 > 0$. Then, a sequence of $(\eta_t)_{t=1}^T$ is a *SPA learning rate* if it has a form of

$$\beta_1 > 0\,, \quad \beta_{t+1} = \beta_t + \frac{c_1 z_t}{\sqrt{c_2 + \bar{z}_t h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}}\,, \qquad \text{and} \qquad \eta_t = \frac{1}{\beta_t}\,. \tag{5.4}$$

Remark. To the best of our knowledge, this is the first learning rate that depends on both the stability and penalty components. Note that when we set the penalties to their worst-case value, that is, $h_t = h_1$ for all $t \in [T]$ (recalling $h_t \leq h_1$), the SPA learning rate in (5.4) becomes equivalent to the standard type of the learning rate, which depends only on the stability and has the form of $\beta_t = 1/\eta_t \simeq \frac{c_1}{\sqrt{h_1}} \sqrt{\bar{z}_1 + \sum_{s=1}^{t-1} z_s}$. On the other hand, when we set the stabilities to be their worst-case value, that is, $z \geq \max_{t \in [T]} z_t$, the SPA learning rate in (5.4) corresponds to the learning rate dependent only on the penalty in Ito et al. (2022a); Tsuchiya et al. (2023a).

Using the learning rate (η_t) in (5.4), we can bound $\widehat{\mathsf{Reg}}_T^{\mathsf{SP}}$ as follows.

Theorem 5.1 (Stability-penalty-adaptive regret bound)**.** *Let $(\eta_t)_{t=1}^T$ be a SPA learning rate in Definition 5.2. Then $\widehat{\mathsf{Reg}}_T^{\mathsf{SP}}$ in* (5.3) *is bounded as follows:*
(I) *If $((h_t, z_t, \bar{z}_t))_{t=1}^T$ in (η_t) satisfies*

$$\frac{\sqrt{c_2 + \bar{z}_t h_1}}{c_1} (\beta_1 + \beta_t) \geq \epsilon + z_t \quad \textit{for all } t \in [T]$$

for some $\epsilon > 0$ (stability condition (S1)*), then*

$$\widehat{\mathsf{Reg}}_T^{\mathsf{SP}} \leq 2 \left(c_1 + \frac{\lambda}{c_1} \log \left(1 + \sum_{u=1}^{T} \frac{z_u}{\epsilon} \right) \right) \sqrt{c_2 + \bar{z}_t h_1 + \sum_{t=1}^{T} z_t h_{t+1}}\,. \tag{5.5}$$

(II) *If $h_t = h_1$ for all $t \in [T]$, $c_2 = 0$, and $((h_t, z_t, \bar{z}_t))_{t=1}^T$ in (η_t) satisfies $\beta_t \geq \frac{a c_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^t z_s}$ for some $a > 0$ (stability condition* (S2)*), then*

$$\widehat{\mathsf{Reg}}_T^{\mathsf{SP}} \leq 2 \left(c_1 + \frac{\lambda}{a c_1} \right) \sqrt{h_1 \sum_{t=1}^{T} z_t}\,.$$

The proof is given in Section 5.7.3. Note that the component $\widehat{\mathsf{Reg}}_T^{\mathsf{SP}}$ of the regret often becomes dominant when we use the Shanon entropy regularizer, and bounding it immediately leads to the regret bound in our framework. In Part (I) of Theorem 5.1, we can see that $\widehat{\mathsf{Reg}}_T^{\mathsf{SP}}$ is bounded by $\sqrt{\sum_{t=1}^T z_t h_{t+1}}$, which will enable us to obtain BOBW and data-dependent bounds simultaneously.

5.5 Sparsity-dependent Bound

This section establishes several sparsity-dependent bounds. We use the FTRL framework in (5.1) with the inverse weighted estimator $\widehat{y}_t \in \mathbb{R}^k$ given by $\widehat{y}_{ti} = \ell_{ti} \mathbb{1}[A_t = i]/p_{ti}$. This estimator is common in the literature and is useful for its unbiasedness, *i.e.*, $\mathbb{E}_{A_t \sim p_t}[\widehat{y}_t \mid p_t] = \ell_t$. We first propose algorithms that achieve sparsity-dependent bounds using stability-dependent learning rates in Section 5.5.1 as preliminaries for the subsequent section. Following that, in Section 5.5.2, we establish a BOBW algorithm with a sparsity-dependent bound based on the SPA learning rate. More specific steps are summarized as follows.

- Section 5.5.1.1 discusses the case of $\ell_t \in [0,1]^k$ and shows that appropriately choosing z_t in the SPA learning rate (5.4) with the Shannon entropy regularizer and $\tilde{\Theta}((kT)^{-2/3})$ uniform exploration achieves a $O(\sqrt{L_2 \log k})$ regret for $\ell_t \in [0,1]^k$ without knowing L_2.
- Section 5.5.1.2 considers the case of $\ell_t \in [-1,1]^k$, which is known to be more challenging than $\ell_t \in [0,1]^k$. We show that the time-invariant log-barrier enables us to choose a "tighter" z_t in (5.4), which removes the uniform exploration used in Section 5.5.1.1. This not only results in the bound of $O(\sqrt{L_2 \log k})$ for $\ell_t \in [-1,1]^k$ but also becomes one of the key properties to achieve BOBW.
- Section 5.5.2 presents a BOBW algorithm with a sparsity-dependent bound using the technique developed in Section 5.5.1 and Theorem 5.1. While Theorem 5.1 itself is a strong tool leading directly to the result for PM (Section 5.6), its application does not lead to the desired bounds. In particular, in this setting the $\tilde{O}\big(\sqrt{\sum_{t=1}^T z_t h_{t+1}}\big)$ term derived through Theorem 5.1 does not immediately imply a BOBW guarantee with a sparsity-dependent bound. To solve this problem, we develop a novel technique to analyze *the variation in FTRL outputs* q_t *in response to the change in a regularizer (Lemma 5.8)*, and prove a BOBW bound with a sparsity-dependent bound of $O(\sqrt{L_2 \log k \log T})$.

5.5.1 Parameter-agnostic Sparsity-dependent Bounds

This section establishes s-agnostic algorithms to achieve sparsity-dependent bounds for the adversarial regime, which are preliminaries for Section 5.5.2.

5.5.1.1 L_2-agnostic algorithm with $O(\sqrt{L_2 \log k})$ bound for $\ell_t \in [0,1]^k$

Here, we use $p_t = \mathcal{T}_t(q_t)$ for $\mathcal{T}_t(q) = (1-\gamma)q + \frac{\gamma}{k}\mathbf{1}$ and $\gamma = \frac{k^{1/3}(\log k)^{1/3}}{T^{2/3}}$ and assume $\gamma \in [0,1/2]$ (this holds when $T \geq \sqrt{8k \log k}$), which implies $2p_t \geq q_t$. We use the Shannon entropy regularizer $\psi_t(p) = -\frac{1}{\eta_t}H(p) = \frac{1}{\eta_t}\psi^{\mathsf{nS}}(p) = \frac{1}{\eta_t}\sum_{i=1}^k p_i \log p_i$ with the following learning rate:

$$\beta_1 = \frac{2c_1}{\sqrt{h_1}}\sqrt{\frac{k}{\gamma}} \text{ and } \beta_{t+1} = \beta_t + \frac{c_1 \omega_t}{\sqrt{\log k}\sqrt{\frac{k}{\gamma} + \sum_{s=1}^{t-1}\omega_s}} \text{ for } \omega_t := \frac{\ell_{tA_t}^2}{p_{tA_t}}, \quad (5.6)$$

which corresponds to the learning rate in Definition 5.2 with $h_t \leftarrow H(q_1) = \log k$, $z_t \leftarrow \omega_t$, $\bar{z}_t \leftarrow k/\gamma$, and $c_2 \leftarrow 0$. The uniform exploration is used to satisfy stability condition (S2) in Theorem 5.1, the amount of which is determined by balancing the regret coming from the uniform exploration and stability condition (S2). Theorem 5.1 immediately gives the following bound.

Corollary 5.1. *When $T \geq \sqrt{8k \log k}$, the above algorithm with $c_1 = 1/\sqrt{2}$ achieves $\mathsf{Reg}_T \leq 2\sqrt{2}\sqrt{L_2 \log k} + (2\sqrt{2}+1)(kT \log k)^{1/3}$ without knowing L_2. In particular, when $T \geq 7k^2/s^3$,*

$$\mathsf{Reg}_T \leq (4\sqrt{2}+1)\sqrt{sT \log k}\,.$$

The proof is given in Section 5.7.5. The most striking feature of the algorithm is its L_2 (or s)-agnostic property. This is essentially made possible by the learning rate using the data-dependent quantity ω_t in (5.6), which satisfies

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \omega_t}\right] \leq \sqrt{\sum_{t=1}^{T} \mathbb{E}[\omega_t]} = \sqrt{\sum_{t=1}^{T}\sum_{i=1}^{k} \ell_{ti}^2} = \sqrt{L_2}\,.$$

The leading constant of the bound is better than the existing bounds, as shown in Table 5.1, despite its agnostic property.

Remark. If $\ell_t \in [0,1]^k$, the first-order bound by Wei and Luo (2018) implies sparsity bounds. This, however, does not hold when $\ell_t \in [-1,1]^k$. In fact, let us consider the case where ℓ_t is a zero vector except that only one arm's loss is -1 for some $t \in [T]$. Then the sparsity-dependent bound becomes $O(\sqrt{T})$. On the other hand, the first-order bound in Wei and Luo (2018) is not directly applicable, and we need to transform losses to range $[0,1]$. This implies that the first-order bound becomes $O(\sqrt{kT})$, which is worse than the sparsity-dependent bound.

We will see in Section 5.5.1.2 that this assumption can be totally removed by adding a time-invariant log-barrier regularization.

5.5.1.2 L_2-agnostic algorithm with $O(\sqrt{L_2 \log k} + k \log T)$ bound for $\ell_t \in [-1,1]^k$

Here, we consider the case of $\ell_t \in [-1,1]^k$. It is worth noting that the negative loss cannot be handled by simply shifting the loss since it removes the sparsity from the losses (ℓ_t); see Kwon and Perchet (2016); Bubeck et al. (2018) for further details. We directly use the output q_t as p_t, that is, $p_t = q_t$. We use the hybrid regularizer consisting of the negative Shannon entropy and the log-barrier function, $\psi_t(p) = \frac{1}{\eta_t}\psi^{\mathsf{nS}}(p) + 2\psi^{\mathsf{LB}}(p)$, where $\psi^{\mathsf{LB}}(p) = \sum_{i=1}^{k} \log(1/x_i)$. We use the learning rate given by

$$\beta_1 = \frac{c_1^2}{8h_1} \text{ and } \beta_{t+1} = \beta_t + \frac{c_1 \nu_t}{\sqrt{\log k}\sqrt{\nu_t + \sum_{s=1}^{t-1} \nu_s}} \text{ for } \nu_t := \omega_t \min\left\{1, \frac{p_{tA_t}}{2\eta_t}\right\} \quad (5.7)$$

where ω_t is defined in (5.6). Learning rate (5.7) corresponds to that in Definition 5.2 with $h_t \leftarrow H(q_1) = \log k$, $z_t \leftarrow \nu_t$, $\bar{z}_t \leftarrow \nu_t$, and $c_2 \leftarrow 0$. We then have the following bound:

Corollary 5.2. *If we run the above algorithm with $c_1 = \sqrt{2}$,*

$$\mathsf{Reg}_T \leq 4\sqrt{2}\sqrt{L_2 \log k} + 2k \log T + k + 1/4\,,$$

which implies that $\mathsf{Reg}_T \leq 4\sqrt{2}\sqrt{sT \log k} + 2k \log T + k + 1/4$.

The proof is given in Section 5.7.6. Corollary 5.2 removes the assumption of $T \geq 7k^2/s^3$ in Corollary 5.1, and it also improves the leading constant of the regret in Bubeck et al. (2018). Note that one can prove a bound of the same order, but with a worse leading constant, by setting $\beta_1 \geq 15k$ and combining the analysis similar to that in Section 5.5.1.1 and the stability bound in Bubeck et al. (2018). We successfully remove the assumption of $T \geq 7k^2/s^3$ thanks to the following lemma, which serves as one of the key elements in achieving a BOBW guarantee with a sparsity-dependent bound (The proof is given in Section 5.7.6.):

Lemma 5.3 (Stability bound for negative losses)**.** *Let $\ell_t \in [-1,1]^k$ and $\widehat{y}_t = \ell_{ti}\mathbb{1}[A_t = i]/p_{ti}$ be the inverse weighted estimator. Assume that $q_t \leq \delta p_t$ for some $\delta \geq 1$. Then the stability term of FTRL with the hybrid regularizer $\psi_t = \frac{1}{\eta_t}\psi^{\mathsf{nS}} + 2\delta\,\psi^{\mathsf{LB}}$ is bounded as*

$$\left\langle q_t - q_{t+1}, \widehat{\ell}_t \right\rangle - D_{\psi_t}(q_{t+1}, q_t) \leq \delta\eta_t \frac{\ell_{tA_t}^2}{p_{tA_t}} \min\left\{1, \frac{p_{tA_t}}{2\eta_t}\right\} = \delta\eta_t\nu_t \,.$$

Remark. We can observe from Lemma 5.3 that the stability term is bounded in terms of ν_t, and the most important observation is that this ν_t is bounded by the inverse of the learning rate $1/(2\eta_t) = \beta_t/2$, *i.e.*, $\nu_t \leq \beta_t/2$. This enables us to guarantee the stability condition (S2) in Theorem 5.1 without needing to mix the $\tilde{\Theta}((kT)^{-2/3})$ uniform exploration used in Section 5.5.1.1. Moreover, this will be a key property to prove a BOBW with a sparsity-dependent bound in the next section.

As a minor contribution, by directly bounding the stability component, the RHS of Lemma 5.3 has a smaller leading constant than the bound obtained by using the bound in Bubeck et al. (2018).

5.5.2 Best-of-both-worlds Guarantee with Sparsity-dependent Bound

Finally, we are ready to establish a BOBW algorithm with a sparsity-dependent bound and derive its regret bound. We use $p_t = \mathcal{T}_t(q_t) = (1-\gamma)q_t + \frac{\gamma}{k}\mathbf{1}$ with $\gamma = \frac{k}{T}$ (*i.e.*, $\Theta(1/T)$ uniform exploration) and assume $\gamma \in [0, 1/2]$, which implies $2p_t \geq q_t$ and $\nu_t \leq T$. We use the hybrid regularizer $\psi_t = \frac{1}{\eta_t}\psi^{\mathsf{nS}} + 4\psi^{\mathsf{LB}}$ with the following learning rate depending on both the stability and penalty components:

$$\beta_1 = 15k \;\text{ and }\; \beta_{t+1} = \beta_t + \frac{c_1\nu_t}{\sqrt{81c_1^2 + \nu_t h_{t+1} + \sum_{s=1}^{t-1}\nu_s h_{s+1}}} \;\text{ with }\; h_t = \frac{1}{1-\frac{k}{T}}H(p_t)\,, \tag{5.8}$$

where ν_t is defined in (5.7). This corresponds to the SPA learning rate in Definition 5.2 with $z_t \leftarrow \nu_t$, $\bar{z}_t \leftarrow \nu_t h_{t+1}/h_1$, and $c_2 \leftarrow 81c_1^2$. One can see that stability assumption (S1) in Part (I) of Theorem 5.1 are satisfied thanks to $\nu_t \leq \beta_t$ (see Section 5.7.7 for the proof). We then have the following bound.

Theorem 5.2 (BOBW with sparsity-dependent bound)**.** *Suppose that $T \geq 2k$. Then the above algorithm with $c_1 = \sqrt{2\log(1 + T/\beta_1)}$ (Algorithm 5.2 in Section 5.7.7) achieves*

$$\mathsf{Reg}_T = O\left(\frac{k\log(T)\log(kT)}{\Delta_{\min}} + \sqrt{\frac{Ck\log(T)\log(kT)}{\Delta_{\min}}}\right)$$

for the adversarial regime with a (Δ, C, T) self-bounding constraint, and

$$\mathsf{Reg}_T \leq 4\sqrt{L_2\log(k)\log(1+T)} + O(k\log T)$$

for the adversarial regime.

The proof is given in Section 5.7.7. This is the first BOBW bound with the sparsity-dependent bound. The bound in the stochastic regime is suboptimal in two respects: its dependence on $\Delta_{\min}$ and $(\log T)^2$. Concurrently, two separate studies improve each suboptimality (Jin et al. 2023 for $\Delta_{\min}$ and Dann et al. 2023 for $(\log T)^2$), but it is highly uncertain if we can prove a BOBW with *a sparsity-dependent bound* based on their approach, and it is an important future work to investigate this problem.

Key elements of the proof In the following, we describe some key elements of the proof of Theorem 5.2. We need to solve one remaining technical issue. Using Part (I) of Theorem 5.1, we can show that the regret is roughly bounded by $\mathbb{E}\left[\sqrt{\sum_{t=1}^T \nu_t h_{t+1}}\right] \leq \sqrt{\sum_{t=1}^T \mathbb{E}[\nu_t h_{t+1}]}$. However, this quantity cannot be straightforwardly bounded since h_{t+1} depends on ν_t.

To address this issue, we analyze the behavior of arm selection probabilities when the regularizer changes. In particular, we first prove in Lemma 5.8 that $h_{t+1} \leq h_t + k\left({}^{\beta_{t+1}}/_{\beta_t} - 1\right) h_{t+1}$. This lemma can be proven by a novel analysis evaluating the changes of the FTRL outputs when the learning rate varies (given Sections 5.7.7.1 and 5.7.7.2), which is not considered and required when we use a time-invariant learning rate (*e.g.*, Bubeck et al. 2018). Using the last inequality, we have

$$\begin{aligned}
\sqrt{\sum_{t=1}^T \mathbb{E}[\nu_t h_{t+1}]} &\lesssim \sqrt{\sum_{t=1}^T \mathbb{E}[\nu_t h_t] + k \sum_{t=1}^T \mathbb{E}\left[\nu_t \left(\frac{\beta_{t+1}}{\beta_t} - 1\right) h_{t+1}\right]} \\
&\lesssim \sqrt{\sum_{t=1}^T \mathbb{E}[\nu_t h_t] + k \sum_{t=1}^T \mathbb{E}[(\beta_{t+1} - \beta_t)\, h_{t+1}]} \\
&\lesssim \sqrt{\sum_{t=1}^T \mathbb{E}[\nu_t h_t] + k \sum_{t=1}^T \mathbb{E}\left[\sqrt{\sum_{t=1}^T \nu_t h_{t+1}}\right]} \lesssim \sqrt{\sum_{t=1}^T \mathbb{E}[\nu_t h_t]} + k\,,
\end{aligned}$$

which holds thanks again to $\nu_t \leq \beta_t$ and based on the fact that $x \leq \sqrt{a + bx}$ for $a, b, x > 0$ implies $x \lesssim \sqrt{b} + a$ (here we ignore some logarithmic factors). This combined with the self-bounding technique leads to a BOBW guarantee in the stochastic regime.

Implementation One may wonder how to compute β_{t+1} satisfying (5.8) since $h_{t+1} = h_{t+1}(\beta_{t+1})$ depends on β_{t+1}. In fact, this can be computed by defining $F_t : [\beta_t, \beta_t{+}T] \to \mathbb{R}$ as $F_t(\alpha) = \alpha - \left(\beta_t + {}^{c_1 \nu_t}/\sqrt{81 c_1^2 + \nu_t h_{t+1}(\alpha) + \sum_{s=1}^{t-1} \nu_s h_{s+1}}\right)$ and setting β_{t+1} to be a root of $F_t(\alpha) = 0$. Such α can be computed using the bisection method because F_t is continuous (proved in Proposition 5.2 in Section 5.7.7.3). The detailed discussion can be found in Section 5.7.7.3.

5.6 Best-of-both-worlds with Game-dependent Bound for Partial Monitoring

This section discusses the result of a BOBW guarantee with a game-dependent bound for PM. We also consider full information (FI) and MAB as well as non-degenerate locally observable PM (PM-local), and let $\mathcal{M}$ be a such underlying model. The desired bound is obtained by direct application of the SPA learning rate and Theorem 5.1, which highlights the usefulness of the SPA learning rate.

5.6.1 Exploration by Optimization and its Extension

Exploration-by-optimization We rely on the Exploration by Optimization (EbO) by Lattimore and Szepesvári (2020b). which is a strong technique to bound the regret in PM with local observability. The key idea behind EbO is to minimize a part of a regret upper bound of the FTRL with the Shannon entropy. Recall that $\mathcal{H}$ is the set of all functions from $[k] \times \Sigma$ to $\mathbb{R}^d$. Then in EbO we consider the choice of $(p_t, G_t) \in \mathcal{P}_k \times \mathcal{H}$ to minimize the sum of the stability and transformation terms for the worst-case outcome given as follows:

$$\mathsf{ST}(p, G; q_t, \eta_t) = \max_{x \in [d]} \left[\frac{(p - q_t)^\top \mathcal{L} e_x}{\eta_t} + \frac{\mathrm{bias}_{q_t}(G; x)}{\eta_t} + \frac{1}{\eta_t^2} \sum_{a=1}^{k} p_a \Psi_{q_t} \left(\frac{\eta_t G(a, \Phi_{ax})}{p_a} \right) \right] . \tag{5.9}$$

Note that the first and third terms in (5.9) corresponds to the stability and transformation terms (divided by the learning rate η_t), respectively. We define the optimal value of the optimization problem by $\mathrm{opt}_q(\eta) = \min_{(p,G) \in \mathcal{P}_k \times \mathcal{H}} \mathsf{ST}(p, G; q_t, \eta)$ and its truncation at round t by $V_t = \max\{0, \mathrm{opt}_{q_t}(\eta_t)\}$ (appeared in Table 5.2). Note that this optimization problem is convex and can be solved numerically by using standard solvers (Lattimore and Szepesvári, 2020b).

Extending exploration-by-optimization While the vanilla EbO is a strong tool to derive a regret bound in PM-local, it only has a guarantee for the adversarial regime. Recall that in the self-bounding technique, we require a lower bound of the regret expressed in terms of q_t (see Lemma 5.1). However, when we use the vanilla EbO, it may make a certain action selection probability p_{ta} for some action a become extremely small even when the output of FTRL q_{ta} is far from zero (Lattimore and Szepesvári, 2020b), which makes it impossible for us to use the self-bounding technique.

To solve this problem, the vanilla EbO was recently extended so that it is applicable to the stochastic regime (and the adversarial regime with a self-bounding constraint) for PM-local in Chapter 4. We define $\mathcal{P}'_k(q, \mathcal{M})$ for a class of games $\mathcal{M}$ by

$$\mathcal{P}'_k(q, \mathcal{M}) = \{p \in \mathcal{P}_k : \mathsf{cond}(q, \mathcal{M})\} \quad \text{with} \quad \mathsf{cond}(q, \mathcal{M}) = \begin{cases} p = q & \text{if } \mathcal{M} \text{ is FI or MAB} , \\ p \geq q/(2k) & \text{if } \mathcal{M} \text{ is PM-local} . \end{cases}$$

We then consider the following optimization problem, which can be seen as a slight generalization of the approach developed in Chapter 4:

$$(p_t, G_t) = \mathop{\arg\min}_{p \in \mathcal{P}'_k(q_t, \mathcal{M}),\, G \in \mathcal{H}} \mathsf{ST}(p, G; q_t, \eta_t) , \tag{5.10}$$

where the feasible region $\mathcal{P}_k$ of p is replaced with $\mathcal{P}'_k(q, \mathcal{M})$. We define the optimal value of (5.10) by $\mathrm{opt}'_q(\eta, \mathcal{M})$ and its truncation at round t by $V'_t(\mathcal{M}) = \max\{0, \mathrm{opt}'_{q_t}(\eta_t, \mathcal{M})\}$. We will abbreviate $\mathcal{M}$ when it is clear from a context. The following proposition shows that the component of the regret in (5.9) can be made small enough even if the feasible region is restricted to $\mathcal{P}'_k(q, \mathcal{M}) \subset \mathcal{P}_k$.

Proposition 5.1. *Let $\mathcal{M}$ be an underlying model. If $\mathcal{M}$ is FI, MAB, or PM-local with $\eta \leq 1/(2mk^2)$,*

$$\mathrm{opt}'_*(\eta) := \sup_{q \in \mathcal{P}_k} \mathrm{opt}'_q(\eta) \leq \bar{V}(\mathcal{M}) := \begin{cases} 1/2 & \textit{if } \mathcal{M} \textit{ is FI} \\ k/2 & \textit{if } \mathcal{M} \textit{ is MAB} \\ 3m^2k^3 & \textit{if } \mathcal{M} \textit{ is PM-local} . \end{cases}$$

One can immediately obtain this result by following the same lines as (Lattimore and Szepesvári, 2020b, Propositions 11 and 12) and Lemma 5 in Chapter 4.

Algorithm 5.1: BOBW algorithm with a game-dependent bound for locally observable games

```
input: $B$
for $t = 1, 2, \ldots$ do
    Compute $q_t$ using (5.1).
    Solve (5.10) to determine $V'_t = \max\{0, \mathrm{opt}'_{q_t}(\eta_t)\}$ and the corresponding solution $p_t$ and $G_t$.
    Sample $A_t \sim p_t$ and observe feedback $\sigma_t \in \Sigma$.
    Compute $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{tA_t}$ and update $\beta_t$ using (5.11).
```

5.6.2 Algorithm

We use the negative Shanon entropy regularizer $\psi_t = \frac{1}{\eta_t}\psi^{\mathsf{nS}}$ for (5.1) with a learning rate given by

$$\beta_1 = B\sqrt{\frac{\log(1+T)}{\log k}} \quad \text{and} \quad \beta_{t+1} = \beta_t + \frac{c_1 V'_t}{\sqrt{\bar{V}h_1 + \sum_{s=1}^{t-1} V'_s h_{s+1}}}\,, \tag{5.11}$$

with $B = 1/2$ for FI, $B = k/2$ for MAB, and $B = 2mk^2$ for PM-local, which corresponds to the learning rate in Definition 5.2 with $h_t \leftarrow H(q_t)$, $z_t \leftarrow V'_t$, $\bar{z}_t \leftarrow 0$, and $c_2 \leftarrow 0$. Algorithm 5.1 summarizes the proposed algorithm.

5.6.3 Main Result

Let $r_{\mathcal{M}}$ be 1 if $\mathcal{M}$ is FI or MAB, and $2k$ if $\mathcal{M}$ is PM-local. Then we have the following bound.

Corollary 5.3. *Let $\mathcal{M}$ be FI, MAB, or PM-local. Then the above algorithm with $c_1 = \sqrt{\log(1+T)/2}$ (Algorithm 5.1) achieves*

$$\mathsf{Reg}_T \le \mathbb{E}\left[\sqrt{2\sum_{t=1}^{T} V'_t \log(k)\log(1+T)}\right] + O(B\sqrt{\log(k)\log(T)})$$

for the adversarial regime, and

$$\mathsf{Reg}_T = O\left(\frac{r_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}} + \sqrt{\frac{Cr_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}}} + B\sqrt{\log(k)\log(T)}\right)$$

for the adversarial regime with a (Δ, C, T) self-bounding constraint.

The bound for the adversarial regime with a self-bounding constraint with $C = 0$ yields the bound in the stochastic regime, which is optimal up to logarithmic factors in FI and MAB, and has the same order as the bounds Theorem 6 in Chapter 4.

The bound for the adversarial regime has a form similar to Lattimore and Szepesvári (2020b) and is game-dependent in the sense that it can be bounded by the empirical difficulty V'_t of the current game. In addition, we can also obtain the worst-case bound by replacing V'_t with its upper bound $\bar{V}$. This bound is optimal up to $\log(T)$ factor in FI and $\log(k)\log(T)$ factor in MAB, and is a factor of $\sqrt{\log T}$ worse than the best known bound in Lattimore and Szepesvári (2020b), which can be seen as the cost for the BOBW guarantee (see also Table 5.2).

5.7 Deferred Discussion and Proofs

5.7.1 Notation

Table 5.3 summarizes the symbols used in this chapter.

Table 5.3: Notation

Symbol	Meaning
$\mathcal{P}_k$	$(k-1)$-dimensional probability simplex
$T \in \mathbb{N}$	time horizon
$k \in \mathbb{N}$	number of arms (or actions)
$A_t \in [k]$	arm (or action) chosen by learner at round t
$s \leq k$	$\max_t \lVert \ell_t \rVert_0$, sparsity level of losses
L_2	$\sum_{t=1}^T \lVert \ell_t \rVert^2$
$\mathcal{L} \in [0,1]^{k \times d}$	loss matrix
Σ	set of feedback symbols
$\Phi \in \Sigma^{k \times d}$	feedback matrix
$d \in \mathbb{N}$	number of outcomes
$m \in \mathbb{N}$	maximum number of distinct symbols in a single row of Φ
$x_t \in [d]$	outcome chosen by opponent at round t
$q_t \in \mathcal{P}_k$	output of FTRL at round t
$p_t \in \mathcal{P}_k$	arm selection probability at round t
$\psi_t \colon \mathcal{P}_k \to \mathbb{R}$	regularizer of FTRL at round t
$\eta_t = 1/\beta_t > 0$	learning rate of FTRL at round t
$\psi^{\mathsf{nS}} \colon \mathbb{R}_+^k \to \mathbb{R}$	$\sum_{i=1}^k x_i \log x_i$, negative Shannon entropy
$\psi^{\mathsf{LB}} \colon \mathbb{R}_+^k \to \mathbb{R}$	$\sum_{i=1}^k \log(1/x_i)$, log-barrier
$\phi^{\mathsf{nS}} \colon \mathbb{R}_+ \to \mathbb{R}$	$x \log(1/x)$
$\phi^{\mathsf{LB}} \colon \mathbb{R}_+ \to \mathbb{R}$	$\log(1/x)$
h_t	penalty component at round t
z_t	stability component at round t
ω_t	stability component z_t introduced in (5.6) (Section 5.5.1.1)
ν_t	stability component z_t introduced in (5.7) (Section 5.5.1.2)
V_t'	stability component z_t introduced in (5.11) (Section 5.6)
$C \geq 0$	corruption level

5.7.2 Additional Related Work

This section provides additional related work, some of which has never mentioned in this chapter.

Multi-armed bandits In the stochastic regime, it is known that the optimal regret is approximately expressed as $\mathsf{Reg}_T = O(k \log T / \Delta_{\min})$ (Lai and Robbins, 1985). In the adversarial regime (a.k.a. non-stochastic regime), it is known that the Online Mirror Descent (OMD) framework with the (negative) Tsallis entropy regularizer achieves $O(\sqrt{kT})$ regret bounds (Audibert and Bubeck, 2009; Abernethy et al., 2015), which match the lower bound of $\Omega(\sqrt{kT})$ (Auer et al., 2002b).

In the adversarial MAB, algorithms with various data-dependent regret bounds have been developed. Typical examples of such bounds are first-order bounds dependent on the cumulative loss and second-order bounds depending on sample variances in losses. Allenberg et al. (2006) provided an algorithm with a first-order regret bound of $O(\sqrt{kL^* \log k})$ for $L^* = \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \langle \ell_t, a \rangle$. Second-order regret bounds have been shown in some studies, *e.g.,* by Hazan and Kale (2011); Wei and Luo (2018); Bubeck et al. (2018), In particular, Bubeck et al. (2018) provided the regret bound of $O(\sqrt{Q_2 \log k})$ for $Q_2 = \sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|^2$. Other examples of data-dependent bounds include path-length bounds in the form of $O(\sqrt{kV_1 \log T})$ for $V_1 = \sum_{t=1}^{T-1} \|\ell_t - \ell_{t+1}\|_1$ as well as a sparsity-dependent bound, which have been investigated by Kwon and Perchet (2016); Bubeck et al. (2019b, 2018); Wei and Luo (2018); Zheng et al. (2019).

The study on a sparsity-dependent bound was initiated by Kwon and Perchet (2016), who showed that when $\ell_t \in [0,1]^k$, the OMD with Tsallis entropy can achieve the bound of $\mathsf{Reg}_T \leq 2\sqrt{\mathrm{e}}\sqrt{sT \log(k/s)})$ and prove the matching (up to logarithmic factor) lower bound of $\mathsf{Reg}_T = \Omega(\sqrt{sT})$ when $T \geq k^3/(4s^2)$. Bubeck et al. (2018) also showed that OMD with a hybrid-regularizer consisting of the Shannon entropy and a log-barrier can achieve $\mathsf{Reg}_T \leq 10\sqrt{L_2 \log k} + 20k \log T$ when $\ell_t \in [-1,1]^k$. Zheng et al. (2019) investigated the sparse MAB problem in the context of the switching regret. Although their result is not directly related to our study, they show that sparsity is useful in some cases. Note that all of these algorithms assume the knowledge of the sparsity level and do not have a BOBW guarantee.

The stability-dependent learning rate is quite ubiquitous (see Orabona 2019 and the references therein). To our knowledge, the literature on the penalty-dependent bound is quite scarce in bandits and considered in the context of BOBW algorithms (Ito et al., 2022a; Tsuchiya et al., 2023a), both of which consider the Shannon entropy regularizer.

Best-of-both-worlds Since the seminal study by Bubeck and Slivkins (2012), BOBW algorithms have been developed for many online decision-making problems. Although they have been investigated mainly in the context of an MAB (Seldin and Lugosi, 2017; Zimmert and Seldin, 2021), other settings have also been investigated, Gaillard et al. (2014); Wei and Luo (2018); Jin et al. (2021), to name a few.

FTRL and OMD are now one of the most common approaches to achieving a BOBW guarantee owing to the usefulness of the self-bounding technique (Gaillard et al., 2014; Wei and Luo, 2018; Zimmert and Seldin, 2021), while the first (Bubeck and Slivkins, 2012) and earlier work (Seldin and Slivkins, 2014; Seldin and Lugosi, 2017) on BOBW do not rely on the technique. Most of the recent algorithms beyond the MAB are also based on FTRL (to name a few, Wei and Luo 2018; Jin et al. 2021; Saha and Gaillard 2022).

Our BOBW algorithm with the sparsity-dependent bound can be seen as one of the studies that aim to achieve BOBW and data-dependent bound simultaneously. There is not so much existing research, and we are only aware of Wei and Luo (2018); Ito (2021c); Ito et al. (2022b); Tsuchiya et al. (2023b). They consider first-, second-order, and path-length bound, and we are the first to investigate the sparsity-dependent bound in this line of work.

Log-barrier regularizer and hybrid regularizer The log-barrier regularizer has been used in various studies (to name a few, Foster et al. 2016; Wei and Luo 2018; Luo et al. 2018; Pogodin and Lattimore 2020; Erez and Koren 2021). The time-invariant log-barrier (a.k.a. constant amount of log-barrier Lee et al. 2020), whose properties are extensively exploited in this chapter, was invented by Bubeck et al. (2018) and has been used in several subsequent studies (Zheng et al., 2019; Lee et al., 2020).

Partial monitoring Starting from the work by Rustichini (1999), PM has been investigated in many works in the literature (Piccolboni and Schindelhauer, 2001; Cesa-Bianchi et al., 2006; Bartók et al., 2011). It is known that all PM games can be classified into four classes based on their minimax regrets (Bartók et al., 2011; Lattimore and Szepesvári, 2019b). In particular, all PM games fall into trivial, easy, hard, and hopeless games, for which its minimax regrets is 0, $\Theta(\sqrt{T})$, $\Theta(T^{2/3})$, and $\Theta(T)$, respectively. PM has also been investigated in both the adversarial and stochastic regimes as for MAB. In the stochastic regime, there are relatively small amount of works (Bartók et al., 2012; Vanchinathan et al., 2014; Komiyama et al., 2015a; Tsuchiya et al., 2020), some of which are proven to achieve an instance-dependent $O(\log T)$ regrets for locally or globally observable games. In the adversarial regime, since the development of the FeedExp3 algorithm (Piccolboni and Schindelhauer, 2001; Cesa-Bianchi et al., 2006), many algorithms achieving the minimax optimal regret have been developed (Bartók et al., 2011; Foster and Rakhlin, 2012; Bartók, 2013; Lattimore and Szepesvári, 2020a).

5.7.3 Proof of Theorem 5.1

Proof. We first prove (5.5) in Part (I).

(Penalty) First, we consider the penalty term $\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) h_{t+1}$. By the definition of β_t in (5.4),

$$\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) h_{t+1} = \sum_{t=1}^T \left(\beta_{t+1} - \beta_t\right) h_{t+1} = \sum_{t=1}^T \frac{c_1 z_t h_{t+1}}{\sqrt{c_2 + \bar{z}_t h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}}$$
$$\leq c_1 \sum_{t=1}^T \frac{z_t h_{t+1}}{\sqrt{\sum_{s=1}^t z_s h_{s+1}}} \leq c_1 \int_0^{\sum_{t=1}^T z_t h_{t+1}} \frac{1}{\sqrt{x}} \, \mathrm{d}x = 2c_1 \sqrt{\sum_{t=1}^T z_t h_{t+1}} \,, \tag{5.12}$$

where the first inequality follows from $\bar{z}_t h_1 \geq z_t h_{t+1}$ and the second inequality follows by Lemma 5.9 given in Section 5.7.9.

(Stability) Next, we consider the stability term $\sum_{t=1}^T \eta_t z_t$. Using the definition of β_t in (5.4) and defining $U_t = \sqrt{c_2 + \bar{z}_t h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}$ for $t \in \{0\} \cup [T]$, we bound β_t from below as

$$\beta_t = \beta_1 + \sum_{u=1}^{t-1} \frac{c_1 z_u}{\sqrt{c_2 + \bar{z}_u h_1 + \sum_{s=1}^{u-1} z_s h_{s+1}}} = \beta_1 + \sum_{u=1}^{t-1} \frac{c_1 z_u}{U_u} \geq \beta_1 + \frac{c_1}{U_T} \sum_{u=1}^{t-1} z_u \,,$$

where the inequality follows since (U_t) is non-decreasing. Using the last inequality, we can bound $\sum_{t=1}^T \eta_t z_t$ as

$$\sum_{t=1}^T \eta_t z_t = 2 \sum_{t=1}^T \frac{z_t}{\beta_t + \beta_t} \leq 2 \sum_{t=1}^T \frac{z_t}{\beta_t + \beta_1 + \frac{c_1}{U_T} \sum_{s=1}^{t-1} z_s} = \frac{2U_T}{c_1} \sum_{t=1}^T \frac{z_t}{\frac{U_T}{c_1}\left(\beta_t + \beta_1\right) + \sum_{s=1}^{t-1} z_s} \,. \tag{5.13}$$

Since we have $\frac{U_T}{c_1}(\beta_1 + \beta_t) \geq \frac{\sqrt{c_2 + \bar{z}_t h_1}}{c_1}(\beta_1 + \beta_t) \geq \epsilon + z_t$ by the assumption, a part of the last inequality is further bounded as

$$\sum_{t=1}^T \frac{z_t}{\frac{U_T}{c_1}\left(\beta_t + \beta_1\right) + \sum_{s=1}^{t-1} z_s} \leq \sum_{t=1}^T \frac{z_t}{\epsilon + \sum_{s=1}^t z_s} \leq \int_\epsilon^{\epsilon + \sum_{t=1}^T z_t} \frac{1}{x} \, \mathrm{d}x \leq \log\left(1 + \sum_{t=1}^T \frac{z_t}{\epsilon}\right) \,, \tag{5.14}$$

where the second inequality follows by Lemma 5.9. Combining (5.13) and (5.14) yields

$$\sum_{t=1}^{T} \eta_t z_t \le \frac{2U_T}{c_1} \log\left(1 + \sum_{t=1}^{T} \frac{z_t}{\epsilon}\right) = \frac{2}{c_1} \log\left(1 + \sum_{t=1}^{T} \frac{z_t}{\epsilon}\right) \sqrt{c_2 + \bar{z}_T h_1 + \sum_{t=1}^{T} z_t h_{t+1}} \, . \tag{5.15}$$

Combining (5.12) and (5.15) completes the proof of (5.5) in Part (I).

We next prove Part (II). For the penalty term, setting $h_t = h_1$ for all $t \in [T]$ in (5.12) gives

$$\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) h_{t+1} \le 2c_1 \sqrt{h_1 \sum_{t=1}^{T} z_t} \, .$$

For the stability term, since there exists $a > 0$ such that $\beta_t \ge \frac{ac_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^{t} z_s}$ for any $t \in [T]$ by the assumption,

$$\sum_{t=1}^{T} \eta_t z_t = \sum_{t=1}^{T} \frac{z_t}{\beta_t} \le \frac{\sqrt{h_1}}{ac_1} \sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} z_s}} \le \frac{2}{ac_1} \sqrt{h_1 \sum_{t=1}^{T} z_t} \, .$$

Summing up the above arguments completes the proof of Part (II). □

5.7.4 Basic Facts to Bound Stability Terms

Here, we introduce basic facts, which is useful to bound the stability term. We have

$$\xi(x) := \exp(-x) + x - 1 \le \begin{cases} \frac{1}{2}x^2 & \text{for } x \ge 0 \\ x^2 & \text{for } x \ge -1 \, , \end{cases} \tag{5.16}$$

$$\zeta(x) := x - \log(1 + x) \le x^2 \quad \text{for } x \in \left[-\frac{1}{2}, \frac{1}{2}\right] \, . \tag{5.17}$$

We also have the following inequalities for $\phi^{\mathsf{nS}}(x) = x \log x$ and $\phi^{\mathsf{LB}}(x) = \log(1/x)$, which are components of the negative Shannon entropy and log-barrier function:

$$\max_{y \in \mathbb{R}} \left\{ a(x - y) - D_{\phi^{\mathsf{nS}}}(y, x) \right\} = x\xi(a) \qquad \text{for } a \in \mathbb{R} \, , \tag{5.18}$$

$$\max_{y \in \mathbb{R}} \left\{ a(x - y) - D_{\phi^{\mathsf{LB}}}(y, x) \right\} = \zeta(ax) \qquad \text{for } a \ge -\frac{1}{x} \, . \tag{5.19}$$

It is easy to prove these facts by the standard calculus and you can find the proofs of (5.18) and (5.19) in Lemma 15 in Chapter 4 and Ito et al. (2022b, Lemma 5), respectively.

5.7.5 Proof of Corollary 5.1

Let $\mathsf{Reg}_T(a) = \mathbb{E}\big[\sum_{t=1}^{T} (\ell_{tA_t} - \ell_{ta})\big]$ for $a \in [k]$. Here we provide the complete proof of Corollary 5.1.

Proof of Corollary 5.1. Fix $i^* \in [k]$. Since $p_t = (1 - \gamma) q_t + \frac{\gamma}{k}\mathbf{1}$, it holds that

$$\begin{aligned} \mathsf{Reg}_T(i^*) &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{tA_t} - \sum_{t=1}^{T} \ell_{ti^*}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell_t, p_t - e_{i^*} \rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell_t, q_t - e_{i^*} \rangle\right] + \mathbb{E}\left[\gamma \sum_{t=1}^{T} \left\langle \ell_t, \frac{1}{k}\mathbf{1} - q_t \right\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \langle \widehat{y}_t, q_t - e_{i^*} \rangle\right] + \gamma T \, , \end{aligned}$$

where the last inequality follows by $\mathbb{E}[\widehat{y}_t \mid q_t] = \ell_t$ and the Cauchy-Schwarz inequality. Then, using the standard analysis of the FTRL described in Section 5.3, the first term in the last inequality is bounded as

$$\sum_{t=1}^{T} \langle \widehat{y}_t, q_t - e_{i^*} \rangle \le \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^{T} \left(\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t) \right) .$$

By (5.16) and (5.18) given in Section 5.7.4, the stability term $\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t)$ in the last inequality is bounded as

$$\begin{aligned}
&\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t) = \langle q_t - q_{t+1}, \widehat{y}_t \rangle - \frac{1}{\eta_t} D_{\psi^{\mathsf{nS}}}(q_{t+1}, q_t) \\
&= \sum_{i=1}^{k} \left(\widehat{y}_{ti}(q_{ti} - q_{t+1,i}) - \frac{1}{\eta_t} D_{\phi^{\mathsf{nS}}}(q_{t+1,i}, q_{ti}) \right) \\
&\le \sum_{i=1}^{k} \frac{1}{\eta_t} q_{ti}\, \xi\left(\eta_t \widehat{y}_{ti}\right) \le \frac{1}{2} \eta_t \sum_{i=1}^{k} q_{ti} \widehat{y}_{ti}^2 \le \eta_t \omega_t \,,
\end{aligned}$$

where the first inequality follows from (5.18), the second inequality follows by (5.16) with $\widehat{y}_t \ge 0$, and the last inequality holds since $\sum_{i=1}^{k} q_{ti} \widehat{y}_{ti}^2 \le \sum_{i=1}^{k} 2 p_{ti} \widehat{y}_{ti}^2 = 2\omega_t$.

We will confirm that the assumptions for Part (II) of Theorem 5.1 are indeed satisfied. Using the definition of β_t in (5.6), We have

$$\begin{aligned}
\beta_t &= \beta_1 + \frac{1}{\sqrt{h_1}} \sum_{u=1}^{t-1} \frac{c_1 \omega_u}{\sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u-1} \omega_s}} = \beta_1 + \frac{2c_1}{\sqrt{h_1}} \sum_{u=1}^{t-1} \frac{\omega_u}{\sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u-1} \omega_s} + \sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u-1} \omega_s}} \\
&\ge \beta_1 + \frac{2c_1}{\sqrt{h_1}} \sum_{u=1}^{t-1} \frac{\omega_u}{\sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u} \omega_s} + \sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u-1} \omega_s}} \\
&= \beta_1 + \frac{2c_1}{\sqrt{h_1}} \sum_{u=1}^{t-1} \left(\sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u} \omega_s} - \sqrt{\frac{k}{\gamma} + \sum_{s=1}^{u-1} \omega_s} \right) \\
&= \beta_1 + \frac{2c_1}{\sqrt{h_1}} \left(\sqrt{\frac{k}{\gamma} + \sum_{s=1}^{t-1} \omega_s} - \sqrt{\frac{k}{\gamma}} \right) \ge \frac{2c_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^{t} \omega_s} \,,
\end{aligned}$$

where the last inequality follows since $\beta_1 = \frac{2c_1}{\sqrt{h_1}} \sqrt{\frac{k}{\gamma}}$ and $\frac{k}{\gamma} \ge \omega_t$. Hence, stability condition (S2) in Theorem 5.1 is satisfied with $a = 2$, and one can see that the other assumptions are trivially satisfied. Hence, by Part (II) of Theorem 5.1,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \widehat{y}_t, q_t - e_{i^*} \rangle \le \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(q_{t+1}) + \sum_{t=1}^{T} \eta_t \omega_t + \frac{H(q_1)}{\eta_1} \\
&\le 2 \left(c_1 + \frac{1}{2c_1} \right) \sqrt{h_1 \sum_{t=1}^{T} \omega_t} + \frac{\log k}{\eta_1} \,,
\end{aligned}$$

where in the last inequality we used $h_1 \le \log k$. Now,

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \omega_t} \right] \le \sqrt{\sum_{t=1}^{T} \mathbb{E}[\omega_t]} = \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\frac{\ell_{tA_t}^2}{p_{tA_t}} \right]} = \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{k} \ell_{ti}^2} = \sqrt{\sum_{t=1}^{T} \|\ell_t\|_2^2} = \sqrt{L_2} \,.$$

Summing up the above arguments and setting $c_1 = 1/\sqrt{2}$, we have

$$\mathsf{Reg}_T \leq 2\sqrt{2}\sqrt{L_2 \log k} + \frac{\log k}{\eta_1} + \gamma T = 2\sqrt{2}\sqrt{L_2 \log k} + (\sqrt{2}+1)(kT\log k)^{1/3}\,,$$

which completes the proof of Corollary 5.1. □

5.7.6 Proof of Corollary 5.2

We first prove Lemma 5.3.

Proof of Lemma 5.3. Recall that $\psi_t(p) = \frac{1}{\eta_t}\psi^{\mathsf{nS}}(p) + 2\delta\psi^{\mathsf{LB}}(p)$. Since $D_{\psi_t} = \frac{1}{\eta_t}D_{\psi^{\mathsf{nS}}} + 2\delta D_{\psi^{\mathsf{LB}}}$ and $D_{\psi^{\mathsf{nS}}}(x,y) = \sum_{i=1}^k D_{\phi^{\mathsf{nS}}}(x_i, y_i)$ and $D_{\psi^{\mathsf{LB}}}(x,y) = \sum_{i=1}^k D_{\phi^{\mathsf{LB}}}(x_i, y_i)$, we can bound the stability term as

$$\begin{aligned}
&\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t) \\
&\leq \langle q_t - q_{t+1}, \widehat{y}_t \rangle - \max\left\{ \frac{1}{\eta_t} D_{\psi^{\mathsf{nS}}}(q_{t+1}, q_t),\, 2\delta D_{\psi^{\mathsf{LB}}}(q_{t+1}, q_t) \right\} \\
&\leq \sum_{i=1}^k \left(\widehat{y}_{ti}(q_{ti} - q_{t+1,i}) - \max\left\{ \frac{1}{\eta_t} D_{\phi^{\mathsf{nS}}}(q_{t+1,i}, q_{ti}),\, 2\delta D_{\phi^{\mathsf{LB}}}(q_{t+1,i}, q_{ti}) \right\} \right) \\
&\leq \sum_{i=1}^k \min\left\{ \frac{1}{\eta_t} q_{ti}\, \xi\left(\eta_t \widehat{y}_{ti}\right),\, 2\delta\, \zeta\left(\frac{1}{2\delta} q_{ti}\widehat{y}_{ti} \right) \right\}\,, \qquad (5.20)
\end{aligned}$$

where in the last inequality we used (5.18) and (5.19) with

$$\frac{\widehat{y}_{ti}}{2\delta} \geq -\frac{1}{2\delta p_{ti}} \geq -\frac{1}{2\delta(q_{ti}/\delta)} \geq -\frac{1}{q_{ti}}\,,$$

where the first inequality follows by the definition of $\widehat{y}_t$ and the second inequality follows by $p_{ti} \geq q_{ti}/\delta$.

Next, we will prove that for any $i \in [k]$,

$$\min\left\{ \frac{1}{\eta_t} q_{ti}\, \xi\left(\eta_t \widehat{y}_{ti}\right),\, 2\delta\, \zeta\left(\frac{1}{2\delta} q_{ti}\widehat{y}_{ti} \right) \right\} \leq \delta\eta_t \frac{\ell_{ti}^2}{p_{ti}} \min\left\{ 1, \frac{p_{ti}}{2\eta_t} \right\} \mathbb{1}[A_t = i] \quad (5.21)$$

Fix $i \in [k]$. By $q_{ti} \leq \delta p_{ti}$,

$$\frac{1}{2\delta} q_{ti}\widehat{y}_{ti} = \frac{1}{2} p_{ti}\widehat{y}_{ti} \leq \frac{1}{2}\,.$$

Using this and $\zeta(x) \leq x^2$ for $x \in [-\frac{1}{2}, \frac{1}{2}]$ in (5.17), it holds for any $p_{ti} \in [0,1]$ that

$$2\delta\, \zeta\left(\frac{1}{2\delta} q_{ti}\widehat{y}_{ti} \right) \leq 2\delta \left(\frac{1}{2\delta} q_{ti}\widehat{y}_{ti} \right)^2 \leq \frac{\delta}{2} \ell_{ti}^2 \mathbb{1}[A_t = i]\,, \quad (5.22)$$

where in the last inequality we used $q_{ti} \leq \delta p_{ti}$. In particular, when $p_{ti} \leq \eta_t$, *i.e.,* the probability of selecting arm i is small to some extent, the last inequality can be further bounded as

$$2\delta\, \zeta\left(\frac{1}{2\delta} q_{ti}\widehat{y}_{ti} \right) \leq \frac{\eta_t}{p_{ti}} \frac{\delta}{2} \ell_{ti}^2 \mathbb{1}[A_t = i] \leq \delta\eta_t \frac{\ell_{ti}^2}{p_{ti}} \mathbb{1}[A_t = i]\,. \quad (5.23)$$

On the other hand when $p_{ti} > \eta_t$, we have $\eta_t \widehat{y}_{ti} \geq -1$. Hence, using the inequality $\xi(x) \leq x^2$ for $x \geq -1$ in (5.16), we have

$$\frac{1}{\eta_t} q_{ti} \xi\left(\eta_t \widehat{y}_{ti}\right) \leq \frac{1}{\eta_t} \delta p_{ti} (\eta_t \widehat{y}_{ti})^2 = \delta\eta_t \frac{\ell_{ti}^2}{p_{ti}} \mathbb{1}[A_t = i]\,. \quad (5.24)$$

Hence, combining (5.22), (5.23), and (5.24) completes the proof of (5.21). Finally, by combining (5.20) and (5.21) we completes the proof of Lemma 5.3. □

Remark. When ℓ_t can be negative, the Shannon entropy regularizer alone cannot bound the stability term if the arm selection probability is small, *i.e.,* $p_{ti} \leq \eta_t$. Introducing a time-invariant log-barrier regularizer enables us to bound the stability term even when the arm selection probability is small. This idea was proposed by Bubeck et al. (2018), who analyzed the variation of arm selection probability for the change of cumulative losses. Unlike their analysis, our proof directly analyses the stability term, enabling us to obtain the tighter regret bound. More importantly, we will utilize the property $\nu_t \leq O(1/\eta_t)$ many times, which directly follows from Lemma 5.3 in the subsequent sections to prove the BOBW guarantee with the sparsity-dependent bound.

Now, we are ready to prove Corollary 5.2.

Proof of Corollary 5.2. Fix $i^* \in [k]$. Define $p^* \in \mathcal{P}_k$ by

$$p^* = \left(1 - \frac{k}{T}\right) e_{i^*} + \frac{1}{T}\mathbf{1} \,.$$

Then, using the definition of the algorithm,

$$\begin{aligned}
\mathsf{Reg}_T(i^*) &= \mathbb{E}\left[\sum_{t=1}^T \ell_{tA_t} - \sum_{t=1}^T \ell_{ti^*}\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - e_{i^*}\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - p^*\rangle\right] + \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p^* - e_{i^*}\rangle\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \langle \widehat{y}_t, p_t - p^*\rangle\right] + k \,,
\end{aligned}$$

where the inequality follows from the definition of p^* and the Cauchy-Schwarz inequality. By the standard analysis of the FTRL, described in Section 5.3,

$$\begin{aligned}
\sum_{t=1}^T \langle \widehat{y}_t, p_t - p^*\rangle \leq \sum_{t=1}^T \Big(\psi_t(p_{t+1}) - \Phi_{t+1}(p_{t+1})\Big) + \Phi_{t+1}(p^*) - \Phi_1(p_1) \\
+ \sum_{t=1}^T \Big(\langle p_t - p_{t+1}, \widehat{y}_t\rangle - D_{\psi_t}(p_{t+1}, p_t)\Big) \,.
\end{aligned}$$

For the penalty term, since $\psi_t(p) = \frac{1}{\eta_t}\psi^{\mathsf{nS}}(p) + 2\psi^{\mathsf{LB}}(p)$,

$$\begin{aligned}
&\sum_{t=1}^T \Big(\psi_t(p_{t+1}) - \Phi_{t+1}(p_{t+1})\Big) + \Phi_{t+1}(p^*) - \Phi_1(p_1) \\
&\leq \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(p_{t+1}) + \frac{H(p_1)}{\eta_1} + 2\sum_{i=1}^k \log\left(\frac{1}{p_i^*}\right) \\
&\leq \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(p_{t+1}) + \frac{\log k}{\eta_1} + 2k\log T \,,
\end{aligned}$$

where in the last inequality we used the fact that $p_i^* \geq 1/T$ for all $i \in [k]$.

For the stability term, by Lemma 5.3 with $\delta = 1$ (since $p_t = q_t$),

$$\sum_{t=1}^T \Big(\langle p_t - p_{t+1}, \widehat{y}_t\rangle - D_{\psi_t}(p_{t+1}, p_t)\Big) \leq \sum_{t=1}^T \eta_t \nu_t \,.$$

We will confirm that the assumptions for Part (II) of Theorem 5.1 are indeed satisfied. By the definition of the learning rate in (5.7),

$$\begin{aligned}\beta_t &= \beta_1 + \sum_{u=1}^{t-1} \frac{c_1 \nu_u}{\sqrt{h_1}\sqrt{\sum_{s=1}^{u} \nu_s}} \geq \beta_1 + \frac{c_1}{\sqrt{h_1}} \sum_{u=1}^{t-1} \frac{\nu_u}{\sqrt{\sum_{s=1}^{u} \nu_s} + \sqrt{\sum_{s=1}^{u-1} \nu_s}} \\ &\geq \beta_1 + \frac{c_1}{\sqrt{h_1}} \sum_{u=1}^{t-1} \left(\sqrt{\sum_{s=1}^{u} \nu_s} - \sqrt{\sum_{s=1}^{u-1} \nu_s} \right) = \beta_1 + \frac{c_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^{t-1} \nu_s} \,.\end{aligned}$$

Using this inequality, β_t is bounded from below as

$$2\beta_t = \beta_t + \beta_t \geq 2\nu_t + \beta_1 + \frac{c_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^{t-1} \nu_s} \geq 2\sqrt{2\beta_1 \nu_t} + \frac{c_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^{t-1} \nu_s} \geq \frac{c_1}{\sqrt{h_1}} \sqrt{\sum_{s=1}^{t} \nu_s} \,,$$

where the first inequality follows by $\nu_t \leq \beta_t/2$ and the above inequality, the second inequality follows by the AM-GM inequality, and the last inequality follows from $2\sqrt{2\beta_1} \geq \frac{c_1}{\sqrt{h_1}}$ and $\sqrt{x} + \sqrt{y} \geq \sqrt{x+y}$ for $x, y \geq 0$. Dividing the both sides by 2, we can see stability condition (S2) in Theorem 5.1 is satisfied with $a = 1/2$. One can also see that the other assumptions are trivially satisfied. Hence, by Part (II) of Theorem 5.1,

$$\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(p_{t+1}) + \sum_{t=1}^{T} \eta_t \nu_t \leq 2 \left(c_1 + \frac{2}{c_1} \right) \sqrt{h_1 \sum_{t=1}^{T} \nu_t} \,.$$

Using the last inequality with $\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \nu_t}\right] \leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \omega_t}\right] \leq \sqrt{L_2}$, and setting $c_1 = \sqrt{2}$, we have

$$\begin{aligned}\mathsf{Reg}_T(i^*) &\leq \mathbb{E}\left[2 \left(c_1 + \frac{2}{c_1} \right) \sqrt{h_1 \sum_{t=1}^{T} \nu_t} + \frac{\log k}{\eta_1} + 2k \log T + k \right] \\ &\leq 4\sqrt{2}\sqrt{L_2 \log k} + 2k \log T + k + \frac{1}{4} \,,\end{aligned}$$

which completes the proof of Corollary 5.2. □

5.7.7 Proof of Results in Section 5.5.2

Section 5.7.7.1 provides preliminary results, which will be used to quantify the difference between q_t and q_{t+1} in Section 5.7.7.2 and will be used to prove the continuity of F_t in Section 5.7.7.3. Section 5.7.7.2 proves Theorem 5.2 and Section 5.7.7.3 discusses the bisection method to compute β_{t+1}.

5.7.7.1 Some stability results

Before proving Theorem 5.2, we prove several important lemmas. Consider the following three optimization problems:

$$\begin{aligned} p &\in \operatorname*{arg\,min}_{p' \in \mathcal{P}_k} \left\langle L - \xi e_1, p' \right\rangle + \beta \psi(p') \,, \\ q &\in \operatorname*{arg\,min}_{q' \in \mathcal{P}_k} \left\langle L, q' \right\rangle + \beta \psi(q') \,, \\ r &\in \operatorname*{arg\,min}_{r' \in \mathcal{P}_k} \left\langle L, r' \right\rangle + \beta' \psi'(r') \end{aligned} \tag{5.25}$$

with $L \in \mathbb{R}_+^k$, $0 \leq \xi \leq \min_{i \in [k]} L_i$, and $\beta, \beta' > 0$ satisfying $\beta' \geq \beta$,

$$\psi(q) = \sum_{i=1}^k (q_i \log q_i - q_i) - \frac{c}{\beta} \sum_{i=1}^k \log q_i \quad \text{and} \quad \psi'(q) = \sum_{i=1}^k (q_i \log q_i - q_i) - \frac{c}{\beta'} \sum_{i=1}^k \log q_i$$

for $c > 0$. Note that the outputs of FTRL with $\psi(q)$ and with $-H(q) - (c/\beta) \sum_{i=1}^k \log q_i$ are identical since adding a constant to ψ does not change the output of the above optimization problems.

In the following lemma, we bound $H(q)$ by $H(p)$.

Lemma 5.4. *Consider p and q in* (5.25). *Then, under $\eta := 1/\beta \leq \frac{1}{15k}$,*

$$H(q) \leq 3H(p)\,.$$

Proof. By Bubeck et al. (2018, Lemma 8) we have $q_i \leq 3p_i$ for all $i \in [k]$. Using this inequality and the concavity of entropy, $H(q)$ is bounded by a linear approximation as

$$\begin{aligned} H(q) \leq H(p) + \langle \nabla H(p), q - p \rangle &= \sum_{i=1}^k \left(p_i \log\left(\frac{1}{p_i}\right) + \left(\log\left(\frac{1}{p_i}\right) - 1\right)(q_i - p_i)\right) \\ &= H(p) + \sum_{i=1}^k (q_i - p_i) \log\left(\frac{1}{p_i}\right) \leq H(p) + \sum_{i=1}^k (3p_i - p_i) \log\left(\frac{1}{p_i}\right) = 3H(p)\,, \end{aligned}$$

where the last inequality follows by $q_i \leq 3p_i$. □

In the following lemma, we investigate the relation between q and r in (5.25).

Lemma 5.5. *Consider q and r in* (5.25). *Then,*

$$r_i \leq q_i^{\beta/\beta'}\,.$$

Proof. From the KKT conditions there exist $\mu, \mu' \in \mathbb{R}$ such that

$$L + \beta \nabla \psi(q) + \mu \mathbf{1} = 0 \quad \text{and} \quad L + \beta' \psi'(r) + \mu' \mathbf{1} = 0\,,$$

which implies, by $(\nabla \psi(q))_i = \log q_i - \frac{c}{\beta q_i}$, that

$$L_i + \beta \log q_i - \frac{c}{q_i} + \mu = 0 \quad \text{and} \quad \eta' L_i + \beta \log r_i - \frac{c}{r_i} + \mu' = 0 \tag{5.26}$$

for all $i \in [k]$. This is equivalent to

$$q_i = \exp\left(-\frac{1}{\beta}\left(L_i - \frac{c}{q_i} + \mu\right)\right) \quad \text{and} \quad r_i = \exp\left(-\frac{1}{\beta'}\left(L_i - \frac{c}{r_i} + \mu'\right)\right)\,.$$

Removing L_i from these equalities yields that

$$r_i = q_i^{\beta/\beta'} \exp\left(\frac{c}{\beta'}\left(\frac{1}{r_i} - \frac{1}{q_i}\right)\right) \exp\left(\frac{1}{\beta'}(\mu - \mu')\right)\,. \tag{5.27}$$

We will prove $\frac{\mathrm{d}\mu}{\mathrm{d}\beta} > 0$. Taking derivative with respect to β of (5.26), we have

$$\log q_i + \left(\frac{1}{q_i} + \frac{c}{q_i^2}\right) \frac{\mathrm{d}q_i}{\mathrm{d}\beta} + \frac{\mathrm{d}\mu}{\mathrm{d}\beta} = 0\,.$$

Multiplying $\left(\frac{1}{q_i}+\frac{c}{q_i^2}\right)^{-1}$ and summing over $i \in [k]$ in the last equality, we have

$$-\left(\frac{1}{q_i}+\frac{c}{q_i^2}\right)^{-1}\log(1/q_i)+\sum_{i=1}^{k}\frac{\mathrm{d}q_i}{\mathrm{d}\beta}+\left(\frac{1}{q_i}+\frac{c}{q_i^2}\right)^{-1}\frac{\mathrm{d}\mu}{\mathrm{d}\beta}=0\,,$$

which with the fact $\sum_{i=1}^{k}\frac{\mathrm{d}q_i}{\mathrm{d}\beta}=0$ implies $\frac{\mathrm{d}\mu}{\mathrm{d}\beta}>0$. Hence, since $\beta \le \beta'$ we have $\mu \le \mu'$.

When $r_i \le q_i$, it is obvious that we get $r_i \le q_i^{\beta/\beta'}$.

When $r_i > q_i$, using (5.27) with the inequalities $\beta \le \beta'$ and $\mu \le \mu'$,

$$r_i = q_i^{\beta/\beta'}\exp\left(\frac{c}{\beta'}\left(\frac{1}{r_i}-\frac{1}{q_i}\right)\right)\exp\left(\frac{1}{\beta'}(\mu-\mu')\right) \le q_i^{\beta/\beta'}\,,$$

which is the desired bound. □

Lemma 5.6. *Consider p, q, and r in* (5.25). *Then, under* $\eta := 1/\beta \le \frac{1}{15k}$, *we have*

$$r_i \le 3p_i^{\beta/\beta'}\,.$$

Proof. By Bubeck et al. (2018, Lemma 8) we have $q_i \le 3p_i$ for all $i \in [k]$. Using this with Lemma 5.5, we have

$$r_i \le q_i^{\beta/\beta'} \le 3q_i^{\beta/\beta'}\,.$$

□

5.7.7.2 Proof of Theorem 5.2

In this section, we will provide the proof of Theorem 5.2. We first see that the ratio β_t/β_{t+1} is close to one to some extent.

Lemma 5.7. *The learning rate* β_t *in* (5.8) *satisfies*

$$1-\frac{\beta_t}{\beta_{t+1}} \in (0, 1/10]\,.$$

Proof. Recall that $\beta_t = \beta_1 + \sum_{u=1}^{t-1} b_u$ with $b_u = \frac{c_1\nu_u}{U_u}$ and $U_t = \sqrt{c_2 + \bar{z}_t h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}$ for $t \in \{0\} \cup [T]$. It suffices to show

$$\frac{\beta_t}{\beta_{t+1}} = \frac{\beta_t}{\beta_t + b_t} \ge \frac{9}{10} \Leftrightarrow \beta_t \ge 9b_t\,.$$

This indeed follows since using $\nu_t \le \beta_t/2$ we have

$$b_t = \frac{c_1\nu_t}{\sqrt{81c_1^2 + \sum_{s=1}^{t-1} z_s h_s + z_t h_{t+1}}} \le \frac{c_1\nu_t}{\sqrt{81c_1^2}} = \frac{\nu_t}{9} \le \frac{\beta_t}{9}\,.$$

□

Finally, we are ready to prove one of the key lemmas for proving the BOBW regret bound with the sparsity-dependent bound. Recall that we have $p_t = \left(1-\frac{k}{T}\right)q_t + \frac{1}{T}\mathbf{1}$ and $h_t = \frac{1}{1-\frac{k}{T}}H(p_t)$. Using the result in Section 5.7.7.1, we will show that h_{t+1} is bounded in terms of h_t.

Lemma 5.8. *Suppose that β_t is defined as* (5.8)*. Then,*

$$h_{t+1} \le 3h_t + \frac{20k}{9}\left(\frac{\beta_{t+1}}{\beta_t} - 1\right)\log\left(\frac{T}{k}\right) h_{t+1} \,.$$

Proof. Let us recall that q_t and q_{t+1} are defined as

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{P}_k} \left\langle \sum_{s=1}^{t-1} \widehat{y}_s,\, q \right\rangle + \psi_t(q) \quad \text{and} \quad q_{t+1} \in \operatorname*{arg\,min}_{q \in \mathcal{P}_k} \left\langle \sum_{s=1}^{t} \widehat{y}_s,\, q \right\rangle + \Phi_{t+1}(q) \,,$$

which corresponds to optimization problems (5.25) with $p = q_t$, $L = \sum_{s=1}^t \widehat{y}_s$, $\xi = \widehat{y}_{tA_t}$, $\psi = \psi_t/\beta_t$, $\eta = 1/\beta_t$, $r = q_{t+1}$, $\psi' = \Phi_{t+1}/\beta_{t+1}$, and $\eta' = 1/\beta_{t+1}$.

Since H is concave, by $p_{ti} = (1 - \frac{k}{T})q_{ti} + \frac{1}{T}$ and Jensen's inequality we have

$$\left(1 - \frac{k}{T}\right) h_t = H(p_t) \ge \left(1 - \frac{k}{T}\right) H(q_t) + \frac{k}{T} H\left(\frac{1}{k}\mathbf{1}\right) \ge \left(1 - \frac{k}{T}\right) H(q_t) \,,$$

which implies $h_t \ge H(q_t)$. By Lemma 5.6 we also have $q_{t+1,i} \le 3q_{ti}^{\beta_t/\beta_{t+1}}$, which implies that

$$p_{t+1,i} = \left(1 - \frac{k}{T}\right) q_{t+1,i} + \frac{1}{T} \le \left(1 - \frac{k}{T}\right) 3q_{ti}^{\beta_t/\beta_{t+1}} + \frac{1}{T} \le 6p_{ti}^{\beta_t/\beta_{t+1}} \,.$$

The last inequality follows since when $\left(1 - \frac{k}{T}\right) 3q_{ti}^{\beta_t/\beta_{t+1}} \le \frac{1}{T}$,

$$\left(1 - \frac{k}{T}\right) 3q_{ti}^{\beta_t/\beta_{t+1}} + \frac{1}{T} \le \frac{2}{T} \le 2\left(\frac{1}{T}\right)^{\beta_t/\beta_{t+1}} \le 2\left(\left(1 - \frac{k}{T}\right) q_{ti} + \frac{1}{T}\right)^{\beta_t/\beta_{t+1}} = 2p_{ti}^{\beta/\beta_{t+1}} \,,$$

and otherwise

$$\left(1 - \frac{k}{T}\right) 3q_{ti}^{\beta_t/\beta_{t+1}} + \frac{1}{T} \le 6\left(1 - \frac{k}{T}\right)^{\beta_t/\beta_{t+1}} q_{ti}^{\beta_t/\beta_{t+1}} \le 6p_{ti}^{\beta/\beta_{t+1}} \,.$$

Using these inequalities, we have

$$\begin{aligned}
h_{t+1} - 3h_t &= \frac{1}{1 - \frac{k}{T}} \left(H(p_{t+1}) - 3H(p_t)\right) \\
&\le \frac{1}{1 - \frac{k}{T}} \left(H(p_t) + \langle \nabla H(p_t), p_{t+1} - p_t \rangle - 3H(p_t)\right) \\
&= \frac{1}{1 - \frac{k}{T}} \sum_{i=1}^{k} \left(p_{t+1,i} - 3p_{ti}\right) \log\left(\frac{1}{p_{ti}}\right) \\
&\le \sum_{i=1}^{k} \left(q_{t+1,i} - 3q_{ti}\right) \log\left(\frac{1}{p_{ti}}\right) \,, \qquad (5.28)
\end{aligned}$$

where the first inequality follows by the concavity of H, the second inequality follows since $p_{t+1,i} - 3p_{ti} \le \left(1 - \frac{k}{T}\right)\left(q_{t+1,i} - q_{ti}\right)$. Defining $\mathcal{Q}_t = \{i \in [k] : q_{t+1,i} - 3q_{ti} \ge 0\}$, (5.28) is further bounded as

$$\begin{aligned}
&\sum_{i=1}^{k} (q_{t+1,i} - 3q_{ti}) \log\left(\frac{1}{p_{ti}}\right) \\
&= \sum_{i \in \mathcal{Q}_t} (q_{t+1,i} - 3q_{ti}) \log\left(\frac{1}{p_{ti}}\right) + \sum_{i \notin \mathcal{Q}_t} (q_{t+1,i} - 3q_{ti}) \log\left(\frac{1}{p_{ti}}\right) \\
&\le \frac{\beta_{t+1}}{\beta_t} \sum_{i \in \mathcal{Q}_t} (q_{t+1,i} - 3q_{ti}) \log\left(\frac{1}{p_{t+1,i}}\right) + 0 \\
&\le \frac{10}{9} \sum_{i \in \mathcal{Q}_t} (q_{t+1,i} - 3q_{ti}) \log\left(\frac{1}{p_{t+1,i}}\right) \\
&\le \frac{10}{9} \sum_{i \in \mathcal{Q}_t} \left(q_{t+1,i} - q_{t+1,i}^{\beta_{t+1}/\beta_t}\right) \log\left(\frac{1}{p_{t+1,i}}\right) \\
&= \frac{10}{9} \sum_{i \in \mathcal{Q}_t} q_{t+1,i} \left(1 - q_{t+1,i}^{\frac{\beta_{t+1}}{\beta_t} - 1}\right) \log\left(\frac{1}{p_{t+1,i}}\right) \,, \qquad (5.29)
\end{aligned}$$

where the first inequality follows by $p_{t+1,i} \le 6p_t^{\beta_t/\beta_{t+1}}$, the second follows by Lemma 5.7, and the last inequality follows by $q_{t+1,i} \le 3q_{ti}^{\beta_t/\beta_{t+1}}$. Since for any $\epsilon > 0$, $x \in [0,1]$, and $\gamma \in [0,1]$, it holds that

$$\begin{aligned}
x(1 - x^{\epsilon}) &\le x \log\left(\frac{1}{x^{\epsilon}}\right) = \epsilon x \log\left(\frac{1}{x}\right) \\
&\le \epsilon \left(\left(\log \frac{1}{\gamma} - 1 \right) (x - r) + \gamma \log \frac{1}{\gamma} \right) \le \epsilon \log\left(\frac{1}{\gamma}\right) (\gamma + (1 - \gamma) x) \,, \qquad (5.30)
\end{aligned}$$

setting $\gamma = k/T$ in (5.30) implies that the RHS of (5.29) is further bounded as

$$\begin{aligned}
&h_{t+1} - 3h_t \\
&\le \frac{10}{9} \sum_{i \in \mathcal{Q}_t} \left(\frac{\beta_{t+1}}{\beta_t} - 1\right) \log(T/k) \left(\frac{k}{T} + \left(1 - \frac{k}{T}\right) q_{t+1,i}\right) \log\left(\frac{1}{p_{t+1,i}}\right) \\
&\le \frac{10k}{9} \left(\frac{\beta_{t+1}}{\beta_t} - 1\right) \log(T/k) \sum_{i=1}^{k} \left(\frac{1}{T} + \left(1 - \frac{k}{T}\right) q_{t+1,i}\right) \log\left(\frac{1}{p_{t+1,i}}\right) \\
&\le \frac{20k}{9} \left(\frac{\beta_{t+1}}{\beta_t} - 1\right) \log(T/k) h_{t+1} \,,
\end{aligned}$$

where the second inequality follows by Lemma 5.7 and the last inequality follows by the definition of h_{t+1}. □

Finally we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. Fix $i^* \in [k]$ and define $p^* \in \mathcal{P}_k$ by

$$p^* = \left(1 - \frac{k}{T}\right) e_{i^*} + \frac{1}{T} \mathbf{1} \,.$$

Then, using the definition of the algorithm,

$$\begin{aligned}
\mathsf{Reg}_T(i^*) &= \mathbb{E}\left[\sum_{t=1}^T \ell_{tA_t} - \sum_{t=1}^T \ell_{ti^*}\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - e_{i^*}\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, q_t - e_{i^*}\rangle\right] + \mathbb{E}\left[\gamma \sum_{t=1}^T \left\langle \ell_t, \frac{1}{k}\mathbf{1} - q_t \right\rangle\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, q_t - p^*\rangle\right] + \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p^* - e_{i^*}\rangle\right] + \gamma T \\
&\le \mathbb{E}\left[\sum_{t=1}^T \langle \widehat{y}_t, q_t - p^*\rangle\right] + 2k\,,
\end{aligned}$$

where the first inequality follows since $p_t = (1-\gamma)q_t + \frac{\gamma}{k}\mathbf{1}$ and the last inequality follows by the definition of p^* and $\gamma = \frac{k}{T}$. By the standard analysis of the FTRL described in Section 5.3,

$$\begin{aligned}
\sum_{t=1}^T \langle \widehat{y}_t, q_t - p^*\rangle \le \sum_{t=1}^T \Big(\psi_t(q_{t+1}) - \Phi_{t+1}(q_{t+1})\Big) + \Phi_{t+1}(p^*) - \Phi_1(q_1) \\
+ \sum_{t=1}^T \Big(\langle q_t - q_{t+1}, \widehat{y}_t\rangle - D_{\psi_t}(q_{t+1}, q_t)\Big)\,.
\end{aligned}$$

We first consider the penalty term. Since $\psi_t = \frac{1}{\eta_t}\psi^{\mathsf{nS}} + 4\psi^{\mathsf{LB}}$,

$$\begin{aligned}
&\sum_{t=1}^T \Big(\psi_t(q_{t+1}) - \Phi_{t+1}(q_{t+1})\Big) + \Phi_{t+1}(p^*) - \Phi_1(q_1) \\
&\le \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + 4\sum_{i=1}^k \log\left(\frac{1}{p_i^*}\right) \\
&\le \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{\log k}{\eta_1} + 4k\log T\,,
\end{aligned}$$

where in the last inequality we used the fact that $p_i^* \ge 1/T$ for all $i \in [k]$.

For the stability term, by Lemma 5.3 with $\delta = 2$,

$$\sum_{t=1}^T \Big(\langle q_t - q_{t+1}, \widehat{y}_t\rangle - D_{\psi_t}(q_{t+1}, q_t)\Big) \le 2\sum_{t=1}^T \eta_t \nu_t\,.$$

We will confirm that the assumptions for Part (I) of Theorem 5.1 are indeed satisfied. By the definition of the learning rate in (5.8) and $\nu_t \le \beta_t/2$,

$$\frac{\sqrt{c_2}}{c_1}(\beta_1 + \beta_t) \ge 9(\beta_1 + \nu_t) \ge \beta_1 + \nu_t\,.$$

Hence stability condition (S1) of Theorem 5.1 is satisfied and one can also see that the other assumptions are trivially satisfied. Hence, by Part (I) of Theorem 5.1,

$$\begin{aligned}\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)+2\sum_{t=1}^{T}\eta_t\nu_t &\le 2\left(c_1+\frac{2}{c_1}\log\left(1+\sum_{s=1}^{T}\frac{\nu_s}{\beta_1}\right)\right)\sqrt{c_2+\sum_{t=1}^{T+1}\nu_t h_{t+1}}\\ &\le 2\left(c_1+\frac{2}{c_1}\log\left(1+\frac{T^2}{\beta_1}\right)\right)\sqrt{c_2+\sum_{t=1}^{T+1}\nu_t h_{t+1}}\,,\\ &= 2\sqrt{2\log\left(1+\frac{T^2}{\beta_1}\right)}\sqrt{c_2+\sum_{t=1}^{T+1}\nu_t h_{t+1}}\,,\end{aligned}$$

where in the last inequality we used $\nu_t \le T$ and in the equality we set $c_1 = \sqrt{2\log\left(1+\frac{T^2}{\beta_1}\right)}$.

By summing up the above arguments and Jensen's inequality, we have

$$\begin{aligned}\mathsf{Reg}_T(i^*) &\le \mathbb{E}\left[2\sqrt{2\log\left(1+\frac{T^2}{\beta_1}\right)}\sqrt{c_2+\sum_{t=1}^{T+1}\nu_t h_{t+1}}\right]+2k+4k\log T+\frac{\log k}{\eta_1}\\ &\le 2\sqrt{2\log\left(1+\frac{T^2}{\beta_1}\right)}\sqrt{c_2+\mathbb{E}\left[\sum_{t=1}^{T+1}\nu_t h_{t+1}\right]}+2k+4k\log T+15k\log k\\ &\le 2\sqrt{2\log\left(1+\frac{T^2}{\beta_1}\right)}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\nu_t h_{t+1}\right]}+O(k\log T)\,. \qquad (5.31)\end{aligned}$$

(Adversarial regime) We first consider the adversarial regime. Recall that $\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\nu_t}\right] \le \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\omega_t}\right] \le \sqrt{L_2}$ as was done in the proof of Corollary 5.2. Hence (5.31) with $h_t \le 2\log k$ (since $T \ge 2k$) yields that

$$\mathsf{Reg}_T \le 4\sqrt{L_2\log(k)\log\left(1+\frac{T^2}{\beta_1}\right)}+O(k\log T)\,.$$

(Adversarial regime with a self-bounding constraint) Next we consider the adversarial regime with a self-bounding constraint. We will bound a component of (5.31). By Lemma 5.8, $\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\nu_t h_{t+1}\right]}$ is bounded as

$$\begin{aligned}X_t &:= \sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\nu_t h_{t+1}\right]}\\ &\le \sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T}\nu_t h_t\right]+\frac{20k}{9}\log\left(\frac{T}{k}\right)\mathbb{E}\left[\sum_{t=1}^{T}\nu_t\left(\frac{\beta_{t+1}}{\beta_t}-1\right)h_{t+1}\right]}\\ &\le \sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T}\nu_t h_t\right]+\frac{10k}{9}\log\left(\frac{T}{k}\right)\mathbb{E}\left[\sum_{t=1}^{T}(\beta_{t+1}-\beta_t)\,h_{t+1}\right]}\\ &= \sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T}\nu_t h_t\right]+\frac{10k}{9}\log\left(\frac{T}{k}\right)\mathbb{E}\left[\sum_{t=1}^{T}\frac{c_1\nu_t h_{t+1}}{\sqrt{c_2+\sum_{t=1}^{T}\nu_s h_{s+1}}}\right]}\end{aligned}$$

Algorithm 5.2: BOBW algorithm with a sparsity-dependent bound in Section 5.5.2

```
for $t = 1, 2, \dots, T$ do
    Compute $q_t$ using (5.1).
    Sample $A_t \sim p_t$, observe $\ell_{tA_t} \in [-1, 1]$, and compute $\widehat{y}_t$.
    Update $\beta_t$ using (5.8) based on the bisection method (Algorithm 5.3).
```

$$
\begin{aligned}
&\leq \sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T} \nu_t h_t\right] + \frac{20k}{9} \log\left(\frac{T}{k}\right) \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \nu_t h_{t+1}}\right]} \\
&= \sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T} \nu_t h_t\right] + \frac{20k}{9} \log\left(\frac{T}{k}\right) X_t}\,,
\end{aligned}
$$

where the first inequality follows by Lemma 5.8, the second inequality follows by $\nu_t \leq \beta_t/2$, the last inequality follows by Lemma 5.9. Since $x \leq \sqrt{a + bx}$ for $x > 0$ implies $x \leq 2\sqrt{a} + b$, we have

$$
\begin{aligned}
X_t &\leq 2\sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T} \nu_t h_t\right]} + \frac{20k}{9} \log\left(\frac{T}{k}\right) = 2\sqrt{3\,\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\nu_t \,|\, p_t] h_t\right]} + \frac{20k}{9} \log\left(\frac{T}{k}\right) \\
&\leq 2\sqrt{6k\,\mathbb{E}\left[\sum_{t=1}^{T} H(p_t)\right]} + \frac{20k}{9} \log\left(\frac{T}{k}\right) .
\end{aligned}
$$

We consider the case of $P(a^*) \geq \mathrm{e}$, since otherwise Lemma 5.2 implies $\sum_{t=1}^{T} H(p_t) \leq \mathrm{e} \log(kT)$ and thus the desired bound is trivially obtained. When $P(a^*) \geq \mathrm{e}$, Lemma 5.2 implies that $\sum_{t=1}^{T} H(p_t) \leq P(a^*) \log(kT)$. Then from the self-bounding technique, for any $\lambda \in (0, 1]$ it holds that

$$
\begin{aligned}
\mathsf{Reg}_T &= (1 + \lambda)\mathsf{Reg}_T - \lambda \mathsf{Reg}_T \\
&\leq \mathbb{E}\Big[(1 + \lambda) O\left(\sqrt{k \log(T) \log(kT) P(a^*)}\right) - \lambda \Delta_{\min} P(a^*)\Big] + \lambda C + O(k \log T) \\
&\leq O\left(\frac{(1 + \lambda)^2 k \log(T) \log(kT)}{\lambda \Delta_{\min}} + \lambda C\right) \\
&= O\left(\frac{k \log(T) \log(kT)}{\Delta_{\min}} + \lambda\left(\frac{k \log(T) \log(kT)}{\Delta_{\min}} + C\right) + \frac{1}{\lambda} \frac{k \log(T) \log(kT)}{\Delta_{\min}}\right),
\end{aligned}
$$

where the first inequality follows by Lemma 5.1 and the second inequality follows from $a\sqrt{x} - bx/2 \leq a^2/(2b)$ for $a, b, x \geq 0$. Setting $\lambda \in (0, 1]$ to

$$
\lambda = \sqrt{\frac{k \log(T) \log(kT)}{\Delta_{\min}} \Big/ \left(\frac{k \log(T) \log(kT)}{\Delta_{\min}} + C\right)}
$$

gives the desired bound for the adversarial regime with a self-bounding constraint. □

5.7.7.3 Discussion on Bisection Method for Computing β_{t+1}

This section describes the bisection method to compute β_{t+1} described in Section 5.5.2. Recall that $F_t : [\beta_t, \beta_t + T] \to \mathbb{R}$ is defined by the difference of the both sides of the

Algorithm 5.3: Bisection method for computing β_{t+1}

```
input: $F_t$
left ← $\beta_t$, right ← $\beta_t + T$
while true do
    center ← (left + right)/2
    if $F_t$(center) < 0 then
        left ← center
    else if $F_t$(center) > 0 then
        right ← center
    else
        break
return center
```

update rule of (β_t) in (5.8):

$$F_t(\alpha) = \alpha - \left(\beta_t + \frac{c_1 \nu_t}{\sqrt{c_2 + \nu_t h_{t+1}(\alpha) + \sum_{s=1}^{t-1} \nu_s h_{s+1}}}\right), \qquad (5.32)$$

where $h_{t+1}(\alpha) = \frac{1}{1-\frac{k}{T}} H(p_{t+1}(\alpha))$, and $p_{t+1}(\alpha)$ is the FTRL output with the regularizer $\psi_t = \alpha\psi^{\mathsf{nS}} + 4\psi^{\mathsf{LB}}$. Note that $c_1\nu_t / \sqrt{c_2 + \nu_t h_{t+1}(\alpha) + \sum_{s=1}^{t-1} \nu_s h_{s+1}} \le c_1\nu_t/c_2 \le T/9$ since $\nu_t \le T$.

Assume that F_t is continuous. Then we can see that there exists $\alpha \in [\beta_t, \beta_t + T]$ such that $F_t(\alpha) = 0$. In fact, if $p_{tA_t} = 0$ then $\beta_{t+1} = \beta_t$, and otherwise, we have $F_t(\beta_t) \le 0$ and $F_t(\beta_t + T) > 0$. Using the intermediate value theorem with the assumption that F_t is continuous, there indeed exists $\alpha \in [\beta_t, \beta_t + T]$ satisfying $F_t(\alpha) = 0$. We can compute such α by the bisection method. In particular, we first set the range of α to $[\beta_t, \beta_t + T]$, and then iteratively halve it by evaluating the value of F_t at the middle point. Such bisection method (binary search) are also used in Wei et al. (2016), although the computed target is different. The whole BOBW algorithm with the sparsity-dependent bound in Section 5.5.2 is given in Algorithm 5.2, and the concrete procedure of the bisection given in Algorithm 5.3.

Now, all that remains is to show that F_t is continuous. To prove this, it suffices to prove that $h_{t+1}(\alpha) = \frac{1}{1-\frac{k}{T}} H(p_{t+1}(\alpha))$ is continuous with respect to α.

Proposition 5.2. *F_t in* (5.32) *is continuous with respect to α.*

Proof of Proposition 5.2. Take any $\alpha \in [\beta_t, \beta_t + T]$ and then consider the following optimization problem:

$$q_{t+1}(\alpha) = \operatorname*{arg\,min}_{q \in \mathcal{P}_k} \left\langle \sum_{s=1}^{t} \widehat{y}_s, q \right\rangle + \Phi_{t+1}(q),$$

where $\Phi_{t+1} = \alpha\psi^{\mathsf{nS}} + 4\psi^{\mathsf{LB}}$. Now using Corollary 8.1 of Hogan (1973) with the fact that the solution of the above optimization problem is unique, $q_{t+1}(\alpha)$ is continuous with respect to α. This completes the proof since p_{t+1} is continuous with respect to q_{t+1}, $1/T \le p_{t+1,i}(\alpha) \le 1 - k/T$, and $H(p)$ is continuous in a neighborhood of $p = p_{t+1}(\alpha)$. □

5.7.8 Proof of Corollary 5.3

This section proves Corollary 5.3. Recall that $B = 1/2$ for FI, $B = k/2$ for MAB, and $B = 2mk^2$ for PM-local, and $r_{\mathcal{M}}$ is 1 if $\mathcal{M}$ is FI or MAB, and $2k$ if $\mathcal{M}$ is PM-local, which are appeared in Section 5.6. Let $\mathsf{Reg}_T(a) = \mathbb{E}\big[\sum_{t=1}^T \big(\mathcal{L}_{A_t x_t} - \mathcal{L}_{a x_t}\big)\big] = \mathbb{E}\big[\sum_{t=1}^T \langle \ell_{A_t} - \ell_a, e_{x_t}\rangle\big]$ for $a \in [k]$.

Proof. Fix $i^* \in [k]$. From Lemma 7 in Chapter 4, if $\eta_t > 0$, we have

$$\mathsf{Reg}_T(i^*) \le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \eta_t V_t'\right] . \tag{5.33}$$

We will confirm that the assumptions for Part (I) of Theorem 5.1 are indeed satisfied. Since

$$\frac{\sqrt{c_2 + \bar{z}_t h_1}}{c_1}(\beta_t + \beta_1) \ge \sqrt{\frac{2\bar{V}\log k}{\log(1+T)}} \cdot 2B\sqrt{\frac{\log(1+T)}{\log k}} \ge \sqrt{2}\left(\bar{V} + \bar{V}_t\right) ,$$

stability condition (S1) is satisfied. One can also see that the other conditions are trivially satisfied. Hence, using Part (I) of Theorem 5.1, we can bound the RHS of (5.33) as

$$\begin{aligned}
\mathsf{Reg}_T(i^*) &\le \mathbb{E}\left[\left(2c_1 + \frac{1}{c_1}\log\left(1 + \sum_{u=1}^T \frac{V_u'}{\bar{V}}\right)\right)\sqrt{\bar{V}H(q_1) + \sum_{t=1}^T V_t' H(q_{t+1})}\right] + \frac{H(q_1)}{\eta_1} \\
&\le \left(2c_1 + \frac{1}{c_1}\log(1+T)\right)\sqrt{\mathbb{E}\left[\sum_{t=1}^T V_t' H(q_{t+1})\right]} \\
&\qquad + O\left(\sqrt{\bar{V}\log(k)\log(T)} + B\sqrt{\log(k)\log T}\right) \\
&= \sqrt{2\log(1+T)}\sqrt{\mathbb{E}\left[\sum_{t=1}^T V_t' H(q_{t+1})\right]} + O\left(B\sqrt{\log(k)\log(T)}\right) , \qquad (5.34)
\end{aligned}$$

where the second inequality follows from $V_u'/\bar{V} \le 1$ and in the equality we set $c_1 = \sqrt{\frac{\log(1+T)}{2}}$ and used $\sqrt{\bar{V}} \le B$.

(Adversarial regime) For the adversarial regime, since $H(q_t) \le \log k$, (5.34) immediately implies

$$\mathsf{Reg}_T \le \mathbb{E}\left[\sqrt{2\sum_{t=1}^T V_t' \log(k)\log(1+T)} + O\left(B\sqrt{\log(k)\log(T)}\right)\right] ,$$

which is the desired bound.

(Adversarial regime with a self-bounding constraint) Next, we consider the adversarial regime with a self-bounding constraint. We consider the case of $Q(a^*) \ge \mathrm{e}$, since otherwise Lemma 5.2 implies $\sum_{t=1}^T H(p_t) \le \mathrm{e}\log(kT)$ and thus the desired bound is trivially obtained. When $Q(a^*) \ge \mathrm{e}$, Lemma 5.2 implies that $\sum_{t=1}^T H(q_t) \le$

$Q(a^*)\log(kT)$. Then from the self-bounding technique, for any $\lambda \in (0,1]$

$$
\begin{aligned}
&\mathsf{Reg}_T = (1+\lambda)\mathsf{Reg}_T - \lambda\mathsf{Reg}_T \\
&\leq \mathbb{E}\left[(1+\lambda)O\left(\sqrt{\bar{V}\log(T)\log(kT)Q(a^*)}\right) - \frac{\lambda\Delta_{\min}Q(a^*)}{r_{\mathcal{M}}}\right] + \lambda C \\
&\leq (1+\lambda)O\left(\sqrt{\bar{V}\log(T)\log(kT)\bar{Q}(a^*)}\right) - \frac{\lambda\Delta_{\min}\bar{Q}(a^*)}{r_{\mathcal{M}}} + \lambda C \\
&\leq O\left(\frac{(1+\lambda)^2 r_{\mathcal{M}}\log(T)\log(kT)}{\lambda\Delta_{\min}} + \lambda C\right) \\
&= O\left(\frac{r_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}} + \lambda\left(\frac{r_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}} + C\right) + \frac{1}{\lambda}\frac{r_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}}\right),
\end{aligned}
$$

where the first inequality follows by (5.34) and Lemma 5.1 with $c' = r_{\mathcal{M}}$ and the second inequality follows from $a\sqrt{x} - bx/2 \leq a^2/(2b)$ for $a, b, x \geq 0$. Setting $\lambda \in (0,1]$ to

$$
\lambda = \sqrt{\frac{r_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}} \Big/ \left(\frac{r_{\mathcal{M}}\bar{V}\log(T)\log(kT)}{\Delta_{\min}} + C\right)}
$$

gives the desired bound for the adversarial regime with a self-bounding constraint. □

5.7.9 Basic Lemma

Lemma 5.9 (Orabona 2019, Lemma 4.13)**.** *Let $a_0 \geq 0$, $(a_t)_{t=1}^T$ be non-negative reals and $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-increasing function. Then,*

$$
\sum_{t=1}^{T} a_t f\left(a_0 + \sum_{s=1}^{t} a_s\right) \leq \int_{a_0}^{\sum_{t=0}^{T} a_t} f(x)\mathrm{d}x\,.
$$

We include the proof for the completeness.

Proof. Let $A_t = \sum_{s=0}^{t} a_s$. Then summing the following inequality over t completes the proof:

$$
a_t f\left(a_0 + \sum_{s=1}^{t} a_s\right) = a_t f(A_t) = \int_{A_{t-1}}^{A_t} f(A_t)\mathrm{d}x \leq \int_{A_{t-1}}^{A_t} f(x)\mathrm{d}x\,.
$$

□

5.8 Conclusion

In this chapter, we established the framework for designing a learning rate that jointly depends on the stability and penalty components (Theorem 5.1). This result combined with a novel stability analysis enables us to achieve BOBW and data-dependent bounds (sparsity- and game-dependent bounds) simultaneously in MAB and PM. There are some remaining questions. First, while we only used the Shannon entropy as a dominant regularizer, it might be interesting to consider the other regularizers, such as negative Tsallis entropy. Second, we only considered the PM game with local observability, and it is an interesting open question to investigate if game-dependent bound with BOBW guarantee is possible in PM only with global observability.

Chapter 6

Further Adaptive Best of Both Worlds Algorithm for Combinatorial Semi-Bandits

There are numerous sequential decision-making problems with combinatorial structures in our daily lives, such as the online advertisement placement problem and the online shortest path problem. While most of these problems can be formulated as a partial monitoring game, such a formulation sometime fails to sufficiently model the combinatorial structure inherent in these problems. In this chapter, we investigate the combinatorial semi-bandit problem that involves such a combinatorial structure in its formulation. We first present a new algorithm with a best-of-both-worlds regret guarantee, in which the regrets are near-optimally bounded in the stochastic and adversarial regimes. In the stochastic regime, we prove a variance-dependent regret bound depending on the tight suboptimality gap introduced by Kveton et al. (2015) with a good leading constant. In the adversarial regime, we show that the same algorithm simultaneously obtains various data-dependent regret bounds. Our algorithm is based on the follow-the-regularized-leader framework with a refined regularizer and adaptive learning rate. Finally, we numerically test the proposed algorithm and confirm its superior or competitive performance over existing algorithms, including Thompson sampling under most settings.

6.1 Introduction

The combinatorial semi-bandit problem is an online decision-making problem, and it includes many practical problems such as multi-task bandits (Cesa-Bianchi and Lugosi, 2012), crowdsourcing (ul Hassan and Curry, 2016), learning spectrum allocations (Gai et al., 2012), shortest path problem (Gai et al., 2012), and recommender systems (Qin et al., 2014). In combinatorial semi-bandits, the learner and environment play the game sequentially. The learner is given an action set $\mathcal{A} \subset \{0,1\}^d$, where $d \in \mathbb{N}$ is the dimension of the action set. For every round $t \in [T] \coloneqq \{1, \dots, T\}$, the environment chooses a loss $\ell(t) \in [0,1]^d$, and the learner then chooses an action $a(t) \in \mathcal{A}$ (also called a *super-arm*), incurs a loss $\langle \ell(t), a(t) \rangle$, and observes $\ell_i(t)$ for all $i \in [d]$ such that $a_i(t) = 1$. We refer to each index $i \in [d]$ as *base-arm* i. The goal of the learner is to minimize their cumulative loss over all rounds. The performance of the learner is evaluated based on regret Reg_T defined as the difference between the cumulative losses of the learner and the single optimal action a^* fixed in terms of the expected cumulative loss, *i.e.,* $a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}[\sum_{t=1}^T \langle \ell(t), a \rangle]$ and

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \langle \ell(t), a(t) - a^* \rangle\right],$$

where the expectation is taken w.r.t. the randomness of $\ell(t)$ and the internal randomness of the algorithm.

The combinatorial semi-bandit problem, or more broadly, a variety of online-decision making problems, have been investigated within mainly two regimes: *stochastic* and *adversarial* regimes. In the stochastic regime, the sequence of losses $(\ell(t))_{t=1}^{T}$ is sampled from a fixed distribution $\mathcal{D}$ in an i.i.d. manner with mean $\mu = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell]$. In the adversarial regime, the losses are arbitrarily decided from $[0,1]^d$ (Kveton et al., 2015; Neu, 2015; Wang and Chen, 2018) or more generally from S^d for some bounded $S \subset \mathbb{R}$ (Wei and Luo, 2018; Zimmert et al., 2019) possibly depending on the past history of learner's actions.

There have been a considerable number of studies on combinatorial semi-bandits for both adversarial and stochastic regimes. In the adversarial regime, the regret bound of $O(\sqrt{mdT})$ was proved for $m = \max_{a \in \mathcal{A}} \|a\|_1$ (Audibert et al., 2014), which matches the lower bound of $\Omega(\sqrt{mdT})$ (Audibert et al., 2014). In the stochastic regime, many algorithms have been shown to achieve logarithmic regrets depending on the minimum suboptimality gap, which is defined by $\Delta = \min\{\mu^\top(a - a^*) : a \in \mathcal{A} \setminus \{a^*\}\}$. Kveton et al. (2015) and Wang and Chen (2018) derived gap-dependent regret bounds given by $O(dm \log(T)/\Delta)$ for general action sets and $O((d-m)\log(T)/\Delta)$ for matroid semi-bandits. Furthermore, Kveton et al. (2015) derived a refined bound given by $O(\sum_{i:a_i^*=1}(m/\Delta_{i,\min})\log T)$ depending on $\Delta_{i,\min} = \min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 1\} \geq \Delta$ of each base-arm i rather than on Δ.

It is unclear which regime's algorithms are better suited to practical applications. Algorithms specialized for the stochastic regime occasionally suffer a linear regret, whereas algorithms for the adversarial regime work poorly in the stochastic regime. Because it is difficult to know it in practice, it is desirable to obtain a near-optimal performance both for the stochastic and adversarial regimes *without* knowing the underlying environment.

To this end, particulary in the classical multi-armed bandits, the Best-of-Both-Worlds (BOBW) algorithm has been developed, which performs near-optimally both in the stochastic and adversarial regimes. In a seminal study, Bubeck and Slivkins (2012) developed the first BOBW algorithm, and the celebrated Tsallis-INF algorithm was recently proposed by Zimmert and Seldin (2021). For combinatorial semi-bandits, we are aware of the works by Wei and Luo (2018), Zimmert et al. (2019), and Ito (2021a). Some BOBW algorithms achieve favorable regret guarantees also in the *stochastic regime with adversarial corruptions* (Lykouris et al., 2018), which is an intermediate regime between the stochastic and adversarial regimes. This intermediate regime is advantageous in practice since the stochastic assumption on losses often fails to hold, whereas the adversarial assumption is excessively pessimistic.

Adaptive algorithms that exploit the characteristics of a sequence of losses have been actively developed for both the adversarial and stochastic regimes. In the adversarial regime, *data-dependent regret bounds* have been recently investigated to enhance the adaptivity of the algorithm to a given structure of the loss data. Well-known examples are the first-order bounds depending on the cumulative loss $L^* = \min_{a \in \mathcal{A}} \mathbb{E}[\sum_{t=1}^{T} \langle \ell(t), a \rangle]$, second-order bounds depending on the empirical variations of losses $Q_2 = \mathbb{E}[\sum_{t=1}^{T} \|\ell(t) - \bar{\ell}\|^2]$ defined with $\bar{\ell} = T^{-1}\mathbb{E}[\sum_{t=1}^{T} \ell(t)]$, and path-length bounds depending on the variation of losses $V_1 = \mathbb{E}[\sum_{t=1}^{T-1} \|\ell(t) - \ell(t+1)\|_1]$. For the semi-bandit problem, Wei and Luo (2018) presented the first-order regret bound of $O(\sqrt{dL^* \log T})$, second-order bound of $O(\sqrt{dQ_2 \log T})$, and the path-length bound of $O(\sqrt{dV_1 \log T})$. Note that the data-dependent bounds developed by Wei and Luo (2018) cannot be achieved using the same algorithm. Table 6.1 summarizes notation used in this chapter.

In the stochastic regime, one of the most promising approaches to making an algorithm more adaptive is to estimate and use distributional information. In the multi-armed bandit problem, algorithms that exploit the *variance* of losses have been developed (Audibert et al., 2007; Ito et al., 2022b), and (co)variance-aware algorithms for semi-bandits

Table 6.1: Notation used in this chapter

Symbol	Meaning
$\mathcal{A} \in \{0,1\}^d$	Action set
$d \in \mathbb{N}$	Dimensionality of action set
$m \le d$	$m = \max_{a\in\mathcal{A}} \|a\|_1$
$a^* \in \mathcal{A}$	Optimal action
$I^* \subset [d]$	$\{i \in [d] : a^*_i = 1\}$, set of optimal base-arms
$J^* \subset [d]$	$[d] \setminus I^*$, set of sub-optimal base-arms
$\mu_i \in [0,1]$	$\mathbb{E}[\ell_i]$, mean of base-arm i
$\sigma_i^2 \in [0,1/4]$	$\mathbb{E}[(\ell_i - \mu_i)^2]$, variance of base-arm i
$\Delta \in (0,m]$	$\min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\}\}$
$\Delta_{i,\min} \ge \Delta$	$\min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 1\}$
$\Delta'_{i,\min} \ge \Delta$	$\min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 0\}$
$w(\mathcal{A}) \le m$	Action-set-dependent constant (Section 6.5)
L^*	$\min_{a\in\mathcal{A}} \mathbb{E}[\sum_{t=1}^T \langle \ell(t), a\rangle]$
Q_2	$\mathbb{E}[\sum_{t=1}^T \|\ell(t) - \bar{\ell}\|^2]$ $(\bar{\ell} = T^{-1}\mathbb{E}[\sum_{t=1}^T \ell(t)])$
V_1	$\mathbb{E}[\sum_{t=1}^{T-1} \|\ell(t) - \ell(t+1)\|_1]$
$C \in [0,T]$	$\mathbb{E}[\sum_{t=1}^T \|\ell(t) - \ell'(t)\|_\infty]$, corruption level

have also been investigated (Komiyama et al., 2015b; Degenne and Perchet, 2016; Merlis and Mannor, 2019; Perrault et al., 2020; Vial et al., 2022; Liu et al., 2022). The variance-aware algorithm is highly advantageous in real-world applications since the variances of losses for each base-arm i, $\sigma_i^2 = \mathbb{E}_{\ell\sim\mathcal{D}}[(\ell_i - \mu_i)^2] \in [0,1/4]$, are extremely small in many real-world applications, whereas the variance can be $1/4$ in the worst case scenario. For example, for a search engine, the click-through rate is usually below 0.05 (Komiyama et al., 2017), implying that the variance of the base-arm is much smaller than $1/4$. Additionally, in the shortest path problem (György et al., 2007), the congestion level of the road does not change substantially in many cases, and hence the variance is expected to be much smaller than in the worst-case scenario also of this problem. Indeed, variance-aware algorithms are known to be highly effective in the problem of online eco-routing for electric vehicles (Chen et al., 2022b). Accordingly, we aim to achieve a variance-dependent regret bounds in the stochastic regime with multiple data-dependent regret bounds simultaneously in the adversarial regime by the same algorithm.

Contribution of This Chapter In this study, we establish a new BOBW algorithm for the combinatorial semi-bandit problem. The proposed algorithm is based on the Optimistic Follow-the-Regularized-Leader (OFTRL) framework (McMahan, 2011; Rakhlin and Sridharan, 2013b,a) with a refined regularizer and adaptive learning rate inspired by Ito et al. (2022b). Let $I^* = \{i \in [d] : a^*_i = 1\}$ and $J^* = [d] \setminus I^*$. OFTRL has a component called an optimistic prediction and the proposed algorithm considers two methods for its estimation: the Least Square (LS) and Gradient Descent (GD) based on past observations. Let $w(\mathcal{A}) \le m$ be an action-set-dependent constant defined in Section 6.5. We drop $\mathcal{A}$ when it is clear from context. The regret of the proposed algorithm with the LS and GD is then bounded as follows.

Theorem 6.1 (Informal)**.** *For the stochastic regime, the proposed algorithm with LS achieves*

$$\mathsf{Reg}_T \le \left(\sum_{i\in J^*} \max\left\{4\frac{w\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1 + \frac{w\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T$$
$$+\, o(\log T) =: \mathcal{R}^{\mathrm{LS}}\,,$$

Table 6.2: Regret upper bounds for combinatorial semi-bandits. $w = w(\mathcal{A}) \leq m$ is an action-set-dependent constant. Some terms are omitted due to the space constraint.

Reference	Regime	Regret bound
Audibert et al. (2014)	Adv.	$O(\sqrt{dmT})$
Kveton et al. (2015)	Stoc.	$534 \sum_{i \in J^*} \frac{m}{\Delta_{i,\min}} \log T + O(dm)$
Zimmert et al. (2019)	Adv.	$O(\sqrt{dmT})$
	Stoc.	$O\left(\frac{dm}{\Delta} \log T\right) =: \mathcal{R}^{\mathrm{ZLS}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{ZLS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{ZLS}}})$
Ito (2021a)	Adv.	$O(\sqrt{d \min\{L^*, Q_2, V_1\} \log T})$
	Stoc.	$O\left(\frac{dm}{\Delta} \log T\right) =: \mathcal{R}^{\mathrm{I}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{I}} + \sqrt{Cm\mathcal{R}^{\mathrm{I}}}$
Proposed (LS)	Adv.	$\sqrt{4d \min\{L^*, mT - L^*, Q_2\} \log T}$
	Stoc.	$\left(\sum_{i \in J^*} \max\left\{ \frac{4w\sigma_i^2}{\Delta_{i,\min}} + o\left(\frac{w\sigma_i^2}{\Delta_{i,\min}}\right),\, O(\epsilon) \right\}\right) \log T =: \mathcal{R}^{\mathrm{LS}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{LS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}})$
Proposed (GD)	Adv.	$\sqrt{\frac{4d}{1-2\eta}\left(\min\left\{L^*, mT - L^*, Q_2, \frac{2V_1}{\eta}\right\} + \frac{d}{\eta}\right) \log T}$
	Stoc.	$\frac{1}{1-2\eta}\left(\sum_{i \in J^*} \max\left\{ \frac{4w\sigma_i^2}{\Delta_{i,\min}} + o\left(\frac{w\sigma_i^2}{\Delta_{i,\min}}\right), O(\epsilon) \right\}\right) \log T =: \mathcal{R}^{\mathrm{GD}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{GD}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{GD}}})$

where $\epsilon \in (0, 1/2]$ is an input parameter for the algorithm and $c = O((\log \epsilon^{-1})^2)$. Further, for the adversarial regime, the algorithm achieves

$$\begin{aligned} \mathsf{Reg}_T \leq\ & \sqrt{4d \min\left\{L^*, mT - L^*, Q_2\right\} \log T} \\ & + O(d \log T) + d^2 + d(1 + 2\delta)\,. \end{aligned}$$

Additionally, for the stochastic regime with adversarial corruptions, we have $\mathsf{Reg}_T \leq \mathcal{R}^{\mathrm{LS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}})$.

Theorem 6.2 (Informal)**.** *For the stochastic regime, the proposed algorithm with GD estimations with a step size $\eta \in (0, 1/2)$ achieves*

$$\begin{aligned} \mathsf{Reg}_T \leq \frac{1}{1-2\eta}\left(\sum_{i \in J^*} \max\left\{4\frac{w\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1 + \frac{w\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right) \log T \\ + o(\log T) =: \mathcal{R}^{\mathrm{GD}}\,. \end{aligned}$$

For the adversarial regime, the algorithm achieves

$$\begin{aligned} \mathsf{Reg}_T \leq & \sqrt{\frac{4d}{1-2\eta}\left(\min\left\{L^*, mT - L^*, Q_2, \frac{2V_1}{\eta}\right\} + \frac{d}{\eta}\right) \log T} \\ & + O(d \log T) + d^2 + d(1 + 2\delta)\,. \end{aligned}$$

Additionally, for the stochastic regime with adversarial corruptions, we have $\mathsf{Reg}_T \leq \mathcal{R}^{\mathrm{GD}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{GD}}})$.

A comparison with existing bounds is given in Section 6.5.

The proposed algorithm is inspired by the algorithm proposed by Ito (2021a); however, their bound depends on Δ and *not* either on σ_i^2 or on $\Delta_{i,\min}$. The proposed algorithm takes care of the characteristics of the instances, and specifically, we modify the regularizer and optimistic prediction in OFTRL and refine the analysis. As a result, the bounds of the proposed algorithm depend on σ_i^2 and $\Delta_{i,\min}$, and a leading constant of our bounds are at least 81 times better than their bound. The resulting regret upper bound in Theorem 6.1 is at most approximately twice as large as the achievable lower bounds (Section 6.5). Note that one can prove the same order of upper bounds as in Theorem 6.2 for the algorithm in Ito (2021a) by using the analysis given in Section 6.5. Table 6.2 lists the regret bounds provided in this study and summarizes comparisons with existing work.

Our regret bounds are favorable compared to those reported in existing studies in that enjoying following properties:

1. Our algorithm enjoys BOBW guarantees and works well even in the stochastic regime with adversarial corruptions.

2. The leading constant of the regret bound in Theorem 6.1 (resp. Theorem 6.2) for the stochastic regime is only twice (resp. $2/(1-\eta)$) as large as an achievable lower bound.

3. The regret bounds in the stochastic regime depend on the tighter suboptimality gap $\Delta_{i,\min}$ rather than the minimal suboptimality gap Δ.

4. The regret bounds in the stochastic regime depend on the variances of base-arms, which can be tremendously small value under certain practical scenarios.

5. The regret in the adversarial regime enjoys data-dependent regret bounds.

Note that the first and fifth properties are already realized in existing studies, (*e.g.,* Zimmert et al. 2019; Ito 2021a.) We consider using a self-bounding technique (Zimmert and Seldin, 2021) to obtain BOBW guarantees. In the self-bounding technique, we first derive upper and lower bounds of the regret using a variable depending on the (base-)arm selection probability, and we then derive a regret bound by combining the upper and lower bounds. For bounding the regret with the tight suboptimality gap $\Delta_{i,\min}$ in the stochastic regime, we derive a new regret lower bound.

To prove the variance-dependent regret upper bound, we consider an algorithm inspired by the learning rate and regularizer developed by Ito et al. (2022b), which focuses on the classical multi-armed bandit problem. However, their theoretical analysis is based on the fact that the sum of the arm selection probabilities equals 1, which does not hold in the semi-bandit problem. Our analysis uses a new approach to handle this problem by deriving a regret upper bound that collaborates well with the new regret lower bound.

Further, we empirically investigate the performance of the proposed algorithm, whereas experiments are often missing in studies on the BOBW algorithm such as Wei and Luo (2018), Lee et al. (2021), and Ito (2021a). The results of this study show that the proposed algorithm empirically works the best in the adversarial regime and as well as Thompson sampling in the practical stochastic regime.

6.2 Related Work

György et al. (2007) and Uchiya et al. (2010) initiated research on the combinatorial semi-bandit problem for the adversarial regime, and since then, many algorithms with

$O(\sqrt{T})$-regret bounds have been developed (*e.g.,* Neu and Bartók 2013; Audibert et al. 2014; Neu 2015; Wei and Luo 2018).

Combinatorial semi-bandits have been also investigated in the stochastic regime, and algorithms in the literature are significantly different from those in the adversarial regime. Most are based on *index-based approaches*, where the algorithm estimates the loss means for each base-arm and *pessimistically* predicts the true value of the losses. Kveton et al. (2015) and Wang and Chen (2018) prove gap-dependent regret bounds depending on $\Delta_{i,\min}$ rather than Δ, and they also consider special action sets such as the size-invariant and matroid semi-bandits.

Since the seminal study conducted by Bubeck and Slivkins (2012), BOBW algorithms have been developed for many online-decision making problems beyond the multi-armed bandits (Zimmert and Seldin, 2021; Seldin and Lugosi, 2017; Rouyer and Seldin, 2020; Huang et al., 2022): the problem of prediction with expert advice (Gaillard et al., 2014; Luo and Schapire, 2015), dueling bandits (Saha and Gaillard, 2022), online learning with feedback graphs (Erez and Koren, 2021; Ito et al., 2022a), linear bandits (Lee et al., 2021), and episodic Markov decision processes (Jin and Luo, 2020; Jin et al., 2021). For combinatorial semi-bandits, we are aware of the works by Wei and Luo (2018), Zimmert et al. (2019), and Ito (2021a).

6.3 Preliminaries

This section introduces the preliminaries for this study. Let $\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ be the Euclidian, ℓ_1, and ℓ_∞-norms for vector x, respectively, and $\mathbf{1}$ be the all-one vector.

6.3.1 Combinatorial Semi-Bandits

We consider the combinatorial semi-bandit problem with action set $\mathcal{A} \subset \{0,1\}^d$, where each element $a \in \mathcal{A}$ is called an action. We assume that for all $i \in [d]$, there exists $a \in \mathcal{A}$ such that $a_i = 1$. Define $m = \max_{a \in \mathcal{A}} \|a\|_1$.

In the combinatorial semi-bandit problem, the learner observes entry-wise bandit feedback. At each step $t \in [T]$, when the learner takes action $a(t) \in \mathcal{A}$, they observe the elements in $I_t = \{i \in [d] \colon a_i(t) = 1\}$, whereas the elements in $J_t = [d] \setminus I_t$ are not observed. We assume that $T \geq \max\{d, 55\}$.

This study also considers the special cases of action sets: *size-invariant semi-bandits* and *matroid semi-bandits*. For size-invariant semi-bandits, the size of action $\|a\|_1$ is fixed to m, *i.e.,* $\mathcal{A} \subset \{a \in \{0,1\}^d \colon \|a\|_1 = m\}$. For the matroid semi-bandits, a special case of size-invariant semi-bandits, an action set $\mathcal{A}$ corresponds to the bases of a matroid. The well-known *m-set semi-bandits*, in which $\mathcal{A} = \{a \in \{0,1\}^d \colon \|a\|_1 = m\}$, is an example of the matroid semi-bandit problem.

In this study, we assume that there exists a unique optimal action $a^* \in \mathcal{A}$. This assumption has been employed by many studies aiming at the development of BOBW algorithms (Gaillard et al., 2014; Luo and Schapire, 2015; Wei and Luo, 2018; Zimmert and Seldin, 2021).

6.3.2 Considered Regimes

We consider three regimes as the assumptions for the losses. In the *stochastic regime*, the loss vectors $(\ell(t))$ follow an unknown distribution $\mathcal{D}$ in an i.i.d. manner for all $t \in [T]$. We define the expectation of the losses by $\mu = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell]$.

By contrast, the *adversarial regime* does not assume any stochastic structure for the losses and the losses can be chosen in an arbitrarily manner. In this regime, the environment can choose $\ell(t)$ depending on the past history until the $(t-1)$-th round, *i.e.,* $\{(\ell(s), a(s))\}_{s=1}^{t-1}$.

We also consider an intermediate regime between the stochastic and adversarial regimes. One of the most representative intermediate regimes is the *stochastic regime with adversarial corruptions*. In this regime, a temporary loss $\ell'(t) \in [0,1]^d$ is sampled from an unknown distribution $\mathcal{D}$, and then the adversary corrupts $\ell'(t)$ to $\ell(t)$. We define the corruption level by $C = \mathbb{E}\big[\sum_{t=1}^T \|\ell(t) - \ell'(t)\|_\infty\big] \geq 0$. If $C = 0$, this regime coincides with the stochastic regime, and if $C = T$, this regime corresponds to the adversarial regime. We will see that the proposed algorithm works without the knowledge of the corruption level C.

6.3.3 Optimistic Follow-the-Regularized-Leader

We establish the algorithm based on the *Optimistic follow-the-regularized-leader (OFTRL)* framework, which has occasionally been used in the development of BOBW algorithms (Wei and Luo, 2018; Ito, 2021c). Let $\mathcal{X} = \mathrm{conv}(\mathcal{A})$ be the convex hull of the action set $\mathcal{A}$. OFTRL maintains $x(t) \in \mathcal{X}$, and it then chooses $a(t) \in \mathcal{A}$ so that $\mathbb{E}[a(t)|x(t)] = x(t)$. The OFTRL update rule is expressed as

$$x(t) \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\langle m(t) + \sum_{s=1}^{t-1} \widehat{\ell}(s), x \right\rangle + \psi_t(x)\,, \tag{6.1}$$

where $m(t) \in [0,1]^d$ corresponds to an optimistic prediction (also known as a hint vector) of the true loss vector $\ell(t)$, the vector $\widehat{\ell}(t) \in \mathbb{R}^d$ is an unbiased estimator of $\ell(t)$, and ψ_t is a convex regularizer function over $\mathcal{X}$.

6.4 Proposed Algorithm

This section describes details of the proposed algorithm (Logarithmic Barrier Implicit Normalized Forecaster considering Variances for semi-bandits; `LBINFV`) by specifying the optimistic prediction $m(t)$, estimator $\widehat{\ell}(t)$, and convex regularization ψ_t in (6.1).

We consider two different methods for estimating optimistic predictions; these methods result in regret upper bounds that differ by a constant factor in the stochastic regime and have different data-dependent bounds. One method is a *least square* (LS) estimation based on the losses thus far, *i.e.,* we define $m(t) = (m_1(t), \ldots, m_d(t))^\top \in [0,1]^d$ by

$$m_i(t) = \frac{1}{1 + N_i(t-1)} \left(\frac{1}{2} + \sum_{s=1}^{t-1} a_i(s)\, \ell_i(s) \right)\,, \tag{6.2}$$

where $N_i(t)$ is the number of times the base-arm i is chosen until the t-th round, *i.e.,* $N_i(t) = |\{s \in [t] : a_i(t) = 1\}|$. The other method is based on the *gradient descent* (GD), where we define $m(t)$ by $m_i(1) = 1/2$ and

$$m_i(t+1) = \begin{cases} (1-\eta) m_i(t) + \eta \ell_i(t) & \text{if } i \in I(t) \\ m_i(t) & \text{otherwise} \end{cases} \tag{6.3}$$

for $i \in [d]$ with a step size $\eta \in (0, 1/2)$.

Let $a(t) \in \mathcal{A}$ be an action selected at round t and $I(t) = \{i \in [d] : a_i(t) = 1\}$ be the set of base-arms selected at round t. Note that $\{a_i(t) = 1\}$ is equivalent to $\{i \in I(t)\}$ and $\Pr[i \in I(t) | x_i(t)] = \Pr[a_i(t) = 1 | x_i(t)] = x_i(t)$.

The design of LS is to reduce the leading constant in the regret, and GD is to derive a path-length bound. LS was developed by Ito et al. (2022b). The original idea of GD comes from online learning literature (Herbster and Warmuth, 2001), and Ito (2021a) developed the idea in semi-bandits.

We use an unbiased estimator $\widehat{\ell}(t) = (\widehat{\ell}_1(t), \ldots, \widehat{\ell}_d(t))^\top \in \mathbb{R}^d$ of $\ell(t)$ given by

$$\widehat{\ell}_i(t) = m_i(t) + \frac{a_i(t)}{x_i(t)}(\ell_i(t) - m_i(t)) \tag{6.4}$$

for $i \in [d]$. This is indeed an unbiased estimator of $\ell(t)$ since $\mathbb{E}[\widehat{\ell}_i(t)|x(t)] = m_i(t) + \frac{x_i(t)}{x_i(t)}(\ell_i(t) - m_i(t)) = \ell_i(t)$. The optimistic prediction $m(t)$ in (6.4) plays a role in reducing the variance of $\widehat{\ell}(t)$; the better $m(t)$ predicts $\ell(t)$, the smaller the variance in $\widehat{\ell}(t)$ becomes.

The regularizer function $\psi_t : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\psi_t(x) = \sum_{i=1}^{d} \beta_i(t)\phi(x_i)\,, \tag{6.5}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\phi(z) = z - 1 - \log z + \gamma\,(z + (1-z)\log(1-z)) \tag{6.6}$$

with $\gamma = \log T$ and regularization parameters $\beta_i(t) \geq 0$. Our regularizer in (6.5) comprises the logarithmic barrier $-\log x_i$ and the (negative) Shannon entropy $(1 - x_i)\log(1 - x_i)$ for the complement of $x_i \in [0, 1]$. Such a regularizer is called a *hybrid* regularizer, and this type of regularizer was employed in existing studies for bounding a component of the regret (Zimmert et al., 2019; Ito et al., 2022b,a). The affine part of the regularizer in (6.6), $z - 1 + \gamma z$, is introduced to simplify the analysis and yields smaller constant factors, which is also used by Ito et al. (2022b).

Regularization parameters $\beta_i(t)$ are defined as

$$\beta_i(t) = \sqrt{(1+\epsilon)^2 + \frac{1}{\gamma}\sum_{s=1}^{t-1}\alpha_i(s)}\,, \tag{6.7}$$

where $\epsilon \in (0, 1/2]$ is an input parameter and

$$\alpha_i(t) = a_i(t)(\ell_i(t) - m_i(t))^2 \min\left\{1, \frac{2(1 - x_i(t))}{x_i(t)^2\gamma}\right\}\,. \tag{6.8}$$

We design $\alpha_i(t)$ in (6.8) so that it corresponds to an upper bound of the component of regret, which appears when we use a standard analysis of (O)FTRL with regularizer (6.5). We can introduce a $2(1 - x_i(t))/(x_i(t)^2\gamma)$ part in $\alpha_i(t)$ thanks to the Shannon entropy part in regularizer (6.5). This part allows us to bound the regret corresponding to optimal base-arms. The $(\ell_i(t) - m_i(t))^2$ part of $\alpha_i(t)$ comes from the use of optimistic predictions and can be related to the base-arm variances by using the LS and GD methods to estimate $m(t)$. Algorithm 6.1 summarizes the proposed algorithms.

From the intuitive viewpoint, $\alpha_i(t)$ determines the strength of the regularization, and as $\alpha_i(t)$ increases, the algorithm further explores base-arm i. Since $(\ell_i(t) - m_i(t))^2$ in (6.8) represents the squared error of the optimistic prediction, the algorithm becomes more explorative when the loss is unpredictable or has a high variance. Also note that $\mu_i \simeq 1$ corresponds to the base-arm with the almost worst expected loss with the least variance. The factor $(1 - x_i(t))$ in (6.8) contributes to a fast elimination of such a base-arm since the regularization does not become strong when $x_i(t) = 1$ is observed.

Algorithm 6.1: LBINFV for semi-bandits

```
input: action set $\mathcal{A}$, time horizon $T$
for $t = 1, 2, \ldots, T$ do
    Compute $x(t) \in \mathcal{X}$ by (6.1) with $\widehat{\ell}(t)$ in (6.4) and $\psi_t$ in (6.5).
    Sample $a(t)$ such that $\mathbb{E}[a(t)|x(t)] = x(t)$.
    Take action $a(t)$ and observe feedback $\ell_i(t)$ for $i$ such that $a_i(t) = 1$.
    Update the regularization parameters $\beta_i(t)$ in (6.7) and optimistic prediction
      $m_i(t)$ using (6.2) or (6.3).
```

6.5 Regret Analysis

This section derives the regret upper bounds of the proposed algorithm. We define the minimum suboptimality gaps that contain and do not contain base-arm i by

$$\begin{aligned}\Delta_{i,\min} &= \min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 1\}\,; \\ \Delta'_{i,\min} &= \min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 0\}\,.\end{aligned}$$

We define constants $v(\mathcal{A})$ and $w(\mathcal{A})$ depending on the action set $\mathcal{A}$ by

$$v(\mathcal{A}) = \begin{cases} 2 & \mathcal{A} \text{ is a matroid} \\ 2\min\{|I^*|, d - m\} & \text{otherwise} \end{cases}$$

and

$$w(\mathcal{A}) = \begin{cases} 2 & \mathcal{A} \text{ is a matroid} \\ 2\min\{m, d - m\} & \mathcal{A} \text{ is size-invariant} \\ 2\min\{m, |J^*|\} & \text{otherwise}\,. \end{cases}$$

6.5.1 Regret Upper Bounds

This section introduces regret upper bounds of the proposed algorithm for each optimistic prediction method.

Theorem 6.3 (Formal version of Theorem 6.1)**.** *Consider Algorithm 6.1 using the least square method in* (6.2) *for optimistic predictions. Then, for the stochastic regime,*

$$\begin{aligned}\mathsf{Reg}_T \le{}& \left(\sum_{i \in J^*} \max\Big\{ 4\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}} + c \log\Big(1 + \frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}\Big),\, 2(1+\epsilon) \Big\} + 2(1+\epsilon)|I^*| \right) \log T \\ &+ O\left(\sum_{i \in I^*} \frac{v(\mathcal{A})}{\Delta'_{i,\min}} \sqrt{\log T} \right) + o(\sqrt{\log T})\,, \qquad (6.9)\end{aligned}$$

where $\epsilon \in (0, 1/2]$ is an input parameter for the algorithm and $c = O((\log \epsilon^{-1})^2)$. Further, for the adversarial regime,

$$\begin{aligned}\mathsf{Reg}_T \le{}& \sqrt{4d \min\{L^*, mT - L^*, Q_\infty\} \log T} \\ &+ O(d \log T) + d^2 + d(1 + 2\delta)\,. \qquad (6.10)\end{aligned}$$

Additionally, in the stochastic regime with adversarial corruptions, we have $\mathsf{Reg}_T \le \mathcal{R}^{\mathrm{LS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}})$, where $\mathcal{R}^{\mathrm{LS}}$ is the RHS of (6.9)*.*

Theorem 6.4 (Formal version of Theorem 6.2)**.** *Consider Algorithm 6.1 using the gradient descent method with a step size $\eta \in (0, 1/2)$ in* (6.3) *for optimistic predictions. Then, for the stochastic regime,*

$$\mathsf{Reg}_T \leq \frac{1}{1-2\eta}\left(\sum_{i \in J^*} \max\left\{4\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1+\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T$$
$$+ O\left(\left(\sum_{i \in I^*} \frac{v(\mathcal{A})}{\Delta'_{i,\min}} + \frac{d}{\sqrt{\eta(1-2\eta)}}\right)\sqrt{\log T}\right) + o(\sqrt{\log T})\,. \tag{6.11}$$

Further, for the adversarial regime,

$$\mathsf{Reg}_T \leq \sqrt{\frac{4d}{1-2\eta}\left(\min\left\{L^*, mT - L^*, Q_2, \frac{2V_1}{\eta}\right\} + \frac{d}{\eta}\right)\log T}$$
$$+ O(d\log T) + d^2 + d(1+2\delta)\,. \tag{6.12}$$

Additionally, in the stochastic regime with adversarial corruptions, we have $\mathsf{Reg}_T \leq \mathcal{R}^{\mathrm{GD}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{GD}}})$*, where* $\mathcal{R}^{\mathrm{GD}}$ *is the RHS of* (6.11).

Note that the proposed algorithm does not require any prior knowledge on σ_i^2, Δ_i, L^*, Q_∞, and C. Theorem 6.4 indicates that the leading constant worsens by a factor of $1/(1-2\eta)$ in the stochastic regime compared to the bound in Theorem 6.3. This is at the expense of the path-length bound depending on V_1 in the adversarial regime.

6.5.2 Comparison with Existing Regret Bounds

The regret upper bounds for the stochastic regime in Theorems 6.3 and 6.4 improve on the existing regret upper bounds in three aspects: (i) dependence on the tight suboptimality gap $\Delta_{i,\min}$, (ii) the dependence on the variance of base-arms σ_i^2, and (iii) the leading constants particularly in the stochastic regime. For the suboptimality gap, our upper bounds are of the same order as the regret upper bound by Kveton et al. (2015), which is an algorithm specialized for the stochastic regime, and our bounds are up to d times better than the regret upper bounds by Zimmert et al. (2019) and Ito (2021a). For the variance dependency, in the stochastic regime, bounds in Theorems 6.3 and 6.4 improve the results in Ito (2021a) by replacing a constant in their bound with variance σ_i^2, which can be considerably small under certain practical scenarios such as ad allocations. Finally, it is worth noting that the leading constants are also significantly improved. The leading constant of our bounds are at least 81 times better than the bound by Ito (2021a). Moreover, the resulting regret upper bound in Theorem 6.3 and (resp. Theorem 6.4) are approximately at most twice (resp. $2/(1-2\eta)$) as large as the achievable lower bounds, which can be confirmed by comparing the bounds with Ito et al. (2022b, Proposition 1).

6.5.3 Key Technique and Analysis

To obtain the regret bound depending on $\Delta_{i,\min}$ in the stochastic regime and the stochastic regime with adversarial corruptions, we prove the following regret *lower* bound.

Lemma 6.1. *In the stochastic regime with adversarial corruptions, for any algorithm and any action set $\mathcal{A}$, the regret is bounded from below as*

$$\mathsf{Reg}_T \geq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{v(\mathcal{A})}\sum_{i \in I^*}\Delta'_{i,\min}(1-a_i(t)) + \frac{1}{w(\mathcal{A})}\sum_{i \in J^*}\Delta_{i,\min}a_i(t)\right)\right] - 2Cm\,.$$

Note that if $a^* \in \mathcal{A}$ is unique, $i \in I^*$ implies that $\Delta'_{i,\min} > 0$, and $i \in J^*$ implies that $\Delta_{i,\min} > 0$. This regret lower bound improves Ito (2021a, Lemma 4) for general action sets.

To prove the variance-dependent regret bounds, we make use of the learning rate inspired by Ito et al. (2022b), in which the classical multi-armed bandit problem is considered. However, their theoretical analysis is based on the fact that the sum of the arm selection probabilities equals 1, which does not hold in the semi-bandits. To handle this problem, we introduce a technique developed in Ito (2021a) and sophisticate the analysis to derive a regret upper bound that collaborates well with the regret lower bound.

In the following, we provide a sketch of analysis commonly used to prove Theorems 6.3 and 6.4, and see that that the regret lower bound in Lemma 6.1 indeed helps us obtain the desired regret bound. In the subsequent analysis, we will mainly focus on terms that are dominant for sufficiently large T, and will not include the other terms. Let $\gamma = \log T$. Using the similar analysis given by Ito et al. (2022b), we first show in Lemma 6.3 that the regret of the proposed algorithm is roughly bounded as

$$\mathsf{Reg}_T = O\left(\gamma \sum_{i=1}^{d} \mathbb{E}\left[\beta_i(T+1)\right]\right) = O\left(\sum_{i=1}^{d} \sqrt{\mathbb{E}\left[\gamma \sum_{t=1}^{T} \alpha_i(t)\right]}\right) .$$

Define (P_i) and (Q_i) by

$$P_i = \mathbb{E}\left[\sum_{t=1}^{T} x_i(t)\right] , \quad Q_i = \mathbb{E}\left[\sum_{t=1}^{T} (1 - x_i(t))\right] ,$$

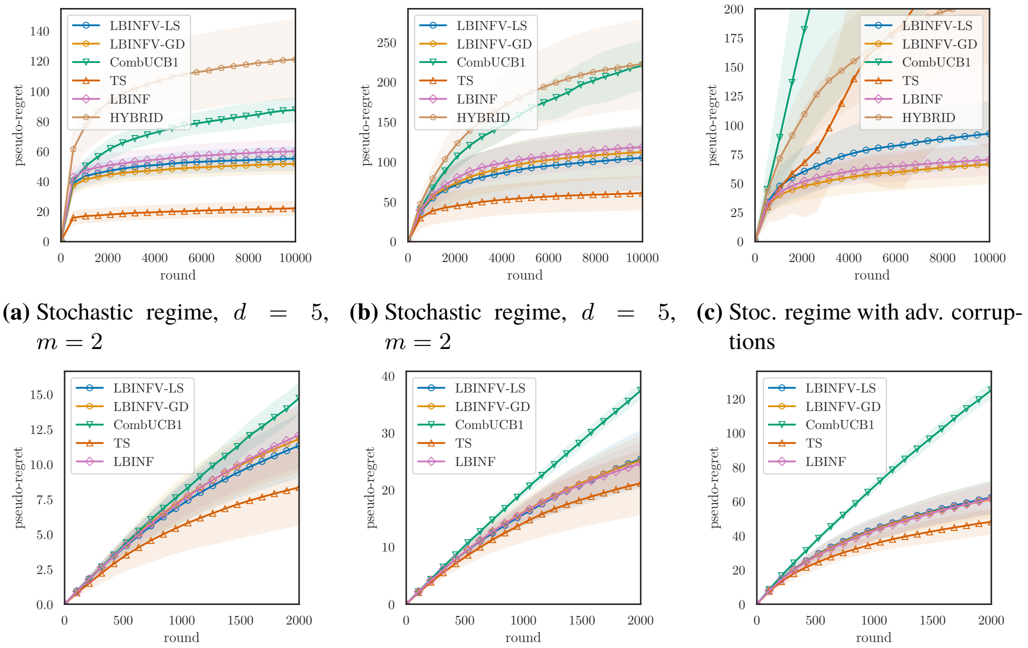
which will be used in the self-bounding argument in the following. Using this and combining the analysis given by Ito et al. (2022b) and Ito (2021a), we can show that the regret is further bounded as

$$\frac{\mathsf{Reg}_T}{\gamma} = O\left(\sum_{i \in J^*} \sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} + \sum_{i^* \in I^*} \sqrt{\frac{Q_i}{\gamma^{3/2}}}\right) . \tag{6.13}$$

For the stochastic regime, using the upper bound (6.13) and lower bound (Lemma 6.1 with $C = 0$), the regret can be further roughly bounded as

$$\begin{aligned}
\frac{\mathsf{Reg}_T}{\gamma} &= 2\frac{\mathsf{Reg}_T}{\gamma} - \frac{\mathsf{Reg}_T}{\gamma} \\
&\leq O\left(\sum_{i \in J^*} \sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} + \sum_{i \in I^*} \sqrt{\frac{Q_i}{\gamma^{3/2}}}\right) - \frac{1}{\gamma}\left(\frac{1}{v(\mathcal{A})} \sum_{i \in I^*} \Delta'_{i,\min} Q_i + \frac{1}{w(\mathcal{A})} \sum_{i \in J^*} \Delta_{i,\min} P_i\right) \\
&= O\Bigg(\sum_{i \in J^*}\left(\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \frac{\Delta_{i,\min}}{w(\mathcal{A})}\frac{P_i}{\gamma}\right) \sum_{i \in I^*}\left(\sqrt{\frac{Q_i}{\gamma^{3/2}}} - \frac{\Delta'_{i,\min}}{v(\mathcal{A})}\frac{Q_i}{\gamma}\right)\Bigg) \\
&\leq O\Bigg(\sum_{i \in J^*} \frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}} + |I^*| \frac{1}{\sqrt{\gamma}} \frac{v(\mathcal{A})}{\Delta'_{i,\min}}\Bigg) ,
\end{aligned}$$

where the first inequality follows by (6.13) and Lemma 6.1 with $C = 0$, and in the last inequality we considered the worst case in terms of $(P_i)_{i \in J^*}$ and $(Q_i)_{i \in I^*}$. This result corresponds to the desired bounds in Theorems 6.3 and 6.4. A more complete and detailed analysis are deferred to the following sections.



(a) Stochastic regime, $d = 5$, $m = 2$

(b) Stochastic regime, $d = 5$, $m = 2$

(c) Stoc. regime with adv. corruptions

(d) Stochastic regime, $d = 6$, $m = 3$

(e) Stochastic regime, $d = 8$, $m = 3$

(f) Stochastic regime, $d = 10$, $m = 3$

Figure 6.1: Regret-round plots of algorithms used for synthetic and semi-synthetic data. The solid lines indicate the average over 20 independent trials. The thin fillings represent the standard error.

6.6 Experiments

This section presents the results of the numerical investigation of the empirical performance of the proposed `LBINFV` algorithm with $\epsilon = 0.2$. The proposed algorithm with LS and GD (with $\eta = 1/4$) estimations for the optimistic predictions are denoted by `LBINFV-LS` and `LBINFV-GD`, respectively. We use the following baselines. The algorithms for the stochastic regime are `CombUCB1` (Kveton et al., 2015) and `Thompson sampling (TS)` (Komiyama et al., 2015b; Wang and Chen, 2018). The algorithms with BOBW guarantees are `HYBRID` (Zimmert et al., 2019) and `LBINF` (Ito, 2021a).

To compare the performance, we consider the m-set semi-bandits with $T = 10^4$. In the m-set semi-bandit setting, it is known that we can sample $a(t)$ satisfying $\mathbb{E}[a(t)|x(t)] = x(t)$ at an $O(d \log d)$ computational cost (Zimmert et al., 2019, Appendix B.2), and we employ this sampling technique. We repeat the simulations 20 times.

6.6.1 Setup

Synthetic data In the synthetic data experiments, we set $d = 5$ and $m = 2$ and consider the stochastic regime and stochastic regime with adversarial corruptions. In the stochastic regime, we consider two instances, where each base-arm is associated with a Bernoulli distribution. We set expectations μ for each instance to $(0.5, 0.5, 0.9, 0.9, 0.9)$ and $(0.5, 0.5, 0.6, 0.6, 0.6)$, respectively. In the stochastic regime with adversarial corruptions, we consider an instance considered by Zimmert et al. (2019). The environment alternates between two stochastic settings, (i) and (ii), where the losses are sampled from Bernoulli distributions with the following time-varying loss means. In setting (i), the expected losses are 0 for the optimal base-arms $i \in I^*$, and Δ' for the suboptimal base-arms $i \in J^*$. In setting (ii), the expected losses are $1 - \Delta'$ for the optimal base-arms, and 1 for the suboptimal arms. We set $\Delta' = 0.1$. The number of rounds between alternations

Table 6.3: Reward means for the semi-synthetic data.

Instance	d	m	Reward means $\mathbf{1} - \mu$
(d)	6	3	$(0.0315, 0.0208, 0.0193, 0.0182, 0.0179, 0.0177)$
(e)	8	3	$(0.0370, 0.0275, 0.0266, 0.0266, 0.0231, 0.0192, 0.0143, 0.0107)$
(f)	10	3	$(0.0774, 0.0709, 0.0669, 0.0631, 0.0430, 0.0393, 0.0296, 0.0217, 0.00797, 0.00219)$

increases exponentially with a factor of 1.6 after each alternation. Note that this instance also belongs to the stochastically constrained adversarial regime (Wei and Luo, 2018; Zimmert and Seldin, 2021).

Semi-synthetic data In semi-synthetic data experiments, we consider the stochastic regime. We used the KDD Cup 2012 track 2 dataset (Tencent Inc., 2012), which was used in the studies on multiple-play bandit problem (Komiyama et al., 2015b; Lagrée et al., 2016; Komiyama et al., 2017), which is equivalent to the m-set semi-bandit problem. The dataset includes session logs of the Tencent search engine, soso.com. We use the estimated *reward* means of Komiyama et al. (2017) although the rewards therein are estimated under a different context, where the reward mean for base-arm i is defined by $1 - \mu_i$ corresponding to the click-through rate for example. Table 6.3 lists the reward means used in the experiments. One characteristic of this type of dataset is that the reward mean for each base-arm is extremely small (smaller than 0.05 in most cases). Hence, each σ_i^2 is supposed to be extremely small, and algorithms with adaptivity to variances are desirable.

6.6.2 Numerical Results

Figure 6.1 shows an empirical comparison of the proposed algorithm against the baselines. The experimental results from the synthetic data in (a) and (b) indicate that the proposed `LBINFV-LS` and `LBINFV-GD` algorithms achieve the best performance in the stochastic regime, except for `Thompson sampling`. Further, under the setting in (a), where the variances of the base-arms are small, the proposed algorithm shows a significant improvement compared to `HYBRID`. Additionally, these figures also confirm that `LBINFV-LS` performs better in the stochastic regime than `LBINF`. This indicates that the modification of the regularizer and the optimistic prediction contribute not only to the better leading constant of the regret upper bound but also to the empirical performance.

The proposed algorithm achieves the best performance in the adversarial regime, whereas `CombUCB1` and `Thompson sampling` highly degrade their performance. We can also see from (a) and (b) that the performance of `LBINFV-GD` becomes slightly worse than that of `LBINF-LS` in most cases, as suggested by the theoretical results, whereas in (c) the performance of `LBINFV-GD` is better than that of `LBINFV-LS`, which seemingly occurs because the adversarial instance in this experiment is a regime with a small path-length and the former algorithm has the path-length bound.

The experimental results using the semi-synthetic data in (d)–(f) indicate that `LBINFV-LS` and `LBINFV-GD` perform comparably well to `Thompson sampling`. These results can be attributed to the fact that the variance is small for semi-synthetic data. Furthermore, (d)–(f), where the variances of the base-arms are extremely small, indicates that `CombUCB1` performs significantly worse than the other variance-aware algorithms. This observation indicates the importance of variance-aware algorithms in practical applications.

6.7 Deferred Proofs

6.7.1 Common Analysis

In this section, we provide preliminary and common analysis used in the subsequent sections.

6.7.1.1 General Regret Upper Bound

Define $\beta_0 = 1 + \epsilon$. Let D_t be the Bregman divergence induced by ψ_t, *i.e.,*

$$D_t(y, x) = \psi_t(y) - \psi_t(x) - \langle \nabla \psi_t(x), y - x \rangle \ .$$

Then, the regret for OFTRL is bounded as follows.

Lemma 6.2. *If $x(t)$ is given by the OFTRL update* (6.1)*, for any $x^* \in \mathcal{X} \cap \mathbb{R}^d_+$ we have*

$$\sum_{t=1}^{T} \left\langle \widehat{\ell}(t), x(t) - x^* \right\rangle \le \underbrace{\psi_{T+1}(x^*) - \psi_1(y(1)) + \sum_{t=1}^{T} (\psi_t(x(t+1)) - \psi_{t+1}(x(t+1)))}_{\text{penalty term}} + \underbrace{\sum_{t=1}^{T} \left(\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \right)}_{\text{stability term}} , \tag{6.14}$$

where we define $y(t) \in \arg\min_{x \in \mathcal{X}} \left\{ \left\langle \sum_{s=1}^{t-1} \widehat{\ell}(s), x \right\rangle + \psi_t(x) \right\}$.

This lemma is standard in the literature and can be found *e.g.,* in Orabona (2019, Chapter 7) and Ito et al. (2022b, Lemma 2). In the RHS of the above inequality (6.14), we refer to the sum of the first three terms as the *penalty term* and the remaining term as the *stability term.*

First, we prove the following lemma.

Lemma 6.3. *The regret of the proposed algorithm is bounded as*

$$\mathsf{Reg}_T \le \gamma \sum_{i=1}^{d} \mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)} \right] + d^2 + d(1 + 2\delta) , \tag{6.15}$$

where $\delta > 0$ is defined by

$$\delta = (1+\epsilon)^3 \log \frac{1+\epsilon}{\epsilon} - (1+\epsilon)^2 - \frac{1+\epsilon}{2} \le \frac{27}{8} \log \frac{3}{2\epsilon} - \frac{3}{2} = O\left(\log \frac{1}{\epsilon} \right) .$$

Proof. Using $x_0 \in \mathcal{X}$ such that $(x_0)_i \ge 1/d$ for all $i \in [d]$, let

$$x^* = \left(1 - \frac{d}{T} \right) a^* + \frac{d}{T} x_0 \, .$$

Using this and the equality $\mathbb{E}[\widehat{\ell} | x_t] = \ell$, we have

$$\begin{aligned} \mathsf{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell(t), x(t) - a^* \rangle \right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell(t), x(t) - x^* \rangle + \sum_{t=1}^{T} \langle \ell(t), x^* - a^* \rangle \right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}(t), x(t) - x^* \right\rangle + \frac{d}{T} \sum_{t=1}^{T} \langle \ell(t), x_0 - a^* \rangle \right] \le \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}(t), x(t) - x^* \right\rangle \right] + d^2 , \end{aligned} \tag{6.16}$$

where in the last inequality we used $\sum_{t=1}^{T} \langle \ell(t), x_0 - a^* \rangle \leq T\|x_0 - a^*\|_1 \leq Td$.

The first term in (6.16) is bounded by (6.14) in Lemma 6.2, the components of which we will bound in the following. We first consider the penalty term. The remaining part of the proof follows a similar argument as that in Ito et al. (2022b), and we include the argument for completeness.

Bounding the penalty term in (6.14) Using the definition of the regularizer $\psi_t(x) = \sum_{i=1}^{d} \beta_i(t)\phi(p_i)$ defined in (6.5), we have

$$\psi_t(x^*) = \sum_{i=1}^{d} \beta_i(t)\phi(x_i^*) \leq \sum_{i=1}^{d} \beta_i(t) \max_{x \in [1/T, 1]} \phi(x) \leq \sum_{i=1}^{d} \beta_i(t) \max\{\phi(1/T), \phi(1)\}\,, \tag{6.17}$$

where the first inequality follows since the definition of x^* implies $x_i^* \geq \frac{d}{T}(x_0)_i \geq 1/T$ for $i \in [d]$ and the second inequality holds since ϕ is a convex function. Further, from the definition of ϕ in (6.6), we have

$$\begin{aligned}\max\{\phi(1/T), \phi(1)\} &= \max\left\{\frac{1}{T} - 1 + \log T + \gamma\left(\frac{1}{T} + \left(1 - \frac{1}{T}\right)\log\left(1 - \frac{1}{T}\right)\right), \gamma\right\} \\ &\leq \max\left\{\frac{1+\gamma}{T} - 1 + \log T, \gamma\right\} = \gamma\,,\end{aligned}$$

where the last inequality follows from $\gamma = \log T$. From this and (6.17), we have

$$\psi_{T+1}(x^*) \leq \gamma \sum_{i=1}^{d} \beta_i(T+1)\,. \tag{6.18}$$

Further, as we have $\beta_i(t) \leq \beta_i(t+1)$ from (6.7) and $\phi(x) \geq 0$ for any $x \in (0, 1]$, we have

$$\begin{aligned}&-\psi_1(y(1)) + \sum_{t=1}^{T}(\psi_t(y(t+1)) - \psi_{t+1}(y(t+1))) \\ &= -\sum_{i=1}^{d}\left(\beta_i(1)\phi(y_i(1)) + \sum_{t=1}^{T}(\beta_i(t+1) - \beta_i(t))\phi(y_i(t+1))\right) \leq 0\,. \end{aligned}\tag{6.19}$$

Combining (6.18) and (6.19), we can bound the penalty term in (6.14) as

$$\begin{aligned}&\psi_{T+1}(x^*) - \psi_1(y(1)) + \sum_{t=1}^{T}(\psi_t(y(t+1)) - \psi_{t+1}(y(t+1))) \\ &\leq \gamma \sum_{i=1}^{d} \beta_i(T+1)\,.\end{aligned}\tag{6.20}$$

Bounding the stability term in (6.14) The Bregman divergence $D_t(x, y)$ is expressed as

$$\begin{aligned}D_t(x, y) &= \sum_{i=1}^{d}\left(\beta_i(t)D^{(1)}(x_i, y_i) + \beta_i(t)\gamma D^{(2)}(x_i, y_i)\right) \\ &\geq \sum_{i=1}^{d} \max\left\{\beta_i(t)D^{(1)}(x_i, y_i), \beta_i(t)\gamma D^{(2)}(x_i, y_i)\right\}\,,\end{aligned}$$

where $D^{(1)}$ and $D^{(2)}$ are Bregman divergences induced by $\phi^{(1)}(x) = -\log x$ and $\phi^{(2)}(x) = (1-x)\log(1-x)$, respectively. We hence have

$$
\begin{aligned}
&\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \\
&\leq \sum_{i=1}^{d} \Bigg((\widehat{\ell}_i(t) - m_i(t))(x_i(t) - y_i(t+1)) \\
&\qquad - \beta_i(t) \max\left\{ D^{(1)}(y_i(t+1), x_i(t)), \gamma D^{(2)}(y_i(t+1), x_i(t)) \right\} \Bigg) \\
&\leq \sum_{i=1}^{d} \left(\min\left\{ \beta_i(t) g\left(\frac{p_i(t)(\widehat{\ell}_i(t) - m_i(t))}{\beta_i(t)} \right), \beta_i(t)\gamma(1 - x_i(t)) h\left(\frac{\widehat{\ell}_i(t) - m_i(t)}{\gamma\beta_i(t)} \right) \right\} \right),
\end{aligned}
\tag{6.21}
$$

where the last inequality follows from the standard technique to boudn the staiblity term (see *e.g.,* Ito et al. 2022b, Lemma 5), and g and h are defined as

$$
g(x) = x - \log(x+1) \leq \frac{1}{2}x^2 + \delta|x|^3 \quad \left(x \geq -\frac{1}{\beta_0} \right), \tag{6.22}
$$

$$
h(x) = \exp(x) - x - 1 \leq x^2 \quad (x \leq 1)\,. \tag{6.23}
$$

Note that $g(0) = h(0) = 0$ and it holds from (6.4) that

$$
\widehat{\ell}_j(t) - m_j(t) = \begin{cases} (\ell_j(t) - m_j(t))/x_j(t) & \text{if } j \in I(t) \\ 0 & \text{otherwise}\,. \end{cases} \tag{6.24}
$$

Therefore, the LHS of (6.21) is further bounded as

$$
\begin{aligned}
&\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \\
&\leq \sum_{j \in I(t)} \min\left\{ \beta_j(t)\, g\left(\frac{\ell_j(t) - m_j(t)}{\beta_j(t)} \right), \beta_j(t)\gamma(1 - x_j(t))\, h\left(\frac{\ell_j(t) - m_j(t)}{\gamma\beta_j(t)x_j(t)} \right) \right\} \\
&\leq \begin{cases} \sum_{j \in I(t)} \left(\frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2} \right) & \text{if } \gamma x_j(t) \leq 1 \\ \sum_{j \in I(t)} \min\left\{ \frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2}, \frac{(1 - x_j(t))(\ell_j(t) - m_j(t))^2}{\gamma x_j(t)^2\beta_j(t)} \right\} & \text{otherwise} \end{cases} \\
&\leq \sum_{j \in I(t)} \min\left\{ \frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2}, \frac{(1 - x_j(t))(\ell_j(t) - m_j(t))^2}{\gamma x_j(t)^2\beta_j(t)} \right\} \\
&\leq \sum_{j \in I(t)} \left(\frac{1}{2\beta_j(t)} + \frac{\delta}{\beta_j(t)^2} \right) (\ell_j(t) - m_j(t))^2 \min\left\{ 1, \frac{2(1 - x_j(t))}{\gamma x_j(t)^2} \right\} \\
&= \sum_{i=1}^{d} \left(\frac{1}{2\beta_i(t)} + \frac{\delta}{\beta_i(t)^2} \right) \alpha_i(t)\,,
\end{aligned}
\tag{6.25}
$$

where the first inequality follows from (6.21) and (6.24), the second inequality follows from (6.22), (6.23), and the fact that $|\frac{\ell_j(t) - m_j(t)}{\beta_j(t)}| \leq \frac{1}{\beta_0} \leq 1$, and the third inequality holds since $\gamma x_j(t) \leq 1$ means $\frac{1 - x_j(t)}{\gamma x_j(t)^2} \geq \frac{1 - 1/\gamma}{\gamma(1/\gamma)^2} = \gamma - 1 \geq \frac{1}{2} + \delta$, which implies

$$
\frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2} \leq \frac{(1 - x_j(t))(\ell_j(t) - m_j(t))^2}{\gamma x_j(t)^2\beta_j(t)}\,.
$$

We hence have

$$
\sum_{t=1}^{T} \left(\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \right) \leq \sum_{i=1}^{d} \sum_{t=1}^{T} \left(\frac{1}{2\beta_i(t)} + \frac{\delta}{\beta_i(t)^2} \right) \alpha_i(t)\,. \tag{6.26}
$$

We can show that a part of (6.26) is bounded as

$$\sum_{t=1}^{T} \frac{\alpha_i(t)}{2\beta_i(t)} \le \gamma \left(\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{t=1}^{T} \alpha_i(t)} - \sqrt{\beta_0^2 - \frac{1}{\gamma}} \right) \le \gamma \left(\beta_i(T+1) - \beta_0\right) \tag{6.27}$$

The first inequality in (6.27) holds since

$$\begin{aligned}
&\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t} \alpha_i(s)} - \sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)} \\
&= \frac{\alpha_i(t)}{\gamma \left(\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t} \alpha_i(s)} + \sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)} \right)} \\
&\ge \frac{\alpha_i(t)}{2\gamma \sqrt{\beta_0^2 + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)}} = \frac{\alpha_i(t)}{2\gamma\beta_i(t)} \,,
\end{aligned}$$

where the inequality follows by $\alpha_i(t) \le 1$. The second inequality in (6.27) follows from

$$\begin{aligned}
\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{t=1}^{T} \alpha_i(t)} - \sqrt{\beta_0^2 - \frac{1}{\gamma}} &\le \sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{t=1}^{T} \alpha_i(t)} - \beta_0 + \frac{1}{\gamma} \\
&\le \beta_i(T+1) - \beta_0 + \frac{1}{\gamma} \,,
\end{aligned}$$

where the first inequality follows from $\sqrt{x} - \sqrt{x-y} \le y/\sqrt{x}$ for $x \ge y \ge 0$ and $\beta_0 \ge 1$. Similarly, we can show

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\alpha_i(t)}{\beta_i(t)^2} &= \sum_{t=1}^{T} \frac{\alpha_i(t)}{\beta_0^2 + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)} = \gamma \sum_{t=1}^{T} \frac{\alpha_i(t)}{\gamma\beta_0^2 + \sum_{s=1}^{t-1} \alpha_i(s)} \\
&\le \gamma \log \left(1 + \frac{1}{\gamma\beta_0^2 - 1} \sum_{t=1}^{T} \alpha_i(t) \right) \le 2\gamma \log \frac{\beta_i(T+1)}{\beta_i(1)} + 2 \,. \qquad (6.28)
\end{aligned}$$

The first inequality in (6.28) follows since

$$\begin{aligned}
&\log \left(1 + \frac{1}{\gamma\beta_0^2 - 1} \sum_{s=1}^{t} \alpha_i(s) \right) - \log \left(1 + \frac{1}{\gamma\beta_0^2 - 1} \sum_{s=1}^{t-1} \alpha_i(s) \right) \\
&= -\log \left(1 - \frac{\alpha_i(t)}{\gamma\beta_0^2 - 1 + \sum_{s=1}^{t} \alpha_i(s)} \right) \ge -\log \left(1 - \frac{\alpha_i(t)}{\gamma\beta_0^2 + \sum_{s=1}^{t-1} \alpha_i(s)} \right) \\
&\ge \frac{\alpha_i(t)}{\gamma\beta_0^2 + \sum_{s=1}^{t-1} \alpha_i(s)} \,,
\end{aligned}$$

where the first inequality follows from $\alpha_i(t) \le 1$ and the last inequality follows from $-\log(1-x) \ge x$ for $x < 1$. The second inequality in (6.28) follows from

$$\begin{aligned}
&\log \left(1 + \frac{1}{\gamma{\beta_0}^2 - 1} \sum_{t=1}^{T} \alpha_i(t) \right) < \log \left(1 + \frac{1}{\gamma{\beta_0}^2} \sum_{t=1}^{T} \alpha_i(t) \right) + \log \frac{\gamma{\beta_0}^2}{\gamma\beta_0^2 - 1} \\
&= \log \left(\frac{\beta_i(T+1)^2}{\beta_0^2} \right) + \log \left(1 + \frac{1}{\gamma\beta_0^2 - 1} \right) \le 2 \log \frac{\beta_i(T+1)}{\beta_0} + \frac{2}{\gamma} \,,
\end{aligned}$$

where the last inequality follows from $\log(1+1/(x-1)) \geq 2/x$ for $x \geq 3/2$. Bounding the RHS of (6.25) with (6.27) and (6.28) yields

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \right) \\
&\leq \gamma \sum_{i=1}^{d} \left(\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)} \right) + d(1+2\delta)\,. \qquad (6.29)
\end{aligned}$$

Finally, by bounding the RHS of (6.14) by sequentially using (6.16), (6.20) and (6.29), we have

$$\mathsf{Reg}_T \leq \gamma \sum_{i=1}^{d} \mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)} \right] + d^2 + d(1+2\delta)\,,$$

which completes the proof. □

6.7.1.2 Proof of Lemma 6.1

Proof. We can bound the regret from below as

$$\begin{aligned}
\mathsf{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell(t), a(t) - a^* \rangle \right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell'_t, a(t) - a^* \rangle + \sum_{t=1}^{T} \langle \ell(t) - \ell'_t, a(t) - a^* \rangle \right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \mu, a(t) - a^* \rangle - \sum_{t=1}^{T} \|\ell(t) - \ell'_t\|_\infty \|a(t) - a^*\|_1 \right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \mu, a(t) - a^* \rangle - 2m \sum_{t=1}^{T} \|\ell(t) - \ell'_t\|_\infty \right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \mu, a(t) - a^* \rangle \right] - 2mC\,, \qquad (6.30)
\end{aligned}$$

where the first inequality follows from the Hölder's inequality and $\mathbb{E}[\ell'_t] = \mu$, the second inequality follows since $\|a(t) - a^*\|_1 \leq 2m$, and the last inequality follows from the definition of $C = \sum_{t=1}^{T} \|\ell(t) - \ell'_t\|_\infty$. We then bound $\mathbb{E}\left[\sum_{t=1}^{T} \langle \mu, a(t) - a^* \rangle \right]$.

We consider the case of general action sets and recall that $I^* = \{i \in [d] : a^*_i = 1\}$ and $J^* = [d] \setminus I^*$. Since $\sum_{i \in I^*}(1 - a_i(t)) \leq \min\{|I^*|, d-m\}$ and $\sum_{i \in J^*} a_i(t) \leq \min\{|J^*|, m\}$, we have

$$\begin{aligned}
\langle \mu, a(t) - a^* \rangle &= \frac{1}{2} \langle \mu, a(t) - a^* \rangle + \frac{1}{2} \langle \mu, a(t) - a^* \rangle \\
&\geq \frac{1}{2\min\{|I^*|, d-m\}} \sum_{i \in I^*} (1 - a_i(t)) \langle \mu, a(t) - a^* \rangle \\
&\quad + \frac{1}{2\min\{|J^*|, m\}} \sum_{i \in J^*} a_i(t) \langle \mu, a(t) - a^* \rangle \\
&\geq \frac{1}{2\min\{|I^*|, d-m\}} \sum_{i \in I^*} \Delta'_{i,\min}(1 - a_i(t)) \\
&\quad + \frac{1}{2\min\{m, |J^*|\}} \sum_{i \in J^*} \Delta_{i,\min} a_i(t)\,,
\end{aligned}$$

where the last inequality follows since for any $i \in I^*$ we have $\langle \mu, a(t) - a^* \rangle \geq \Delta'_{i,\min}$, and for any $i \in J^*$ we have $\langle \mu, a(t) - a^* \rangle \geq \Delta_{i,\min}$. Combining this inequality with (6.30) completes the proof. □

Note that in the stochastic regime with adversarial corruptions, from Lemma 6.1 it holds that

$$\begin{aligned}\mathsf{Reg}_T &\geq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{v(\mathcal{A})}\sum_{i\in I^*}\Delta'_{i,\min}(1-a_i(t))+\frac{1}{w(\mathcal{A})}\sum_{i\in J^*}\Delta_{i,\min}a_i(t)\right)\right]-2Cm\\ &=\frac{1}{v(\mathcal{A})}\sum_{i\in I^*}\Delta'_{i,\min}Q_i+\frac{1}{w(\mathcal{A})}\sum_{i\in J^*}\Delta_{i,\min}P_i-2Cm\,, \end{aligned} \tag{6.31}$$

where the equality follows from the law of iterated expectations.

6.7.2 Proof of Theorem 6.3

6.7.2.1 Preliminaries

Before proving the regret upper bounds in Theorem 6.3, we prepare some lemmas. We bound the sum over $i \in [d]$ in (6.15) by considering different upper bounds for the optimal and sub-optimal base-arms. Recall that $\alpha_i(t)$ and $m_i(t)$ are given by (6.7) and (6.2), respectively. We use a following lemma to bound $\sum_{t=1}^{T}\alpha_i(t)$ for sub-optimal base-arms $i \in J^*$.

Lemma 6.4. *It holds for any* $i \in [d]$ *and* $m_i^* \in [0,1]$ *that*

$$\sum_{t=1}^{T}\alpha_i(t) \leq \sum_{t=1}^{T}a_i(t)(\ell_i(t)-m_i(t))^2 \leq \sum_{t=1}^{T}a_i(t)(\ell_i(t)-m_i^*)^2+\log(1+N_i(T))+\frac{5}{4}\,.$$

To prove this lemma, we use the following lemma.

Lemma 6.5 (Ito et al. 2022b, Lemma 8)**.** *Suppose* $\ell(s) \in [0,1]$ *for any* $s \in [t]$ *and define* $m(t) \in [0,1]$ *by*

$$m(t)=\frac{1}{t}\left(\frac{1}{2}+\sum_{s=1}^{t-1}\ell(s)\right)\,.$$

Then, for any $m^* \in [0,1]$ *we have*

$$\sum_{t=1}^{T}((\ell(t)-m(t))^2-(\ell(t)-m^*)^2) \leq \frac{5}{4}+\log T\,.$$

Proof of Lemma 6.4. From the definition of $\alpha_i(t)$, we have

$$\begin{aligned}\sum_{t=1}^{T}\alpha_i(t) &\leq \sum_{t=1}^{T}a_i(t)(\ell_i(t)-m_i(t))^2\\ &\leq \sum_{t=1}^{T}a_i(t)(\ell_i(t)-m_i^*)^2+\frac{5}{4}+\log\left(1+\sum_{t=1}^{T}a_i(t)\right)\\ &=\sum_{t=1}^{T}a_i(t)(\ell_i(t)-m_i^*)^2+\frac{5}{4}+\log\left(1+N_i(T)\right)\,,\end{aligned}$$

where the second inequality follows from Lemma 6.5 and the definition of $m_i(t)$ given in (6.2). □

From Lemma 6.4, in the stochastic regime it holds that

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_i(t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}x_i(t)\sigma_i^2 + \log(1+N_i(T))\right] + \frac{5}{4} \leq \sigma_i^2 P_i + \log(1+P_i) + \frac{5}{4}, \tag{6.32}$$

where the first inequality follows from Lemma 6.4 with $m_i^* = \mu_i$ and in the last inequality we define the expected number of times that the base-arm i is chosen by

$$P_i = \mathbb{E}\left[N_i(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}[i \in I(t)]\right] = \mathbb{E}\left[\sum_{t=1}^{T}a_i(t)\right] = \mathbb{E}\left[\sum_{t=1}^{T}x_i(t)\right]. \tag{6.33}$$

On the other hand, for the analysis of the optimal base-arms $i^* \in I^*$, we give a bound on $\sum_{t=1}^{T}\alpha_i(t)$ using the following lemma.

Lemma 6.6. *It holds for any* $i^* \in [d]$ *that*

$$\mathbb{E}[\alpha_{i^*}(t)] \leq 2\mathbb{E}\left[\min\left\{x_{i^*}(t),\ \frac{1-x_{i^*}(t)}{\sqrt{\gamma}}\right\}\right] \leq 2\mathbb{E}\left[\frac{1-x_{i^*}(t)}{\sqrt{\gamma}}\right].$$

Proof. From the definition of $\alpha_i(t)$ in (6.7), we have

$$\begin{aligned}
\mathbb{E}[\alpha_i(t)|x_i(t)] &= \mathbb{E}\left[a_i(t)(\ell_i(t)-m_i(t))^2\min\left\{1, \frac{2(1-x_i(t))}{\gamma x_i(t)^2}\right\}\ \middle|\ x_i(t)\right] \\
&\leq \mathbb{E}\left[a_i(t)\min\left\{1, \frac{2(1-x_i(t))}{\gamma x_i(t)^2}\right\}\ \middle|\ x_i(t)\right] \\
&= \min\left\{x_i(t), \frac{2(1-x_i(t))}{\gamma x_i(t)}\right\} \\
&\leq \begin{cases} x_i(t) & (x_i(t) < \frac{1}{\sqrt{\gamma}}) \\ \frac{2(1-x_i(t))}{\sqrt{\gamma}} & (x_i(t) \geq \frac{1}{\sqrt{\gamma}}) \end{cases} \leq \frac{2}{\sqrt{\gamma}}(1-x_i(t)),
\end{aligned}$$

where the first inequality follows from the condition of $\ell_i(t), m_i(t) \in [0,1]$ and the last inequality is due to $\sqrt{\gamma} \geq 2$ that follows from the assumption of $T \geq 55$. □

6.7.2.2 Proof for the Stochastic Regime

Proof of (6.9) **in Theorem 6.3.** We bound the RHS of (6.15) separately considering sub-optimal and optimal base-arms.

Sub-optimal base-arms side First, we let $i \in J^*$ be a sub-optimal base-arm. From (6.32), the component of the RHS of (6.15) is bounded as

$$
\begin{aligned}
&\mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)}\right] \\
&= \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T} \alpha_i(t)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)\right)\right] \\
&\le 2\sqrt{\beta_0^2 + \frac{1}{\gamma}\left(\sigma_i^2 P_i + \log(1+P_i) + \frac{5}{4}\right)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\left(\sigma_i^2 P_i + \log(1+P_i) + \frac{5}{4}\right)\right) \\
&\le 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} + \frac{1}{\gamma\beta_0}\left(\log(1+P_i) + \frac{5}{4}\right) - \beta_0 \\
&\qquad + \delta \log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\delta}{\gamma\beta_0^2}\left(\log(1+P_i) + \frac{5}{4}\right) \\
&= 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta \log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right), \qquad (6.34)
\end{aligned}
$$

where the first inequality follows from (6.32), the second inequality follows from $\sqrt{x+y} \le \sqrt{x} + \frac{y}{2\sqrt{x}}$ that holds for any $x > 0$ and $y \ge 0$, $\log(1+x+y) \le \log(1+x) + y$ that holds for any $x, y \ge 0$, and in the last equality we define $\xi = \frac{1}{\beta_0} + \frac{\delta}{\beta_0^2} = \frac{1}{1+\epsilon} + \frac{\delta}{(1+\epsilon)^2}$.

Optimal base-arm side Next, we let $i \in I^*$ be an optimal base-arm. We define the complement version of P_i by

$$
Q_i = \mathbb{E}\left[\sum_{t=1}^{T}(1 - x_i(t))\right]
$$

for $i \in [d]$. Then from Lemma 6.6 we have

$$
\begin{aligned}
&\mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)}\right] \\
&= \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)\right)\right] \\
&\le \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)} - \beta_0 + 2\delta\left(\sqrt{1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)} - 1\right)\right] \\
&= 2\left(\beta_0 + \delta\right)\mathbb{E}\left[\sqrt{1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)} - 1\right] + \beta_0 \\
&\le 2(\beta_0 + \delta)\left(\sqrt{1 + \frac{2}{\gamma^{3/2}\beta_0^2}\mathbb{E}\left[\sum_{t=1}^{T}(1 - x_i(t))\right]} - 1\right) + \beta_0 \,. \\
&\le 2(\beta_0 + \delta)\sqrt{\frac{2}{\gamma^{3/2}\beta_0^2}\mathbb{E}\left[\sum_{t=1}^{T}(1 - x_i(t))\right]} + \beta_0 \,. \\
&\le 2(1 + \delta)\sqrt{\frac{2}{\gamma^{3/2}}Q_i} + \beta_0 \,, \qquad (6.35)
\end{aligned}
$$

where the first inequality follows from the inequality of $\log(1+x) \leq 2(\sqrt{1+x}-1)$ for $x > 0$, the second inequality follows from Lemma 6.6, the third inequality follows from $\sqrt{1+x}-1 \leq \sqrt{x}$ for $x \geq 0$, and the last inequality follows from $\beta_0 \geq 1$.

Putting together the upper and lower bounds and applying a self-bounding technique Bounding the RHS of (6.15) using (6.34) and (6.35) yields the regret upper bound depending on $(P_i)_{i \in J^*}$ and $(Q_i)_{i \in I^*}$ as

$$\begin{aligned}\frac{\mathsf{Reg}_T}{\gamma} &\leq \sum_{i \in J^*} \left(2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta \log\left(1 + \frac{\sigma_i^2 P_i}{\gamma \beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right)\right) \\ &\quad + 2\left(1+\delta\right) \sum_{i \in I^*} \sqrt{\frac{2}{\gamma^{3/2}} Q_i} + \beta_0 |I^*| + \frac{d^2 + d(1+2\delta)}{\gamma} \\ &= \sum_{i \in J^*} \bar{f}_i\left(\frac{P_i}{\gamma}\right) + 2\left(1+\delta\right) \sum_{i \in I^*} \sqrt{\frac{2}{\gamma^{3/2}} Q_i} + \beta_0 |I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi |J^*|\right) ,\end{aligned} \tag{6.36}$$

where we define convex function $\bar{f}_i : \mathbb{R}_+ \to \mathbb{R}$ by

$$\bar{f}_i(x) = 2\sqrt{\beta_0^2 + \sigma_i^2 x} + \delta \log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) + \frac{\xi}{\gamma} \log(1 + \gamma x) - \beta_0 \,. \tag{6.37}$$

In the stochastic regime, setting $C = 0$ in (6.31) yields the regret lower bound depending on $(P_i)_{i \in J^*}$ and $(Q_i)_{i \in I^*}$ as

$$\mathsf{Reg}_T \geq \frac{1}{v(\mathcal{A})} \sum_{i \in I^*} \Delta'_{i,\min} Q_i + \frac{1}{w(\mathcal{A})} \sum_{i \in J^*} \Delta_{i,\min} P_i \,. \tag{6.38}$$

Combining (6.36) and (6.38), we have

$$\begin{aligned}&\frac{\mathsf{Reg}_T}{\log T} = \frac{\mathsf{Reg}_T}{\gamma} = 2\frac{\mathsf{Reg}_T}{\gamma} - \frac{\mathsf{Reg}_T}{\gamma} \\ &\leq 2\frac{\mathsf{Reg}_T}{\gamma} - \frac{1}{\gamma}\left(\frac{1}{v(\mathcal{A})} \sum_{i \in I^*} \Delta'_{i,\min} Q_i + \frac{1}{w(\mathcal{A})} \sum_{i \in J^*} \Delta_{i,\min} P_i\right) \\ &\leq \sum_{i \in J^*} \left(2\bar{f}_i\left(\frac{P_i}{\gamma}\right) - \frac{\Delta_{i,\min}}{w(\mathcal{A})}\frac{P_i}{\gamma}\right) + \sum_{i \in I^*}\left(4(1+\delta)\sqrt{\frac{2}{\gamma^{1/2}}\frac{Q_i}{\gamma}} - \frac{\Delta'_{i,\min}}{v(\mathcal{A})}\frac{Q_i}{\gamma}\right) \\ &\quad + 2\beta_0 |I^*| + \frac{2}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) \\ &\leq \sum_{i \in J^*} \max_{x \geq 0}\left\{2\bar{f}_i(x) - \frac{\Delta_{i,\min}}{w(\mathcal{A})} x\right\} + \sum_{i \in I^*} \max_{x \geq 0}\left\{4(1+\delta)\sqrt{\frac{2}{\gamma^{1/2}} x} - \frac{\Delta'_{i,\min}}{v(\mathcal{A})} x\right\} \\ &\quad + 2\beta_0 |I^*| + \frac{2}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) \\ &\leq \sum_{i \in J^*} \max_{x \geq 0}\left\{2\bar{f}_i(x) - \frac{\Delta_{i,\min}}{w(\mathcal{A})} x\right\} + \sum_{i \in I^*} \frac{16(1+\delta)^2 v(\mathcal{A})}{\sqrt{\gamma}\Delta'_{i,\min}} \\ &\quad + 2\beta_0 |I^*| + \frac{2}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) ,\end{aligned} \tag{6.39}$$

where the second inequality follows from (6.36) and the last inequality follows from $a\sqrt{x} - bx \leq a^2/(2b)$ for $a, b, x \geq 0$.

In the following, we evaluate the first term of (6.39).

Bounding the first term of (6.39) We will prove the following statement:

$$\max_{x\geq 0}\left\{2\bar{f}_i(x)-\frac{\Delta_{i,\min}}{w(\mathcal{A})}x\right\}\leq h\left(w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}\right)+O\left(\frac{\log(1+\gamma)}{\gamma}\right)\,, \tag{6.40}$$

where $h:\mathbb{R}_+\to\mathbb{R}$ is defined as

$$h(z)=\begin{cases}2\beta_0 & \text{if } 0\leq z\leq \frac{\beta_0}{2(1+\delta/\beta_0)}\,,\\[2ex] 2z\bigl(1+\sqrt{1+2\frac{\delta}{z}}\bigr)-2\delta+4\delta\bigl(\log\frac{z}{\beta_0}+\log(1+\sqrt{1+2\frac{\delta}{z}})\bigr)+\frac{\beta_0^2}{z}-2\beta_0 & \text{if } z>\frac{\beta_0}{2(1+\delta/\beta_0)}\,.\end{cases} \tag{6.41}$$

Let $\bar{\Delta}_i=\Delta_{i,\min}/w(\mathcal{A})$ for the notational simplicity. As f_i is concave, the maximum of $2f_i(x)-\bar{\Delta}_i x$ is attained by $x_i^*\in\mathbb{R}$ satisfying $2f'(x_i^*)=\bar{\Delta}_i$. Define $\tilde{x}_i\geq 0$ by

$$\tilde{x}_i\coloneqq\max\left\{\left(\frac{4\sigma_i}{\bar{\Delta}_i}\right)^2,\frac{8\delta}{\bar{\Delta}_i},\frac{16\xi}{\gamma\bar{\Delta}_i}\right\}\,.$$

We then have

$$2f_i'(\tilde{x}_i)\leq\frac{2\sigma_i}{\sqrt{\left(\frac{4\sigma_i}{\bar{\Delta}_i}\right)^2}}+\frac{2\delta\sigma_i^2}{{\beta_0}^2+\sigma_i^2\frac{8\delta}{\bar{\Delta}_i}}+\frac{2\xi}{1+\gamma\frac{16\xi}{\gamma\bar{\Delta}_i}}\leq\frac{\bar{\Delta}_i}{2}+\frac{\bar{\Delta}_i}{4}+\frac{\bar{\Delta}_i}{8}<\bar{\Delta}_i\,,$$

which implies $\tilde{x}_i\geq x_i^*$. Hence, we have

$$\begin{aligned}&\max_{x\geq 0}\left\{2f_i(x)-\bar{\Delta}_i x\right\}=2f_i(x_i^*)-\bar{\Delta}_i x_i^*\\&=4\sqrt{\beta_0^2+\sigma_i^2x_i^*}+2\delta\log\left(1+\frac{\sigma_i^2x_i^*}{\beta_0^2}\right)+2\frac{\xi}{\gamma}\log(1+\gamma x_i^*)-\bar{\Delta}_i x_i^*-2\beta_0\\&\leq 4\sqrt{\beta_0^2+\sigma_i^2x_i^*}+2\delta\log\left(1+\frac{\sigma_i^2x_i^*}{\beta_0^2}\right)+2\frac{\xi}{\gamma}\log(1+\gamma\tilde{x}_i)-\bar{\Delta}_i x_i^*-2\beta_0\\&\leq\max_{x\geq 0}\left\{4\sqrt{\beta_0^2+\sigma_i^2x}+2\delta\log\left(1+\frac{\sigma_i^2x}{\beta_0^2}\right)-\bar{\Delta}_i x\right\}+2\frac{\xi}{\gamma}\log(1+\gamma\tilde{x}_i)-2\beta_0\\&=\max_{x\geq 0}\left\{g_i(x)-\bar{\Delta}_i x\right\}-2\beta_0+O\left(\frac{\log(1+\gamma)}{\gamma}\right)\,,\end{aligned} \tag{6.42}$$

where we define

$$g_i(x)=4\sqrt{\beta_0^2+\sigma_i^2x}+2\delta\log\left(1+\frac{\sigma_i^2x}{\beta_0^2}\right)\,.$$

From (6.42) and (6.39), we have

$$\limsup_{T\to\infty}\frac{R_T}{\log T}\leq\sum_{i\in J^*}\left(\max_{x\geq 0}\left\{g_i(x)-\bar{\Delta}_i x\right\}-2\beta_0\right)+2\beta_0|I^*|\,.$$

In the following, we write $z_i=\frac{\sigma_i^2}{\bar{\Delta}_i}$. As we have

$$g_i'(x)=\frac{2\sigma_i^2}{\sqrt{\beta_0^2+\sigma_i^2x}}+\frac{2\delta\sigma_i^2}{\beta_0^2+\sigma_i^2x}\leq 2\sigma_i^2\left(\frac{1}{\beta_0}+\frac{\delta}{\beta_0^2}\right)\,,$$

If $z_i = \frac{\sigma_i^2}{\bar{\Delta}_i} \leq \frac{1}{2(1/\beta_0+\delta/\beta_0^2)} = \frac{\beta_0}{2(1+\delta/\beta_0)}$, the maximum of $g_i(x) - \bar{\Delta}_i x$ is attained by $x = 0$, implying

$$\max_{x \geq 0} \left\{ g_i(x) - \bar{\Delta}_i x \right\} = g_i(0) = 4\beta_0 \quad \text{if} \quad z_i := \frac{\sigma_i^2}{\bar{\Delta}_i} \leq \frac{\beta_0}{2(1+\delta/\beta_0)} . \tag{6.43}$$

Otherwise, we have

$$\begin{aligned} g_i(x) - \bar{\Delta}_i x &= 4\beta_0 \sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}} + 2\delta \log \left(1 + \frac{\sigma_i^2 x}{\beta_0^2} \right) - \frac{\beta_0^2 \bar{\Delta}_i}{\sigma_i^2} \left(1 + \frac{\sigma_i^2 x}{\beta_0^2} \right) + \frac{\beta_0^2}{z_i} \\ &= 4\beta_0 \sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}} + 4\delta \log \left(\sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}} \right) - \frac{\beta_0^2}{z_i} \left(\sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}} \right)^2 + \frac{\beta_0^2}{z_i} . \end{aligned}$$

From this, by setting $y = \sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}}$, we obtain

$$\max_{x \geq 0} \left\{ g_i(x) - \bar{\Delta}_i x \right\} \leq \max_{y \geq 0} \left\{ 4\beta_0 y + 4\delta \log y - \frac{\beta_0^2}{z_i} y^2 \right\} + \frac{\beta_0^2}{z_i} . \tag{6.44}$$

We here use the following:

$$\max_{y \geq 0} \left\{ ay + b \log y - cy^2 \right\} = \frac{1}{2} \left(\frac{a}{4c} \left(a + \sqrt{a^2 + 8bc} \right) - b \right) + b \log \frac{a + \sqrt{a^2 + 8bc}}{4c} ,$$

which holds for any $a, b, c > 0$. We hence have

$$\begin{aligned} &\max_{y \geq 0} \left\{ 4\beta_0 y + 4\delta \log y - \frac{\beta_0^2}{z_i} y^2 \right\} \\ &= \frac{1}{2} \left(\frac{4\beta_0 z_i}{4\beta_0^2} \left(4\beta_0 + \sqrt{(4\beta_0)^2 + 32 \frac{\delta \beta_0^2}{z_i}} \right) - 4\delta \right) + 4\delta \log \frac{4\beta_0 + \sqrt{(4\beta_0)^2 + 32\delta\beta_0^2/z_i}}{4\beta_0^2/z_i} \\ &= 2 \left(z_i \left(1 + \sqrt{1 + 2\frac{\delta}{z_i}} \right) - \delta \right) + 4\delta \left(\log \frac{z_i}{\beta_0} + \log \left(1 + \sqrt{1 + 2\frac{\delta}{z_i}} \right) \right) . \end{aligned} \tag{6.45}$$

Combining (6.42) with (6.43), (6.44), and (6.45), we obtain

$$\max_{x \geq 0} \left\{ 2f_i(x) - \bar{\Delta}_i x \right\} \leq h \left(\frac{\sigma_i^2}{\bar{\Delta}_i} \right) + O \left(\frac{\log(1+\gamma)}{\gamma} \right) = h \left(w(\mathcal{A}) \frac{\sigma_i^2}{\Delta_{i,\min}} \right) + O \left(\frac{\log(1+\gamma)}{\gamma} \right) , \tag{6.46}$$

where $h : \mathbb{R}_+ \to \mathbb{R}$ is defined by (6.41). From (6.39) and (6.46), we complete the proof of (6.40).

Bounding h For $z > \frac{\beta_0}{2(1+\delta/\beta_0)}$, $h(z)$ in (6.41) is bounded as

$$\begin{aligned} h(z) &\leq 2z \left(1 + 1 + \frac{\delta}{z} \right) - 2\delta + 4\delta \left(\log z + \log \left(1 + \sqrt{1 + 2\frac{\delta}{z}} \right) \right) + \frac{\beta_0^2}{\beta_0} \cdot 2 \left(1 + \frac{\delta}{\beta_0} \right) - 2\beta_0 \\ &= 4z + 4\delta \left(\log z + \log \left(1 + \sqrt{1 + 2\frac{\delta}{z}} \right) + \frac{1}{2} \right) \\ &\leq 4z + c \log(1+z) \quad \left(c = O\left(\delta^2\right) = O\left(\left(\log \epsilon^{-1} \right)^2 \right) \right) , \end{aligned}$$

where the last inequality follows from $\log(1+z) = \Omega(1/\delta)$ that holds for $z > \frac{\beta_0}{2(1+\delta/\beta_0)}$. Hence, for any $z \geq 0$, $h(z)$ is bounded as

$$h(z) \leq \max\{4z + c\log(1+z), 2\beta_0\} \ . \tag{6.47}$$

From this and (6.46), recalling that $\beta_0 = 1+\epsilon$, we obtain

$$\begin{aligned}\mathsf{Reg}_T \leq & \left(\sum_{i \in J^*} \max\left\{4\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1 + \frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T \\ & + \sum_{i \in I^*} \frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T} + o(\sqrt{\log T})\,,\end{aligned}$$

which completes the proof of (6.9) in Theorem 6.3. □

6.7.2.3 Proof for the Stochastic Regime with Adversarial Corruptions

We here show a regret bound for the stochastic regime with adversarial corruptions given in Theorem 6.3, which is the following regret bound:

$$\mathsf{Reg}_T \leq \mathcal{R}^{\mathrm{LS}} + O\left(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}}\right)\,,$$

where $\mathcal{R}^{\mathrm{LS}}$ is the RHS of (6.9) and C is the corruption level defined in Section 6.3.

Proof. In stochastic regimes with adversarial corruptions, using Lemma 6.4 with $m_i^* = \mu_i$ we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \alpha_i(t)\right] &\leq \mathbb{E}\left[\sum_{t=1}^T a_i(t)(\ell_i(t) - \mu_i)^2 + \log(1 + N_i(T))\right] + \frac{5}{4} \\ &= \mathbb{E}\left[\sum_{t=1}^T x_i(t)(\ell_i(t) - \ell'_i(t) + \ell'_i(t) - \mu_i)^2 + \log(1 + N_i(T))\right] + \frac{5}{4} \\ &= \mathbb{E}\left[\sum_{t=1}^T x_i(t)\left((\ell_i(t) - \ell'_i(t))^2 + \sigma_i^2\right) + \log(1 + N_i(T))\right] + \frac{5}{4} \\ &\leq \sigma_i^2 P_i + \log(1 + P_i) + \frac{5}{4} + P'_i\,,\end{aligned} \tag{6.48}$$

where we define

$$P'_i = \mathbb{E}\left[\sum_{t=1}^T x_i(t)(\ell_i(t) - \ell'_i(t))^2\right]\,. \tag{6.49}$$

Hence, by a similar argument to that of showing (6.34), by using (6.48) instead of (6.32), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)}\right] \\
&= \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)\right)\right] \\
&\le 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta \log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right) + 2\sqrt{\frac{P_i'}{\gamma}} + \delta\log\left(1 + \frac{P_i'}{\gamma\beta_0^2}\right) \\
&\le 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta \log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right) + \left(2 + \frac{\delta}{\beta_0}\right)\sqrt{\frac{P_i'}{\gamma}}\,,
\end{aligned}
$$

where the last inequality follows from $\log(1+x) \le \sqrt{x}$ for $x \ge 0$. Combining this with (6.15) and (6.35), via a similar argument to that of showing (6.39), we have

$$
\frac{\mathsf{Reg}_T}{\gamma} \le \sum_{i\in J^*} \bar{f}_i\left(\frac{P_i}{\gamma}\right) + \beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) + \left(2 + \frac{\delta}{\beta_0}\right)\sum_{i\in J^*}\sqrt{\frac{P_i'}{\gamma}}\,, \tag{6.50}
$$

where we recall that $\bar{f}_i$ is defined in (6.37) by

$$
\bar{f}_i(x) = 2\sqrt{\beta_0^2 + \sigma_i^2 x} + \delta\log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) + \frac{\xi}{\gamma}\log(1+\gamma x) - \beta_0\,.
$$

We further have

$$
\begin{aligned}
&\sum_{i\in J^*}\sqrt{\frac{P_i'}{\gamma}} \le \sqrt{\frac{|J^*|}{\gamma}\sum_{i\in J^*}P_i'} = \sqrt{\frac{|J^*|}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in J^*} x_i(t)(\ell_i(t) - \ell_i'(t))^2\right]} \\
&\le \sqrt{\frac{m|J^*|}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell(t) - \ell'(t)\|_\infty^2\right]} \le \sqrt{\frac{m|J^*|}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell(t) - \ell'(t)\|_\infty\right]} = \sqrt{\frac{m|J^*|}{\gamma}C}\,,
\end{aligned} \tag{6.51}
$$

where the first inequality follows from the Cauchy-Schwarz inequality, the first equality follows from the definition of P_i' in (6.49), and the second inequality follows from the fact that $\sum_{i\in J^*} x_i(t) \le m$. Combining (6.50) and (6.51), we obtain

$$
\frac{\mathsf{Reg}_T}{\gamma} \le \sum_{i\in J^*} \bar{f}_i\left(\frac{P_i}{\gamma}\right) + \beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) + \left(2 + \frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C}\,. \tag{6.52}
$$

From (6.52) and Lemma 6.1, for any $\lambda \in (0,1]$, letting $\bar{\Delta}_i = \Delta_{i,\min}/w(\mathcal{A})$ we have

$$
\begin{aligned}
\frac{\mathsf{Reg}_T}{\log T} &= (1+\lambda)\frac{\mathsf{Reg}_T}{\gamma} - \lambda\frac{\mathsf{Reg}_T}{\gamma} \\
&\le \sum_{i\in J^*}\max_{x\ge 0}\left\{(1+\lambda)\bar{f}_i(x) - \lambda\bar{\Delta}_i x\right\} + \sum_{i\in I^*}\frac{(1+\lambda)^2}{\lambda}\frac{4(1+\delta)^2 v(\mathcal{A})}{\sqrt{\gamma}\Delta_{i,\min}'} \\
&\quad + 2\left(2 + \frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C} + \frac{2\lambda C m}{\gamma} \\
&\quad + (1+\lambda)\left(\beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right)\right)\,,
\end{aligned} \tag{6.53}
$$

which can be shown in a way similar to the argument of (6.39). Further, we have

$$\begin{aligned}
&\max_{x\geq 0}\left\{(1+\lambda)\bar{f}_i(x)-\lambda\bar{\Delta}_i x\right\}=\frac{1+\lambda}{2}\max_{x\geq 0}\left\{2\bar{f}_i(x)-\frac{2\lambda\bar{\Delta}_i}{1+\lambda}x\right\}\\
&\leq\frac{1+\lambda}{2}h\left(\frac{(1+\lambda)\sigma_i^2}{2\lambda\bar{\Delta}_i}\right)+O\left(\frac{\log(1+\gamma)}{\gamma}\right)\\
&\leq\max\left\{\frac{(1+\lambda)^2}{\lambda}\frac{\sigma_i^2}{\bar{\Delta}_i}+c\log\left(1+\frac{\sigma_i^2}{\lambda\bar{\Delta}_i}\right),(1+\lambda)\beta_0\right\}+O\left(\frac{\log(1+\gamma)}{\gamma}\right)\\
&\leq\max\left\{4\frac{\sigma_i^2}{\bar{\Delta}_i}+c\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right),2\beta_0\right\}+(1+c)\left(\frac{1}{\lambda}-1\right)\frac{\sigma_i^2}{\bar{\Delta}_i}+O\left(\frac{\log(1+\gamma)}{\gamma}\right),
\end{aligned}\tag{6.54}$$

where $h(z)$ is defined as (6.41), the first inequality follows from (6.46), the second inequality comes from (6.47) and $\lambda\in(0,1]$, and the last inequality follows from

$$\begin{aligned}
\frac{(1+\lambda)^2}{\lambda}&=\lambda+2+\frac{1}{\lambda}\leq 3+\frac{1}{\lambda}=4+\left(\frac{1}{\lambda}-1\right),\\
\log\left(1+\frac{\sigma_i^2}{\lambda\bar{\Delta}_i}\right)&\leq\frac{1}{\lambda}\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right)\leq\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right)+\left(\frac{1}{\lambda}-1\right)\frac{\sigma_i^2}{\bar{\Delta}_i}.
\end{aligned}$$

Using (6.53), (6.54), and $\lambda\leq 1$, we obtain

$$\begin{aligned}
\frac{\mathsf{Reg}_T}{\log T}\leq&\sum_{i\in J^*}\max\left\{4\frac{\sigma_i^2}{\bar{\Delta}_i}+c\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right),2\beta_0\right\}+2\beta_0|I^*|\\
&+2\left(2+\frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C}+2\lambda\frac{Cm}{\gamma}+(1+c)\left(\frac{1}{\lambda}-1\right)\sum_{i\in J^*}\frac{\sigma_i^2}{\bar{\Delta}_i}\\
&+\sum_{i\in I^*}\frac{(1+\lambda)^2}{\lambda}\frac{4(1+\delta)^2v(\mathcal{A})}{\sqrt{\gamma}\Delta'_{i,\min}}+O\left(\frac{\log(1+\gamma)}{\gamma}\right).
\end{aligned}\tag{6.55}$$

By choosing $\lambda=\sqrt{\frac{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)+2Cm}}$, we have

$$\begin{aligned}
\lambda&\leq\sqrt{\frac{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}{2Cm}}\\
\frac{1}{\lambda}-1&=\sqrt{1+\frac{2Cm}{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}}-1\leq\sqrt{\frac{2Cm}{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}},
\end{aligned}$$

which imply that

$$2\left(2+\frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C}+\frac{2\lambda Cm}{\gamma}+(1+c)\left(\frac{1}{\lambda}-1\right)\sum_{i\in J^*}\frac{\sigma_i^2}{\bar{\Delta}_i}=O\left(\sqrt{\frac{Cm}{\gamma}\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}\right).$$

From this and (6.55), recalling that $\gamma=\log T$, $\beta_0=1+\epsilon$ and $\bar{\Delta}_i=\Delta_{i,\min}/w(\mathcal{A})$, we obtain

$$\begin{aligned}\mathsf{Reg}_T \le & \left(\sum_{i\in J^*} \max\left\{4w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1+w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}\right), 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T \\ & + O\left(\sqrt{Cm\sum_{i\in J^*}\left(w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}+1\right)\log T}\right) \\ & + \sum_{i\in I^*}\frac{(1+\lambda)^2}{\lambda}\frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T} + o(\sqrt{\log T})\,,\end{aligned}$$

which completes the proof for the stochastic regime with adversarial corruptions. □

6.7.2.4 Proof for the Adversarial Regime

Proof of (6.10) **in Theorem 6.3.** First, we prove $\mathsf{Reg}_T \le \sqrt{4dQ_2\log T} + O(d\log T) + d^2 + d(1+2\delta)$. For any $m^* \in [0,1]^d$, bounding the RHS of Lemma 6.3 we have

$$\begin{aligned}\mathsf{Reg}_T &\le \gamma\sum_{i=1}^d \mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta\log\frac{\beta_i(T+1)}{\beta_i(1)}\right] + d^2 + d(1+2\delta) \\ &\le 2\gamma\sum_{i=1}^d \mathbb{E}\left[\beta_i(T+1)\right] + O(d\gamma + d^2) \\ &= 2\gamma\sum_{i=1}^d \mathbb{E}\left[\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^T \alpha_i(t)}\right] + O(d\gamma+d^2) \\ &\le 2\gamma\sum_{i=1}^d \mathbb{E}\left[\sqrt{\beta_0^2 + \frac{1}{\gamma}\left(\sum_{t=1}^T a_i(t)(\ell_i(t)-m_i^*)^2 + \log(1+N_i(T)) + \frac{5}{4}\right)}\right] + O(d\gamma+d^2) \\ &\le 2\gamma\sum_{i=1}^d \mathbb{E}\left[\sqrt{\frac{1}{\gamma}\sum_{t=1}^T a_i(t)(\ell_i(t)-m_i^*)^2}\right] + O(d\gamma+d^2) \\ &\le 2\mathbb{E}\left[\sqrt{d\gamma\sum_{i=1}^d\sum_{t=1}^T a_i(t)(\ell_i(t)-m_i^*)^2}\right] + O(d\gamma+d^2) \qquad (6.56) \\ &\le 2\mathbb{E}\left[\sqrt{d\gamma\sum_{t=1}^T \|\ell(t)-m^*\|_2^2}\right] + O(d\gamma+d^2)\,,\end{aligned}$$

where the second inequality follows from $\beta_i(T+1) = O(T)$, the third inequality follows from Lemma 6.4, and the fifth inequality follows from the Cauchy-Schwarz inequality. Since m^* is arbitrary, we obtain the desired results by $m^* = \bar{\ell}$.

Next, we prove $\mathsf{Reg}_T \le \sqrt{4dL^*\log T} + O(d\log T) + d^2 + d(1+2\delta)$. By setting $m^* = 0$ in (6.56), we have

$$\begin{aligned}\mathsf{Reg}_T &\le 2\mathbb{E}\left[\sqrt{d\gamma\sum_{t=1}^T\sum_{i\in I(t)} \ell_i(t)^2}\right] + O(d\gamma+d^2) \\ &\le 2\mathbb{E}\left[\sqrt{d\gamma\sum_{t=1}^T\sum_{i\in I(t)} \ell_i(t)}\right] + O(d\gamma+d^2)\end{aligned}$$

$$
\begin{aligned}
&= 2\mathbb{E}\left[\sqrt{d\gamma\sum_{t=1}^{T}\ell(t)^\top a(t)}\right] + O(d\gamma + d^2)\\
&= 2\mathbb{E}\left[\sqrt{d\gamma\left(\sum_{t=1}^{T}(\ell(t)^\top a(t) - \ell(t)^\top a^*) + \sum_{t=1}^{T}\ell(t)^\top a^*\right)}\right] + O(d\gamma + d^2)\\
&\leq 2\sqrt{d\gamma\left(\mathbb{E}\left[\sum_{t=1}^{T}(\ell(t)^\top a(t) - \ell(t)^\top a^*)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\ell(t)^\top a^*\right]\right)} + O(d\gamma + d^2)\\
&= 2\sqrt{d\gamma\left(\mathsf{Reg}_T + L^*\right)} + O(d\gamma + d^2)\,,
\end{aligned}
$$

where the third inequality follows from Jensen's inequality. By solving this inequation in Reg_T, we obtain

$$
\mathsf{Reg}_T \leq 2\sqrt{d\gamma L^*} + O(d\gamma + d^2)\,,
$$

which is the desired bound.

Finally, we prove $\mathsf{Reg}_T \leq \sqrt{4d(mT - L^*)\log T} + O(d\log T) + d^2 + d(1+2\delta)$. By setting $m^* = \mathbf{1}$ in (6.56) and repeating a similar argument as for proving $\mathsf{Reg}_T \leq \sqrt{4dL^*\log T} + O(d\log T) + d^2 + d(1+2\delta)$ we have

$$
\begin{aligned}
\mathsf{Reg}_T &\leq 2\mathbb{E}\left[\sqrt{d\gamma\sum_{t=1}^{T}\sum_{i\in I(t)}(\ell_i(t) - 1)^2}\right] + O(d\gamma + d^2)\\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma\sum_{t=1}^{T}\sum_{i\in I(t)}(1 - \ell_i(t))}\right] + O(d\gamma + d^2)\\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma\left(mT - \sum_{t=1}^{T}\ell(t)^\top a^* - \sum_{t=1}^{T}\langle\ell(t), a(t) - a^*\rangle\right)}\right] + O(d\gamma + d^2)\\
&\leq 2\sqrt{d\gamma\left(mT - L^* - \mathsf{Reg}_T\right)} + O(d\gamma + d^2)\,,
\end{aligned}
$$

where the third inequality follows since $\|a_i(t)\|_1 \leq m$ and the forth inequality follows from Jensen's inequality. By solving this inequation in Reg_T, we obtain

$$
\mathsf{Reg}_T \leq 2\sqrt{d\gamma(mT - L^*)} + O(d\gamma + d^2)\,,
$$

which completes the proof. □

6.7.3 Proof of Theorem 6.4

We can prove Theorem 6.4 by using a similar argument as for Theorem 6.3. We first discuss the key lemma for this argument, the very similar argument of which is given in Ito (2021c).

6.7.3.1 Preliminary

Here, we present the key lemma for proving Theorem 6.4.

Lemma 6.7. *Assume that* $m_i(t)$ *is given by* (6.3). *Then for any* $i \in [d]$ *and* $u_i(1), \ldots, u_i(T) \in [0, 1]$ *we have*

$$\sum_{t=1}^{T} \alpha_i(t) \leq \sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i(t))^2$$
$$\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} a_i(t)(\ell_i(t) - u_i(t))^2 + \frac{1}{\eta(1-2\eta)} \left(\frac{1}{4} + 2 \sum_{t=1}^{T-1} |u_i(t+1) - u_i(t)| \right) .$$

Proof. Take $i \in [d]$ satisfying $a_i(t) = 1$. Then it holds that

$$\begin{aligned}
&(\ell_i(t) - m_i(t))^2 - (\ell_i(t) - u_i(t))^2 \\
&\leq 2(\ell_i(t) - m_i(t))(u_i(t) - m_i(t)) \\
&= 2(\ell_i(t) - m_i(t))(m_i(t+1) - m_i(t)) + 2(\ell_i(t) - m_i(t))(u_i(t) - m_i(t+1)) \\
&= 2\eta(\ell_i(t) - m_i(t))^2 + \frac{2}{\eta}(m_i(t+1) - m_i(t))(u_i(t) - m_i(t+1)) \\
&\leq 2\eta(\ell_i(t) - m_i(t))^2 + \frac{1}{\eta}\left((u_i(t) - m_i(t))^2 - (u_i(t) - m_i(t+1))^2\right) ,
\end{aligned}$$

where the inequalities follow from $y^2 - x^2 = 2y(y-x) - (x-y)^2 \leq 2y(y-x)$ for $x, y \in \mathbb{R}$ and the last equality follows from the definition of $m(t)$ in (6.3). Hence, we have

$$(\ell_i(t) - m_i(t))^2 \leq \frac{1}{1-2\eta} \left((\ell_i(t) - u_i(t))^2 + \frac{1}{\eta}\left((u_i(t) - m_i(t))^2 - (u_i(t) - m_i(t+1))^2\right) \right) . \tag{6.57}$$

From the definition of $\alpha_i(t)$ in (6.8) and (6.57), we have

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_i(t) \\
&\leq \sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i(t))^2 \\
&\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} (\ell_i(t) - u_i(t))^2 + \frac{1}{\eta(1-2\eta)} \sum_{t=1}^{T} \left\{ (u_i(t) - m_i(t))^2 - (u_i(t) - m_i(t+1))^2 \right\} \\
&= \frac{1}{1-2\eta} \sum_{t=1}^{T} (\ell_i(t) - u_i(t))^2 \\
&\quad + \frac{1}{\eta(1-2\eta)} \left(\sum_{t=1}^{T} \left\{ (u_i(t+1) - m_i(t+1))^2 - (u_i(t) - m_i(t+1))^2 \right\} + (u_i(1) - m_i(1))^2 \right) \\
&\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} (\ell_i(t) - u_i(t))^2 \\
&\quad + \frac{1}{\eta(1-2\eta)} \left(\sum_{t=1}^{T} (u_i(t+1) + u_i(t) - 2m_i(t+1))(u_i(t+1) - u_i(t)) + \frac{1}{4} \right) \\
&\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} a_i(t)(\ell_i(t) - u_i(t))^2 + \frac{1}{\eta(1-2\eta)} \left(\frac{1}{4} + 2 \sum_{t=1}^{T-1} |u_i(t+1) - u_i(t)| \right) ,
\end{aligned}$$

which completes the proof. □

6.7.3.2 Proof for the Stochastic Regime

Proof of (6.11) **in Theorem 6.4.** From Lemma 6.7, setting $u_i(t) = \mu_i$ for all $i \in [d]$ and $t \in [T]$ in Lemma 6.7 and taking the expectation yield that

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_i(t)\right] \le \frac{1}{1-2\eta}\mathbb{E}\left[\sum_{t=1}^{T} a_i(t)(\ell_i(t)-\mu_i)^2\right] + \frac{1}{4\eta(1-2\eta)} = \frac{1}{1-2\eta}\sigma_i^2 P_i + \frac{1}{4\eta(1-2\eta)},$$

where P_i is defined in (6.33). By using this inequality instead of (6.32) and repeating the same argument as that in Section 6.7.2.2, we obtain

$$\begin{aligned}\mathsf{Reg}_T \le{}& \frac{1}{1-2\eta}\left(\sum_{i\in J^*} \max\left\{4\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1+\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T\\ &+ O\left(d\sqrt{\frac{\log T}{\eta(1-2\eta)}}\right) + \sum_{i\in I^*}\frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T} + o(\sqrt{\log T}),\end{aligned}$$

which is the desired bound. □

6.7.3.3 Proof for the Stochastic Regime with Adversarial Corruptions

Here we show a regret bound for the stochastic regime with adversarial corruptions given in Theorem 6.4:

$$\mathsf{Reg}_T \le \mathfrak{R}^{\mathrm{GD}} + O\left(\sqrt{Cm\mathfrak{R}^{\mathrm{GD}}}\right).$$

Proof. Letting $u_i(t) = \mu_i$ for all $i \in [d]$ and $t \in [T]$ in Lemma 6.7 and taking the expectation yield that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \alpha_i(t)\right] &\le \frac{1}{1-2\eta}\mathbb{E}\left[\sum_{t=1}^{T} a_i(t)(\ell_i(t)-\mu_i)^2\right] + \frac{1}{4\eta(1-2\eta)}\\ &\le \frac{1}{1-2\eta}\sigma_i^2 P_i + P_i' + \frac{1}{4\eta(1-2\eta)},\end{aligned}$$

where P_i is defined in (6.33) and the last inequality is obtained by a similar argument as for (6.48). By using this inequality instead of (6.32) and repeating a similar argument as that in Section 6.7.2.3, we obtain

$$\begin{aligned}\mathsf{Reg}_T \le{}& \frac{1}{1-2\eta}\left(\sum_{i\in J^*} \max\left\{4\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1+\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T\\ &+ O\left(d\sqrt{\frac{\log T}{\eta(1-2\eta)}}\right) + O\left(\sqrt{Cm\sum_{i\in J^*}\left(\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}+1\right)\log T}\right)\\ &+ \sum_{i\in I^*}\frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T} + o(\sqrt{\log T}),\end{aligned}$$

which completes the proof. □

6.7.3.4 Proof for the Adversarial Regime

Proof of (6.12) **in Theorem 6.4.** From Lemma 6.7, we immediately obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t) \leq \frac{1}{1-2\eta}\sum_{t=1}^{T}\sum_{i=1}^{d}a_i(t)(\ell_i(t)-u_i(t))^2 + \frac{1}{\eta(1-2\eta)}\left(\frac{d}{4}+2\sum_{t=1}^{T-1}\|u(t+1)-u(t)\|_1\right) \tag{6.58}$$

for any $u(t) = (u_1(t), \dots, u_d(t))^\top \in [0,1]^d$.

First, we prove $\mathsf{Reg}_T \leq \sqrt{\frac{\gamma}{\eta(1-2\eta)}(d+8V_1)} + O(d\gamma + d^2)$. Letting $u(t) = \ell(t)$ in (6.58) we can bound the regret as

$$\begin{aligned}
\mathsf{Reg}_T &\leq 2\gamma\sum_{i=1}^{d}\mathbb{E}\left[\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)}\right] + O(d\gamma + d^2)\\
&\leq 2\mathbb{E}\left[\sqrt{\gamma\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t)}\right] + O(d\gamma + d^2)\\
&\leq \frac{2}{\sqrt{\eta(1-2\eta)}}\mathbb{E}\left[\sqrt{\gamma\left(\frac{d}{4}+2\sum_{t=1}^{T-1}\|\ell(t+1)-\ell(t)\|_1\right)}\right] + O(d\gamma + d^2)\\
&\leq \sqrt{\frac{\gamma}{\eta(1-2\eta)}(d+8V_1)} + O(d\gamma + d^2)\,,
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality, the third inequality follows by setting $u_i(t) = \ell_i(t)$ for all $i \in [d]$ and $t \in [T]$ in (6.58), and the last inequality follows from Jensen's inequality. This becomes the desired path-length bound.

Next, we prove we prove $\mathsf{Reg}_T \leq \sqrt{\frac{\gamma}{1-2\eta}\min\{L^*, mT - L^*, Q_2\}} + O(d\gamma + d^2)$. For any $m^* \in [0,1]^d$, letting $u(t) = m^*$ for all $t \in [T]$ in (6.58), we have

$$\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t) \leq \frac{1}{1-2\eta}\sum_{t=1}^{T}\sum_{i=1}^{d}a_i(t)(\ell_i(t)-m_i^*)^2 + \frac{d}{4\eta(1-2\eta)}\,.$$

Using this inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t)\right] &\leq \frac{1}{1-2\eta}\min_{m^*\in[0,1]^d}\left\{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}a_i(t)(\ell_i(t)-m_i^*)^2\right]\right\} + \frac{d}{4\eta(1-2\eta)}\\
&\leq \frac{1}{1-2\eta}\min\{\mathsf{Reg}_T + L^*, mT - L^* - \mathsf{Reg}_T, Q_2\} + \frac{d}{4\eta(1-2\eta)}\,,
\end{aligned}$$

where in the last inequality we set $m^* = 0$ and (resp. $m^* = 1$) and use the same argument as that in Section 6.7.2.4 for deriving the term with $\mathsf{Reg}_T + L^*$ (resp. $mT - L^* - \mathsf{Reg}_T$), and $m^* = \bar{\ell}$ for deriving the term with Q_2, and this complete the proof. □

6.8 Conclusion

In this chapter, we considered the combinatorial semi-bandit problem and presented the new BOBW algorithm with various adaptive guarantees. The new algorithm enjoys a variance-dependent regret bound depending on the tight suboptimality gap with a good leading constant in the stochastic regime and multiple data-dependent regret bounds. We numerically investigated the performance of the proposed algorithm and confirmed that the proposed algorithm performs competitively to Thompson sampling and achieve the best results in the adversarial regime.

One limitation of the proposed algorithm lies in its computational complexity: (i) sampling action $a(t)$ based on $x(t)$ and (ii) efficiently computing $x(t)$ in (6.1). Limitation (i) has long been a problem in semi-bandits using the (O)FTRL framework. Although polynomial-time algorithms exist (*e.g.,* Schrijver 1998, Corollary 14.1g), they are not very practical. For limitation (ii), it is not easy to efficiently compute $x(t)$ in existing studies, where Shannon entropy regularization for $1 - x_i$ is combined with the typical regularizers. If we can safely remove the Shannon entropy regularization for $1 - x_i(t)$, $x(t)$ then has a closed form, and an analysis for such a variant is important future work.

Chapter 7

Conclusion and Future Direction

In this chapter, we summarize this dissertation and discuss important future research prospects.

7.1 Summary

This dissertation was devoted to establish various adaptive algorithms mainly for structured bandits. This technological development allows us to choose more structure-capturing and adaptive algorithms for solving real-world problems. The following summarizes the contributions of each chapter.

Chapters 3 and 4 focused on partial monitoring. In particular, in Chapter 3, we investigated if we can construct a Thompson-sampling-based algorithm, which is known to perform significantly well in practice, with a distribution-dependent bound. We answered this question affirmatively by establishing the new Thompson-sampling-based algorithm that theoretically achieves the distribution-dependent regret bound in the stochastic regime in the linearized version of the game with local observability. Moreover, the proposed algorithm significantly outperformed existing algorithms in the experiments. In Chapter 4, we investigated if such a logarithmic guarantee for the stochastic regime can be obtained while at the same time having the theoretical guarantee for the adversarial regime, *i.e.,* the BOBW guarantee. We provided a positive response to this question: we developed algorithms based on the follow-the-regularized-leader framework by extending the framework of exploration by optimization and adaptive learning rate for online learning with feedback graphs. In Chapter 5, we aimed to further improve adaptivity of follow-the-regularized-leader. This was accomplished by constructing the learning rate so that it adapts simultaneously to stability and penalty components, whereas existing learning rates so far are adaptive to only one of the two components. This allowed us to achieve BOBW and data-dependent bounds simultaneously, or more specifically, the sparsity-dependent bound for multi-armed bandits and the game-dependent bound for partial monitoring. Finally, in Chapter 6, we targeted combinatorial semi-bandits and constructed the follow-the-regularized-leader-based algorithm that simultaneously achieves BOBW and several data-dependent bounds (first-, second-order and path-length bounds) simultaneously with the tight suboptimality gaps by developing the novel adaptive learning rate taking the underlying variances into account and deriving the tighter regret lower bound.

7.2 Future Direction

Throughout history and through the contributions in this dissertation, BOBW algorithms have rapidly advanced beyond vanilla multi-armed bandits and this progression has led to a vast accumulation of knowledge. Still, there are several important research questions to be addressed. The following section describes the future direction that will be important for the further spread of BOBW algorithms, especially for real-world applications.

7.2.1 Improving Theoretical Understanding

One of the important future directions arises from the fact that the “best” of the BOBW algorithms are in fact not truly the best in the stochastic regime. We discuss this from two main perspectives in the following.

What is the achievable leading constant of a regret upper bound in the stochastic regime while preserving a guarantee in the adversarial regime? In this dissertation, Definitions 2.1 and 2.2 are adopted as the definition of a BOBW algorithm in the sequential decision-making problem. However, this definition only considers the order with respect to T and does not consider the leading constant values. In fact, it is still unknown whether the existing BOBW algorithms can achieve optimal leading constants in the stochastic regime. For example, one algorithm in Zimmert and Seldin (2021) can achieve regret upper bounds of $\mathsf{Reg}_T \leq 2\sqrt{kT}$ in the adversarial regime and $\mathsf{Reg}_T \leq \sum_{a \neq a^*} \frac{\log T}{\Delta_a} + 28k \log T + o(\log T)$ in the stochastic regime. One can see that the leading constant of the bound in the stochastic regime is approximately twice as worse as that of the optimal regret in the stochastic regime (see Theorem 2.1).

An important question then arises: can we achieve the regret upper bound with optimal leading constants in multi-armed bandits in the stochastic regime while guaranteeing the regret upper bound of $\tilde{O}(\sqrt{kT})$ in the adversarial regime? If not, how small a leading constant can we obtain? This investigation is important for implementing BOBW algorithms in real-world problems with almost stochastic environments. As we can see from the experimental results of Zimmert and Seldin (2021) and Chapter 6, the BOBW algorithms actually perform worse than Thompson sampling in the *truly* stochastic regime, and thus it is difficult to determine whether to use Thompson sampling or BOBW algorithms for real-world problems with a very high degree of stochasticity.

To what extent can we exploit distributional information in the stochastic regime while maintaining guarantees in the adversarial regime? Another important aspect is how much distributional information in the stochastic regime can be exploited. In Chapter 6, we discussed variance-dependent regret upper bound with this motivation and found that a significant performance improvement can indeed be obtained solely by taking variances into account. However, the truly optimal regret bound is expressed using the information on higher-order moments of an underlying distribution (Burnetas and Katehakis, 1996), and whether this can be achieved while preserving the performance guarantee in an adversarial regime is an important research question.

7.2.2 Performance Evaluation through Comprehensive Studies and Improvement of User-Friendliness

Important future directions for the use of the BOBW algorithms in real-world problems are not limited to theoretical aspects. There are two important future directions from application aspects: comprehensive numerical studies and evaluation of existing algorithms in real-world problems, and improving their user-friendliness.

Comprehensive numerical studies and evaluation of existing algorithms on real-world problems It had been rare for BOBW algorithms to be compared by numerical experiments until the paper of the celebrated Tsallis-INF algorithm by Zimmert and Seldin (2019, 2021). They conducted extensive numerical experiments on multi-armed bandits and demonstrated the usefulness of the Tsallis-INF algorithm, motivating further investigation of BOBW algorithms. Subsequently, as discussed in Chapter 2, many BOBW algorithms have been developed for various structured bandits and in various setups. In the most recent example, at the time of writing this dissertation, a framework that achieves a $O(\log T)$ regret upper bound in a stochastic regime for a wide range of sequential decision-making problems, such as only learning with feedback graphs, linear bandits, and episodic Markov decision processes, was established (Dann et al., 2023). However, most of the recent BOBW algorithms for structured bandits do not include numerical experimental evaluation, possibly because they are too computationally intensive or do not have sufficient numerical performance. Therefore, it is desirable to conduct comprehensive numerical experiments to investigate the numerical properties of each algorithm that would shed light on what algorithms to use in practice.

In particular, the performance evaluation in the stochastic regime with adversarial corruptions is an important challenge. As discussed in Chapter 2, this regime is a very practical regime for real-world problems, and one of the major advantages of the BOBW algorithm is in its adaptivity to this regime. However, numerical experiments in the adversarial regime of existing BOBW methods are limited to the stochastically constrained adversarial regime. In this regime, we only need to determine the expected difference between the losses of arms and can evaluate the performance of the algorithm with the same degree of freedom as in the stochastic regime. In contrast, in the stochastic regime with adversarial corruptions, we need to determine how the adversarial noise is added, which makes it difficult to evaluate the performance of the algorithm in a proper manner.

Improving user-friendliness of best-of-both-worlds algorithms Algorithms often used in real-world problems tend to be the UCB algorithm or Thompson sampling. There appear to be several reasons for this, but the most likely reason is that they are relatively easy to implement and do not involve highly complex optimization problems. As introduced in Chapter 2, the UCB algorithm just needs to compute the UCB-index and selects the arm with the maximum index, and Thompson sampling samples from the posterior distribution and selects the arm with the lowest sampled losses. In contrast, most existing BOBW algorithms involve solving some form of convex optimization problem. For instance, this occurs when computing an output of FTRL or when selecting an arm such that the probability of that arm being selected aligns with the output of FTRL. Furthermore, the UCB algorithm and Thompson sampling are widely well-known, and there are already a large number of books and web pages with their explanations and implementation instructions. In contrast, although many theoretically and empirically superior bandit algorithms besides UCB and Thompson sampling have been developed, it cannot be said that they are sufficiently applied in the industry. In light of such a situation, it seems that several obstacles need to be overcome for BOBW algorithms to be accepted and applied in the industry. Building an accessible environment for those who work on sequential decision-making problems through survey articles, online explanations, and the development of a library of BOBW algorithms will be an important research topic, as well as evaluation through the comprehensive numerical experiments described above.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320, 2011.

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, volume 28, pages 2197–2205, 2015.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *the 25th Annual Conference on Learning Theory*, volume 23, pages 39.1–39.26, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31, pages 99–107, 2013a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *the 30th International Conference on Machine Learning*, volume 28, pages 127–135, 2013b.

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In José L. Balcázar, Philip M. Long, and Frank Stephan, editors, *Algorithmic Learning Theory*, pages 229–243, 2006.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 23–35, 2015.

Yoann Altmann, Steve McLaughlin, and Nicolas Dobigeon. Sampling from a multivariate gaussian distribution truncated on a simplex: A review. In *2014 IEEE Workshop on Statistical Signal Processing*, pages 113–116, 2014.

Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, and Roi Livni. Prediction with corrupted expert advice. In *Advances in Neural Information Processing Systems*, volume 33, pages 14315–14325, 2020.

Idan Amir, Guy Azov, Tomer Koren, and Roi Livni. Better best of both worlds bounds for bandits with switching costs. In *Advances in Neural Information Processing Systems*, volume 35, pages 15800–15810, 2022.

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.I.D. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, volume 7, pages 1–122, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165, 2007.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory*, volume 49, pages 116–120, 2016.

Peter Auer, Nicolò 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.

Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 696–710, 2013.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. In *Algorithmic Learning Theory*, pages 224–238, 2010.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19, pages 133–154, 2011.

Gábor Bartók, Navid Zolghadr, and Csaba Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *the 29th International Conference on Machine Learning*, pages 1–20, 2012.

Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. *Theoretical Computer Science*, 324(2):137–146, 2004.

Sébastien Bubeck, Nicoló Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 41.1–41.14, 2012.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 42.1–42.23, 2012.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 111–127, 2018.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 508–528, 2019a.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference on Learning Theory*, pages 508–528, 2019b.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.

George Casella, Christian P. Robert, and Martin T. Wells. *Generalized accept-reject sampling schemes*, volume 45 of *Lecture Notes–Monograph Series*, pages 342–347. Institute of Mathematical Statistics, 2004.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012. JCSS Special Issue: Cloud Computing 2011.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2011.

Cheng Chen, Canzhe Zhao, and Shuai Li. Simultaneously learning stochastic and adversarial bandits under the position-based model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6202–6210, 2022a.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 151–159, 2013.

Xiaowei Chen, Jiawei Xue, Zengxiang Lei, Xinwu Qian, and Satish V. Ukkusuri. Online eco-routing for electric vehicles using combinatorial multi-armed bandit with estimated covariance. *Transportation Research Part D: Transport and Environment*, 111: 103447, 2022b.

Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 210–227, 2013.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 208–214, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, 2008.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. *arXiv preprint arXiv:2302.09739*, 2023.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37): 1281–1316, 2014.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, volume 29, pages 2972–2980, 2016.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 34, pages 28511–28521, 2021.

Dean Foster and Alexander Rakhlin. No internal regret via neighborhood watch. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 382–390, 2012.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, volume 29, pages 4734–4742, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 176–196, 2014.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(79): 2369–2403, 2007.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Mark Herbster and Manfred K Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

William W Hogan. Point-to-set maps in mathematical programming. *SIAM review*, 15 (3):591–603, 1973.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. In *the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 375–383, 2014.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9173–9200, 2022.

Ruitong Huang, Tor Lattimore, András György, and Csaba Szepesvári. Following the leader and fast rates in online linear prediction: Curved constraint sets and other regularities. *Journal of Machine Learning Research*, 18(145):1–31, 2017.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 2654–2667, 2021a.

Shinji Ito. On optimal robustness to adversarial corruption in online decision problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 7409–7420, 2021b.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 2552–2583, 2021c.

Shinji Ito. Revisiting online submodular minimization: Gap-dependent regret bounds, best of both worlds and adversarial robustness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9678–9694, 2022.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 1421–1422, 2022b.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB : An optimal exploration algorithm for multi-armed bandits. In *The 27th Conference on Learning Theory*, pages 423–439, 2014.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems*, volume 33, pages 16557–16566, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. In *Advances in Neural Information Processing Systems*, volume 34, pages 20491–20502, 2021.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. *arXiv preprint arXiv:2302.13534*, 2023.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213, 2012b.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2328–2369, 2020.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision-making: Algorithms, regret bounds and applications. *arXiv preprint arXiv:2302.03683*, 2023.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 594–605, 2003.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 28, pages 1792–1800, 2015a.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1152–1161, 2015b.

Junpei Komiyama, Junya Honda, and Akiko Takeda. Position-based multiple-play bandit problem with unknown position bias. In *Advances in Neural Information Processing Systems*, volume 30, pages 4998–5008, 2017.

Fang Kong, Yichi Zhou, and Shuai Li. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 11473–11482, 2022.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems 26*, pages 1448–1456, 2013.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 535–543, 09–12 May 2015.

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016.

Paul Lagrée, Claire Vernade, and Olivier Cappe. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems*, volume 29, pages 1597–1605, 2016.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *the 32nd Annual Conference on Learning Theory*, volume 99, pages 2111–2139, 2019a.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 529–556, 2019b.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 2111–2139, 2019c.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020a.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2488–2515, 2020b.

Chung-Wei Lee, Haipeng Luo, and Mengxiao Zhang. A closer look at small-loss bounds for bandits with graph feedback. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2516–2564, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6142–6151, 2021.

Tian Lin, Bruno Abrahao, Robert Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 901–909, 2014.

Xutong Liu, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John Lui, and Wei Chen. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. In *Advances in Neural Information Processing Systems*, volume 35, pages 14904–14916, 2022.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems*, volume 30, pages 3258–3266, 2017.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1286–1304, 2015.

Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, volume 31, pages 8235–8245, 2018.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. volume 24, 2011.

Shie Mannor and Nahum Shimkin. On-line learning with imperfect monitoring. In *Learning Theory and Kernel Machines*, pages 552–566. Springer, 2003.

Shie Mannor, Vianney Perchet, and Gilles Stoltz. Set-valued approachability and online learning with partial monitoring. *Journal of Machine Learning Research*, 15(94): 3247–3295, 2014.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the tsallis-inf algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 3330–3350, 2021.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 11752–11762, 2022.

Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 525–533, 2011.

Nadav Merlis and Shie Mannor. Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 2465–2489, 2019.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20(83):1–28, 2019.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1360–1375, 2015.

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 234–248, 2013.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Aldo Pacchiano, Christoph Dann, and Claudio Gentile. Best of both worlds model selection. In *Advances in Neural Information Processing Systems*, volume 35, pages 1883–1895, 2022.

Vianney Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677, 2011.

Pierre Perrault, Michal Valko, and Vianney Perchet. Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3152–3184, 2020.

My Phan, Yasin Abbasi Yadkori, and Justin Domke. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems*, volume 32, pages 8804–8813, 2019.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss (extended abstract). In *Computational Learning Theory*, pages 208–223, 2001.

Roman Pogodin and Tor Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pages 894–904, 2020.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469, 2014.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, pages 3066–3074, 2013b.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3227–3249, 2020.

Chloé Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 9127–9135, 2021.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 35035–35048, 2022.

Daniel J. 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.

Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1):224–243, 1999.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19011–19026, 2022.

Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1743–1759, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1287–1295, 2014.

Tencent Inc. KDD Cup - 2012 track 2, Kaggle, 2012. URL `https://www.kaggle.com/c/kddcup2012-track2`.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

Taira Tsuchiya, Junya Honda, and Masashi Sugiyama. Analysis and design of Thompson sampling for stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 33, pages 8861–8871, 2020.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, 2023a.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023b.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. *arXiv preprint arXiv:2305.17301*, 2023c.

Taishi Uchiya, Atsuyoshi Nakamura, and Mineichi Kudo. Algorithms for adversarial bandit problems with multiple plays. In *Algorithmic Learning Theory*, pages 375–389, 2010.

Umair ul Hassan and Edward Curry. Efficient task assignment for spatial crowdsourcing: A combinatorial fractional optimization approach with semi-bandit learning. *Expert Systems with Applications*, 58:36–56, 2016.

Hastagiri P Vanchinathan, Gábor Bartók, and Andreas Krause. Efficient partial monitoring with prior information. In *Advances in Neural Information Processing Systems*, volume 27, pages 1691–1699, 2014.

Daniel Vial, Sujay Sanghavi, Sanjay Shakkottai, and R. Srikant. Minimax regret for cascading bandits. In *Advances in Neural Information Processing Systems*, volume 35, pages 29126–29138, 2022.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5114–5122, 2018.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1263–1291, 2018.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing Systems*, volume 29, pages 3972–3980, 2016.

Jeffrey M Wooldridge. Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese economic journal*, 1(2):117–139, 2002.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, 2018.

Kai Zheng, Haipeng Luo, Ilias Diakonikolas, and Liwei Wang. Equipping experts/bandits with long-term memory. In *Advances in Neural Information Processing Systems*, volume 32, pages 5929–5939, 2019.

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26, pages 1583–1591, 2013.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems 32*, pages 11973–11982, 2019.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 467–475, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7683–7692, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *the Twentieth International Conference on Machine Learning*, pages 928–935, 2003.