Adaptive Algorithms for Online Learning in Non-Stationary Environments

by

Quan M. Nguyen

B.Eng., FPT University, 2014
M.Sc., University of Hamburg, 2018

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Computer Science

We acknowledge and respect the lək ̓əŋən (Songhees and X sepsəm/Esquimalt) Peoples on whose territory the university stands, and the lək ̓əŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Adaptive Algorithms for Online Learning in Non-Stationary Environments

by

Quan M. Nguyen

B.Eng., FPT University, 2014
M.Sc., University of Hamburg, 2018

Supervisory Committee

Dr. Nishant Mehta, Supervisor
(Department of Computer Science)

Dr. Venkatesh Srinivasan, Departmental Member
(Department of Computer Science)

Dr. Cristóbal Gúzman, Outside Member
(Department of Electrical and Computer Engineering)

ABSTRACT

Traditional online learning literature often assumes static environments, where fundamental properties like data distribution or action spaces do not change over time, and the learner competes against a single best action. This framework, however, fails to capture the complexity of many practical scenarios, such as automated diagnostic systems or inventory management, where the optimal course of action is non-stationary and changes sequentially. In such settings, adaptivity is crucial as algorithms must maintain and leverage past information to respond effectively to unforeseen changes. This thesis advances the theory of online learning in non-stationary environments by developing adaptive algorithms with provably strong theoretical guarantees.

Two key non-stationary learning problems are online multi-task reinforcement learning (OMTRL) and multi-armed bandits with sleeping arms. In OMTRL, a learner interacts with a sequence of Markov Decision Processes (MDPs). Each MDP is chosen adversarially from a small collection of MDPs, requiring the learner to efficiently transfer knowledge between tasks. In multi-armed bandits with sleeping arms, the set of available arms varies adversarially across rounds, prompting the learner with unique exploration-exploitation tradeoff methods. A key contribution of this thesis is a number of novel lower bounds and algorithms with near-optimal worst-case regret upper bounds for these two problems. In addition, this thesis applies the new techniques in these new algorithms into deriving improved sample complexity for group distributionally robust optimization (GDRO) and novel data-dependent best-of-both-worlds regret upper bounds for multi-armed bandits.

In summary, this thesis provides mathematically-grounded adaptive algorithms that achieve state-of-the-art performance guarantees in learning from non-stationary and adversarially changing environments in reinforcement learning and multi-armed bandits, as well as showing new, fundamental connections between multi-armed bandits with sleeping arms and robust optimization.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor, Nishant Mehta, for introducing the world of theory research to me and providing me the vital support throughout my PhD journey. When I first came to Victoria in 2021, I did not even know how to effectively read a machine learning theory paper, much less how to do theory research. Thanks to Nishant's support and guidance, I was able to carry out the kind of research that I had only dreamed of before joining his lab. Beyond research, Nishant has been an excellent mentor in helping me navigate the Canadian academia, attend theory conferences, connect with amazing collaborators and prepare for job applications. It has truly been a great privilege to do my PhD with Nishant.

Next, I wish to thank my supervisory committee members – Venkatesh Srinivasan and Cristóbal Gúzman – for their valuable feedback throughout my PhD study. Additionally, I want to thank Cristóbal for also being a great collaborator and for introducing me to the GDRO problem, which has led to two major publications in this thesis. I am also grateful for the opportunities to work with Shinji Ito and Junpei Komiyama, who taught me how to prove best-of-both-worlds bounds for bandits.

I extend my gratitude to the University of Victoria and the Department of Computer Science for their institutional support and administrative assistance. Without your help, I would not have been able to come to Canada during the height of the Covid-19 pandemic. I am also thankful for your financial support that covered parts of my international conference travel expenses, and for providing me with on-campus accommodation in the last five years.

All of my friends at UVic have been instrumental to my PhD journey, who have been a healthy source of distraction whenever my research got stuck. To my friends in Nishant's lab, I fondly remember all of the times that I hung out with Ali, Andrea, Mica, Steve and Bingshan. I also want to thank Ha Nguyen Tuan Dat, Irene Hang Duong, UVic Badminton Club and UVic Competitive Programming Club, for so much fun cooking, hiking and playing together.

Last but not least, I want to express my deepest gratitude to my immediate family, whose unwavering support and belief saw me through my most difficult times. I am especially thankful to my wife, Linh Hoang, who has always been a never-ending source of strength, encouragement and positivity. Even though her name appears only in this Acknowledgements section, it is my belief that her contributions to the completion of this thesis are every bit as significant as my own.

Chapter 1

Introduction

This thesis investigates algorithms for online learning, a set of machine learning problems that belong to the class of sequential decision-making problems. In these problems, data is presented sequentially to a machine learning algorithm over a series of interactive rounds. In each round, the algorithm performs an action based on the current data, receives feedback on its performance, and then proceeds to the next round. Online learning has long been of great interest in theoretical machine learning due to its versatility in modeling a wide range of theoretical frameworks, such as supervised learning, non-convex optimization, and combinatorial optimization, as well as practical applications like weather forecasting, inventory management, and automated diagnostic systems.

The existing literature in online learning traditionally considers learning from static environments whose fundamental properties such as data distribution or action spaces do not change over time. In these static environments, the learner's primary concern is to compete with a single best action over all interactive rounds. This setting fails to cover many practical scenarios. For example, the effectiveness of an automated diagnostic system that deals with a sequence of patients with different underlying conditions must be measured with respect to not just one generic "best" treatment procedure, but with a sequence of appropriate treatment procedures, one of each patient. Another example is in inventory management, where the decision maker chooses which items to be stocked based on the list of available items and their past sales data. Since a particular item may be available only for certain periods of time, the notion of a single best item is not well-defined, and instead optimizing with respect to the sequence of varying available items is more suitable.

In the presence of non-stationarity, *adaptivity* becomes an important factor in the decision making processes. At a very high level, this involves maintaining past information that can help make decisions upon *unforseenable* changes in the environment. In the automated diag-

nostic system example, this past information could be the patient records that help quickly distinguish two different strains of a virus, leading to effective and timely treatment procedures. In the inventory management example, this information could be summary statistics of an item in the periods when it was available: how much revenue the item generates, how better it compared to other items in the same periods, and so on. The question of *how* to use this information to obtain adaptivity depends on the problem setup, the algorithm framework and the optimization objective, all of which are studied in the subsequent chapters.

This PhD thesis is based on a collection of four peer-reviewed publications, all of them in the domain of online learning algorithms and their applications. Each publication studies different notions of non-stationarity, adaptivity and performance measure. They are unified under a central them of developing mathematically-principled approaches for learning from sequential data generated by interacting with non-stationary environments. In particular, we aim to derive novel algorithms that are provably (near-) optimal, where optimality is measured with respect to an omniscient algorithm that can see all the future changes ahead and pick the optimal action in each round. A key focus in this thesis is ensure these new algorithms are *worst-case optimal*, such that under the worst possible changes in the environment, they perform as well as an omniscient algorithm. Furthermore, in certain settings, we obtain worst-case optimal algorithms that are simultaneously best-case optimal as they perform optimally in static environments. In addition to theoretical results, whenever possible, we also provide supporting empirical results to corroborate our theoretical findings.

1.1 Non-Stationary Online Learning Problems

For an integer K, let $[K]$ denote the set $\{1, 2, \ldots, K\}$. Throughout this thesis, we consider two types of non-stationary online learning problems: online multi-task reinforcement learning and multi-armed bandits with adversarially varying sets of arms. We defer the formal background of reinforcement learning and multi-armed bandits to Chapter 2. In this section, we give a brief overview of these two online learning problems. Algorithm 1.1 illustrates the general setup, in which the learner interacts with the environment in T rounds. In round t, the learner first receives a space $\mathcal{X}_t$ of available decisions, then takes a decision x_t from $\mathcal{X}_t$ and observes a feedback $\ell_t(x_t)$ of the chosen decision. The feedback function $\ell_t : \mathcal{X}_t \to \mathbb{R}^H$ The exact structure of $\mathcal{X}_t$ and the dimensionality H of ℓ_t are problem-dependent and determined by the tasks. In particular,

- In episodic reinforcement learning, $\mathcal{X}_t$ is the space of H-step policies for a finite-horizon

Algorithm 1.1 General Non-Stationary Online Learning Setup

for rounds $t = 1, \ldots, T$ **do**
  Learner receives a decision space $\mathcal{X}_t$
  Learner takes a decision $x_t \in \mathcal{X}_t$
  Learner observes a feedback $\ell_t(x_t)$
end

Algorithm 1.2 Online Multi-Task Reinforcement Learning Setup

Adversary has a set of MDPs $\mathcal{M}$ unknown to the learner
for episode $k = 1, \ldots, K$ **do**
  Adversary picks a task $m^k \in \mathcal{M}$
  Learner computes and runs policy π^k on m^k
end

Markov Decision Process (MDP) with horizon $H \geq 1$. The feedback $\ell_t(x_t)$ is the sequence of H reward values collected from running the chosen policy x_t in episode t.

- In multi-armed bandits, $\mathcal{X}_t$ is the finite set of available arms in round t. The feedback $\ell_t(x_t)$ is a scalar value, i.e., $H = 1$, indicating the loss of the learner's chosen arm x_t.

1.1.1 Online Multi-Task Reinforcement Learning

Algorithm 1.2 illustrates the online multi-task reinforcement learning (OMTRL) setting. In this setting, the adversary has a finite set $\mathcal{M} = \{\mathcal{M}_i : i = 1, 2, \ldots, M\}$ of M MDPs $\mathcal{M}_i := (\mathcal{S}, \mathcal{A}, P_i, r, H)$. The MDPs in $\mathcal{M}$ share the state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, horizon H. They differ in the transition kernels $(P_i)_{i=1,2,\ldots,M}$. In episode k, the adversary picks an element $m^k \in \mathcal{M}$ as the task for the learner. In this thesis, we assume that the reward function r is the same for all MDPs in $\mathcal{M}$ and fully known to the learner, but the identity of m^k is hidden.

For a given MDP m, a policy π defines how the learner chooses its action given the past states, actions and rewards while interacting with m. The policy is often represented as a conditional probability distribution over the action space of m, where the condition is the observed states, actions and rewards. The action in each step is drawn from this distribution. The performance of a policy π is measured by its *value functions* $(V_h^{m,\pi})_{h=1,2,\ldots,H}$ (formal definition in Chapter 2). In essence, $V_h^{m,\pi}(s)$ specifies the *expected* total reward that the learner collects over $H - h$ steps by running policy π from step h on m. Let π^* be an optimal policy for m, such that $V_1^{m,\pi^*}(s) \geq V_1^{m,\pi}(s)$ for all policies π. The performance of an online

learner is measured by its *regret* with respect to the sequence of optimal policies over K episodes:

$$\mathrm{MTRegret(K)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{V}_1^{\mathrm{m^k},*}(\mathrm{s}_1^{\mathrm{k}}) - \mathrm{V}_1^{\mathrm{m^k},\pi_\mathrm{k}}(\mathrm{s}_1^{\mathrm{k}}),$$

where s_1^k is the initial state in episode k. Intuitively, this notion of regret measures the expected loss in rewards that the learner incurs due to not knowing the identity of $(m^k)_{k=1,2,\ldots,K}$ and their transition kernels.

Remark 1.1.1. We recall the example of treating a sequence of patients in an automatic diagnostic system. Each patient is an MDP, whose states are the conditions of the patient and the actions are the drugs that a doctor can give the patient at a time period. There are finite number of treatment period ($H < \infty$), and the drug at time h can impact the subsequent states and effectiveness ("reward") in time $h+1, h+2, \ldots, H$, similar to an MDP.

Remark 1.1.2. A related problem is non-stationary reinforcement learning with a varying budget [Mao et al., 2021], where the learner also encounters a sequence of MDPs with different transition kernels. In this problem, the budget $Q \geq 0$ controls the maximum value of the total differences between two consecutive episodes can be, where the differences between two transition kernels are measured by the ℓ_1-norm. This varying budget problem is fundamentally different from OMTRL. In OMTRL, because the varying budget Q can be very large (i.e. linear in K) while the number of tasks is small, it is important to transfer knowledge between episodes in order to be robust to a worst-case adversary. As such, keeping a memory of previous interactions in the past episodes is helpful. In contrast, the varying budget problem generally require Q to be small (i.e. sub-linear in K) to obtain non-trivial regret bounds. Moreover, existing work [e.g Mao et al., 2021] has shown that "forgetting" past interactions and implementing a restart mechanism lead to near-optimal regret bounds in the varying budget problem.

1.1.2 Multi-armed Bandits with Sleeping Arms

A conceptually simpler realization of the general non-stationary online learning setup in Algorithm 1.1 is the multi-armed bandits with sleeping arms problem, illustrated in Algorithm 1.3. In this problem, there are at most K actions ("arms") that the learner can choose

Algorithm 1.3 Sleeping Multi-armed Bandits Setup

for round $t = 1, \ldots, T$ **do**
  Adversary picks a set $\mathcal{A}_t \subseteq [K]$ and reveals this set to the learner
  Adversary picks a hidden set of losses values $\ell_t(a)$ for $a \in \mathcal{A}_t$
  Learner takes an action $a_t \in \mathcal{A}_t$ and incurs $\ell_t(a_t)$
end

from in each round of interaction. After choosing an action a_t, the learner observes a loss value $\ell_t(a_t)$.

In standard multi-armed bandits, all of K arms are available to the learner in every round. The goal of the learner is to minimize the cumulative regret over T rounds:

$$\max_{a\in[K]} \mathrm{Regret_T(a)} = \max_{\mathrm{a\in[K]}} \sum_{\mathrm{t=1}}^{\mathrm{T}} (\ell_\mathrm{t}(\mathrm{a_t}) - \ell_\mathrm{t}(\mathrm{a})). \tag{1.1}$$

A more detailed introduction to the standard multi-armed bandits problem is given in Chapter 2. This thesis considers a variant of the standard multi-armed bandits problem, in which an arm may become inactive in certain rounds. In each round t, the learner is given a set of *active* arms $\mathcal{A}_t$, from which the learner can choose one arm $a_t \in \mathcal{A}_t$. Essentially, for an inactive arm $a \notin \mathcal{S}_t$, its loss value in round t is undefined. Therefore, the notion of regret in Equation (1.1) is no longer well-defined. Instead, another notion called *per-action regret* is used, which compares the cumulative loss of the learner to an arm in the rounds where the arm is active:

$$\max_{a\in[K]} \mathrm{PARegret_T(a)} = \max_{\mathrm{a\in[K]}} \sum_{\mathrm{t=1}}^{\mathrm{T}} \mathbb{1}\{\mathrm{a} \in \mathcal{S}_\mathrm{t}\}(\ell_\mathrm{t}(\mathrm{a_t}) - \ell_\mathrm{t}(\mathrm{a})), \tag{1.2}$$

where $\mathbb{1}\{a \in \mathcal{S}_t\} = 1$ if $a \in \mathcal{S}_t$ and 0 otherwise.

Bandits with Sparse Signed Losses

A closely related problem is bandits with sparse signed losses, which was proposed by Kwon and Perchet [2016]. In this problem, the losses are in the range $[-1, 1]$. In each round t, the adversary picks a hidden subset of $\mathcal{S}_t \in [K]$ arms and chooses non-zero losses for these arms, while other arms have 0 losses. The maximum number of non-zero losses in every round is $\max_t |\mathcal{S}_t| \leq S$. Here, S might be also hidden from the learner. Note that all K arms are still available in each round. The main research question is: *Is there an upper bound for the regret* (1.1) *that grows dominantly with* S *instead of* K*?*

While all K arms can be chosen by the learner in each round, effectively there are only at most S arms with non-trivial losses, from which the learner should choose to obtain useful information for minimizing the regret (1.1). From this point of view, bandits with sparse signed losses is more challenging than sleeping bandits as the learner has no knowledge of what the "active" arms are in each round, thus establishing a bound dependent on S instead of K is difficult.

1.2 Contributions and Organizations

Throughout this thesis, we use $\tilde{O}$ to to hide polylogarithmic factors. The main contributions of this thesis are organized as follows.

- Chapter 2 presents background topics on Markov Decision Processes, multi-armed bandits and group distributionally robust optimization.
- Chapter 3 is based on the following publication [Nguyen and Mehta, 2023]:

  Nguyen, Q. and Mehta, N. A. (2023). Adversarial online multi-task reinforcement learning. In International Conference on Algorithmic Learning Theory (ALT).

  This chapter considers the adversarial online multi-task RL problem, in which the number of tasks M is known to the learner, and the MDPs in $\mathcal{M}$ are *communicating*. The formal definition of communicating MDPs is given in Chapter 2. The learner's objective is to minimize MTRegret(K). Most existing works are limited by strong assumptions such as the sequence of tasks being drawn from a fixed distribution, or that the tasks are separated in every state-action pair. Addressing these main theoretical challenges is important towards understanding this setting.

  Contributions.

  - We prove a minimax lower bound of $\Omega(K\sqrt{DSAH})$ on the regret of any learning algorithm. We then show that for sufficiently small horizon H, running standard single-episode RL algorithms achieves near-matching regret upper bound. This indicates that the horizon H must be sufficiently long for knowledge transfer between episodes.
  - We introduce a new task-separability condition called λ-separability and show that this notion generalizes a large number of prior task-separability notions

from previous works. Assuming that the MDPs in $\mathcal{M}$ are well-separated under λ-separability, we prove an instance-specific lower bound of $\Omega(\frac{K}{\lambda^2})$ in sample complexity for a class of *uniformly good* cluster-then-learn algorithms. This class of algorithms includes state-of-the-art prior approaches for multi-task RL. We use a novel construction called *2-JAO MDP* for proving this instance-specific lower bound.

– The lower bounds are complemented with a polynomial time algorithm that obtains $\tilde{O}(\frac{K}{\lambda^2})$ sample complexity guarantee for the clustering phase and $\tilde{O}(\sqrt{MK})$ regret guarantee for the learning phase, indicating that the dependency on K and $\frac{1}{\lambda^2}$ is tight.

- Chapter 4 is based on the following publication [Nguyen and Mehta, 2024]:

  Nguyen, Q. and Mehta, N. A. (2024). Near-Optimal Per-Action Regret Bounds for Sleeping Bandits. In International Conference on Artificial Intelligence and Statistics (AISTATS).

  This chapter considers the fully-adversarial variant of the sleeping multi-armed bandits problem, where both the set of available arms and their losses in every round are determined by the adversary. In a setting with K total arms and at most A available arms in each round over T rounds, the best known per-action regret upper bound is $O(K\sqrt{TA\ln K})$, obtained indirectly via minimizing the internal sleeping regret [Gaillard et al., 2023]. Compared to the minimax $\Omega(\sqrt{TA})$ lower bound, this upper bound contains an extra multiplicative factor of $K\ln K$.

  Contributions.

  – Using generalized versions of algorithms for non-sleeping MABs with a fixed set of arms, we directly minimize the per-action regret and obtain near-optimal bounds of order $O(\sqrt{TA\ln K})$ and $O(\sqrt{T\sqrt{AK}})$. These upper bounds reduce to their static counterparts if the set of available arms is fixed (i.e. $A = K$).

  – We extend our results to the setting of bandits with advice from sleeping experts, generalizing EXP4 along the way. This leads to new proofs for a number of existing adaptive and tracking regret bounds for standard non-sleeping bandits.

  – Extending our results to the bandit version of experts that report their confidences leads to new bounds for the confidence regret that mirror the confidence regret bounds in the full-information setting.

  - We prove a strongly adaptive per-action regret lower bound, showing that for any minimax optimal algorithm, there exists an arm whose regret is sublinear in T but linear in the number of its active rounds.

- Chapter 5 is based on the following publication [Nguyen et al., 2025b]:

  Nguyen, Q., Mehta, N. A., and Guzmán, C. (2025b). Beyond minimax rates in group distributionally robust optimization via a novel notion of sparsity. In International Conference on Machine Learning (ICML).

  This chapter applies the near-optimal per-action regret bounds from Chapter 4 to design new, adaptive algorithms with problem-dependent sample efficiency for the Group Distributionally Robust Optimization (GDRO) problem. GDRO [Sagawa et al., 2020] is an emerging paradigm for training deep neural networks, focusing on improving robustness of neural network models on imbalanced datasets. A formal introduction to GDRO is deferred to Chapter 2. In essence, GDRO is formulated as a stochastic min-max optimization problem over K sample distributions $(P_i)_{i=1,2,\dots,K}$:

$$\min_{\theta\in\Theta} \max_{i\in[K]} \mathbb{E}_{z\sim P_i}[\ell(z)],$$

  where Θ is a hypothesis set and $\ell : \mathcal{Z} \to \mathbb{R}_+$ is a loss function defined on the sample space $\mathcal{Z}$. Theoretical work on GDRO generally considers a setup in which Θ is a bounded convex set and ℓ is a bounded, Lipschitz and convex function. In this setup, the minimax sample complexity of GDRO has been determined up to a $\log(K)$ factor.

  Contributions.

  - We go beyond the minimax perspective via a novel notion of sparsity that we call (λ, β)-sparsity. This condition states that at any parameter θ, there is a set of at most β groups whose risks at θ are all at least λ larger than the risks of the other groups. We show that this notion of sparsity exists in a large number of important machine learning problems in both theory (e.g. Gaussian linear regression) and practice (e.g. fine-grained object classification).
  - Given a target accuracy ϵ, we show a novel algorithm for finding an ϵ-optimal θ. Our analysis reveals that the ϵ-dependent term in the sample complexity can swap a linear dependence on K for a linear dependence on the potentially much smaller β. This improvement leverages recent progress in sleeping bandits, showing a fundamental connection between the two-player zero-sum game optimization

framework for GDRO and per-action regret bounds in sleeping bandits.

  - Next, we show an adaptive algorithm which, up to logarithmic factors, obtains a sample complexity bound that adapts to the best (λ, β)-sparsity condition. We also show how to obtain a dimension-free semi-adaptive sample complexity bound with a computationally efficient method.
  - Finally, we demonstrate the practicality of the (λ, β)-sparsity condition and the improved sample efficiency of our algorithms on both synthetic and real-life datasets.

- Chapter 6 is based on the following publication [Nguyen et al., 2025a]:

  Nguyen, Q., Ito, S., Komiyama, J., and Mehta, N. A. (2025a). Data-dependent bounds with T-optimal best-of-both-worlds guarantees in multi-armed bandits using stability-penalty matching. In the Conference on Learning Theory (COLT).

  This chapter first considers the multi-armed bandits with sparse signed losses problem. As mentioned earlier, this setting could be seen as a more challenging variant of sleeping bandits in that the set of arms with informative feedback is unknown. Prior results on this problem are either restricted to unrealistic assumptions such as the sparsity level S being known, or sub-optimal bounds with additional $\ln T$ dependency while incurring expensive computational overhead. This chapter fully addresses this gap by proposing a new algorithm that obtains T-optimal and S-agnostic regret bounds. Moreover, this new algorithm is extended to become a general strategy for adaptively tuning the learning rates in a much broader class of online learning problems with bandit feedback,thereby achieving optimal adaptivity to various unknown, problem-dependent quantities of the underlying environment.

  Contributions.

  - We propose real-time stability-penalty matching (SPM), a new method for obtaining regret bounds that are simultaneously data-dependent, best-of-both-worlds and T-optimal for multi-armed bandits problems. Our approach extends the SPM method originally proposed by Ito et al. [2024] to establish regret bounds based on real-time observed losses. The original SPM method relied on the maximum possible value of the losses (e.g. 1) to tune the learning rates, so its adaptivity is limited. Our method overcomes this limitation by using the observed loss of the chosen arm to tune the learning rates. We derive a general procedure for effectively bounding the regret in real-time SPM.

– Next, we apply that general regret bound on bandits with sparse signed losses. In particular, we show that real-time SPM obtains bounds with worst-case guarantees of order $\tilde{O}(\sqrt{TS})$ without knowing S. If the underlying environment is stochastic (details are in Chapter 2), our algorithm simultaneously obtains an $O(\frac{S \ln T \ln(K)}{\Delta_{\min}})$ regret bound in this regime. We extend our approach to obtaining best-of-both-worlds adaptive regret bound for losses with small total variations Q, where the adversarial bound is $O(\sqrt{Q \ln(K)})$. Our results do not require any sophisticated doubling trick to estimate Q.

– We introduce a coordinate-wise version of real-time SPM that maintains a separate learning rate for each arm, similar to AdaGrad [Duchi et al., 2011]. We prove that this version of real-time SPM also obtains T-optimal adaptive bounds in the adversarial regime and an improved, more fine-grained bound in the stochastic regime.

Chapter 2

Background Topics

2.1 Markov Decision Processes

One of the most important mathematical frameworks for modelling sequential decision making problems is Markov Decision Processes (MDPs). MDPs came out of a field called operation research in the 1950s [Puterman, 1994; Sutton and Barto, 2018] and has been one of the most active research topics since then. The popularity of MDPs is a result of its versatility in modelling a broad range of practical applications in fields such as engineering, finance, sciences and medical research.

A finite-horizon stationary MDP is specified by a tuple $m = (\mathcal{S}, \mathcal{A}, P, r, H)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition function where $P(s'|s, a)$ specifies the probability of being in state s' after taking action a at state s, $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the reward function specifying the numerical gain by taking an action in a state, and H is the horizon.

Given an MDP m, a learner interacts with m by taking H actions and collecting rewards after each action. In step $h = 1, 2, \ldots, H$, the learner first observes its state s_h, takes an a_h, then moves to a new state s_{h+1} and receives a reward $r_h := r(s_h, a_h)$ as a result of its action. The total reward of the learner is

$$R = \sum_{h=1}^{H} r_h.$$

The objective of the learner is to maximize the expected value of the total reward $\mathbb{E}_{(s_h, a_h, r_h)_{h\in[H]}}[R]$. Details are given in Algorithm 2.1. The expectation is taken over two sources of randomness. The first source of randomness is the state-transitions: the probability of moving to the new

Algorithm 2.1 Single-Task Episodic Reinforcement Learning

for episode $k = 1, 2, \ldots, K$ **do**
  Adversary chooses an initial state s_1^k
  History $\mathcal{H}^k = \{s_1^k\}$
  Learner computes policy π^k
  for step $h = 1, 2, \ldots, H$ **do**
    Learner takes $a_h^k \sim \pi_h^k(\cdot \mid \mathcal{H}^k)$
    Learner observes state s_{h+1}^k and reward $r_h^k = r(a_h^k, a_h^k)$
    Update $\mathcal{H}^k = \mathcal{H}^k \cup \{a_h^k, r_h^k, s_{h+1}^k\}$
  end
end

state s_{h+1} is determined by $P(s_{h+1} \mid s_h, a_h)$. The second source of randomness is the process of choosing the action a_h in state s_h. Let $A = |\mathcal{A}|$ be the size of the action set. A policy π defines how the learner chooses its action given the past states, actions and rewards. Each policy can be represented as a conditional probability distribution over the action space $\mathcal{A}$, and the action in each step is drawn from this distribution. Let s_h^k, a_h^k denote the state and action of the learner in episode k at time step h. The first action $a_1^k \sim \pi^k(\cdot \mid s_1^k)$ is decided based on the initial state s_1^k. The second action $a_2^k \sim \pi^k(\cdot \mid s_1^k, a_1^k, s_2^k)$ is decided based on the history (s_1^k, a_1^k, s_2^k), and so on.

The performance of a policy π is measured by its *value functions* $(V_h^\pi)_{h=1,2,\ldots,H}$, where

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r'_h \mid s_h = s, a'_h \sim \pi(\cdot \mid (s_i, a_i, r_i)_{i=1,2,\ldots,h'-1}, s_{h'})\right].$$

In essence, $V_h^\pi(s)$ specifies the *expected* total reward that the learner collects over $H - h + 1$ steps by running policy π from step h on m. Let π^* be a policy such that for all policies π, $V_1^{\pi^*}(s) \geq V_1^\pi(s)$ for all $s \in \mathcal{S}$ (such a π^* always exists). We call π^* an optimal policy for m, and write V^* for the optimal value functions.

In (single-task) episodic reinforcement learning, the learner's goal is finding an optimal *policy* π^* when the transition kernels P and/or the reward function r are unknown. The performance of an online learner is measured by its *regret* with respect to the sequence of optimal policies over K episodes:

$$\text{Regret}(K) = \sum_{k=1}^{K} [V_1^* - V_1^{\pi^k}](s_1^k), \tag{2.1}$$

where π^k is the policy of the learner in episode k.

Algorithm 2.2 Multi-armed Bandits Setup

Input: Number of arms $K \geq 1$, number of rounds $T \geq 1$
for round $t = 1, \ldots, T$ **do**
  Adversary picks a hidden loss vector $\ell_t \in \mathbb{R}^K$
  Learner takes an action $I_t \in [K]$ and incurs ℓ_{t,I_t}
end

Remark 2.1.1. For K sufficiently large, Azar et al. [2017] showed that a computationally efficient algorithm called UCBVI obtains a regret upper bound of order $\tilde{O}(H\sqrt{SAK})$, where $\tilde{O}$ hides a logarithmic factor. Except for this logarithmic factor, this upper bound matches the lower bound in Jaksch et al. [2010].

2.2 Multi-armed Bandits

Multi-armed bandits [Lai and Robbins, 1985; Auer et al., 2002b] is a fundamental online learning framework that has found applications in a large number of practical sequential decision making problems such as resource allocation, online advertising, product testing and clinical trials. Algorithm 2.2 illustrates the interactive rounds between the learner and the environment. In each round, there are K arms from which the learner can choose. In rounds $t = 1, 2, \ldots, T$, an adversary (which is part of the environment) chooses a loss value $\ell_{t,a}$ for arm $a = 1, 2, \ldots, K$. The loss vector is $\ell_t \in \mathbb{R}^K$. Initially, all K loss values are hidden from the learner. The learner can choose to pull an arm $I_t \in [K]$ and observe its loss ℓ_{t,I_t}, while the losses of other, non-chosen arms remain hidden. This is a special version of the sleeping bandits problem presented in Algorithm 1.3, where the set of active arms is $\mathcal{A}_t = [K]$ for all rounds t. In other words, the standard multi-armed bandits problem is a "static" problem, in which there are no sleeping arms. The goal of the learner is to minimize its regret with respect to every arm, i.e., bounding

$$\max_{a \in [K]} R_T(a) = \max_a \sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,a}. \tag{2.2}$$

Note that (2.2) is a restatement of (1.1). More importantly, (2.2) is equal to the per-action regret in (1.2) with $\mathcal{A}_t = [K]$, as $\mathbb{1}\{a \in \mathcal{A}_t\} = 1$ for all $a \in [K]$ and $t \in [T]$ (all arms are always active).

Stochastic and Adversarial Bandits. Depending on the statistical assumption of the losses, multi-armed bandits problems are often divided into two types: stochastic and

adversarial. In stochastic bandits, the losses $(\ell_{t,i})_{t\in[T]}$ are generated from a fixed distribution P_i with finite mean μ_i and variance σ_i. Existing work [Lai and Robbins, 1985; Ito et al., 2022] have shown tight $\Theta(\ln T \sum_{i=2}^{K} \frac{\sigma_i}{\mu_i - \mu_1})$ worst-case regret bounds for this setting (assuming, without loss of generality, that arm 1 is the unique best arm).

On the other hand, in adversarial bandits, no statistical assumption is made on how the sequences of losses $(\ell_{t,i})_{t\in[T]}$ is generated. In particular, $\ell_{t,i}$ may even depend on the learner's past choices. Existing work [Auer et al., 2002b; Zimmert and Seldin, 2021a] has shown tight $\Theta(\sqrt{TK})$ worst-case regret bound for this setting.

For stochastic bandits, the optimal worst-case guarantee grows as a logarithmic function in T. On the contrary, the optimal worst-case guarantee in adversarial bandits grows as $\sqrt{T}$. Originally, these bounds were obtained by two very different approaches, both relying on knowing which type of environment the learner is on beforehand. An important line of research in the last 15 years [e.g. Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021a; Dann et al., 2023; Ito, 2021] is deriving best-of-both-worlds algorithms that can simultaneously and automatically achieve the optimal worst-case guarantees on both stochastic and adversarial bandits without knowing which environment the data is coming from. The most successful best-of-both-worlds algorithms so far are based on the follow-the-regularized-leader (FTRL) framework, which proved to be effective in obtaining adaptive $O(\ln(T))$ and $O(\sqrt{T})$ dependency on T [e.g. Zimmert and Seldin, 2021a].

Adapting to Easy Loss Sequences in Adversarial Bandits. Another set of important adaptivity results is adapting to easy loss sequences in adversarial bandits. An extreme example is if the losses in every round are the same: $\ell_{t,i} = \ell_{1,i}$ for all arms $i \in [K]$ and rounds $t \in [T]$, in which case a straightforward algorithm that pulls each arm once for K rounds and then keeps choosing the best arm for the remaining rounds would incur a regret of at most $O(K)$. Another example is if the total loss of the best arm is much smaller than that of other arms, in which case it should also not take long to the learner to realize which arm is the best arm. Intuitively, the sequence of losses represents easy problem instances in both examples, and so worst-case guarantees might be too pessimistic. Constructing algorithms that can adapt to both easy- and worst-case problem instances for adversarial bandits has also been an important research area in the last decade, with recent results showing greater successes in adapting to various quantities of the loss sequences such as total variation, sparsity and small losses [e.g. Cesa-Bianchi and Lugosi, 2006; Hazan and Kale, 2011; Tsuchiya et al., 2023; Ito et al., 2022].

2.3 Group Distributionally Robust Optimization

This section presents the basic setup of the group distributionally robust optimization problem (GDRO) [Sagawa et al., 2020]. In GDRO, given a convex hypothesis set Θ, there are $K \geq 2$ convex risk functions $R_1, R_2, \ldots, R_K$, where $R_i : \Theta \to \mathbb{R}$. The worst-case risk of a hypothesis θ is $\max_{i\in[K]} R_i(\theta)$. The optimal hypothesis θ^* is the one with minimal worst-case risk:

$$\theta^* = \arg\min_{\theta\in\Theta} \max_{i\in[K]} R_i(\theta). \tag{2.3}$$

In practice, it is often not possible to find θ^* exactly (e.g. R_i is a population risk). Instead, a more practical goal is finding a $\bar{\theta}$ whose worst-case risk is *close* to that of θ^*. For a target accuracy ϵ, GDRO aims to find an ϵ-optimal $\bar{\theta}$ such that $\mathrm{err}(\bar{\theta}) \leq \epsilon$, where

$$\mathrm{err}(\bar{\theta}) = \max_{i\in[K]} R_i(\bar{\theta}) - \max_{i\in[K]} R_i(\theta^*) \tag{2.4}$$

is the optimality gap of a $\bar{\theta} \in \Theta$.

Remark 2.3.1. A practical example of GDRO is in drug testing, where Θ is the space of all testable drugs and K is the number of groups of targeted patients. These patients may differ in various demographic factors such as age, existing conditions, gender, etc. Finding θ^* corresponds to finding a drug that is safe for all patient groups, in the sense that its maximum risk (across all groups) is minimized. Thus, GDRO can also be motivated by a fairness measure.

Two-Player Zero-Sum Game Approach.

The two-player zero-sum game approach is the standard approach for solving GDRO problems [Sagawa et al., 2020; Zhang et al., 2023a]. The main idea is to convert the min-max problem in (2.3) to a stochastic convex-linear minimax problem, and then model the optimization process as a game between a min-player and a max-player.

Let $\Delta_K = \{q \in \mathbb{R}^K : q_i \geq 0, \sum_{i=1}^K q_i = 1\}$ be the K-dimensional simplex. For any $q \in \Delta_K$, let $\phi(\theta, q) = \sum_{i=1}^K q_i R_i(\theta)$ be the weighted sum of the risks of θ over K groups. Since ϕ is linear in q and the simplex Δ_K is convex, we have $\max_{i\in[K]} R_i(\theta) = \max_{q\in\Delta_K} \phi(\theta, q)$. It

Algorithm 2.3 Two-player Zero-Sum Game

for round $t = 1, \ldots, T$ **do**
  Min-player plays θ_t
  Max-player plays q_t
  Min-player incurs $\phi(\theta_t, q_t)$
  Max-player incurs $-\phi(\theta_t, q_t)$
end
Return: $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$ and $\bar{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$

follows that θ^* is part of the solution of the following problem:

$$\min_{\theta \in \Theta} \max_{q \in \Delta_K} \phi(\theta, q).$$

Next, let

$$\text{err}(\bar{\theta}, \bar{q}) = \max_{q \in \Delta_K} \phi(\bar{\theta}, q) - \min_{\theta \in \Theta} \phi(\theta, \bar{q})$$

be the duality gap of $\bar{\theta} \in \Theta$ and $\bar{q} \in \Delta_K$. Since $\max_{i \in [K]} R_i(\theta) \geq \phi(\theta, \bar{q})$ for all θ and $\bar{q}$, we have $\text{err}(\bar{\theta}) \leq \text{err}(\bar{\theta}, \bar{q})$. Therefore, to find a ϵ-optimal $\bar{\theta}$, it suffices to find a $\bar{\theta}$ and $\bar{q}$ such that $\text{err}(\bar{\theta}, \bar{\text{q}}) \leq \epsilon$. To this end, the two-player zero-sum game approach runs the game between a min-player and a max-player, illustrated in Algorithm 2.3. The theoretical justification for this approach is given in the following lemma.

Lemma 2.3.2. Define the regrets of the two players in Algorithm 2.3 over T rounds by

$$R_{\mathcal{A}_\theta} = \max_{q \in \Delta_K} \sum_{t=1}^{T} \phi(\theta_t, q) - \phi(\theta_t, q_t),$$
$$R_{\mathcal{A}_q} = \min_{\theta \in \Theta} \sum_{t=1}^{T} \phi(\theta, q_t).$$

Let $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$ and $\bar{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$. Their optimality gap is bounded by

$$\text{err}(\bar{\theta}, \bar{q}) \leq \frac{1}{T}\left(R_{\mathcal{A}_\theta} + R_{\mathcal{A}_q}\right).$$

Proof. By Jensen's inequality, we have

$$\begin{aligned}
\mathrm{err}(\bar{\theta}, \bar{q}) &= \max_{q \in \Delta_K} \phi(\bar{\theta}, q) - \min_{\theta \in \Theta} \phi(\theta, \bar{q}) \\
&= \max_{q \in \Delta_K} \phi\left(\frac{1}{T}\sum_{t=1}^{T} \theta_t, q\right) - \min_{\theta \in \Theta} \phi\left(\theta, \frac{1}{T}\sum_{t=1}^{T} q_t\right) \\
&\leq \frac{1}{T} \max_{q \in \Delta_K} \left(\sum_{t=1}^{T} \phi(\theta_t, q)\right) - \frac{1}{T} \min_{\theta \in \Theta} \left(\sum_{t=1}^{T} \phi(\theta, q_t)\right) \qquad (2.5) \\
&= \frac{1}{T} \left(\max_{q \in \Delta_K} \sum_{t=1}^{T} \phi(\theta_t, q) - \phi(\theta_t, q_t)\right) + \frac{1}{T} \left(\phi(\theta_t, q_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} \phi(\theta, q_t)\right) \\
&= \frac{1}{T} \left(R_{\mathcal{A}_\theta} + R_{\mathcal{A}_q}\right).
\end{aligned}$$

□

As a result, solving the group distributionally robust optimization reduces to solving two adversarial online learning problems.

Chapter 3

Adversarial Online Multi-Task Reinforcement Learning

3.1 Introduction

The majority of theoretical works in online reinforcement learning (RL) have focused on single-task settings in which the learner is given the same task in every episode. In practice, an autonomous agent might face a sequence of different tasks. For example, an automatic medical diagnosis system could be given an arbitrarily ordered sequence of patients who are suffering from an unknown set of variants of a virus. In this example, the system needs to classify and learn the appropriate treatment for each variant of the virus. This example is an instance of the adversarial online multi-task episodic RL setting, an important learning setting for which the theoretical understanding is rather limited. The framework commonly used in existing theoretical works is an episodic setting of K episodes; in each episode an unknown Markov decision process (MDP) from a finite set $\mathcal{M}$ of size M is given to the learner. When $M = 1$, the setting reduces to single-task episodic RL. Most existing algorithms for single-task episodic RL are based on aggregating samples in all episodes to obtain sub-linear bounds on various notions of regret [Azar et al., 2017; Jin et al., 2018; Simchowitz and Jamieson, 2019] or finite (ϵ, δ)-PAC bounds on the sample complexity of exploration [Dann and Brunskill, 2015]. When $M > 1$, without any assumptions on the common structure of the tasks, aggregating samples from different tasks could produce negative transfer [Brunskill and Li, 2013]. To avoid negative transfer, existing works [Brunskill and Li, 2013; Hallak et al., 2015; Kwon et al., 2021] assumed that there exists some notion of task-separability that defines how different the tasks in $\mathcal{M}$ are. Based on this notion of separability, most

existing algorithms followed a two-phase cluster-then-learn paradigm that first attempts to figure out which MDP is being given and then uses the samples from the previous episodes of the same MDP for learning. However, most existing works employ strong assumptions such that the tasks are given stochastically following a fixed distribution [Azar et al., 2013; Brunskill and Li, 2013; Steimle et al., 2021; Kwon et al., 2021] or the task-separability notion allows the MDPs to be distinguished in a small number of exploration steps [Hallak et al., 2015; Kwon et al., 2021]. These strong assumptions become the main theoretical challenges towards understanding this setting.

Our goal in this work is to study the adversarial setting with a more general task-separability notion, in which the aformentioned strong assumptions do not hold. Specifically, the learner makes no statistical assumptions on the sequence of tasks; the task in each episode can be either the same or different from the tasks in any other episodes. Moreover, the difference between the tasks in two consecutive episodes can be large (linear in the length of the episodes) so that algorithms based on a fixed budget for total variation such as RestartQ-UCB [Mao et al., 2021] cannot be applied. The performance of the learner is measured by its regret with respect to an omniscient agent that knows which tasks are coming in every episode and the optimal policies for these tasks. We consider the same cluster-then-learn paradigm of the previous works and focus on the following two questions:

- *Is there a task-separability notion that generalizes the notions from previous works while still enabling tasks to be distinguished by a cluster-then-learn algorithm with polynomial time and sample complexity? If so, what is the optimal sample complexity of clustering under this notion?*

- *Is there a polynomial time cluster-then-learn algorithm that simultaneously obtains near-optimal sample complexity in the clustering phase and near-optimal regret guarantee for the learning phase in the adversarial setting?*

We answer both questions positively. For the first question, we introduce the notion of λ-separability, a task-separability notion that generalizes the task-separability definitions in previous works in the same setting [Brunskill and Li, 2013; Hallak et al., 2015; Kwon et al., 2021]. Definition 3.3.1 formally defines λ-separability. A more informal version of λ-separability has appeared in the discounted setting of Concurrent PAC RL [Guo and Brunskill, 2015] where multiple MDPs are learned concurrently; however the implications on the episodic sequential setting and the tightness of their results were lacking. In essence, λ-separability assumes that between every pair of MDPs in $\mathcal{M}$, there exists some state-action pair whose transition functions are well-separated in ℓ_1-norm. This setting is more

challenging than the one considered by [Hallak et al., 2015] where *all* state-action pairs are well-separated. In Appendix 3.A, we show that λ-separability is more general than the entropy-based separability defined in [Kwon et al., 2021] and thus requires novel approaches to exploring and clustering samples from different episodes. Under this notion of λ-separability, we show an instance-specific lower bound[1] $\Omega(\frac{K}{\lambda^2})$ on both the sample complexity and regret of the clustering phase for a class of cluster-then-learn algorithms that includes most of the existing works.

To answer the second question, we propose a new cluster-then-learn algorithm, AO-MultiRL, which obtains a regret upper bound of $\tilde{O}\left(\frac{K}{\lambda^2} + \sqrt{MK}\right)$ (the $\tilde{O}$ hides logarithmic terms). This upper bound indicates that the linear dependency on K and λ^2 in the lower bounds are tight. The $\tilde{O}(\sqrt{MK})$ upper bound in the learning phase is near-optimal because if the identity of the model is revealed to a learner at the beginning of every episode (so that no clustering is necessary), there exists a straightforward $\Omega(\sqrt{MK})$ lower bound obtained by combining the lower bound for the single-task episodic setting of [Domingues et al., 2021b] and Cauchy-Schwarz inequality. In the stochastic setting, the L-UCRL algorithm [Kwon et al., 2021] obtains $O(\sqrt{MK})$ regret with respect to the optimal policy of a partially observable MDP (POMDP) setting that does not know the identity of the MDPs in each episode; thus their notion of regret is weaker than the one in our work.

Overview of Techniques

- In Section 3.4, we present two lower bounds. The first is a minimax lower bound $\Omega(K\sqrt{SAH})$ on the total regret of any algorithm. This result uses the construction of JAO MDPs in [Jaksch et al., 2010]. The second is a $\Omega\left(\frac{K}{\lambda^2}\right)$ instance-specific lower bound on the sample complexity and regret of the clustering phase for a class of *uniformly good* cluster-then-learn algorithms when both λ and M are sufficiently large. The instance-specific lower bound relies on the novel construction of *2-JAO MDP*, a hard instance combining two JAO MDPs in which one is the minimax lower bound instance and the other satisfies λ-separability. We show that learning 2-JAO MDPs is fundamentally a two-dimensional extension of the problem of finding a biased coin among a collection of fair coins [e.g. Tulsiani, 2014], for which information theoretic techniques of the one-dimensional problem can be adapted.
- In Section 3.5, we show that AOMultiRL obtains a regret upper bound of $\tilde{O}\left(\frac{K}{\lambda^2} + \sqrt{MK}\right)$.

[1] Here and throughout the introduction, we suppress factors related to the MDPs such that the number of states and actions and the horizon length in all the bounds.

The main idea of AOMultiRL is based on the observation that a fixed horizon of order $\Theta(\frac{1}{\lambda^2})$ with a small constant factor is sufficient to obtain a λ-dependent coarse estimate of the transition functions of all state-action pairs. In turn, this coarse estimate is sufficient to have high-probability guarantees for the correctness of the clustering phase. This allows AOMultiRL to have a fixed horizon for the learning phase and be able to apply single-task RL algorithms with theoretical guarantees such as UCBVI-CH [Azar et al., 2017] in the learning phase.

Our paper is structured as follows: Section 3.3 formally sets up the problem. Section 3.4 presents the lower bounds. AOMultiRL and its regret upper bound are shown in Section 3.5. Several numerical simulations are in Section 3.6. The appendix contains formal proofs of all results.

3.2 Related Work

Stochastic Online Multi-task RL. The Finite-Model-RL algorithm [Brunskill and Li, 2013] considered the stochastic setting with infinite-horizon MDPs and focused on deriving a sample complexity of exploration in a (ϵ, δ)-PAC setting. As shown by [Dann et al., 2017], even an optimal (ϵ, δ)-PAC bound can only guarantee a necessarily sub-optimal $O(K_m^{2/3})$ regret bound for each task $m \in [M]$ that appears in K_m episodes, leading to an overall $O(M^{1/3}K^{2/3})$ regret bound for the learning phase in the multi-task setting.

The Contextual MDPs algorithm by [Hallak et al., 2015] is capable of obtaining a $O(\sqrt{K})$ regret bound in the learning phase after the right cluster has been identified; however their clustering phase has exponential time complexity in K. The recent L-UCRL algorithm [Kwon et al., 2021] considered the stochastic finite-horizon setting and reduced the problem to learning the optimal policy of a POMDP. Under a set of assumptions that allow the clusters to be discovered in $O(\text{polylog}(MSA))$, L-UCRL is able to obtain an overall $O(\sqrt{MK})$ regret with respect to a POMDP planning oracle which aims to learn a policy that maximizes the expected single-task return when a task is randomly drawn from a known distribution of tasks. In contrast, our work adopts a stronger notion of regret that encourages the learner to maximize its expected return for a sequence of tasks chosen by an adversary. When the models are bandits instead of MDPs, [Azar et al., 2013] use spectral learning to estimate the mean reward of the arms in all models and obtains an upper bound linear in K.

Lifelong RL. Learning a sequence of related tasks is more well-studied in the lifelong

learning literature. Recent works in lifelong RL [Abel et al., 2018; Lecarpentier et al., 2021] often focus on the setting where tasks are drawn from an unknown distribution of MDPs and there exists some similarity measure between MDPs that support transfer learning. Our work instead focuses on learning the dissimilarity between tasks for the clustering phase and avoiding negative transfer.

Active model estimation The exploration in AOMultiRL is modelled after the active model estimation problem [Tarbouriech et al., 2020], which is often presented in PAC-RL setting. Several recent works on active model estimation are PAC-Explore [Guo and Brunskill, 2015], FW-MODEST [Tarbouriech et al., 2020], $\beta-$curious walking [Sun and Huang, 2020], and GOSPRL [Tarbouriech et al., 2021].

The $\Theta(\tilde{D}|\Gamma^{\alpha}|N)$ bound on the horizon of clustering in Lemma 3.5.3 has the same $O(S^2A)$ dependency on the number of states and actions as the state-of-the-art bound by GOSPRL [Tarbouriech et al., 2021] for the active model estimation problem. The main drawback is that H_0 depends linearly on the hitting time $\tilde{D}$ and not the diameter D of the MDPs. As the hitting time is often strictly larger than the diameter [Jaksch et al., 2010; Tarbouriech et al., 2021], this dependency on $\tilde{D}$ is sub-optimal. On the other hand, AOMultiRL is substantially less computationally expensive than GOSPRL since there is no shortest-path policy computation involved.

3.3 Problem Setup

Our learning setting consists of K episodes. In episode $k = 1, 2, \ldots, K$, an adversary chooses an unknown Markov decision process (MDP) m^k from a set of finite-horizon tabular stationary MDP models $\mathcal{M} = \{(\mathcal{S}, \mathcal{A}, H, P_i, r) : i = 1, 2, \ldots, M\}$ where $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the shared reward function, $\mathcal{S}$ is the set of states with size S, $\mathcal{A}$ is the set of actions with size A, H is the length of each episode, and $P_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition function where $P_i(s'|s, a)$ specifies the probability of being in state s' after taking action a at state s. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are known and shared between all models; however, the transition functions are distinct and unknown. Following a common practice in single-task RL literature [Azar et al., 2017; Jin et al., 2018], we assume that the reward function is known and deterministic, however our techniques and results extend to the setting of unknown stochastic r. Furthermore, the MDPs are assumed to be communicating with a finite diameter D [Jaksch et al., 2010]. A justification for this assumption on the diameter is provided in Section 3.3.1.

The adversary also chooses the initial state s_1^k. The policy π_k of the learner in episode k is a collection of H functions $\pi^k = \{\pi_{k,h} : \mathcal{S} \mapsto \mathcal{A}\}$, which can be non-stationary and history-dependent. The value function of π_k starting in state s at step h is the expected rewards obtained by following π_k for $H - h + 1$ steps $V_h^{\pi_k}(s) = \mathbb{E}[\sum_{h'=h}^{H} r(s_{h'}^k, \pi_{k,h}(s_{h'}^k)) \mid s_h^k = s]$, where the expectation is taken with respect to the stochasticity in m^k and π^k. Let $V_1^{k,*}$ denote the value function of the optimal policy in episode k.

The performance of the learner is measured by its regret with respect to the optimal policies in every episode:

$$\text{Regret}(K) = \sum_{k=1}^{K} [V_1^{k,*} - V_1^{\pi_k}](s_1^k). \tag{3.1}$$

Let $[M] = \{1, 2, \ldots, M\}$. We assume that the MDPs in $\mathcal{M}$ are λ-*separable*:

Definition 3.3.1 (λ-separability)**.** Let $\lambda > 0$ and consider set of MDP models $\mathcal{M} = \{m_1, \ldots, m_M\}$ with M models. For all $(i, j) \in [M] \times [M]$ and $i \neq j$, the λ-distinguishing set for two models m_i and m_j is defined as the set of state-action pairs such that the ℓ_1 distance between $P_i(s, a)$ and $P_j(s, a)$ is larger than λ: $\Gamma_{i,j}^{\lambda} = \{(s, a) \in \mathcal{S} \times \mathcal{A} : \|P_i(s, a) - P_j(s, a)\| \geq \lambda\}$, where $\|\cdot\|$ denotes the ℓ_1-norm and $P_i(s, a) = P_i(\cdot \mid s, a)$.

The set $\mathcal{M}$ is λ-separable if for every two models m_i, m_j in $\mathcal{M}$, the set $\Gamma_{i,j}^{\lambda}$ is non-empty:

$$\forall i, j \in [M], i \neq j : \Gamma_{i,j}^{\lambda} \neq \emptyset.$$

In addition, λ is called a separation level of $\mathcal{M}$, and we say a state-action pair (s, a) is λ-distinguishing for two models m_i and m_j if $\|P_i(s, a) - P_j(s, a)\| > \lambda$.

We use the following notion of a λ-distinguishing set for a collection of MDP models $\mathcal{M}$:

Definition 3.3.2 (λ-distinguishing set)**.** Given a λ-separable set of MDPs $\mathcal{M}$, a λ-distinguishing set of $\mathcal{M}$ is a set of state-action pairs $\Gamma^{\lambda} \subseteq \mathcal{S} \times \mathcal{A}$ such that for all $i, j \in [M], \Gamma_{i,j}^{\lambda} \cap \Gamma^{\lambda} \neq \emptyset$. In particular, the set $\Gamma = \cup_{i,j} \Gamma_{i,j}^{\lambda}$ is a λ-distinguishing set of $\mathcal{M}$.

By definition, a state-action pair can be λ-distinguishing for some pairs of models and not λ-distinguishing for other pairs of models.

3.3.1 Assumption on the finite diameter of the MDPs

In this work, all MDPs are assumed to be communicating. We employ the following formal definition and assumption commonly used in literature [Jaksch et al., 2010; Brunskill and Li, 2013; Sun and Huang, 2020; Tarbouriech et al., 2021]:

Definition 3.3.3. ([Jaksch et al., 2010]) Given an ergodic Markov chain $\mathcal{F}$, let $T^{\mathcal{F}}_{s,s'} = \inf\{t > 0 \mid s_t = s', s_0 = s\}$ be the first passage time for two states s, s' on $\mathcal{F}$. Then the hitting time of a unichain MDP G is $T_G = \max_{s,s' \in S} \max_\pi \mathbb{E}[T^{\mathcal{F}_\pi}_{s,s'}]$, where $\mathcal{F}_\pi$ is the Markov chain induced by π on G. In addition, $T'_G = \max_{s,s' \in S} \min_\pi \mathbb{E}[T^{\mathcal{F}_\pi}_{s,s'}]$ is the diameter of G.

Assumption 3.3.1. The diameter of all MDPs in $\mathcal{M}$ are bounded by a constant D.

While this finite diameter assumption is common in undiscounted and discounted single-task setting [Jaksch et al., 2010; Guo and Brunskill, 2015], it is not necessary in the episodic single-task setting [Jin et al., 2018; Mao et al., 2021]. Therefore, it is important to justify this assumption in the episodic multi-task setting. In the episodic single-task setting, for any initial state s_1, the average time between any pair of states reachable from s_1 is bounded $2H$; hence, H plays the same role as D [Domingues et al., 2021b]. This allows the learner to visit and gather state-transition samples in each state multiple times and construct accurate estimates of the model.

However, in the multi-task setting, the same initial state s_1 in one episode might belong to a different MDP than the state s_1 in the previous episodes. Therefore, the set of reachable states and their state-transition distributions could change drastically. Hence, it is important that the λ-distinguishing state-action pairs be reachable from any initial state s_1 for the learner to recognize which MDP it is in and use the samples appropriately. Otherwise, combining samples from different MDPs could lead to negative transfer. Conversely, if the MDPs are allowed to be non-communicating, the component that makes them λ-separable might be unreachable from other components. In this case, the adversary can pick the initial states in these components and block the learner from accessing the λ-distinguishing state-actions. A construction that formalizes this argument is shown at the end of Section 3.4.

3.4 Minimax and Instance-Dependent Lower Bounds

We first show that if λ is sufficiently small and $M = \Theta(SA)$, then the setting is uninteresting in the sense that one cannot do much better than learning every episode individually without any transfer, leading to an expected regret that grows linearly in the number of episodes K.

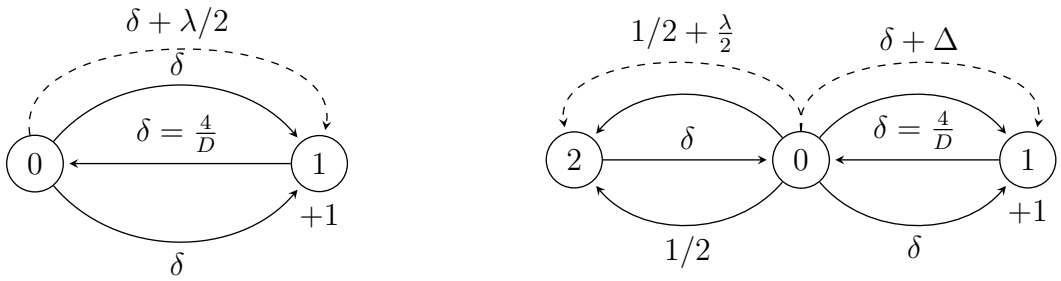

Figure 3.1: A JAO MDP (left) and a 2-JAO MDP (right). Only state 1 has reward +1. The dashed arrows indicate the best actions.

Lemma 3.4.1 (Minimax Lower Bound)**.** Suppose $S, A \geq 10, D \geq 20 \log_A(S)$ and $H \geq DSA$ are given. Let $\lambda = \Theta(\sqrt{\frac{SA}{HD}})$. There exists a set of λ-separable MDPs $\mathcal{M}$ of size $M = \frac{SA}{4}$, each with S states, A actions, diameter at most D and horizon H such that if the tasks are chosen uniformly at random from $\mathcal{M}$, the expected regret of any sequence of policies $(\pi_k)_{k=1,\ldots,K}$ over K episodes is

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(K\sqrt{DSAH}\right).$$

Proof. (Sketch) We construct $\mathcal{M}$ so that each MDP in $\mathcal{M}$ is a JAO MDP [Jaksch et al., 2010] of two states $\{0, 1\}$, $\frac{SA}{4}$ actions and diameter $\frac{D}{4}$. Figure 3.1 (left) illustrates the structure of a JAO MDP. State 0 has no reward, while state 1 has reward +1. Each model has a unique best action a^* that starts from 0 and goes to 1. The pair $(0, a^*)$ is a λ-distinguishing state-action pair.

A JAO MDP can be converted to an MDP with S states, A actions and diameter D, and this type of MDP gives the minimax lower bound proof in the undiscounted setting [Jaksch et al., 2010]. The adversary selects a model from $\mathcal{M}$ uniformly at random, and so previous episodes provide no useful information for the current episode; hence, the regret of any learner is equal to the sum of its K one-episode learning regrets. The one-episode learning regret for JAO MDPs is known to be $\Omega(\sqrt{DSAH})$ when comparing against the optimal infinite-horizon average reward. For JAO MDPs, the optimal infinite horizon policy is also optimal for finite horizon; so, we can use a geometric convergence result from Markov chain theory [Levin et al., 2008] to convert this lower bound to a lower bound of the standard finite-horizon regret of the same order, giving the result. □

Using the same technique in the proof of Lemma 3.4.1, we can show that applying UCRL2 [Jaksch et al., 2010] in every episode individually leads to a regret upper bound of

$O\left(KDS\sqrt{AH\ln H}\right)$. This implies that learning every episode individually already gives a near-optimal regret guarantee.

Remark 3.4.2. Our proof for Lemma 3.4.1 contains a simple yet rigorous proof for the mixing-time argument used in [Mao et al., 2021; Jin et al., 2018]. This argument claims that for JAO MDPs, when the diameter is sufficiently small compared to the horizon, the optimal H-step value function V_1^* in the regret of the episodic setting can be replaced by the optimal average reward $\rho^* H$ in the undiscounted setting without changing the order of the lower bound. To the best of our knowledge, our proof is the first rigorous proof for this argument that applies for any number of episodes including $K = 1$. [Domingues et al., 2021b] provide an alternative proof; however the results therein hold in a different setting where K is sufficiently large and the horizon H can be much smaller than D.

We emphasize that the lower bound in Lemma 3.4.1 holds for *any* learning algorithms. This result motivates the more interesting setting in which λ is a fixed and large constant independent of H. In this case, we are interested in an instance-specific lower bound. For multi-armed bandits, instance-specific lower bounds are constructed with respect to a class of *uniformly good* learning algorithms [Lai and Robbins, 1985]. In our setting, we focus on defining a class of uniformly good algorithms that include the cluster-then-learn algorithms in the previous works for multi-task PAC RL settings such as Finite-Model-RL [Brunskill and Li, 2013] and PAC-EXPLORE [Guo and Brunskill, 2015]. We consider a class of MDPs and a cluster-then-learn algorithm uniformly good if they satisfy an intuitive property: for any MDP in that class, the algorithm should be able to correctly classify whether a cluster of samples is from that MDP or not with an arbitrarily low (but not zero) failure probability, provided that the horizon H is sufficiently long for the algorithm to collect enough samples. The following definition formalizes this idea.

Definition 3.4.1 (PAC identifiability of MDPs)**.** A set of models $\mathcal{M}$ of size M is PAC identifiable if there exists a function $f : (0,1) \mapsto \mathbb{N}$, a sample collection policy π and a classification algorithm $\mathcal{C}$ with the following property: for every $p \in (0,1)$, for each model $1 \leq m \leq M$ in $\mathcal{M}$, if π is run for $f(p)$ steps and the state-transition samples are given to $\mathcal{C}$, then the algorithm $\mathcal{C}$ returns the correct identity of m with probability at least $1-p$, where the probability is taken over all possible sequence of $f(p)$ samples collected by running π on m for $f(p)$ steps. The smallest choice of function $f(p)$ among all possible choices is called the sample complexity of model identification of $\mathcal{M}$.

The clustering algorithm in a cluster-then-learn framework solves a problem different from classification: they only need to tell whether a cluster of samples belong to the same

or different distribution than another cluster of samples, not the identity of the distribution. We can reduce one problem to the other by the following construction: consider the adversary that gives all M models in the first M episodes. After the first M episodes, there are M clusters of samples, each corresponding to one model in $\mathcal{M}$. Once the learner has constructed M different clusters, from the episode $M+1$, the clustering problem is as hard as classification since identifying the right cluster immediately implies the identity of the MDP where the samples come from, and vice versa. Hence, we can apply the sample complexity of classification to that of clustering.

Next, we show the lower bound on the sample complexity of model identification for the class of λ-separable communicating MDPs.

Lemma 3.4.3. For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, \frac{1}{2}]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $\frac{SA}{12}$, each with at most S states, A actions and diameter D such that for any classification algorithm $\mathcal{C}$, if the number of state-transition samples given to $\mathcal{C}$ is less than $\frac{SA}{180\lambda^2}$ then for at least one MDP in $\mathcal{M}$, algorithm $\mathcal{C}$ fails to identify that MDP with probability at least $\frac{1}{2}$.

Proof. (Sketch) The set $\mathcal{M}$ is a set of 2-JAO MDPs, shown in Figure 3.1 (right). Each 2-JAO MDP combines two JAO MDPs with the same number of actions and with diameter in the range $[\frac{D}{2}, D]$; one is λ-separable and one is the hard instance for the minimax lower bound of [Jaksch et al., 2010]. Rewards exist only in the part containing the hard instance. If a learner completely ignores the λ-separable part, by Lemma 3.4.1 the learner cannot do much better than just learning every episode individually. On the other hand, with enough samples from the λ-separable part, the learner can identify the MDP and use the samples collected in the previous episodes of the same MDP to accelerate learning the hard instance part. However, the λ-separable part is also a JAO MDP, for which no useful information from previous episodes can help identify the MDP in the current episode.

Only the actions at state 0 are λ-distinguishing and can be used to identify the MDPs. Taking an action in state 0 can be seen as flipping a coin: heads for transitioning to another state and tails for staying in state 0. Identifying a 2-JAO MDP reduces to the problem of using at most H coin flips to identify, in a $Q \times 2$ matrix of coins, a row j that has coins that are slightly different from the others. The first column has fair coins except in row j, where the success probability is $\frac{1}{2} + \lambda$. The second column coins with success probability of $\delta \leq \frac{1}{4}$ except in row j, where the coin is upwardly biased by $\Delta \leq \lambda$. Lemma 3.B.1 and Corollary 3.B.2 in the appendix show a $\Omega\left(\frac{Q}{\lambda^2}\right)$ lower bound on the number of coin flips on the first column (the left part of the 2-JAO MDP), implying the desired result. □

Lemma 3.4.3 imply that for 2-JAO MDPs, any uniformly good model identification algorithm needs to collect at least $\Omega\left(\frac{SA}{\lambda^2}\right)$ samples from state 0 on the left part. Whenever an action towards state 2 is taken from state 0, the learner may end up in state 2. Once in state 2, the learner needs to get back to state 0 to obtain the next useful sample. The expected number of actions needed to get back to state 0 from state 2 is $\frac{1}{\delta} = \frac{D}{4}$. This implies the following two lower bounds on the horizon of the clustering phase and the total regret of any cluster-then-learn algorithms.

Corollary 3.4.4. For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, 1]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $M = \frac{SA}{12}$, each with S states, A actions and diameter D such that for any uniformly good cluster-then-learn algorithm, to find the correct cluster with probability of at least $\frac{1}{2}$, the expected number of exploration steps needed in the clustering phase is $\Omega(\frac{DSA}{\lambda^2})$. Furthermore, the expected regret over K episodes of the same algorithm is

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(\frac{KDSA}{\lambda^2}\right).$$

Proof. (Sketch) In the lower bound construction, the learner is assumed to know everything about the set of models, including their optimal policies. Hence, after having identified the model in the clustering phase, the learner can follow the optimal policy in the learning phase and incur a small regret of at most $D/2$ in this phase. Therefore, the regret is dominated by the regret in the clustering phase, which is of order $\frac{DSA}{\lambda^2}$. □

Remark 3.4.5. The lower bound in Corollary 3.4.4 holds for a particular class of uniformly good cluster-then-learn algorithms under an adaptive adversary. It remains an open question whether this lower bound holds for any algorithms, not just cluster-then-learn.

Remark 3.4.6. Corollary 3.4.4 implies that, without further assumptions, it is not possible to improve the $\frac{1}{\lambda^2}$ dependency on λ. At the first glance this seems to contradict the existing results in bandits and online learning literature, where the regret bound depends on $\frac{1}{\text{gap}}$ where gap is the the difference in expected reward between the best arm and the sub-optimal arms. However, λ does not play the same role as the gaps in bandits. Observe that on the 2-JAO MDPs, the set of arms with positive reward is only in the right JAO MDP. The lower-bound learner knows this, but chooses to pull the arms on the left JAO MDP (with zero-reward) to collect side information that helps learn the right part faster. In this analogy, λ does not

play the same role as the gaps in bandits, since the learner already knows the arms on the left JAO MDP are suboptimal. The role of λ is in model identification, for which similar $\frac{1}{\lambda^2}$ lower bounds are known [e.g. Tulsiani, 2014].

Finally, we construct a non-communicating variant of the 2-JAO MDP to show that the finite diameter assumption is necessary. Figure 3.4 in Appendix 3.B illustrates this construction. On this variant, all the transitions from state 0 to state 2 are reversed. In addition, no actions take state 0 to state 2, making this MDP non-communicating. A set of these non-communicating MDPs is still λ-separable due to the state-action pairs that start at state 2. However, by setting the initial state to 0, the adversary can force the learner to operate only on the right part, regardless of how large λ is.

3.5 Non-Asymptotic Upper Bounds

We propose and analyze AOMultiRL, a polynomial time cluster-then-learn algorithm that obtains a high-probability regret bound of $\tilde{O}(\frac{KDSA}{\lambda^2} + H^{3/2}\sqrt{MSAK})$. In each episode, the learner starts with the clustering phase to identify the cluster of samples generated in previous episodes that has the same task. Once the right cluster is identified, the learner can use the samples from previous episodes in the learning phase.

A fundamental difference between the undiscounted infinite horizon setting considered in previous works [Guo and Brunskill, 2015; Brunskill and Li, 2013] and the episodic finite horizon in our work is the horizon of the two phases. In previous works, different episodes might have different horizons for the clustering phase depending on whether the learner decides to start exploration at all [Brunskill and Li, 2015] or which state-action pairs are to be explored [Brunskill and Li, 2013]. This poses a challenge for the episodic finite-horizon setting, because a varying horizon for the clustering phase leads to a varying horizon for the learning phase. Thus, standard single-task algorithms that rely on a fixed horizon such as UCBVI [Azar et al., 2017] and StrongEuler [Simchowitz and Jamieson, 2019] cannot be applied directly. From an algorithmic standpoint, for a fixed horizon H, a non-asymptotic bound on the horizon of the clustering phase is necessary so that the learner knows exactly whether H is large enough and when to stop collecting samples.

AOMultiRL alleviates this issue by setting a fixed horizon for the clustering phase, which reduces the learning phase to standard single-task episodic RL. First, we state an assumption on the ergodicity of the MDPs.

Assumption 3.5.1. The hitting times of all MDPs in $\mathcal{M}$ are bounded by a known constant $\tilde{D}$.

The main purpose of Assumption 3.5.1 is simplifying the computation of a non-asymptotic upper bound for the clustering phase in order to focus the exposition on the main ideas. We discuss a method for removing this assumption in Appendix 3.F.

Algorithm 3.3 outlines the main steps of our approach. Given a set Γ^α of α-distinguishing state-action pairs, in the clustering phase the learner employs a history-dependent policy specified by Algorithm 3.1, `ExploreID`, to collect at least N samples for each state-action pair in Γ^α, where N will be determined later. Once all (s, a) in Γ^α have been visited at least N times, Algorithm 3.2, `IdentifyCluster`, computes the empirical means of the transition function of these (s, a) and then compares them with those in each cluster to determine which cluster contains the samples from the same task (or none do, in which case a new cluster is created). For the rest of the episode, the learner uses the UCBVI-CH algorithm [Azar et al., 2017] to learn the optimal policy.

The algorithms and results up to Theorem 3.5.5 are presented for a general set Γ^α. Since Γ^α is generally unknown, Corollary 3.5.6 shows the result for $\alpha = \lambda$ and $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$.

Algorithm 3.1 ExploreID

Input: Episode k, state s, set Γ^α and number N

Set $\mathcal{G}(s) = \left\{ a \in \mathcal{A} : (s, a) \in \Gamma^\alpha, N_{\mathcal{B}_k}(s, a) < N \right\}$

if $\mathcal{G}(s) \neq \emptyset$ **then**

  return $\arg\max_{a \in \mathcal{G}(s)} N_{\mathcal{B}_k}(s, a)$

else

  return $\arg\max_{a \in \mathcal{A}} \sum_{s'=1}^{S} \hat{P}^k(s' \mid s, a) \mathbb{I}\{\mathcal{G}(s') \neq \emptyset\}$

Algorithm 3.2 Identify Cluster

Input: Episode k, set Γ^α, clusters $\mathcal{C}$, and threshold δ

for $c = 1, \ldots, \|\mathcal{C}\|$ **do**
  Initialize `id` $\leftarrow c$
  for $(s, a) \in \Gamma$ **do**
    if $\left\| [\hat{P}_c - \hat{P}^k](s, a) \right\| > \delta$ **then**
      `id` $\leftarrow 0$
      break;
  end
  if `id` $== c$ **then**
    return `id`;
end
return 0;

3.5.1 The Exploration Algorithm

Given a collection $\mathcal{B}$ of tuples (s, a, s'), the empirical transition functions estimated by $\mathcal{B}$ are

$$\hat{P}_{\mathcal{B}}(s' \mid s, a) = \begin{cases} \frac{N_{\mathcal{B}}(s,a,s')}{N_{\mathcal{B}}(s,a)} & \text{if } N_{\mathcal{B}}(s, a) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where

$$N_{\mathcal{B}}(s, a, s') = \sum_{(x,y,z) \in \mathcal{B}} \mathbb{I}\{x = s, y = a, z = s'\},$$
$$N_{\mathcal{B}}(s, a) = \sum_{s' \in \mathcal{S}} N_{\mathcal{B}}(s, a, s')$$

are the number of instances of (s, a, s') and (s, a) in $\mathcal{B}$, respectively.

For each episode k, let P^k denote the transition function of the task m^k and $\mathcal{B}_k$ denote the collection of samples (s_h, a_h, s_{h+1}) collected during the learning phase. The empirical means $\hat{P}^k$ estimated using samples in $\mathcal{B}_k$ are $\hat{P}^k = \hat{P}_{\mathcal{B}_k}$. The value of N can be chosen so that for all $(s, a) \in \Gamma^\alpha$, with high probability $\hat{P}^k(s, a)$ is close to $P^k(s, a)$. Specifically, we find that if N is large enough so that $\hat{P}^k(s, a)$ is within $\lambda/8$ in ℓ_1 norm of the true function $P^k(s, a)$, then the right cluster can be identified in every episode. The exact value of N is given in the following lemma.

Lemma 3.5.2. Suppose the learner is given a constant $p_1 \in (0,1)$ and a α-distinguishing set $\Gamma^\alpha \subseteq \mathcal{S} \times \mathcal{A}$. If each state-action pair in Γ^α is visited at least

$$N = \frac{256}{\lambda^2} \max\{S, \ln\left(\frac{K|\Gamma^\alpha|}{p_1}\right)\}$$

times during the clustering phase of each episode $k = 1, 2, \ldots, K$, then with probability at least $1 - p_1$, the event

$$\mathcal{E}_k^{\Gamma^\alpha} = \left\{\forall (s,a) \in \Gamma^\alpha, \left\| P^k(s,a) - \hat{P}^k(s,a) \right\| \leq \frac{\lambda}{8}\right\} \text{ holds for all } k \in [K].$$

The exploration in AOMultiRL is modelled as an instance of the active model estimation problem [Tarbouriech et al., 2020]. Given the current state s, if there exists an action a such that $(s,a) \in \Gamma^\alpha$ and (s,a) has not been visited at least N times, this action will be chosen (with ties broken by selecting the most chosen action). Otherwise, the algorithm chooses an action that has the highest estimated probability of leading to an under-sampled state-action pair in Γ^α. The following lemma computes the number of steps H_0 in the clustering phase.

Lemma 3.5.3. Consider p_1 and N defined in Lemma 3.5.2. By setting

$$H_0 = 12\tilde{D}|\Gamma^\alpha|N = \frac{3072\tilde{D}|\Gamma^\alpha|}{\lambda^2} \max\{S, \ln\left(\frac{K|\Gamma^\alpha|}{p_1}\right)\},$$

with probability at least $1 - p_1$, Algorithm 3.1 visits each state-action pair in Γ^α at least N times during the clustering phase in each of the K episodes.

3.5.2 The Clustering Algorithm

Denote by $\mathcal{C}$ the set of clusters, $C = |\mathcal{C}|$ the number of clusters and $\mathcal{C}_i$ the i^{th} cluster. Each $\mathcal{C}_i$ is a collection of two multisets $\mathcal{C}_i^{model}, \mathcal{C}_i^{regret} \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ which contain the (s,a,s') samples collected during the clustering and learning phases, respectively. Formally, up to episode k we have

$$\mathcal{C}_i^{model} = \cup_{k'=1}^{k-1}\{(s_h^{k'}, a_h^{k'}, s_{h+1}^{k'}) : h \leq H_0, \mathrm{id}^{k'} = i\},$$
$$\mathcal{C}_i^{regret} = \cup_{k'=1}^{k-1}\{(s_h^{k'}, a_h^{k'}, s_{h+1}^{k'}) : h > H_0, \mathrm{id}^{k'} = i\},$$

where s_h^k and a_h^k are the state and action at time step h of episode k, respectively and $id^{k'}$ is the cluster index returned by Algorithm 3.2 in episode k'.

Let $\hat{P}_i = \hat{P}_{\mathcal{C}_i^{model}}$ denote the empirical means estimated using samples in $\mathcal{C}_i^{model}$. For each episode k, from Lemma 3.5.3 with high probability after the first H_0 steps each state-action pair $(s,a) \in \Gamma^\alpha$ has been visited at least N times. Algorithm 3.2 determines the right cluster for a task by computing the ℓ_1 distance between $\hat{P}^k$ and the empirical transition function $\hat{P}_i$ for each cluster $i = 1, 2, \ldots, C$. If there exists an $(s,a) \in \Gamma^\alpha$ such that the distance is larger than a certain threshold δ, i.e., $\left\| [\hat{P}_i - \hat{P}^k](s,a) \right\| > \delta$, then the algorithm concludes that the task belongs to another cluster. Otherwise, the task is considered to belong to cluster i. We set $\delta = \alpha - \lambda/4$. The following lemma shows that with this choice of δ, the right cluster is identified by Algorithm 3.2 in all episodes.

Lemma 3.5.4. Consider a λ-separable set of MDPs $\mathcal{M}$ and an α-distinguishing set Γ^α where $\alpha \geq \lambda/2$. If the events $\mathcal{E}_k^{\Gamma^\alpha}$ defined in Lemma 3.5.2 hold for all $k \in [K]$, then with the distance threshold $\delta = \alpha - \lambda/4$ Algorithm 3.2 always produces a correct output in each episode: the trajectories of the same model in two different episodes are clustered together and no two trajectories of two different models are in the same cluster.

Once the clustering phase finishes, the learner enters the learning phase and uses the UCBVI-CH algorithm [Azar et al., 2017] to learn the optimal policy for this phase. In principle, almost all standard single-task RL algorithms with a near-optimal regret guarantee can be used for this phase. We chose UCBVI-CH to simplify the analysis and make the exposition clear.

To simulate the standard single-task episodic learning setting, the learner only uses the samples in $\mathcal{C}_i^{regret}$ for regret minimization. The impact of combining samples in two phases for regret minimization is addressed in Appendix 3.G. Theorem 3.5.5 states a regret bound for Algorithm 3.3.

Algorithm 3.3 Adversarial online multi-task RL

Input: Number of models M, number of episodes K, MDPs parameters $\mathcal{S}, \mathcal{A}, H, \tilde{D}, \lambda$, probability p, separation level α and an α-distinguishing set Γ^α.

Compute $p_1 = p/3, N = \frac{256}{\lambda^2}\max\{S, \ln\left(\frac{K|\Gamma^\alpha|}{p_1}\right)\}, \delta = \alpha - \lambda/4, H_0 = 12D|\Gamma^\alpha|N$

Initialize $\mathcal{C} \leftarrow \emptyset$

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

  Initialize $\mathcal{B}_k \leftarrow \emptyset$

  The environment chooses a task m^k

  Observe the initial state s_1 **for** $h = 1, \ldots, H_0$ **do**

    $a_h =$ `ExploreID`(s_h, Γ^α)

    Observe s_{h+1} and r_{h+1}

    Add (s_h, a_h, s_{h+1}) to $\mathcal{B}_k$

  end

  $id \leftarrow$ `IdentifyCluster`$(\mathcal{B}_k, \Gamma^\alpha, \mathcal{C}, \delta)$

  if $id \geq 1$ **then**

    $\mathcal{C}_{id}^{model} = \mathcal{C}_{id}^{model} \cup \mathcal{B}_k$

  end

  else

    $id \leftarrow |\mathcal{C}| + 1$

    $\mathcal{C}_{id}^{model} = \mathcal{B}_k$, $\mathcal{C}_{id}^{regret} = \emptyset$

    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_{id}$

  end

  $\pi_k =$ `UCBVI-CH`$(\mathcal{C}_{id}^{regret})$

  for $h = H_0 + 1, \ldots, H$ **do**

    $a_h = \pi_k(h, s_h)$

    Observe s_{h+1} and r_{h+1}

    $\mathcal{C}_{id}^{regret} = \mathcal{C}_{id}^{regret} \cup (s_h, a_h, s_{h+1})$

  end

end

Theorem 3.5.5. For any failure probability $p \in (0, 1)$, with probability at least $1 - p$ the regret of Algorithm 3.3 is bounded as

$$\text{Regret}(K) \leq 2KH_0 + 67H_1^{3/2}L\sqrt{MSAK} + 15MS^2AH_1^2L^2,$$

where $H_0 = 12\tilde{D}|\Gamma^\alpha|N$, $N = \frac{256}{\lambda^2}\max\{S, \ln\left(\frac{3K|\Gamma^\alpha|}{p}\right)\}$, $H_1 = H - H_0$, and $L = \ln(15SAKHM/p)$.

For $K > MS^3AH$, the first two terms are the most significant. The $2KH_0$ term accounts for the clustering phase and the fact that the exploration policy might lead the learner to an undesirable state after H_0 steps. The $\tilde{O}(\sqrt{K})$ term comes from the fact that the learning phase is equivalent to episodic single-task learning with horizon H_1. When $H \gg H_0$, the sub-linear bound on the learning phase is a major improvement compared to the $O(K\sqrt{HSA})$ bound of the strategy that learns each episode individually.

By setting $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$ and $\alpha = \lambda$, we obtain

Corollary 3.5.6. For any failure probability $p \in (0, 1)$, with probability at least $1 - p$, by setting $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$ with $\alpha = \lambda$, the regret of Algorithm 3.3 is

$$\text{Regret}(K) \leq O\left(\frac{K\tilde{D}SA}{\lambda^2} \ln\left(\frac{KSA}{p}\right) + H^{3/2}L\sqrt{MSAK}\right). \tag{3.2}$$

where $L = \ln(15SAKH_1M/p)$.

Time Complexity The clustering algorithm runs once in each episode, which leads to time complexity of $O(MSA+H)$. When $H \gg H_0$, the overall time complexity is dominated by the learning phase, which is $O(HSA)$ for UCBVI-CH.

Remark 3.5.7. Instead of clustering, a different paradigm involves actively merging samples from different MDPs to learn a model that is an averaged estimate of the MDPs in $\mathcal{M}$. The best regret guarantee in this paradigm, to the best of our knowledge, is $\tilde{O}(S^{1/3}A^{1/3}B^{1/3}H^{5/3}K^{2/3})$, where B is a variation budget, achieved by RestartQ-UCB [Mao et al., 2021, Theorem 3]. In our setting, if the adversary frequently alternates between tasks then $B = \Omega(KH\lambda)$ and therefore this bound becomes $\tilde{O}(\lambda^{1/3}S^{1/3}A^{1/3}H^2K)$, which is larger than the trivial bound KH and worse than the bound in Corollary 3.5.6. If the adversary selects tasks so that B is small i.e. $B = o(K)$ then the bound offered by RestartQ-UCB is better since it is sub-linear in K. Note that this does not contradict the lower bound result in Section 3.4, since the lower bound is constructed with an adversary that selects tasks uniformly at random, and hence B is linear in K.

3.5.3 Learning a distinguishing set when M is small

As pointed out by [Brunskill and Li, 2013], for all $\alpha > 0$, the size of the smallest α-distinguishing set of $\mathcal{M}$ is at most $\binom{M}{2}$. If $M^2 \ll SA$ and such a set is known to the

learner, then the clustering phase only need collect samples from this set instead of the full $\mathcal{S} \times \mathcal{A}$ set of state-action pairs. However, in general this set is not known. We show that if the adversary is weaker so that all models are guaranteed to appear at least once early on, the learner will be able to discover a $\frac{\lambda}{2}$-distinguishing set $\hat{\Gamma}$ of size at most $\binom{M}{2}$. Specifically, we employ the following assumption:

Assumption 3.5.8. There exists an unknown constant $K_1 \geq M$ satisfying $K_1 SA < K$ such that after at most K_1 episodes, each model in $\mathcal{M}$ has been given to the learner at least once.

In order to discover $\hat{\Gamma}$, the learner uses Algorithm 3.4, which consists of two stages:

- Stage 1: the learner starts by running Algorithm 3.3 with the λ-distinguishing set candidate $\mathcal{S} \times \mathcal{A}$ until the number of clusters is M. With high probability, each cluster corresponds to a model. At the end of stage 1, the learner uses the empirical estimates in all clusters $\hat{P}_i$ for $i \in [M]$ to construct a $\lambda/2$-distinguishing set $\hat{\Gamma}$ for $\mathcal{M}$.
- Stage 2: the learner runs Algorithm 3.3 with the distinguishing set $\hat{\Gamma}$ as an input.

Extracting $\lambda/2$-distinguishing pairs: After K_1 episodes, with high probability there are M clusters corresponding to M models. For two clusters i and j, the set $\hat{\Gamma}_{i,j}$ contains the first state-action pair (s, a) that satisfies $\left\| \hat{P}_i(s,a) - \hat{P}_j(s,a) \right\| > 3\lambda/4$. With high probability, every $(s, a) \in \Gamma_{i,j}$ satisfies this condition, hence $\hat{\Gamma}_{i,j} \neq \emptyset$.

Let $i^\star \in [M]$ denote the index of the MDP model corresponding to cluster i. For all $(s, a) \in \hat{\Gamma}_{i,j}$, by the triangle inequality, we have

$$\|P_{i^\star} - P_{j^\star}\| \geq \left\| \hat{P}_i - \hat{P}_j \right\| - \left\| \hat{P}_i - P_{i^\star} + P_{j^\star} - \hat{P}_j \right\| > 3\lambda/4 - (\lambda/8 + \lambda/8) = \lambda/2,$$

where (s, a) is omitted for brevity. It follows that the set $\hat{\Gamma} = \cup_{i,j} \hat{\Gamma}_{i,j}$ is $\lambda/2$-distinguishing and $|\hat{\Gamma}| \leq \binom{M}{2}$. Although $\lambda/2$ is smaller than the λ-separation level of Γ, it is sufficient for the conditions in Lemma 3.5.4 to hold. Thus, with high probability the clustering algorithm in stage 2 works correctly. The next theorem shows the regret guarantee of Algorithm 3.4.

Theorem 3.5.9. Under Assumption 3.5.8, With probability at least $1 - p$, the regret of Algorithm 3.4 is

$$\text{Regret}(K) = O\left(\frac{K\tilde{D}M^2}{\lambda^2} \ln \frac{KM^2}{p} + H^{3/2} L \sqrt{MKSA} \right),$$

where $H_{0,M} = \frac{3072\tilde{D}M^2}{\lambda^2} \max\{S, \ln\left(\frac{3KM^2}{p}\right)\}$ and $L = \ln(15SAKH_1M/p)$.

Algorithm 3.4 AOMultiRL with all models being given at least once

Input: Number of models M, number of episodes K, MDPs parameters $\mathcal{S}, \mathcal{A}, H, \tilde{D}, \lambda$, probability p

Stage 1: Run Algorithm 3.3 with the distinguishing set $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$ and $\alpha = \lambda$ until the number of clusters is M

for $i, j \in [M] \times [M], i \neq j$ **do**
  $\hat{\Gamma}_{i,j} = \emptyset$ **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    if $\left\| \hat{P}_i(s,a) - \hat{P}_j(s,a) \right\| > 3\lambda/4$ **then**
      $\hat{\Gamma}_{i,j} = \hat{\Gamma}_{i,j} \cup (s,a)$ break
    end
  end
end

$\hat{\Gamma} = \cup_{i,j} \hat{\Gamma}_{i,j}$

Stage 2: Run Algorithm 3.3 with distinguishing set $\hat{\Gamma}$ and $\alpha = \lambda/2$ for $K_2 = K - K_1$ episodes.

Compared to Corollary 3.5.6, Theorem 3.5.9 improves the clustering phase's dependency from SA to M^2. This implies that if the number of models is small and all models appear relatively early, we can discover a $\lambda/2$-distinguishing set quickly without increasing the order of the total regret bound.

3.6 Experiments

We evaluate AOMultiRL on a sequence of $K = 200$ episodes, where the task in each episode is taken from a set of $M = 4$ MDPs. Each MDP in $\mathcal{M}$ is a 4×4 grid of $S = 16$ cells with $A = 4$ valid actions: `up, down, left, right`. The state for row r and column c (0-indexed) is represented by the tuple (r, c). The reward is 0 in every state, except for the four corners $(0,0), (0,3), (3,0)$, and $(3,3)$, where the reward is 1. We fix the initial state at $(1,1)$.

To simulate an adversarial sequence of tasks, episodes 100 to 150 and episodes 180 to 200 contain only the MDP m_4. Other episodes chooses m_1, m_2 and m_3 uniformly at random. The hitting time is $D = 7$ and the failure probability is $p = 0.03$. We use the `rlberry` framework [Domingues et al., 2021a] for our implementation.

We construct the transition functions so that each MDP has only one easy-to-reach corner, which corresponds to a unique optimal policy. The separation level λ is 1.2999. Furthermore, there exists state-action pairs that are $\lambda/2$-distinguishing but not λ-distinguishing. More details can be found in Appendix 3.I.

The baseline algorithms include a random agent that chooses actions uniformly at ran-

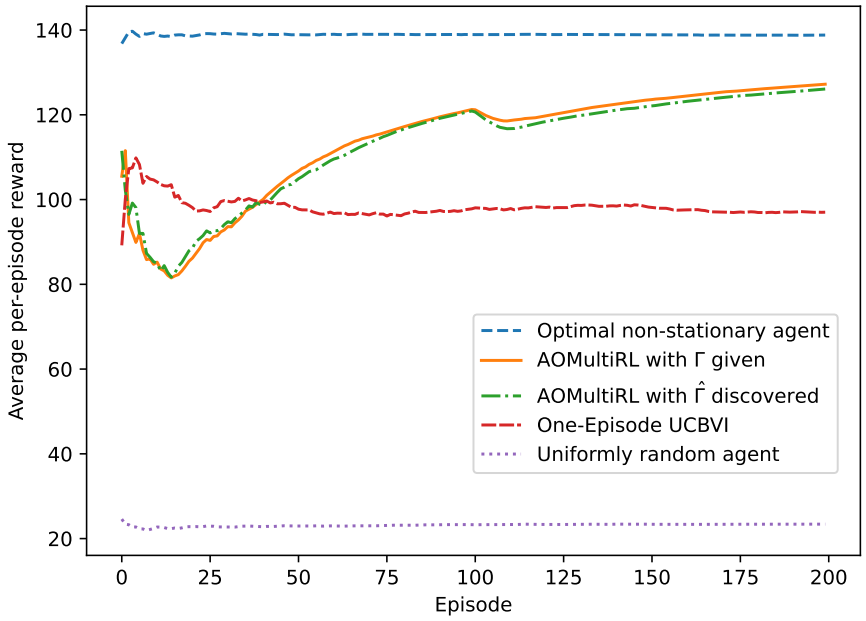

Figure 3.2: Average per-episode reward.

dom, a one-episode UCBVI agent which does not group episodes and learns using only the samples of each episode, and the optimal non-stationary agent that acts optimally in every episode. The first and the last baselines serve as the lower bound and upper bound performance for AOMultiRL, while the second baseline helps illustrate the effectiveness of clustering episodes correctly. We evaluate two instances of AOMultiRL: AOMultiRL1 with a set Γ of $|\Gamma| = 3$ given and AOMultiRL2 without any distinguishing set given. We follow the approach of [Kwon et al., 2021] and evaluate all five algorithms based on their expected cumulative reward when starting at state $(1, 1)$ and following their learned policy for $H_1 = 200$ steps (averaged over 10 runs). While the horizon for the learning phase is much smaller than the horizon for clustering phase of $H_0 \approx 80000$, we ensure the fairness of the comparisons by not using the samples collected in the clustering phase in the learning phase, thus simulating the setting where $H_0 \ll H_1$ without the need to use significantly larger MDPs. We use the average per-episode reward as the performance metric. Figure 3.2 shows the results.

The effectiveness of the clustering on the learning phase. To measure the effectiveness of aggregating samples from episodes of the same task for the learning phase, we compare AOMultiRL1 and the one-episode UCBVI agent. Since for every pair of MDP models, the transition functions are distinct for state-action pairs adjacent to two of the corners, AOMultiRL1 can only learn the estimated model accurately for each MDP model if the clustering phase produces correct clusters in most of the episodes. We can observe in Figure 3.2 that after about thirty episodes, AOMultiRL1 starts outperforming the one-episode UCBVI agent and approaching the performance of the optimal non-stationary agent. The model m_4 appears for the first time in episode 100, which accounts for the sudden drop in performance in that episode. Afterwards, the performance of AOMultiRL1 steadily increases again. This demonstrates that the AOMultiRL1 is able to identify the correct cluster in most of the episodes, which enables the multi-episode UCBVI algorithm in AOMultiRL1 to estimate the MDP models much more accurately than the non-transfer one-episode UCBVI agent. This suggests that for larger MDPs where $H_1 \gg H_0$, spending a number of initial steps on finding the episodes of the same task would yield higher long-term rewards.

Performance of AOMultiRL with the discovered $\hat{\Gamma}$. Next, we examine the performance of AOMultiRL2 when no distinguishing set is given. We run AOMultiRL2 for 204 episodes, in which stage 1 consists of the first four episodes, each containing one of the four MDP models in $\mathcal{M}$. As the identities of the models are not given, the algorithm has to correctly construct four clusters and then compute a $\lambda/2$-distinguishing set after the 4^{th} episode even though each model is seen just once. As mentioned above, the MDPs are set up so that if the AOMultiRL2 correctly identifies four clusters, then the discovered $\hat{\Gamma}$ will

contain at least one state-action pair that is $\lambda/2$-distinguishing but not λ-distinguishing. In stage 2, the horizon of the learning phase is set to the same $H_1 = 200$ used for AOMultiRL1. The performance in stage 2 of AOMultiRL2 approaches that of AOMultiRL1, indicating that the discovered $\hat{\Gamma}$ is as effective as the set Γ given to AOMultiRL1.

3.7 Conclusion

In this paper, we studied the adversarial online multi-task RL setting with the tasks belonging to a finite set of well-separated models. We used a general notion of task-separability, which we call λ-separability. Under this notion, we proved a minimax regret lower bound that applies to all algorithms and an instance-specific regret lower bound that applies to a class of uniformly good cluster-then-learn algorithms. We further proposed AOMultiRL, a polynomial time cluster-then-learn algorithm that obtains a nearly-optimal instance-specific regret upper bound. These results addressed two fundamental aspects of online multi-task RL, namely learning an adversarial task sequence and learning under a general task-separability notion. Adversarial online multi-task learning remains challenging when the diameter and the number of models are unknown; this is left for future work.

3.A The generality of λ-separability notion

In this section, we show that the general separation notion in Definition 3.3.1 defines a broader class of online multi-task RL problems that extends the entropy-based separation assumption in the latent MDPs setting [Kwon et al., 2021]. We start by restating the entropy-based separation condition of [Kwon et al., 2021]:

Definition 3.A.1. Let Π denote the class of all history-dependent and possibly non-Markovian policies, and let $\tau \sim (m, \pi)$ be a trajectory of length H sampled from MDP m by a policy $\pi \in \Pi$. The set $\mathcal{M}$ is well-separated if the following condition holds:

$$\forall m, m' \in \mathcal{M}, m' \neq m, \pi \in \Pi, \Pr_{\tau \sim (m,\pi)} \left(\frac{\Pr_{m',\pi}(\tau)}{\Pr_{m,\pi}(\tau)} > (\epsilon_p / M)^{c_1} \right) < (\epsilon_p / M)^{c_2}, \tag{3.3}$$

where $\epsilon_p \in (0, 1)$ is a target failure probability, $c_1 \geq 4, c_2 \geq 4$ are universal constants and $\Pr_{m,\pi}(\tau)$ is the probability that τ is realized when running policy π on model m.

The following lemma constructs a set $\mathcal{M}$ of just two models that satisfy the λ-separability condition but not the entropy-based separation condition.

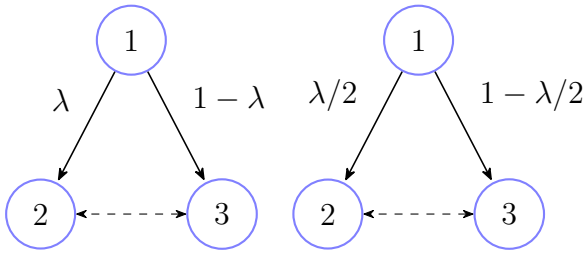

Figure 3.3: An instance of λ-separable LMDPs where Definition 3.A.1 does not apply

Lemma 3.A.1. Given any $\lambda \in (0,1), \epsilon_p \in (0,1), H > 0$ and any constants $c_1, c_2 \geq 4$, there exists a set of MDPs $\mathcal{M} = \{m_1, m_2\}$ with horizon H that is λ-separable but is not well-separated in the sense of Definition 3.A.1.

Proof. Consider the set $\mathcal{M}$ with $M = 2, \mathcal{S} = \{s^1, s^2, s^3\}, \mathcal{A} = \{a^1, a^2\}$ in Figure 3.3. Both m_1 and m_2 have the same transition functions in all state-action pairs except for (s^1, a^1):

$$
\begin{aligned}
P_1(s^2 \mid s^1, a^1) &= \lambda \\
P_1(s^3 \mid s^1, a^1) &= 1 - \lambda \\
P_2(s^2 \mid s^1, a^1) &= \lambda/2 \\
P_2(s^3 \mid s^1, a^1) &= 1 - \lambda/2.
\end{aligned}
$$

It follows that the ℓ_1 distance between $P_1(s^1, a^1)$ and $P_2(s^1, a^1)$ is

$$
\begin{aligned}
\left\| P_1(s^1, a^1) - P_2(s^1, a^1) \right\| &= \left\| P_1(s^2 \mid s^1, a^1) - P_2(s^2 \mid s^1, a^1) \right\| \\
&\quad + \left\| P_1(s^3 \mid s^1, a^1) - P_2(s^3 \mid s^1, a^1) \right\| \\
&= 2(\lambda - \lambda/2) \\
&= \lambda.
\end{aligned}
$$

As a result, this set $\mathcal{M}$ is λ-separable. However, any deterministic policy that takes action a_2 in s_1 and an arbitrary action in s_2 and s_3 will induce the same Markov chain on two MDP models. Thus, the entropy-based separation definition does not apply. An example of such a policy is shown below.

Consider running the following deterministic policy on model m_1:

$$
\begin{aligned}
\pi(s^1) &= a^2 \\
\pi(s^2) &= a^1 \\
\pi(s^3) &= a^1.
\end{aligned}
$$

Consider an arbitrary trajectory τ. The probability that this trajectory is realized with

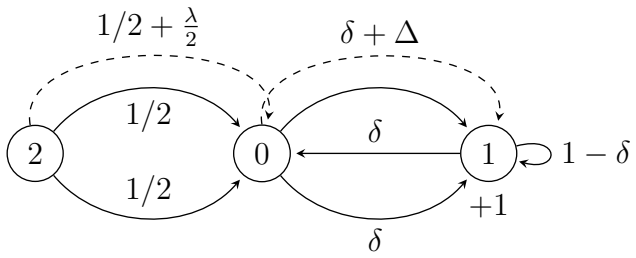

Figure 3.4: A non-communicating 2-JAO MDP. There are no rewards at states 0 and 2, while state 1 has reward +1. We set $\Delta = \Theta(\sqrt{\frac{SA}{HD}})$. The dashed arrows indicate the unique actions with highest transition probabilities on the left and right parts of the MDP. No actions take state 0 to state 2, making this MDP non-communicating.

respect to both models is

$$\Pr_{m_1,\pi}(\tau) = \prod_{t=1}^{H} P_1(s_{t+1} \mid s_t, a_t) \tag{3.4}$$

$$= \prod_{t=1}^{H} P_1(s_{t+1} \mid s_t, \pi(s_t)) \tag{3.5}$$

$$= \prod_{t=1}^{H} P_2(s_{t+1} \mid s_t, a_t) \qquad \text{since } (s_t, \pi(s_t)) \neq (s^1, a^1) \tag{3.6}$$

$$= \Pr_{m_2,\pi}(\tau). \tag{3.7}$$

As a result, for all τ,

$$\frac{\Pr_{m_2,\pi}(\tau)}{\Pr_{m_1,\pi}(\tau)} = 1, \tag{3.8}$$

which implies that

$$\Pr_{\tau\sim m_1,\pi}\left(\frac{\Pr_{m_2,\pi}(\tau)}{\Pr_{m_1,\pi}(\tau)} > (\epsilon_p/M)^{c_1}\right) = \Pr_{\tau\sim m_1,\pi}(1 > (\epsilon_p/M)^{c_1}) = 1, \tag{3.9}$$

which is larger than $(\epsilon_p/M)^{c_2}$. □

3.B Proofs of the lower bounds

Lemma 3.4.1 (Minimax Lower Bound)**.** Suppose $S, A \geq 10, D \geq 20 \log_A(S)$ and $H \geq DSA$ are given. Let $\lambda = \Theta(\sqrt{\frac{SA}{HD}})$. There exists a set of λ-separable MDPs $\mathcal{M}$ of size $M = \frac{SA}{4}$, each with S states, A actions, diameter at most D and horizon H such that if the tasks are chosen uniformly at random from $\mathcal{M}$, the expected regret of any sequence of policies $(\pi_k)_{k=1,\ldots,K}$ over K episodes is

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(K\sqrt{DSAH}\right).$$

Proof. We construct $\mathcal{M}$ in the following way: each MDP in $\mathcal{M}$ is a JAO MDP [Jaksch et al., 2010] of two states and SA actions and diameter $D' = D/4$. The translation from this JAO MDP to an MDP with S states, A actions and diameter D is straightforward [Jaksch et al., 2010]. State 1 has reward +1 while state 0 has no reward. In state 0, for all actions the probability of transitioning to state 1 is δ except for one best action where this probability is $\delta + \lambda/2$. Every MDP in $\mathcal{M}$ has a unique best action: for $i = 1, \ldots, SA$, the i^{th} action is the best action in the MDP m_i. The starting state is always $s_1 = 0$.

We consider a learner who knows all the parameters of models in $\mathcal{M}$, except the identity of the task m^k given in episode k. We employ the following information-theoretic argument from [Mao et al., 2021]: when the task m^k in episode k is chosen uniformly at random from $\mathcal{M}$, no useful information from the previous episodes can help the learner identify the best action in m^k. This is true since all the information in the previous episodes is samples from the MDPs in $\mathcal{M}$, which provide no further information than the parameters of the models in $\mathcal{M}$. Since $\mathcal{M} = SA$, all actions (from state 0) are equally probable to be the best action in m^k. Therefore, the learner is forced to learn m^k from scratch. It follows that the total regret of the learner is the sum of the one-episode-learning regrets in every episode:

$$\text{Regret}(K) = \sum_{k=1}^{K} R^k,$$

where $R^k = V_1^*(s_1) - V_1^{\pi_k}(s_1)$ is the one-episode-learning regret in episode k. The one-episode-learning is equivalent to the learning in the undiscounted setting with horizon H. Applying the lower bound result for the undiscounted setting in [Jaksch et al., 2010, Theorem 5] obtains that for all π_k,

$$\rho^* H - \mathbb{E}_{m^k \sim \mathcal{M}} V_1^{\pi_k}(s_1) \geq \Omega(\sqrt{DSAH}),$$

where $\rho^* = \frac{\delta+\lambda/2}{2\delta+\lambda/2}$ is the average reward of the optimal policy [Jaksch et al., 2010]. Note since only state 1 has reward +1, ρ^* is also the stationary probability that the optimal learner is at state 1.

Next, we show that for all $H \geq 2$ and $m^k \in \mathcal{M}$, it holds that $|V_1^* - \rho^* H| \leq \frac{D}{2}$. The optimal policy on all m^k induces a Markov chain between two states with transition matrix

$$\begin{bmatrix} 1-\delta-\lambda/2 & \delta+\lambda/2 \\ \delta & 1-\delta \end{bmatrix}.$$

Let $P_{m^k}(s_t = 1 \mid s_1 = 0)$ be the probability that the Markov chain is in state 1 after t time steps with the initial state $s_1 = 0$. Let $\Delta_t = P_{m^k}(s_t = 1 \mid s_1 = 0) - \rho^*$. Obviously, $\Delta_1 = -\rho^*$. By [Levin et al., 2008, Equation 1.8], we have $\Delta_t = (1 - 2\delta - \lambda/2)^{t-1}\Delta_1$. It follows that, for the optimal policy,

$$\begin{aligned} V_1^*(s_1) &= \sum_{t=1}^{H} P_{m^k}(s_t = 1 \mid s_1 = 0) && (3.10)\\ &= \sum_{t=1}^{H} (\Delta_t + \rho^*) && (3.11)\\ &= \rho^* H + \sum_{t=1}^{H} \Delta_t && (3.12)\\ &= \rho^* H + \sum_{t=1}^{H} (1 - 2\delta - \lambda/2)^{t-1}\Delta_1 && (3.13)\\ &= \rho^* H + \Delta_1 \frac{1 - (1 - 2\delta - \lambda/2)^H}{2\delta + \lambda/2}. && (3.14) \end{aligned}$$

Hence,

$$|V_1^*(s_1) - \rho^* H| = \left|\Delta_1 \frac{1-(1-2\delta-\lambda/2)^H}{2\delta+\lambda/2}\right| \tag{3.15}$$

$$\leq \left|\frac{\Delta_1}{2\delta+\lambda/2}\right| \tag{3.16}$$

$$= \frac{\rho^*}{2\delta+\lambda/2} \tag{3.17}$$

$$\leq \frac{1}{2\delta+\lambda/2} \tag{3.18}$$

$$\leq \frac{1}{2\delta} \tag{3.19}$$

$$= \frac{D}{2}, \tag{3.20}$$

where the last equality follows from $\delta = \frac{D}{4}$.

For any $H \geq DSA$ and $S, A \geq 2$, we have $\sqrt{HDSA} \geq DSA \geq 4D$, and hence $\sqrt{HDSA} - \frac{D}{2} \geq \frac{\sqrt{HDSA}}{2}$. We conclude that

$$\begin{aligned}
\mathbb{E}[\text{Regret}(K)] &= \sum_{k=1}^{K} \mathbb{E}[R^k] \\
&= \sum_{k=1}^{K} \mathbb{E}[V_1^* - V_1^{\pi_k}](s_1) \\
&\geq \sum_{k=1}^{K} (\rho^* H - \frac{D}{2} - V_1^{\pi_k}(s_1)) \\
&= \Omega(K\sqrt{DHSA}).
\end{aligned}$$

□

The upper bound of UCRL2 can be proved similarly: Theorem 2 in [Jaksch et al., 2010] states that for any $p \in (0,1)$, by running UCRL2 with failure parameter p, we obtain that for any initial state s_1 and any $H > 1$, with probability at least $1-p$,

$$\rho^* H - \sum_{h=1}^{H} r_h \leq O\left(DS\sqrt{AH \ln \frac{H}{p}}\right). \tag{3.21}$$

Setting $p = \frac{1}{H}$ and trivially bound the regret in the failure cases by H to obtain

$$\rho^* H - E[\sum_{h=1}^{H} r_h] \leq O\left(DS\sqrt{AH \ln H^2}\right) + \frac{1}{H} \times H \tag{3.22}$$

$$= O\left(DS\sqrt{AH \ln H}\right). \tag{3.23}$$

This bound holds across all episodes, hence the total regret bound with respect to $\rho^* H$ is $O\left(KDS\sqrt{AH \ln H}\right)$. Combining this with the fact that $V_1^*(s_1) \leq \rho^* H + \frac{D}{2}$, we obtain the upper bound.

Lemma 3.4.3. For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, \frac{1}{2}]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $\frac{SA}{12}$, each with at most S states, A actions and diameter D such that for any classification algorithm $\mathcal{C}$, if the number of state-transition samples given to $\mathcal{C}$ is less than $\frac{SA}{180\lambda^2}$ then for at least one MDP in $\mathcal{M}$, algorithm $\mathcal{C}$ fails to identify that MDP with probability at least $\frac{1}{2}$.

Before showing the proof of Lemma 3.4.3, we consider the following auxiliary problem: Suppose we are given three constants $\delta, \lambda, \epsilon \in (0, \frac{1}{4}]$ and a set of $2Q$ coins. The coins are arranged into a $Q \times 2$ table of Q rows and 2 columns so that each cell contains exactly one coin. The rows are indexed from 1 to Q and the columns are indexed from 1 to 2. In the first column, all coins are fair except for one coin at row θ which is biased with probability of heads equal to $\frac{1}{2} + \lambda$. In the second column, all coins have probability of heads equal to δ except for the coin at row θ which has probability of heads $\delta + \epsilon$. In this setting, row θ is a special row that contains the most biased coins in the two columns. The objective is to find this special row θ after at most H coin flips, where $H > 0$ is a constant representing a fixed budget. Note that if we ignore the second column, then this problem is reduced to the well-known problem of identifying one biased coin in a collection of Q-coins [Tulsiani, 2014].

Let N_1, N_2 be the number of flips an algorithm performs on the first and second column, respectively. For a fixed global budget H, after $\tau = N_1 + N_2 \leq H$ coin flips, the algorithm recommends $\hat{\theta}$ as its prediction for θ. Note that τ is a random stopping time which can depend on any information the algorithm observes up to time τ. Let X_t be the random variable for the outcome of t^{th} flip, and $X_1^\tau = (X_1, X_2, \ldots, X_\tau)$ be the sequence of outcomes after τ flips. For $j \in [Q]$, let P_j denote the probability measure induced by Alg corresponding to the case when $\theta = j$. We first show that if the algorithm fails to flip the coins sufficiently many times in both columns, then for some θ the probability of failure is at least $\frac{1}{2}$.

Lemma 3.B.1. Let $Q \geq 12, C_1 = 40$ and $C_2 = 64$. For any algorithm Alg, if

$$N_1 \leq T_1 := \frac{Q}{4C_1\lambda^2} \quad \text{and} \quad N_2 \leq T_2 := \frac{Q(\delta+\epsilon)}{4C_2\epsilon^2},$$

then there exists a set $J \subseteq [Q]$ with $|J| \geq \frac{Q}{6}$ such that

$$\forall j \in J,\ P_j[\hat{\theta} = j] \leq \frac{1}{2}.$$

The proof uses a reasonably well-known reverse Pinsker inequality [Sason, 2015, Equation 10]:

Let P and Q be probability measures over a common discrete set. Then

$$KL(P \parallel Q) \leq \frac{4\log_2 e}{\min_x Q(x)} \cdot D_{TV}(P \parallel Q)^2. \tag{3.24}$$

where D_{TV} is the total variation distance. In the particular case where P and Q are Bernoulli distributions with success probabilities p and $q \leq \frac{1}{2}$ respectively, we get

$$KL(P \parallel Q) \leq \frac{4\log_2 e}{q} \cdot (p-q)^2. \tag{3.25}$$

Proof. (of Lemma 3.B.1) As reasoned in the proof for the lower bound of multi-armed bandits [Auer et al., 2002b], we can assume that Alg is deterministic[2]. Our proof closely follows the main steps in the proof of [Tulsiani, 2014] for the setting where there is only one column. We will lower bound the probability of mistake of Alg based on its behavior on a hypothetical instance where $\lambda = \epsilon = 0$.

To account for algorithms which do not exhaust both budgets T_1 and T_2, we introduce two "dummy coins" by adding a zero'th row with two identical coins, solely for the analysis. These two coins have the same mean of 1 under all Q models and hence flipping either of them provides no information. An algorithm which wishes to stop in a round $\tau < H$ will simply flip any dummy coin in the remaining rounds $\tau+1, \tau+2, \ldots, H$. This way, we have the convenient option of always working with a sequence of outcomes X_1^H in the analysis.

Let P_0 and $\mathbb{E}_0$ denote the probability and expectation over X_1^H taken on the hypothetical instance with $\lambda = \epsilon = 0$, respectively. Let $a_t = (a_{t,0}, a_{t,1}) \in \{0, 1, \ldots, Q\} \times \{1, 2\}$ be the coin

[2]Deterministic conditional on the random history

that the algorithm flips in step t. Let $x_t \in \{0,1\}$ denote the outcome of a_t where 0 is tails and 1 is heads.

The number of flips the coin in row i, column k is

$$N_{i,k} = \sum_{t=1}^{T} \mathbb{1}\{a_t = (i,k)\}.$$

By the earlier definition of N_k for $k \in \{1,2\}$, we have

$$N_1 = \sum_{i=1}^{Q} N_{i,1},$$
$$N_2 = \sum_{i=1}^{Q} N_{i,2}.$$

We define

$$J_1 := \left\{ i \in [Q] \colon \left(\mathbb{E}_0[N_{i,1}] \le \frac{4T_1}{Q} \right) \wedge \left(\mathbb{E}_0[N_{i,2}] \le \frac{4T_2}{Q} \right) \right\}.$$

Clearly, at most $\frac{Q}{4}$ rows i satisfy $\mathbb{E}_0[N_{i,1}] > \frac{4T_1}{Q}$ and, similarly, at most $\frac{Q}{4}$ rows i satisfy $\mathbb{E}_0[N_{i,2}] > \frac{4T_2}{Q}$. Therefore, $|J_1| \ge Q - 2 \cdot \frac{Q}{4} = \frac{Q}{2}$.

We also define

$$J_2 := \left\{ i \in [Q] \colon P_0(\hat{\theta} = i) \le \frac{3}{Q} \right\}.$$

As at most $\frac{Q}{3}$ arms i can satisfy $P_0(\hat{\theta} = i) > \frac{3}{Q}$, it holds that $|J_2| \ge \frac{2Q}{3}$.

Consequently, defining $J := J_1 \cap J_2$, we have $|J| \ge \frac{Q}{6}$.

For any $j \in J$, we have

$$\begin{aligned} |P_j[c^* = j] - P_0[c^* = j]| &= |\mathbb{E}_j[\mathbb{1}\{c^* = j\}] - \mathbb{E}_0[\mathbb{1}\{c^* = j\}]| && (3.26) \\ &\le \frac{1}{2} \left\| P_0(X_1^H) - P_j(X_1^H) \right\|_1 && (3.27) \\ &\le \frac{1}{2} \sqrt{2 \ln 2 KL(P_0(X_1^H) \parallel P_j(X_1^H))}, && (3.28) \end{aligned}$$

where the first inequality follows from [Auer et al., 2002b, Equation 28] since the final output c^* is a function of the outcomes X_1^H, and the last inequality is Pinsker inequality.

Since Alg is deterministic, the flip a_t at step t is fully determined given the previous out-

comes x_1^{t-1}. Applying the chain rule for KL-divergences [Cover and Thomas, 2006, Theorem 2.5.3] we obtain

$$KL(P_0(X_1^H) \parallel P_j(X_1^H)) = \sum_{t=1}^{H} \sum_{x_1^{t-1}} P_0[x_{1:t-1}] KL(P_0[x_t] \parallel P_j[x_t] \mid x_1^{t-1}).$$

Note that x_t is the result of a single coin flip. When $a_{t,0} \neq j$, the KL-divergence is zero since the two instances have the identical coins on both columns. When $a_{t,0} = j$, the KL-divergence is either $B_1 = KL(\frac{1}{2} \parallel \frac{1}{2} + \lambda)$ or $B_2 = KL(\delta \parallel \delta + \epsilon)$, depending on whether $a_{t,1} = 1$ or $a_{t,1} = 2$, respectively. It follows that

$$\begin{aligned} KL(P_0(X_1^H) \parallel P_j(X_1^H)) &= \sum_{t=1}^{H} \sum_{x_{1:t-1}} P_0[x_{1:t-1}] \left(\mathbb{1}\{a_t = (j,1)\} B_1 + \mathbb{1}\{a_t = (j,2)\} B_2 \right) \\ &= \mathbb{E}_0[N_{j,1}] B_1 + \mathbb{E}_0[N_{j,2}] B_2 \\ &\leq \frac{4T_1}{Q} B_1 + \frac{4T_2}{Q} B_2 \\ &\leq \frac{B_1}{C_1 \lambda^2} + \frac{(\delta + \epsilon) B_2}{C_2 \epsilon^2} \end{aligned}$$

Since $\lambda \leq \frac{1}{4}$ and $\delta + \epsilon \leq \frac{1}{2}$, we can bound $B_1 \leq \frac{5\lambda^2}{2 \ln 2}$ [Tulsiani, 2014] and $B_2 \leq \frac{4 \log_2(e) \epsilon^2}{\delta + \epsilon}$. Consequently,

$$KL(P_0(X_1^H) \parallel P_j(X_1^H)) \leq \frac{5}{(2 \ln 2) C_1} + \frac{4 \log_2(e)}{C_2}$$

Plugging this into Equation 3.28 and applying $Q \geq 12$, we obtain

$$\begin{aligned} P_j[\hat{\theta} = j] &\leq P_0[\hat{\theta} = j] + \frac{1}{2} \sqrt{2 \ln 2 \left(\frac{5}{(2 \ln 2) C_1} + \frac{4 \log_2(e)}{C_2} \right)} \\ &= \frac{3}{Q} + \frac{1}{2} \sqrt{\frac{5}{C_1} + \frac{8}{C_2}} \\ &\leq \frac{3}{12} + \frac{1}{2} \sqrt{\frac{5}{40} + \frac{8}{64}} \\ &= \frac{1}{2}. \end{aligned}$$

□

The next result shows that if ϵ is sufficiently small, then any algorithm has to flip the

coins in the first column sufficiently many times; otherwise the probability of failure is at least $\frac{1}{2}$.

Corollary 3.B.2. Let Q, C_1 and C_2 be the constants defined in Lemma 3.B.1. Let $H > 0$ be the budget for the number of flips on both columns. If $\epsilon = \frac{1}{20}\sqrt{\frac{Q\delta}{H}}$, then for any algorithm Alg, if

$$N_1 \leq \frac{Q}{4C_1\lambda^2},$$

then there exists a set $J \subseteq [Q]$ with $|J| \geq \frac{Q}{6}$ such that

$$\forall j \in J, P_j[\hat{\theta} = j] \leq \frac{1}{2}.$$

Proof. We will show that when $\epsilon = \frac{1}{20}\sqrt{\frac{Q\delta}{H}}$, the inequality $N_2 \leq T_2 = \frac{Q(\delta+\epsilon)}{4C_2\epsilon^2}$ holds trivially for any $N_2 \leq H$ (recall that H is the fixed budget for the total number of coin flips). The result then follows directly from Lemma 3.B.1. We have

$$\begin{aligned} T_2 = \frac{Q(\delta+\epsilon)}{4C_2\epsilon^2} &\geq \frac{Q\delta}{4C_2\epsilon^2} \\ &= \frac{Q\delta}{256\epsilon^2} \qquad \text{since } C_2 = 64 \\ &= \frac{400}{256}H \\ &> H \\ &\geq N_2, \end{aligned}$$

which implies that $N_2 \leq T_2$ always holds for any $N_2 \leq H$.

□

We are now ready to prove Lemma 3.4.3.

Proof. (of Lemma 3.4.3) We construct $\mathcal{M}$ as the set of $\frac{SA}{12}$ 2-JAO MDPs in Figure 3.1 (right). Each MDP has a left part and a right part, where each part is a JAO MDP. The left part of the MDP m_i consists of two states $\{0, 2\}$ and $\frac{SA}{12}$ actions numbered from 1 to $\frac{SA}{12}$, where all actions from state 0 transition to state 2 with probability of $\frac{1}{2}$ or stay at state 0 with probability $\frac{1}{2}$, except for the i^{th} action that transitions to state 2 with probability $\frac{1}{2} + \frac{\lambda}{2}$ and stays at state 0 with probability $\frac{1}{2} - \frac{\lambda}{2}$. The right part of the i^{th} MDP consists of two states

$\{0,1\}$ and also $\frac{SA}{12}$ actions numbered from 1 to $\frac{SA}{12}$, where all actions from state 0 transition to state 1 with probability of $\delta = \frac{4}{D} \leq \frac{1}{4}$ or stays at state 0 with probability $1-\delta$, except for the i^{th} action that transitions to state 2 with probability $\delta + \Delta$ and stays at state 0 with probability $1-\delta-\Delta$. We set $\Delta = \frac{1}{20}\left(\sqrt{\frac{SA}{3HD}}\right)$. We will show the conversion from these 2-JAO MDPs to MDPs with S states and A actions later.

Since each model in $\mathcal{M}$ has a distinct index for the actions on both parts that transitions from 0 to 1 and 2 with probability higher than any other actions, identifying a model in $\mathcal{M}$ is equivalent to identifying this distinct action. Each action on both parts can be seen as a (possibly biased) coin, where the probability of getting tails is equal to the probability of ending up in state 0 when the action is taken. Thus, the problem of identifying this distinct action index reduces to the above auxiliary problem of identifying the row of the most biased coins, where taking an action from state 0 is equivalent to flipping a coin, $Q = \frac{SA}{12} \geq 12, \epsilon = \Delta$ and λ is replaced by $\lambda/2$. Corollary 3.B.2 states that for every algorithm, if the number of coin flips on the first column is less than $\frac{SA}{480\lambda^2}$, then there exists a set of size at least $\frac{SA}{72}$ positions of the row with the most biased coins such that the algorithm fails to find the biased coin with probability at least $\frac{1}{2}$. Correspondingly, for any model classification algorithm, if the number of state-transition samples from state 0 towards state 2 (i.e. the first column) is less than $\frac{SA}{480\lambda^2}$ then the algorithm fails to identify the model for at least $\frac{SA}{72}$ MDPs in $\mathcal{M}$.

Finally, we show the conversion from the 2-JAO MDP to an MDP with S states and A actions. The conversion is almost identical to that of [Jaksch et al., 2010], which starts with an *atomic* 2-JAO MDP of three states and $A' = \frac{A}{2}$ actions and builds an A'-ary tree from there. Assuming A' is an even positive number, each part of the atomic 2-JAO MDP has $\frac{A'}{2}$ actions. We make $\frac{S}{3}$ copies of these atomic 2-JAO MDPs, where only one of them has the best action on the right part. Arranging $\frac{S}{3}$ copies of these atomic 2-JAO MDPs and connecting their states 0 by $A - A'$ connections, we obtain an A'-ary tree which represents a composite MDP with at most S states, A actions and diameter D. The transitions of the $A - A'$ actions on the tree are defined identically to that of [Jaksch et al., 2010]: self-loops for states 1 and 2, deterministic connections to the state 0 of other nodes on the tree for state 0. By having $\delta = \frac{4}{D}$ in each atomic 2-JAO MDP, the diameter of this composite MDP is at most $\frac{2}{\delta} + \log_{A'} \frac{S}{3} \leq D$. This composite MDP is harder to explore and learn than the 2-JAO MDP with three states and $\frac{SA}{6}$ actions, and hence all the lower bound results apply. □

Corollary 3.4.4. For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0,1]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $M = \frac{SA}{12}$, each with S states, A actions and diameter D such that for any uniformly good cluster-then-learn algorithm, to find the correct cluster with

probability of at least $\frac{1}{2}$, the expected number of exploration steps needed in the clustering phase is $\Omega(\frac{DSA}{\lambda^2})$. Furthermore, the expected regret over K episodes of the same algorithm is

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(\frac{KDSA}{\lambda^2}\right).$$

Proof. As argued in Section 3.4, we can apply the sample complexity of the classification algorithm onto that of the clustering algorithm. Using the same set $\mathcal{M}$ of 2-JAO MDPs constructed in the proof of Lemma 3.4.3, for any given MDP $\mathcal{M}$, any PAC classification learner has to be in state 0 and takes at least $Z = \Omega(\frac{SA}{\lambda^2})$ actions from state 0 to state 2. If the learner stays at state 0, then it can take the next action from 0 to 2 in the next time step. However, if the learner transitions to state 2, then it has to wait until it gets back to state 0 to take the next action. Let Z_2 denote the number of times the learner ends up in state 2 after taking Z actions on the left part from state 0. Since every action from 0 to 2 has probability at least $\frac{1}{2}$ of ending up in state 2, we have

$$\mathbb{E}[Z_2 \mid Z] \geq \frac{Z}{2}. \tag{3.29}$$

Since every action from state 2 transitions to state 0 with the same probability of $\delta = \Theta(\frac{1}{D})$, every time the learner is in state 2, the expected number of time steps it needs to get back to state 0 is $\Theta(\frac{1}{\delta}) = \Theta(D)$. Hence, the expected number of time steps the learner needs to get back to state 0 after Z_2 times being in state 2 is $\Theta(Z_2 D)$. We conclude that for any PAC learner, the expected number of exploration steps needed to identify the model with probability of correct at least $\frac{1}{2}$ is at least

$$\mathbb{E}[Z + Z_2 D] \geq \Omega(ZD) = \Omega\left(\frac{DSA}{\lambda^2}\right). \tag{3.30}$$

Next, we lower bound the expected regret of the same algorithm. Let H_0 be the number of time steps the algorithm spends on the left part and H_1 on the right part of each model in $\mathcal{M}$. Note that H_0 and H_1 are random variables. Recall that the right part of each MDP in $\mathcal{M}$ resembles the JAO MDP in the minimax lower bound proof in Lemma 3.4.1, hence we can apply the regret formula of the JAO MDP for 2-JAO MDP and obtain that the regret

in each episode is of the same order as

$$\text{Regret} = \rho^* H - \mathbb{E}[\sum_{h=1}^{H} r(s_h, a_h)] \tag{3.31}$$

$$= \rho^* E[H_0 + H_1] - \mathbb{E}[\mathbb{E}[\sum_{h=1}^{H_0} r(s_h, a_h)] + \mathbb{E}[\sum_{h=H_0+1}^{H} r(s_h, a_h)] \mid H_0, H_1] \tag{3.32}$$

$$= \rho^* E[H_0 + H_1] - \mathbb{E}[\mathbb{E}[\sum_{h=H_0+1}^{H} r(s_h, a_h) \mid H_0, H_1]] \tag{3.33}$$

$$= \rho^* E[H_0] + E\left[\left(\rho^* H_1 - \mathbb{E}[\sum_{h=H_0+1}^{H} r(s_h, a_h)]\right) \mid H_1\right] \tag{3.34}$$

$$\geq \Omega\left(\rho^* E[H_0]\right) - \frac{D}{2} \tag{3.35}$$

$$= \Omega\left(\frac{DSA}{\lambda^2}\right), \tag{3.36}$$

where

- the second equality follows from $H = H_0 + H_1$,
- the third equality follows from the fact that the H_0 time steps spent on the left part of the MDP returns no rewards,
- the fourth equality follows from the linearity of expectation,
- the inequality follows from $H_1 = H - H_0$ and (3.20),
- the last equality follows from $\rho^* = \frac{\delta+\Delta}{2\delta+\Delta} \geq \frac{1}{2}$ for all $\delta, \Delta > 0$ and $E[H_0] \geq \Omega\left(\frac{DSA}{\lambda^2}\right)$.

We conclude that the expected regret over K episodes is at least

$$\Omega(\mathbb{E}[KH_0]) = \Omega\left(\frac{KDSA}{\lambda^2}\right).$$

□

3.C Proofs of the upper bounds

First, we state the following concentration inequality for vector-valued random variables by [Weissman et al., 2003].

Lemma 3.C.1 ([Weissman et al., 2003])**.** Let P be a probability distribution on the set $\mathcal{S} = \{1, \ldots, S\}$. Let $\mathcal{X}^N$ be a set of N i.i.d samples drawn from P. Then, for all $\epsilon > 0$:

$$\Pr\left(\left\|P - \hat{P}_{\mathcal{X}^N}\right\| \geq \epsilon\right) \leq (2^S - 2)e^{-N\epsilon^2/2}.$$

Using Lemma 3.C.1, we can show that $N = O(\frac{S}{\lambda^2})$ samples are sufficient for each $(s, a) \in \Gamma$ so that with high probability, the empirical means of the transition function $\hat{P}_{\mathcal{B}}(\cdot \mid s, a)$ are within $\lambda/8$ of their true values, measured in ℓ_1 distance.

Corollary 3.C.2. Denote $p_1 \in (0, 1)$. If a state-action pair (s, a) is visited at least

$$N = \frac{256}{\lambda^2} \max\{S, \ln(1/p_1)\} \tag{3.37}$$

times, then with probability at least $1 - p_1$,

$$\left\|P(s, a) - \hat{P}_{\mathcal{X}^N}(s, a)\right\| \leq \lambda/8.$$

Proof. We simplify the bound in Lemma 3.C.1 as follows:

$$\Pr\left(\left\|P - \hat{P}_{\mathcal{X}^N}\right\| \geq \epsilon\right) \leq (2^S - 2)e^{-N\epsilon^2/2} \leq e^{S - N\epsilon^2/2}$$

Next, we substitute $\epsilon = \lambda/8$ into the right hand side and solve the following inequality for N:

$$e^{S - N\lambda^2/128} \leq p_1$$

to obtain $N \geq \frac{128}{\lambda^2}(S + \ln(1/p_1))$. Thus $N = \frac{256}{\lambda^2} \max\{S, \ln(1/p_1)\}$ satisfies this condition. □

Taking a union bound of the result in Corollary 3.C.2 over all state-action pairs in the set Γ of all episodes from 1 to K, we obtain Lemma 3.5.2.

Next, we show the proof of Lemma 3.5.3. The proof strategy is similar to that of [Auer and Ortner, 2007; Sun and Huang, 2020].

Lemma 3.5.3. Consider p_1 and N defined in Lemma 3.5.2. By setting

$$H_0 = 12\tilde{D}|\Gamma^\alpha|N = \frac{3072\tilde{D}|\Gamma^\alpha|}{\lambda^2}\max\{S, \ln\left(\frac{K|\Gamma^\alpha|}{p_1}\right)\},$$

with probability at least $1 - p_1$, Algorithm 3.1 visits each state-action pair in Γ^α at least N times during the clustering phase in each of the K episodes.

Proof. The history-dependent exploration policy in Algorithm 3.1 visits an under-sampled state-action pair in Γ^α whenever possible; otherwise it starts a sequence of steps that would lead to such a state-action pair. In the latter case, denote the current state of the learner by s and the number of steps needed to travel from s' to an under-sampled state s by $T(s', s)$. By Assumption 3.3.1 and using Markov inequality, we have

$$\Pr\left(T(s', s) > 2\tilde{D}\right) \leq \frac{E[T(s', s)]}{2\tilde{D}} \leq \frac{\tilde{D}}{2\tilde{D}} = \frac{1}{2}.$$

It follows that $\Pr\left(T(s', s) > 2\tilde{D}\right) \leq 1/2$. In other words, in every interval of $2\tilde{D}$ time steps, the probability of visiting an under-sampled state-action pair in Γ^α is at least $1/2$. Over such n intervals, the expected number of such visits is lower bounded by $n/2$. Fix a $(s, a) \in \Gamma^\alpha$. Let V_n denote number of visits to $(s, a) \in \Gamma^\alpha$ after n intervals. Using a Chernoff bound for Poisson trials, we have

$$\Pr(V_n \geq (1 - \epsilon)n/2) \geq 1 - e^{-\epsilon^2 n/4}$$

for any $\epsilon \in (0, 1)$. Setting $\epsilon = 1 - 2N/m$ and solving

$$e^{-(1-2N/n)^2 n/4} \leq p_1$$

for n, we obtain

$$n \geq 2(N + \ln(1/p_1)) + 2\sqrt{2N\ln(1/p_1) + (\ln(1/p_1))^2}. \tag{3.38}$$

By definition of N,

$$\begin{aligned} 2N\ln(1/p_1) + (\ln(1/p_1))^2 &\le (1 + \frac{512}{\lambda^2})\max\{S, \ln(1/p_1)\}^2 \\ &\le \left(\frac{256}{\lambda}\max\{S, \ln(1/p_1)\}\right)^2 \\ &\le N^2. \end{aligned}$$

We also have $N \ge \ln(1/p)$. Overall, $n = 6N$ satisfies the condition in Equation 3.38. Taking a union bound over all $(s, a) \in \Gamma^\alpha$ and noting that each interval has length $2\tilde{D}$ steps, the total number of identifying steps needed is $H_0 = 2\tilde{D}n|\Gamma^\alpha| = 12\tilde{D}|\Gamma^\alpha|N$. □

To prove Lemma 3.5.4, we state the following auxiliary proposition and its corollary.

Proposition 3.C.3. Suppose we are given a probability distribution P over $\mathcal{S} = 1, \ldots, S$, a constant $\epsilon > 0$ and two set of samples $\mathcal{X} = (X_1, \ldots, X_{N_\mathcal{X}})$ and $\mathcal{Y} = (Y_1, \ldots, Y_{N_\mathcal{Y}})$ drawn from P such that $\left\|P - \hat{P}_\mathcal{X}\right\| \le \epsilon$ and $\left\|P - \hat{P}_\mathcal{Y}\right\| \le \epsilon$. Then,

$$\left\|P - \hat{P}_{\mathcal{X}\cup\mathcal{Y}}\right\| \le \epsilon.$$

Proof. Let $N_\mathcal{X}(s)$ and $N_\mathcal{Y}(s)$ denote the number of samples of $s \in [S]$ in $\mathcal{X}$ and $\mathcal{Y}$, respec-

tively. We have:

$$\left\|P - \hat{P}_{\mathcal{X}\cup\mathcal{Y}}\right\| = \sum_{s=1}^{S} |P(s) - \frac{N_{\mathcal{X}}(s) + N_{\mathcal{Y}}(s)}{N_{\mathcal{X}} + N_{\mathcal{Y}}}| \tag{3.39}$$

$$= \frac{1}{N_{\mathcal{X}} + N_{\mathcal{Y}}} \sum_{s=1}^{S} |N_{\mathcal{X}}P(s) - N_{\mathcal{X}}(s) + N_{\mathcal{Y}}P(s) - N_{\mathcal{Y}}(s)| \tag{3.40}$$

$$\leq \frac{1}{N_{\mathcal{X}} + N_{\mathcal{Y}}} \sum_{s=1}^{S} (|N_{\mathcal{X}}P(s) - N_{\mathcal{X}}(s)| + |N_{\mathcal{Y}}P(s) - N_{\mathcal{Y}}(s)|) \tag{3.41}$$

$$= \frac{1}{N_{\mathcal{X}} + N_{\mathcal{Y}}} \left(N_{\mathcal{X}} \sum_{s=1}^{S} |P(s) - \frac{N_{\mathcal{X}}(s)}{N_{\mathcal{X}}}| \right) \tag{3.42}$$

$$+ \frac{1}{N_{\mathcal{X}} + N_{\mathcal{Y}}} \left(N_{\mathcal{Y}} \sum_{s=1}^{S} |P(s) - \frac{N_{\mathcal{Y}}(s)}{N_{\mathcal{Y}}}| \right) \tag{3.43}$$

$$= \frac{1}{N_{\mathcal{X}} + N_{\mathcal{Y}}} (N_{\mathcal{X}} \left\|P - \hat{P}_{\mathcal{X}}\right\|_1 + N_{\mathcal{Y}} \left\|P - \hat{P}_{\mathcal{Y}}\right\|) \tag{3.44}$$

$$\leq \epsilon \tag{3.45}$$

□

Corollary 3.C.4. Suppose we are given a probability distribution P over $\mathcal{S} = 1, \ldots, S$, a constant $\epsilon > 0$ and a finite number of set of samples $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_t$ such that $\left\|P - \hat{P}_{\mathcal{X}_i}\right\| \leq \epsilon$ for all $i = 1, 2, \ldots, t$. Then,

$$\left\|P - \hat{P}_{\cup_{i=1,\ldots,t}\mathcal{X}_i}\right\| \leq \epsilon. \tag{3.46}$$

Proof. (Of Lemma 3.5.4) The proof is by induction. The claim is trivially true for the first episode ($k = 1$). For an episode $k > 1$, assume that the outputs of the Algorithm 3.2 are correct until the beginning of this episode. We consider two cases:

- When the task m_k has never been given to the learner before episode k.

  Consider an arbitrary existing cluster c. Denote by $i \in [M]$ the identity of the model to which the samples in c belong, $j \in [M]$ the identity of the task m_k, and (s, a) in $\Gamma^{\alpha}_{i,j}$ a state-action pair that distinguishes these two models. Under the definition of $\Gamma^{\alpha}_{i,j}$, the result in Lemma 3.5.2 and the result in Corollary 3.C.4, the following three

inequalities hold true:

$$\begin{aligned} \|[P_i - P_j](s,a)\| &> \alpha \\ \left\|[P_j - \hat{P}_{\mathcal{B}_k}](s,a)\right\| &\leq \lambda/8 \\ \left\|[P_i - \hat{P}_c](s,a)\right\| &\leq \lambda/8. \end{aligned}$$

From here, we omit the (s,a) and write P for $P(s,a)$ when no confusion is possible. Applying the triangle inequality twice, we obtain:

$$\begin{aligned} \left\|\hat{P}_c - \hat{P}_{\mathcal{B}_k}\right\| &\geq \|P_i - P_j\| - (\left\|P_i - \hat{P}_c\right\| + \left\|P_j - \hat{P}_{\mathcal{B}_k}\right\|) \\ &> \alpha - (\lambda/8 + \lambda/8) \\ &= \delta. \end{aligned}$$

It follows that the `break` condition in Algorithm 3.2 is satisfied, and the correct value of 0 is returned. A new cluster is created containing only the samples generated by the new task m_k.

- When the task m_k has been given to the learner before episode k.

  In this case, there exists a cluster c' containing the samples generated from model j. Using a similar argument in the previous part, we have that whenever the iteration in Algorithm 3.2 reaches a cluster c whose identity $i \neq j$, the `break` condition is true for at least one $(s,a) \in \Gamma^\alpha$, and the algorithm moves to the next cluster. When the iteration reaches cluster c', for all $(s,a) \in \tilde{\Gamma}^\alpha$, we have:

$$\begin{aligned} \left\|\hat{P}_{\mathcal{B}_k} - \hat{P}_{c'}\right\| &\leq \left\|\hat{P}_{\mathcal{B}_k} - P_j\right\| + \left\|P_j - \hat{P}_{c'}\right\| \\ &\leq \lambda/8 + \lambda/8 = \lambda/4 \\ &\leq \delta. \end{aligned}$$

  Hence, the `break` condition is false for all $(s,a) \in \Gamma$, and thus the algorithm returns `id` $= c'$ as expected.

By induction, under event $\mathcal{E}_\Gamma$, Algorithm 3.2 always produces correct outputs throughout the K episodes. □

We can now state the regret bound of Algorithm 3.3 where the regret minimization algorithm in every episode is UCBVI-CH [Azar et al., 2017]. For each state-action pair (s,a)

in episode k, UCBVI-CH needs a bonus term defined as

$$b_k(s,a) = 7H_1 L_k \sqrt{\frac{1}{N_k^{regret}(s,a)}},$$

where $L_k = \ln(5SAK_{m_k}H_1/p_1)$, $N_k^{regret}(s,a)$ is the total number of visits to (s,a) in the learning phase before episode k, and K_{m^k} is the total number of episodes in which the model m^k is given to the learner. However, K_{m^k} is unknown to the learner. We instead upper bound K_{m^k} by K and modify the bonus term as

$$b'_k(s,a) = 7H_1 L \sqrt{\frac{1}{N_k^{regret}(s,a)}} \tag{3.47}$$

where $L = \ln(5SAKHM/p_1)$. Since $b'_k \geq b_k$, this algorithm still retain the optimism principle needed for UCBVI-CH. The total regret of each model in $\mathcal{M}$ is bounded by the following result, whose proof is in Appendix 3.D.

Lemma 3.C.5. With probability at least $1-p_1$, applying UCBVI-CH with the bonus term b'_k defined in Equation 3.47, each task m in $\mathcal{M}$ has a total regret of

$$\text{Regret}(m, K_m) \leq K_m(H_0 + D) + 67H_1^{3/2} L\sqrt{SAK_m} + 15S^2A^2H_1^2L^2$$

Theorem 3.5.5. For any failure probability $p \in (0,1)$, with probability at least $1-p$ the regret of Algorithm 3.3 is bounded as

$$\text{Regret}(K) \leq 2KH_0 + 67H_1^{3/2} L\sqrt{MSAK} + 15MS^2AH_1^2L^2,$$

where $H_0 = 12\tilde{D}|\Gamma^\alpha|N$, $N = \frac{256}{\lambda^2}\max\{S, \ln\left(\frac{3K|\Gamma^\alpha|}{p}\right)\}$, $H_1 = H - H_0$, and $L = \ln(15SAKHM/p)$.

Proof. Summing up the regret for all $m \in \mathcal{M}$ and applying the Cauchy-Schwarz inequality, Lemma 3.C.5 together with Lemma 3.5.4 and Lemma 3.5.3 imply that with probability $1-p$, the total regret is bounded by

$$\text{Regret}(K) \leq K(H_0 + D) + 67H_1 L\sqrt{MSAKH_1} + 15MS^2AH_1^2L^2. \tag{3.48}$$

Note that the bound in Equation 3.48 is tighter than the bound in Theorem 3.5.5. To obtain the bound in Theorem 3.5.5, notice that $D \leq \tilde{D} \leq H_0$ and thus $K(H_0 + D) \leq$

$K(H_0 + H_0) = 2KH_0$. □

Theorem 3.5.9. Under Assumption 3.5.8, With probability at least $1 - p$, the regret of Algorithm 3.4 is

$$\text{Regret}(K) = O\left(\frac{K\tilde{D}M^2}{\lambda^2} \ln \frac{KM^2}{p} + H^{3/2}L\sqrt{MKSA}\right),$$

where $H_{0,M} = \frac{3072\tilde{D}M^2}{\lambda^2} \max\{S, \ln\left(\frac{3KM^2}{p}\right)\}$ and $L = \ln(15SAKH_1M/p)$.

Proof. In stage 1, as the distinguishing set has size $|\tilde{\Gamma}| = SA$, the number of time steps needed in the clustering phase is

$$H_{0,1} = 12\tilde{D}|\tilde{\Gamma}|N_1 = 12DSAN_1,$$

where $N_1 = \frac{256}{\lambda^2} \max\{S, \ln\left(\frac{3KSA}{p}\right)\}$.

In stage 2, the length of the clustering phase is

$$H_{0,2} = 12\tilde{D}|\hat{\Gamma}|N_2,$$

where $N_2 = \frac{256}{\lambda^2} \max\{S, \ln\left(\frac{3K|\hat{\Gamma}|}{p}\right)\}$.

Substituting $H_{0,1}$ and $H_{0,2}$ into Theorem 3.5.5, we obtain the regret bound of stage 1 and stage 2:

$$\text{Regret}_{Stage1} \leq 2K_1H_{0,1} + 67(H_{1,1})^{3/2}L_1\sqrt{MSAK_1} + 15MS^2A(H_{1,1})^2L_1^2,$$

where $L_1 = \ln\left(\frac{15MSAKH_{1,1}}{p}\right)$ and $H_{1,1} = H - H_{0,1}$.

$$\text{Regret}_{Stage2} \leq 2K_2H_{0,2} + 67H_{1,2}^{3/2}L_2\sqrt{MSAK_2} + 15MS^2AH_{1,2}^2L_2^2,$$

where $L_2 = \ln\left(\frac{15MSAKH_{1,2}}{p}\right)$ and $H_{1,2} = H - H_{0,2}$.

Since $H_{0,1} \geq H_{0,2}$, we have $H_{1,1} \leq H_{1,2}$. Using the assumption that $K_1SA < K_2$ and the Cauchy-Schwarz inequality for the sum $\sqrt{K_1} + \sqrt{K_2}$, we obtain

$$\begin{aligned}
\text{Regret}(K) &= \text{Regret}_{Stage1} + \text{Regret}_{Stage2} && (3.49)\\
&\leq 4KH_{0,2} + 67H_{1,2}^{3/2}L_2\sqrt{2MSAK} + 30MS^2AH_{1,2}^2L_2^2. && (3.50)
\end{aligned}$$

By having $|\hat{\Gamma}| \leq \binom{M}{2} \leq M^2, H_{1,2} \leq H$ and $\max\{L_1, L_2\} \leq L$, we obtain

$$\text{Regret}(K) \leq 4KH_{0,M} + 67H^{3/2}L\sqrt{2MSAK} + 30MS^2AH^2L^2. \tag{3.51}$$

where $H_{0,M} = \frac{3072\tilde{D}M^2}{\lambda^2}\max\{S, \ln\left(\frac{3KM^2}{p}\right)\}$.

□

3.D Per-model Regret analysis

First, we prove the following lemma which upper bound the per-episode regret as a function of H_0 and the regret of the clustering phase.

Lemma 3.D.1. The regret of Algorithm 3.3 in episode k is

$$\Delta_k = [V_1^{k,*} - V_1^{\pi_k}](s_1^k) \leq H_0 + D + \max_{s\in\mathcal{S}}[V_{H_0+1}^{k,*} - V_{H_0+1}^{\pi_k}](s).$$

Proof. Denote by $\Pr\big(s_h^k = s \mid s_1, \pi\big)$ the probability of visiting state s at time h when the learner follows a (possibly non-stationary) policy π in model m^k starting from state s_1. The regret of task m in a single episode $k \in \mathcal{K}_m$ can be written as

$$\begin{aligned}
\Delta_k &= [V_1^{k,*} - V_1^{\pi_k}](s_1^k) \\
&= E[\sum_{h=1}^{H} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k^*(s_h)] - E[\sum_{h=1}^{H} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k(s_h)] \\
&= \left(E[\sum_{h=1}^{H_0} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k^*(s_h)] + \sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s)\right) \\
&\quad - \left(E[\sum_{h=1}^{H_0} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k(s_h)] + \sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{\pi_k}(s)\right) \\
&\leq H_0 + \sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s) - \sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{\pi_k}(s) \\
&= H_0 + \left(\sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s) - \sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{k,*}(s)\right) \\
&\quad + \sum_{s\in\mathcal{S}} \Pr_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k)[V_{H_0+1}^{k,*} - V_{H_0+1}^{\pi_k}](s) \\
&\leq H_0 + \underbrace{\left(\max_{s\in\mathcal{S}} V_{H_0+1}^{k,*}(s) - \min_{s\in\mathcal{S}} V_{H_0+1}^{k,*}(s)\right)}_{(\clubsuit)} + \max_{s\in\mathcal{S}}[V_{H_0+1}^{k,*} - V_{H_0+1}^{\pi_k}](s).
\end{aligned}$$

The first inequality follows from the assumption that $r(s,a) \in [0,1]$ for all (s,a). The second inequality follows the fact that

$$\begin{aligned}\sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi^*_k) V^{k,*}_{H_0+1}(s) &\leq \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi^*_k) \max_{x\in\mathcal{S}} V^{k,*}_{H_0+1}(x) \\ &= \left(\max_{x\in\mathcal{S}} V^{k,*}_{H_0+1}(x)\right) \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi^*_k) \\ &= \max_{x\in\mathcal{S}} V^{k,*}_{H_0+1}(x),\end{aligned}$$

and

$$\begin{aligned}\sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) V^{k,*}_{H_0+1}(s) &\geq \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) \min_{x\in\mathcal{S}} V^{k,*}_{H_0+1}(x) \\ &= \left(\min_{x\in\mathcal{S}} V^{k,*}_{H_0+1}(x)\right) \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) \\ &= \min_{x\in\mathcal{S}} V^{k,*}_{H_0+1}(x).\end{aligned}$$

Furthermore, since $V^{k,*}_{H_0+1}(s) \geq V^{\pi_k}_{H_0+1}(s)$ for all $s \in \mathcal{S}$, we have

$$\begin{aligned}&\sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k)[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s) \\ &\leq \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) \max_{x\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](x) \\ &= \max_{x\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](x) \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) \\ &= \max_{x\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](x).\end{aligned}$$

For each state s, the value of $V^{k,*}_h(s)$ is the expected total $(H-h)$-step reward of an optimal non-stationary $(H-h)$ step policy starting in state s on the MDP m. Thus, the term (♣) represents the *bounded span* of the finite-step value function in MDP m. Applying equation 11 of [Jaksch et al., 2010], the span of the value function is bounded by the diameter of the MDP. We obtain for all h

$$\max_{s\in\mathcal{S}} V^{k,*}_h(s) - \min_{s\in\mathcal{S}} V^{k,*}_h(s) \leq D.$$

It follows that

$$\Delta_k \leq H_0 + D + \max_{s\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s).$$

□

Denote $\mathcal{K}_m$ the set of episodes where the model m is given to the learner. The total regret of the learner in episodes $\mathcal{K}_m$ is

$$\begin{aligned}\text{Regret}(m, K_m) &= \sum_{k \in \mathcal{K}_m} \Delta_k \\ &\leq K_m(H_0 + D) + \underbrace{\sum_{k \in \mathcal{K}_m} \max_{s \in S} [V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s)}_{(\heartsuit)}.\end{aligned}$$

The policy π_k from time step $H_0 + 1$ to H is the UCBVI-CH algorithm [Azar et al., 2017]. Therefore, the term ($\heartsuit$) corresponds to the total regret of UCBVI-CH in an adversarial setting in which the starting state s_1^k in each episode is chosen by an adversary that maximizes the regret in each episode. In Appendix 3.E, we given a simplified analysis for UCBVI-CH and show that with probability at least $1 - p_1/M$,

$$(\heartsuit) = \sum_{k \in \mathcal{K}_m} \max_{s \in S} [V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s) \leq 67 H_1^{3/2} L \sqrt{SAK_m} + 15 S^2 A^2 H_1^2 L^2. \tag{3.52}$$

The proof of Lemma 3.C.5 is completed by plugging the bound of (2) in Equation 3.52 to obtain

$$\begin{aligned}\text{Regret}(m, K_m) &= \sum_{k \in \mathcal{K}_m} \Delta_k \\ &\leq K_m(H_0 + D) + 67 H_1^{3/2} L \sqrt{SAK_m} + 15 S^2 A^2 H_1^2 L^2.\end{aligned}$$

3.E A simplified analysis for UCBVI-CH

Algorithm 3.5 UCBVI

Input: Failure probability p
Initialize an empty collection $\mathcal{B}$
for episode $k = 1, \ldots, K$: **do**
  $Q_{k,h} =$ `UCB-Q-Values` $(\mathcal{B}, p)$
  for $h = 1, \ldots, H$: **do**
    Take action $a_{k,h} = \arg\max_a Q_{k,h}(s_h^k, a)$ Add $(s_h^k, a_h^k, s_{h+1}^k)$ to $\mathcal{B}$
  end
end

Algorithm 3.6 UCB-Q-Values with Hoeffding bonus

Input: Collection $\mathcal{B}$, probability p
for $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ **do**
  $N_k(s, a, s') = \sum_{(x,a',y)\in\mathcal{B}} \mathbb{I}(x = s, a' = a, y = s')$
  $N_k(s, a) = \sum_{s'\in S} N_k(s, a, s')$
end
for $(s, a) \in \{(s, a) : N_k(s, a) > 0\}$ **do**
  $\hat{P}_k(s' \mid s, a) = \frac{N_k(s,a,s')}{N_k(s,a)}$
  $b_{k,h}(s, a) = 7HL\sqrt{\frac{1}{N_k(s,a)}}$ where $L = \ln(5SAKH/p)$
end
Initialize $V_{k,H+1}(s) = 0$ for all $x \in \mathcal{S}$
for $h = H, H - 1, \ldots, 1$: **do**
  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    if $N_k(s, a) > 0$ **then**
      $Q_{k,h}(s, a) = \min\{H, r(s, a) + \left(\sum_{s'\in\mathcal{S}} \hat{P}_k(s' \mid s, a)V_{k,h+1}(s')\right) + b_{k,h}(s, a)\}$
    else
      $Q_{k,h} = H$
    $V_{k,h}(s) = \max_a Q_{k,h}(s, a)$
  end
end

In section, we construct a simplified analysis for the UCBVI-CH algorithm in [Azar et al., 2017]. The proof largely follows the existing constructions in [Azar et al., 2017], with two differences: the definition of "typical" episodes and the analysis are tailored specifically for the Chernoff-type bonus of UCBVI-CH, without being complicated by handling of the variances for the Bernstein-type bonus of UCBVI-BF in [Azar et al., 2017]. For completeness, the full UCBVI-CH algorithm from [Azar et al., 2017] is shown in Algorithms 3.5 and 3.6.

Notation. In this section, we consider the standard single-task episodic RL setting in [Azar et al., 2017] where the learner is given the same MDP $(\mathcal{S}, \mathcal{A}, H, P, r)$ in K episodes. We assume the reward function $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is deterministic and known. The state and action spaces $\mathcal{S}$ and $\mathcal{A}$ are discrete spaces with size S and A, respectively. Denote by p the failure probability and let $L = \ln(5SAKH/p)$. We assume the product $SAKH$ is sufficiently large that $L > 1$.

Let V_1^* denote the optimal value function and $V_1^{\pi_k}$ the value function of the policy π_k of

the UCBVI-CH agent in episode k. The regret is defined as follows.

$$\text{Regret}(K) = \sum_{k=1}^{K} \delta_{k,1}, \tag{3.53}$$

where $\delta_{k,h} = [V_h^* - V_h^{\pi_k}](s_h^k)$.

Denote by $N_k(s,a)$ the number of visits to the state-action pair (s,a) up to the beginning of episode k.

We call an episode k "typical" if all state-action pairs visited in episode k have been visited at least H times at the beginning of episode k. The set of typical episodes is defined as follows.

$$[K]_{typ} = \{i \in [K] : \forall h \in [H], N_i(s_h^i, a_h^i) \geq H\}. \tag{3.54}$$

Equation 3.53 can be written as

$$\begin{aligned} \text{Regret}(K) &= \sum_{k \notin [K]_{typ}} \delta_{k,1} + \sum_{k \in [K]_{typ}} \delta_{k,1} \\ &\leq \sum_{k \notin [K]_{typ}} H + \sum_{k \in [K]_{typ}} \delta_{k,1} \\ &\leq SAH^2 + \sum_{k \in [K]_{typ}} \delta_{k,1}. \end{aligned} \tag{3.55}$$

The first inequality follows from the trivial upper bound of the regret in an episode $\delta_{k,1} \leq H$. The second inequality comes from the fact that each state-action pair can cause at most H episodes to be non-typical; therefore there are at most SAH non-typical episodes.

Next, we have:

$$\sum_{k \in [K]_{typ}} \delta_{k,1} = \sum_{k}^{K} \delta_{k,1} \mathbb{I}\{k \in [K]_{typ}\}. \tag{3.56}$$

From here we write $\mathbb{I}_k = \mathbb{I}\{k \in [K]_{typ}\}$ for brevity.

Lemma 3 in [Azar et al., 2017] implies that, for all $k \in [K]$,

$$\delta_{k,1} \leq e \sum_{h=1}^{H} \left[\varepsilon_{k,h} + 2\sqrt{L}\bar{\varepsilon}_{k,h} + c_{1,k,h} + b_{k,h} + c_{4,k,h} \right]. \tag{3.57}$$

where $c_{4,k,h} = \frac{4SH^2L}{N_k(s_h^k, a_h^k)}$, $\varepsilon_{k,h}$ and $\bar{\varepsilon}_{k,h}$ are martingale difference sequences which, by Lemma

5 in [Azar et al., 2017], satisfy

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\varepsilon_{k,h} &\leq H\sqrt{KHL}\\ \sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\varepsilon}_{k,h} &\leq \sqrt{KH},\end{aligned} \tag{3.58}$$

and $c_{1,k,h}$ is a confidence interval to be defined later.

Plugging Equation 3.57 into Equation 3.56 and combining with Equation 3.58, we obtain:

$$\begin{aligned}\sum_{k\in[K]_{typ}}\delta_{k,1} &\leq e\sum_{k=1}^{K}\left(\sum_{h=1}^{H}\left[\varepsilon_{k,h}+2\sqrt{L}\bar{\varepsilon}_{k,h}+c_{1,k,h}+b_{k,h}+c_{4,k,h}\right]\right)\mathbb{I}_k\\ &= e\left[\left(\sum_{k=1}^{K}\mathbb{I}_k\sum_{h=1}^{H}(\varepsilon_{k,h}+2\sqrt{L}\bar{\varepsilon}_{k,h})\right)+\left(\sum_{k=1}^{K}\mathbb{I}_k\sum_{h=1}^{H}(b_{k,h}+c_{1,k,h}+c_{4,k,h})\right)\right]\\ &\leq e\left[\left(\sum_{k=1}^{K}\sum_{h=1}^{H}(\varepsilon_{k,h}+2\sqrt{L}\bar{\varepsilon}_{k,h})\right)+\left(\sum_{k=1}^{K}\sum_{h=1}^{H}(b_{k,h}\mathbb{I}_k+c_{1,k,h}\mathbb{I}_k+c_{4,k,h}\mathbb{I}_k)\right)\right]\\ &\leq e\left[\left(H\sqrt{KHL}+2\sqrt{L}\sqrt{KH}\right)+\left(\sum_{k=1}^{K}\sum_{h=1}^{H}(b_{k,h}\mathbb{I}_k+c_{1,k,h}\mathbb{I}_k+c_{4,k,h}\mathbb{I}_k)\right)\right]\\ &= e\left[\left((H+2)\sqrt{KHL}\right)+\left(\sum_{k=1}^{K}\sum_{h=1}^{H}(b_{k,h}\mathbb{I}_k+c_{1,k,h}\mathbb{I}_k+c_{4,k,h}\mathbb{I}_k)\right)\right]\end{aligned}$$

Note that the second inequality follows from the fact that $\mathbb{I}_k \leq 1$, and the last inequality follows directly from Equation 3.58.

Let $\mathbb{I}_{k,h} = \mathbb{I}\{N_k(s_h^k, a_h^k) \geq H\}$. By the definition of a "typical" episode, $\mathbb{I}_k = 1$ implies that $\mathbb{I}_{k,h} = 1$ for all h. It follows that $\mathbb{I}_k \leq \mathbb{I}_{k,h}$. Thus,

$$\sum_{k\in[K]_{typ}}\delta_{k,1} \leq e\left((H+2)\sqrt{KHL}+\sum_{i=1}^{K}\sum_{j=1}^{H}(b'_{k,h}+c'_{1,k,h}+c'_{4,k,h})\right), \tag{3.59}$$

where $b'_{k,h} = b_{k,h}\mathbb{I}_{k,h}$, $c'_{1,k,h} = c_{1,k,h}\mathbb{I}_{k,h}$ and $c'_{4,k,h} = c_{4,k,h}\mathbb{I}_{k,h}$.

Next, we compute $c_{1,k,h}$. In Equation (32) in [Azar et al., 2017], $c_{1,k,h}$ corresponds to the confidence interval of

$$(\hat{P}_h^\pi - P_h^\pi)V_{h+1}^*(s_h^k) = \sum_{s'\in\mathcal{S}}\left[\hat{P}(s' \mid s_h^k, a_h^k) - P_h(s' \mid s_h^k, a_h^k)\right]V_{h+1}^*(s').$$

Equation (9) in [Azar et al., 2017] computes a confidence interval for this term using the Bernstein inequality. Instead, we use the Hoeffding inequality and obtain

$$[(\hat{P}_h^\pi - P_h^\pi)V_{h+1}^*] \leq H\sqrt{\frac{L}{2N_k(s_h^k, a_h^k)}} = c_{1,k,h}. \tag{3.60}$$

Combining Equations 3.60, 3.59 and 3.55, the total regret is bounded as

$$\text{Regret} \leq SAH^2 + e\left((H+2)\sqrt{KHL} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}(b'_{k,h} + c'_{1,k,h} + c'_{4,k,h})}_{(a)}\right) \tag{3.61}$$

where $b'_{k,h} = \frac{7HL\mathbb{I}_{k,h}}{\sqrt{N_k(s_h^k,a_h^k)}}$, $c'_{1,k,h} = \frac{H\sqrt{L}\mathbb{I}_{k,h}}{\sqrt{2N_k(s_h^k,a_h^k)}}$ and $c'_{4,k,h} = \frac{4SH^2L\mathbb{I}_{k,h}}{N_k(s_h^k,a_h^k)}$.

We focus on the third and dominant term (a). As $b_{k,h} \geq c_{1,k,h}$, this term can be upper bounded by

$$\begin{aligned}(a) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\frac{8HL\mathbb{I}_{k,h}}{\sqrt{N_k(s_h^k,a_h^k)}} + \frac{4SH^2L\mathbb{I}_{k,h}}{N_k(s_h^k,a_h^k)}\right] \quad \text{(since } L > 1\text{)} \\ &= 8HL\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}_{k,h}}{\sqrt{N_k(s_h^k,a_h^k)}}}_{(b)} + 4SH^2L\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}_{k,h}}{N_k(s_h^k,a_h^k)}}_{(c)}.\end{aligned} \tag{3.62}$$

We bound (b) and (c) separately.

First, we bound (b). We introduce the following lemma, which is an analogy to Lemma 19 in [Jaksch et al., 2010] in the finite-horizon setting.

Lemma 3.E.1. Let $H \geq 1$. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as

$$\begin{aligned} Z_0 &\geq H \\ Z_k &= Z_{k-1} + z_k \qquad \text{for } k \geq 1. \end{aligned}$$

Then, for all $n \geq 1$,

$$\sum_{k=1}^{n}\frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2}+1)\sqrt{Z_n}.$$

Using Lemma 3.E.1, we can bound (b) by Lemma 3.E.2.

Lemma 3.E.2. Denote $v_i(s,a) = \sum_{j=1}^{H} \mathbb{I}(a_{i,j} = a, s_{i,j} = s)$ the number of times the state-action pair (s,a) is visited during episode i, and let $\tau(s,a) = \arg\min_{k\in[K]}\{N_k(s,a) \geq H\}$ be the first episode where the state-action pair (s,a) is visited at least H times. Then,

$$(b) \leq (\sqrt{2}+1)\sqrt{SAKH}. \tag{3.63}$$

Proof. By definition, $N_i(s,a) = \sum_{k=1}^{i-1} v_k(s,a)$. Regrouping the sum in (b) by (s,a), we have

$$\begin{aligned}
(b) &= \sum_{s,a}\sum_{i=1}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}}\mathbb{I}\{N_i(s,a) \geq H\} \\
&= \sum_{s,a}\left(\sum_{i=1}^{\tau(s,a)-1} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}}\mathbb{I}\{N_i(s,a) \geq H\} + \sum_{i=\tau(s,a)}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}}\right) \\
&= \sum_{s,a}\sum_{i=\tau(s,a)}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}} \\
&\leq \sum_{s,a}(\sqrt{2}+1)\sqrt{N_K(s,a)+v_K(s,a)} \\
&\leq (\sqrt{2}+1)\sqrt{SAKH}.
\end{aligned}$$

where the last two inequalities follow from Lemma 3.E.1, the Cauchy-Schwarz inequality and the fact that $\sum_{s,a} N_K(s,a) \leq KH$. □

In order to bound the term (c) in Equation 3.62, we use the following lemma, which is a variant of Lemma 3.E.1 and was stated in [Azar et al., 2017] without proof.

Lemma 3.E.3. Let $H \geq 1$. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as

$$\begin{aligned}
Z_0 &\geq H \\
Z_k &= Z_{k-1} + z_k \qquad \text{for } k \geq 1.
\end{aligned}$$

Then, for all $n \geq 1$,

$$\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq \sum_{j=1}^{Z_n - Z_0} \frac{1}{j} \leq \ln(Z_n - Z_0) + 1.$$

Proof. The second half follows immediately from existing results for the partial sum of the harmonic series. We prove the first half of the inequality by induction. By definition of the two sequences, $Z_k \geq H \geq 1$ and $z_k \leq H \leq Z_{k-1}$ for all k. At $n = 1$, if $z_1 = 0$ then the inequality trivially holds. If $z_1 > 0$, then $Z_1 - Z_0 = z_1$ and

$$\frac{z_1}{Z_0} \leq \frac{z_1}{H} = \left(\underbrace{\frac{1}{H} + \cdots + \frac{1}{H}}_{z_1 \text{ terms}} \right) \leq 1 + \frac{1}{2} + \cdots + \frac{1}{z_1}$$

since $z_1 \leq H$.

For $n > 1$, by the induction hypothesis, we have

$$\begin{aligned}
\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} &= \sum_{k=1}^{n-1} \frac{z_k}{Z_{k-1}} + \frac{z_n}{Z_{n-1}} \\
&\leq \left(\sum_{j=1}^{Z_{n-1}-Z_0} \frac{1}{j} \right) + \frac{z_n}{Z_{n-1}} \\
&= \left(\sum_{j=1}^{Z_{n-1}-Z_0} \frac{1}{j} \right) + \left(\underbrace{\frac{1}{Z_{n-1}} + \cdots + \frac{1}{Z_{n-1}}}_{z_n \text{terms}} \right) \\
&\leq \left(\sum_{j=1}^{Z_{n-1}-Z_0} \frac{1}{j} \right) + \left(\frac{1}{Z_{n-1} - Z_0 + 1} + \cdots + \frac{1}{Z_{n-1} - Z_0 + z_n} \right) \\
&= \sum_{j=1}^{Z_n - Z_0} \frac{1}{j},
\end{aligned}$$

where the last inequality follows from $z_n \leq Z_0$. Therefore, the induction hypothesis holds for all $n \geq 1$. □

Using Lemma 3.E.3, the term (c) can be bounded similarly to term (b) as follows:

Lemma 3.E.4. With $v_i(s, a)$ and $\tau(s, a)$ defined in Lemma 3.E.2, we have

$$(c) \leq SAL + SA.$$

Proof. We write (c) as

$$\begin{aligned}
(c) &= \sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}\{N_i(s,a)\geq H\}}{N_i(s_{i,j},a_{i,j})} \\
&= \sum_{s,a}\sum_{i=1}^{K}\frac{v_i(s,a)}{N_i(s,a)}\mathbb{I}\{N_i(s,a)\geq H\} \\
&\leq \sum_{s,a}\left(\sum_{i=1}^{\tau(s,a)-1}\frac{v_i(s,a)}{N_i(s,a)}\mathbb{I}\{N_i(s,a)\geq H\} + \sum_{i=\tau(s,a)}^{K}\frac{v_i(s,a)}{N_i(s,a)}\right) \\
&= \sum_{s,a}\sum_{i=\tau(s,a)}^{K}\frac{v_i(s,a)}{N_i(s,a)} \\
&\leq \sum_{s,a}\left(\ln\left(N_K(s,a)+v_K(s,a)-N_{\tau(s,a)}(s,a)\right)+1\right)
\end{aligned}$$

where the last inequality follows from Lemma 3.E.3. Trivially bounding the logarithm term by $\ln(KH)$, we obtain

$$(c) \leq SA\ln(KH) + SA \leq SAL + SA.$$

□

Combining Lemma 3.E.2 and Lemma 3.E.4, we obtain

$$\begin{aligned}
(a) &\leq 8HL((\sqrt{2}+1)\sqrt{SAKH}) + 4SH^2L(SAL+SA) \\
&\leq 20HL\sqrt{SAKH} + 5S^2AH^2L^2.
\end{aligned}$$

Substituting this into Equation 3.61, we obtain

$$\begin{aligned}
\text{Regret} &\leq SAH^2 + e(H+2)\sqrt{KHL} + e20HL\sqrt{SAKH} + e5S^2AH^2L^2 \\
&\leq 67HL\sqrt{SAKH} + 15S^2AH^2L^2.
\end{aligned}$$

3.F Removing the assumption on the hitting time

GOSPRL [Tarbouriech et al., 2021, Lemma 3] guaranteed that in the undiscounted infinite horizon setting, with $H_0 = O(\frac{DS^2A}{\lambda^2})$, Lemma 3.5.3 holds with high probability. Thus, in the episodic finite horizon setting, by setting $H_0 = c\frac{DS^2A}{\lambda^2}$ for some appropriately large constant $c > 0$ and applying GOSPRL in each episode we obtain a tight bound in the

dependency of K and λ for communicating MDPs. One difficulty in this approach is both c and D are unknown. One possible way to overcome this is to apply the doubling-trick as following: at the beginning of episode k, we set $H_0 = c_k \frac{S^2 A}{\lambda^2}$, where $c_1 = 1$. If the learner successfully visits every state-action pair at least N times after H_0 steps, we set $c_{k+1} = c_k$. Otherwise, $c_{k+1} = 2c_k$. There are at most $\log_2(cD)$ episodes with failed exploration until c_k is large enough so that with high probability, all the subsequent episodes will have successful explorations. Moreover, the horizons of the clustering and learning phases change at most $\log_2(cD)$ times. The full analysis of this approach is not in the scope of this paper and is left to future work.

3.G Using samples in both phases for regret minimization

One of the results from previous works on the stochastic infinite-horizon multi-task setting [Brunskill and Li, 2013] is that in the cluster-then-learn paradigm, the samples collected in the their first stage (before all models have been seen at least once) can be used to accelerate the learning in their second stage (after all models have been seen at least once). In this work, we study the similar effects at the phase level. Specifically, in the finite horizon setting, the clustering phase is always followed by the learning phase; therefore it is desirable to use the samples collected in the clustering phase to improve the regret bound of the learning phase.

Our goal is to improve the regret of stage 1 in Algorithm 3.4. The reason that we focus on Stage 1 is two-fold:

- In case Assumption 3.5.8 does not hold, i.e. K_1 is close to K, the total regret is dominated by the regret of stage 1. Given that the length of the clustering phase H_0 is already of the same order $O(S^2A)$ with respect to the state-of-the-art bound of the recently proposed GOSPRL algorithm [Tarbouriech et al., 2021], without further assumptions we conjecture that it is difficult to improve H_0 substantially, and thus we focus on improving the learning phase.

- In stage 1, every state-action pair is uniformly visited at least N times before the learning phase. This uniformity allows us to study their impact in a systematic way without any further assumptions.

Algorithm 3.7 UCBVI-CH with external samples

Input: Number of episode K, horizon H, failure probability p, number of external samples for each state-action pair N

Initialize two empty collections $\mathcal{H}$ and $\mathcal{B}$
for episode $k = 1, 2, \ldots, K$ **do**
  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    for *counter* $= 1, 2, \ldots, N$ **do**
      The oracle draws s' from $P(\cdot \mid s, a)$
      Add (s, a, s') to $\mathcal{B}$
    end
  end
  $\pi_k = \texttt{UCBVI-CH}(\mathcal{H} \cup \mathcal{B})$
  Observe the starting state s_1
  for $h = 1, 2, \ldots, H$ **do**
    Learner takes action $a_h = \pi_k(s_h)$
    Observe state s_{h+1}
    Add (s_h, a_h, s_{h+1}) to $\mathcal{H}$
  end
end

Using samples collected in both phases for the learning phase in Algorithm 3.3 is equivalent to using the policy

$$\pi_k = \texttt{UCBVI-CH}(\mathcal{C}_{id})$$

for the learning phase, since $\mathcal{C}_{id}$ contains both $\mathcal{C}_{id}^{model}$ and $\mathcal{C}_{id}^{regret}$.

The regret minimization process in the learning phase is now equivalent to learning single-task episodic RL where at the beginning of each episode, the learner is given SAN more (s, a, s') samples, in which the transition function $P(\cdot \mid s, a)$ of each (s, a) is sampled i.i.d. N times. We extend the UCBVI-CH algorithm in [Azar et al., 2017] to this new setting and obtain Algorithm 3.7. The bonus function of episode k in UCBVI-CH is set to

$$b_k(s, a) = 7HL_N \sqrt{\frac{1}{N_k(s, a) + kN}}, \tag{3.64}$$

where $L_N = \ln(5SAK(H + N)/p)$.

The regret of this algorithm is bounded in the following theorem (proved in Appendix 3.H).

Theorem 3.G.1. Given a constant $p \in (0, 1)$. With probability at least $1 - p$, the regret of

Algorithm 3.7 is bounded by

$$\begin{aligned}&\text{Regret}(K)\\&\leq \frac{SAH^2}{N+1} + e(H+1)\sqrt{KL_N} + 60\sqrt{\frac{2H-1}{N+2H-1}}H^{3/2}L_N\sqrt{SAK}\\&\quad + 15\frac{2H-1}{N+2H-1}S^2AH^2L_N^2.\end{aligned}$$

It can be observed that when $N = 0$, this bound recovers the bound of UCBVI-CH (up to a constant factor). Intuitively, when N is small compared to H, then the regret should still be of order $O(H\sqrt{SAKH})$ since most of the useful information for learning still comes from exploring the environment. As N increases, since the logarithmic term L_N increases much slower compared to $O(1/\sqrt{N})$, the dominant term $O(\sqrt{\frac{2H-1}{N+2H-1}}H^{3/2}L_N\sqrt{SAK})$ converges to 0.

Using Theorem 3.G.1 and $H_1 \leq H$, we can directly bound the regret of each model m that is given in $\mathcal{K}_m$:

Lemma 3.G.2. The stage-1 regret of each model m is

$$\begin{aligned}\text{Regret}_{Stage1}(m, K_m) \leq &\frac{SAH_1^2}{N+1} + e(H_1+1)\sqrt{K_mL_N}\\&+ 60\sqrt{\frac{2H_1-1}{N+2H_1-1}}H_1^{3/2}L_N\sqrt{SAK_m} + 15\frac{2H_1-1}{N+H_1-1}S^2AH_1^2L_N^2.\end{aligned}$$

where $L_N = \ln(5SAK(H+N)/p)$.

Adding up the bound in Lemma 3.G.2 for all models $m \in \mathcal{M}$ and applying the Cauchy-Schwarz inequality, we obtain the total regret bound of Stage 1:

Theorem 3.G.3.

$$\begin{aligned}\text{Regret}_{Stage1} \leq &K_1H_0 + \frac{MSAH_1^2}{N+1} + e(H_1+1)\sqrt{MKL_N}\\&+ 60\sqrt{\frac{2H_1-1}{N+2H_1-1}}H_1^{3/2}L_N\sqrt{MSAK} + 15M\frac{2H_1-1}{N+H_1-1}S^2AH_1^2L_N^2.\end{aligned}$$

In our setting, recall that $N = O\left(\frac{S}{\lambda^2}\right)$ and $H_0 = O(DSAN) = O(DS^2A/\lambda^2)$. Since we assumed that $SA << H$, we also have $N \ll H_1 = H - H_0$, and thus the bound in

Theorem 3.G.3 is an improvement from the bound for stage 1 in the proof of Theorem 3.5.9, albeit the order stays the same. Intuitively, this means that the length of the learning phase is much larger than the length of the clustering phase, and therefore the learner spends more time on learning the optimal policy. When the length of the learning phase is small compared to N, then the samples collected in the clustering phase significantly reduce the regret bound of the learning phase. Therefore, Algorithm 3.7 also accelerates the learning phase after the exploration phase, which is consistent with findings on the stochastic infinite-horizon multi-task setting in [Brunskill and Li, 2013].

3.H Proofs for Appendix 3.G

We analyze the regret of the UCBVI-CH algorithm with external samples, where at the beginning of each episode, each state-action pair receives $N \geq 1$ additional samples drawn i.i.d from the transition function $P(\cdot \mid s, a)$.

Adapting from Equation 3.61, the regret of E-UCBVI-CH can be bounded by

$$\begin{aligned}\text{Regret}(K) \leq \frac{SAH^2}{N+1} + e(H+1)\sqrt{KHL_N} \\ + e\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\left[\frac{8HL_N\mathbb{I}_{i,j}}{\sqrt{kN+N_i(s_{i,j},a_{i,j})}} + \frac{4SH^2L_N\mathbb{I}_{i,j}}{kN+N_i(s_{i,j},a_{i,j})}\right]}_{(a)},\end{aligned} \tag{3.65}$$

where $\mathbb{I}_{i,j} = \mathbb{I}\{N_i(s_{i,j}, a_{i,j}) \geq H\}$ as defined in Appendix 3.E.

The first term $\frac{SAH^2}{N+1}$ bounds the total regret of episodes where a state-action pair is visited less than H times: in each episode where a pair (s, a) is visited at least once there are at least $N+1$ more samples of this pair, and therefore there can be at most $\frac{SAH}{N+1}$ such episodes.

Similar to Appendix 3.E, we bound (a) by bounding its two components (b) and (c) where

$$(a) = 8HL_N\left(\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}_{i,j}}{\sqrt{kN+N_i(s,a)}}}_{(b)}\right) + 4SH^2L_N\left(\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}_{i,j}}{kN+N_i(s,a)}}_{(c)}\right).$$

In order to bound (b), we first prove the following technical lemma, which quantifies the

fraction of the regret that is reduced when using external samples.

Lemma 3.H.1. Suppose two constants $N \geq 1, H \geq 1$ are given. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as

$$\begin{aligned} Z_0 &\leq 2H - 1 \\ Z_k &= Z_{k-1} + z_k \qquad \text{for } k \geq 1 \end{aligned}$$

Then, for all k,

$$\frac{z_k}{\sqrt{kN + Z_{k-1}}} \leq \sqrt{\frac{(k+1)H - 1}{kN + (k+1)H - 1}} \frac{z_k}{\sqrt{Z_{k-1}}}.$$

Proof. If $z_k = 0$, then the claim is trivially true. For $z_k > 0$, the claim is equivalent to

$$\begin{aligned} & \frac{1}{\sqrt{kN + Z_{k-1}}} \leq \sqrt{\frac{(k+1)H - 1}{kN + (k+1)H - 1}} \frac{1}{\sqrt{Z_{k-1}}} \\ \Leftrightarrow & \sqrt{(kN + (k+1)H - 1)}\sqrt{Z_{k-1}} \leq \sqrt{(k+1)H - 1}\sqrt{kN + Z_{k-1}} \\ \Leftrightarrow & \qquad Z_{k-1} \leq (k+1)H - 1, \end{aligned}$$

which is true, since $Z_{k-1} = Z_0 + \sum_{i=1}^{k-1} z_k \leq Z_0 + \sum_{i=1}^{k-1} H \leq 2H - 1 + (k-1)H = (k+1)H - 1$. □

Corollary 3.H.2. Suppose two constants $N \geq 1, H \geq 1$ are given. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as

$$\begin{aligned} 1 \leq Z_0 &\leq 2H - 1 \\ Z_k &= Z_{k-1} + z_k \qquad \text{for } k \geq 1 \end{aligned}$$

Then, for all $n \geq 1$,

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{kN + Z_{k-1}}} \leq \sum_{k=1}^{n} \sqrt{\frac{(k+1)H - 1}{kN + (k+1)H - 1}} \frac{z_k}{\sqrt{Z_{k-1}}} \leq \sqrt{\frac{2H - 1}{N + 2H - 1}} \sum_{k=1}^{n} \frac{z_k}{\sqrt{kN + Z_{k-1}}}.$$

Proof. The first half of the claim is true, following Lemma 3.H.1. We now show that the second half is true. Consider the following function

$$f(x) = \frac{(x+1)H - 1}{xN + (x+1)H - 1}$$

The derivative is $f'(x) = \frac{N(1-H)}{(xN+(x+1)H-1)^2}$. Since $H \geq 1$, we have $f'(x) \leq 0 \ \forall x$, and therefore $f(x)$ is decreasing. It follows that for $k \geq 1$,

$$f(k) = \frac{(k+1)H - 1}{kN + (k+1)H - 1} \leq f(1) = \sqrt{\frac{2H-1}{N+2H-1}}.$$

□

Using Corollary 3.H.2, we can bound (b) as following.

Lemma 3.H.3. With $v_i(s,a)$ and $\tau(s,a)$ defined in Lemma 3.E.2, we have

$$(b) \leq \sqrt{\frac{2H-1}{N+2H-1}}(\sqrt{2}+1)\sqrt{SAKH}.$$

Proof. We can write (b) as follows

$$(b) = \sum_{s,a} \sum_{i=\tau(s,a)}^{K} \frac{v_i(s,a)}{\sqrt{iN + N_i(s,a)}}.$$

By definition of $\tau(s,a)$:

$$N_{\tau(s,a)} = N_{\tau(s,a)-1} + v_{\tau(s,a)-1} \leq H - 1 + H = 2H - 1.$$

□

Applying Corollary 3.H.2 and Lemma 3.E.2 we obtain

$$\begin{aligned}(b) &\leq \sqrt{\frac{2H-1}{N+2H-1}} \sum_{s,a} \sum_{i=\tau(s,a)}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}} \\ &\leq \sqrt{\frac{2H-1}{N+2H-1}}(\sqrt{2}+1)\sqrt{SAKH}.\end{aligned}$$

Next, we bound (c). Using similar techniques in Lemma 3.H.1 and Corollary 3.H.2, we can show that the following claims are true.

Lemma 3.H.4. Given two constants $N \geq 0, H \geq 1$. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as

$$\begin{aligned} 1 \leq Z_0 &\leq 2H-1 \\ Z_k &= Z_{k-1} + z_k \qquad \text{for } k \geq 1 \end{aligned}$$

Then, for all k,

$$\frac{z_k}{kN + Z_{k-1}} \leq \frac{(k+1)H - 1}{kN + (k+1)H - 1} \frac{z_k}{Z_{k-1}}.$$

And for all $n \geq 1$,

$$\sum_{k=1}^{n} \frac{z_k}{kN + Z_{k-1}} \leq \frac{2H-1}{N+2H-1} \sum_{k=1}^{n} \frac{z_k}{Z_{k-1}}.$$

Consequently, (c) is bounded in the following corollary.

Corollary 3.H.5. With $v_i(s,a)$ and $\tau(s,a)$ defined in Lemma 3.E.2, we have

$$(c) \leq \frac{2H-1}{N+2H-1}(SAL_N + SA).$$

Combining Corollaries 3.H.2 and 3.H.5 we obtain

$$\begin{aligned} (a) &\leq 8HL\sqrt{\frac{2H-1}{N+2H-1}}((\sqrt{2}+1)\sqrt{SAKH}) + 4SH^2L_N\frac{2H-1}{N+2H-1}(SAL_N + SA) \\ &\leq \sqrt{\frac{2H-1}{N+2H-1}}20HL_N\sqrt{SAKH} + \frac{2H-1}{N+2H-1}5S^2AH^2L_N^2. \end{aligned}$$

and the total regret is

$$\begin{aligned} \text{Regret}(K) &\leq \frac{SAH^2}{N+1} + e(H+1)\sqrt{KL_N} && (3.66) \\ &\quad + e\left(\sqrt{\frac{2H-1}{N+2H-1}}20HL_N\sqrt{SAKH} + \frac{2H-1}{N+2H-1}5S^2AH^2L_N^2\right). && (3.67) \end{aligned}$$

3.I Experimental Details

Transition functions. Figure 3.5 illustrate the 4×4 gridworld environment of the four MDPs in $\mathcal{M}$. The rows are numbered top to bottom from 0 to 3. The columns are numbered left to right from 0 to 3. The starting state s_1 is at position $(1, 1)$. In every state, the probability of success of all actions is 0.85. When an action is unsuccessful, the probability of being in one of the other adjacent cells is equally divided from the remaining probability of 0.15. There are several exceptions:

- In the four corners, if the agent takes an action in the direction of the border then with probability of 0.7 it will stay in the same corner, and with probability 0.3 it will end up in the cell in the opposite direction. For example, if the agent is at $(0, 0)$ and takes action `up`, then with probability 0.3 it will actually goes down to the cell $(1, 0)$.
- Each of four MDPs have an easy-to-reach corner and three hard-to-reach corners. The easy-to-reach corners in models m_1, m_2, m_3 and m_4 are $(0, 0), (0, 3), (3, 0)$ and $(3, 3)$, respectively. In each of these model, the probability of success of an action that leads to one of the hard-to-reach corners is 0.2, except for the $(3, 3)$ corner where this probability is 0.3. For example, in model m_1, taking action `right` in cell $(0, 2)$ has probability of success equal to 0.2 while taking the action `down` in cell $(2, 3)$ has probability of success equal to 0.3.
- On the four edges, any action that takes the agent out of the grid has probability of success equal to 0, and the agent ends up in one of the three adjacent cells with equal probability of $\frac{1}{3}$. For each example, taking action `up` in position $(0, 1)$ will take the agent to one of the three positions $(0, 0), (0, 1)$ and $(1, 1)$ with probability $\frac{1}{3}$.

Under this construction, the seperation level is $\lambda = 1.2999$. One example of a λ-distinguishing set of optimal size is $\Gamma = \{(1, 0), (8, 3), (2, 1)\}$. One example of a $\lambda/2$-distinguishing but not λ-distinguishing is $\Gamma^{\lambda/2} = \{(11, 3), (4, 2), (13, 0)\}$.

Performance metric. At the end of each episode, the two AOMultiRL agents and the one-episode UCBVI agent obtain their estimated model $\hat{P}$. The estimated optimal policy computed based on $\hat{P}$ is run for $H_1 = 200$ steps starting from $(1, 1)$. The average per-episode reward (APER) in episode $k = 1, 2, \ldots, K$ of an agent is defined as

$$\text{APER}(k) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{H_1} r_{i,j}}{k} \tag{3.68}$$

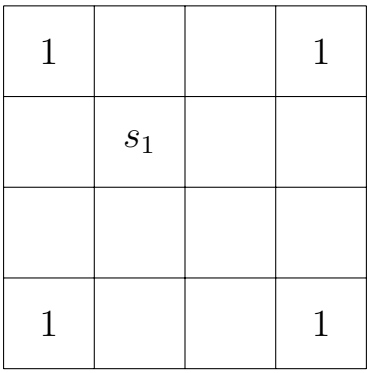

Figure 3.5: A 4×4 gridworld MDP with start state at $(1, 1)$ and reward of 1 in four corners

where $r_{i,j} = r(s_j^i, a_j^i)$ the reward this agent received in step j of episode i.

Horizon settings For AOMultiRL2, the horizons of the clustering phase in two stages are different since the distinguishing sets in the two stages are different. In order to make a fair comparison with other algorithms, the horizon of the learning phase is set to $H_1 = 0$ in stage 1 and $H_1 = 200$ in stage 2. Since we assumed that stage 2 is dominant, the goal of the experiment is to examine whether a $\lambda/2$-distinguishing set can be discovered and how effective that set can be. We observe that AOMultiRL2 is able to discover the same $\lambda/2$-distinguishing set $\{(14, 1), (7, 2), (13, 0)\}$ in all 10 runs. Since this set also has an optimal size of 3, in stage 2 the clustering phase's horizon H_0 of AOMultiRL2 is identical to that of AOMultiRL1.

Chapter 4

Near-Optimal Per-Action Regret Bounds for Sleeping Bandits

4.1 Introduction

The multi-armed bandit (MAB) framework and its variants have been widely used for practical applications in various domains such as clinical trials, finance and recommender systems [Bouneffouf et al., 2020]. In the standard MAB framework, a learner interacts with K arms over T rounds. In each round, the learner chooses to observe the loss of one of the arms. While the losses of the arms in each round are unknown to the learner, the number of arms K is assumed to be fixed in every round. However, this assumption does not always hold in practice. For example, in drug testing where each arm is a drug type, certain types of drugs can only be tested in some certain rounds, or new and more effective drugs might be available only in later rounds. In such scenarios, it is important to have algorithms capable of learning with time-varying sets of available arms. This is the *sleeping bandits* setting [Kleinberg et al., 2010], where in each round $t = 1, \ldots, T$, only a subset $\mathbb{A}_t \subseteq \{1, \ldots, K\}$ of active arms are accessible to the learner.

The *sleeping experts* setting (also known as the specialist setting) [Blum, 1997; Freund and Schapire, 1997] is the full-information feedback variant of this problem, in which the losses of the active arms are revealed at the end of each round. Prior works on sleeping experts have mainly used two different notions of regret to measure the performance of a learner, namely per-action regret [Blum and Mansour, 2007; Gaillard et al., 2014; Luo and Schapire, 2015] and ordering regret [Kleinberg et al., 2010; Kanade and Steinke, 2014; Neu and Valko, 2014]. Besides the notions of regrets, an important characteristic of the setting is

the stochastic or adversarial nature of the sets $\mathbb{A}_t$ and the arms' losses. [Kanade and Steinke, 2014] indicated that obtaining a sublinear ordering regret bound is computationally hard when both $\mathbb{A}_t$ and losses are adversarial. As a result, subsequent works on sleeping bandits usually assume at least one component to be stochastic [Slivkins, 2013; Neu and Valko, 2014; Slivkins, 2014; Saha et al., 2020]. Recently, [Gaillard et al., 2023] developed a new notion of regret for sleeping bandits called *sleeping internal regret*, which can be minimized efficiently in the fully adversarial setting with adversarial $\mathbb{A}_t$ and adversarial losses.

Our work focuses on minimizing the per-action regret in the fully adversarial setting. This notion of regret compares the cumulative loss of the learner to that of a single best arm in hindsight during the rounds in which that arm was active. To the best of our knowledge, in the fully adversarial setting, no prior work has focused on directly deriving optimal per-action regret bounds. We are interested in obtaining more fine-grained bounds that depend on the maximum number of active arms in any round A, where $A = \max_{t=1,\dots,T} |\mathbb{A}_t| \leq K$. The smallest existing bound is the $O(K\sqrt{TA\ln K})$ bound by [Gaillard et al., 2023], obtained indirectly from minimizing the internal sleeping regret. This bound can be much larger than an $\Omega(\sqrt{TA})$ minimax lower bound implied by suitably adapting an existing minimax lower bound construction for standard bandits [Auer et al., 2002b]. Moreover, as we show in this work, the factor of K outside the square root can be eliminated entirely.

Another motivation for bounding the per-action regret in sleeping bandits is its implication on the adaptive and tracking regrets in standard non-sleeping bandits. Adaptive regret (also known as interval regret) [Hazan and Seshadhri, 2009; Luo et al., 2018] is the regret against a fixed arm on a time interval, while tracking regret (also known as shifting or switching regret) [Herbster and Warmuth, 1998] is the regret against a sequence of arms over T rounds. Previous work obtained adaptive and tracking regret bounds for standard non-sleeping experts via a reduction to regret bounds for sleeping experts [Freund et al., 1997; Adamskiy et al., 2016]. In bandits, instead of the reduction to sleeping bandits, $\tilde{O}(\sqrt{T})$ bounds[1] on tracking and adaptive regret have been obtained via Fixed Share [Auer et al., 2002b; Herbster and Warmuth, 1998; Luo et al., 2018]. Our work shows that the reduction to sleeping bandits also leads to $\tilde{O}(\sqrt{T})$ adaptive and tracking bounds.

Overview of Main Results and Techniques

We extend the EXP3 [Auer et al., 2002b], EXP3-IX [Neu, 2015], Follow-The-Regularized-Leader (FTRL) with Tsallis entropy [Audibert and Bubeck, 2009; Abernethy et al., 2015]

[1] $\tilde{O}$ hides terms in K, number of switches S and $\ln T$.

Table 4.1: A Summary of Bounds on Per-Action Regret. Hyphens indicate bounds that are either not comparable to a per-action regret bound or unavailable.

Algorithms	**Adversarial?**	**Pseudo**	**High-Prob**
AUER [Kleinberg et al., 2010]	No (sto. losses)	$\sqrt{TK \ln T}$	-
Sleeping-EXP3 [Saha et al., 2020]	No (sto. $\mathbb{A}_t$)	-	-
SR_MAB [Blum and Mansour, 2007]	Yes	$K^2\sqrt{TA \ln K}$	-
SI-EXP3 [Gaillard et al., 2023]	Yes	$K\sqrt{TA \ln K}$	-
SB-EXP3 (this work)	Yes	$\sqrt{TA \ln K}$	$\sqrt{TA \ln(K/\delta)}$
FTARL (this work)	Yes	$\sqrt{T\sqrt{AK}}$	$\sqrt{T\sqrt{AK}} + \sqrt{TA \ln\left(\frac{K}{\delta}\right)}$

and EXP4 [Auer et al., 2002b] algorithms for standard bandits to sleeping bandits, obtaining new bounds that strictly generalize the existing bounds. Our results lead to new proofs for $\tilde{O}(\sqrt{T})$ adaptive and tracking regret bounds for standard bandits. The generalized algorithms and analyses are adapted to the bandit-feedback version of the experts that report their confidences setting [Blum and Mansour, 2007]. A summary of our contributions in comparison to prior works is in Table 4.1. All of our results hold for both pseudo-regret and high probability regret bounds. Our paper is organized as follows (all proofs are in the appendix):

- Section 4.3 introduces the $O(\sqrt{TA \ln K})$ and $O(\sqrt{T\sqrt{AK}})$ regret bounds for sleeping bandits. These bounds improve the best existing $O(K\sqrt{TA \ln K})$ bound, as well as recover the near-optimal $O(\sqrt{TK \ln K})$ and minimax $O(\sqrt{TK})$ bounds in non-sleeping bandits. Section 4.3.1 shows a novel algorithm called SB-EXP3 and its $O(\sqrt{TA \ln G_T})$ regret bound guarantee for sleeping bandits, where $G_T \leq K$ is the number of arms that were active at least once after T rounds. Its analysis relies on a new technique for bounding the growth of the potential function by decomposing the potential at round $t+1$ based on the set of active arms in round t. In Section 4.3.2, the $O(\sqrt{T\sqrt{AK}})$ bound is obtained by the Follow-the-Active-and-Regularized-Leader (FTARL) algorithm, an adaptation of FTRL with Tsallis entropy to sleeping bandits. Section 4.3.3 considers the bandit-feedback version of the experts that report their confidences setting. Applying SB-EXP3 to this setting leads to new regret bounds which replace the dependence on T and A by the cumulative confidence over T rounds.

- Section 4.4 studies the *bandits with advice from sleeping experts* setting and presents SE-EXP4, a generalized version of EXP4 algorithm. The analysis developed for SB-EXP3

also works for SE-EXP4, leading to the same $O(\sqrt{TK \ln M})$ regret bound of EXP4 with M experts. For standard bandits, this implies an $O(\sqrt{TK \ln(KT)})$ bound on adaptive regret. This bound is the same as the one obtained by [Luo et al., 2018], but with a different proof based on sleeping bandits instead of Fixed Share. This also implies both the $O(S\sqrt{KT \ln(KT)})$ and $O(\sqrt{SKT \ln(KT)})$ tracking regret bounds [Auer et al., 2002b; Neu, 2015] for unknown and known number of arm-switches S, respectively, where the latter is obtained via restarting SE-EXP4 after every T/S rounds.

- Section 4.5 defines the per-action strongly adaptive regret bound as a bound that depends only on T_a for every action a, where T_a is the number of active rounds of arm a. Extending the construction of [Daniely et al., 2015] for non-sleeping bandits to sleeping bandits, we show a linear $\Omega(T_a)$ per-action strongly adaptive lower bound. This implies that no algorithm can simultaneously guarantee an optimal per-action regret and sublinear $o(T_a)$ per-action regret for all arms.

4.2 Preliminaries

We consider the adversarial multi-armed bandit problem with K underlying arms, where K might be unknown. Let $[K] = \{1, 2, \ldots, K\}$. In round $t = 1, 2, \ldots, T$, a (possibly non-oblivious) adversary selects and reveals a set $\mathbb{A}_t \subseteq [K]$ of active arms to the learner. Let $I_{i,t} = 1$ (resp. $I_{i,t} = 0$) indicates that arm i is active (resp. inactive) in round t. Then, for each arm $i \in \mathbb{A}_t$, the adversary selects a (hidden) loss value $\ell_{i,t} \in [0, 1]$. The learner pulls one active arm $i_t \in \mathbb{A}_t$ and observes loss $\ell_{i_t,t}$.

The learner's goal is to compete with the best arm in hindsight. For an arm $a \in [K]$, the regret of the learner with respect to arm a is the difference in the cumulative loss of the learner and that of arm a over its active rounds:

$$R(a) = \sum_{t=1}^{T} I_{a,t}(\ell_{i_t,t} - \ell_{a,t}). \tag{4.1}$$

We prove two types of regret bounds. The first is

$$\max_{a \in [K]} \mathbb{E}_{i_1,\ldots,i_T}[R(a)] \leq \epsilon, \tag{4.2}$$

where the expectation is taken over the sequence of the learner's selected arms. In standard non-sleeping bandits, this corresponds to the notion of *pseudo-regret* [Auer et al., 2002b]. If

the adversary is oblivious, the pseudo-regret is equivalent to the expected regret. The second type of bound is

$$\Pr\left(\max_{a\in[K]} R(a) \leq \epsilon\right) \geq 1-\delta, \tag{4.3}$$

where the probability is taken over the sequence of the learner's selected arms.

Notations. Let $A_t = |\mathbb{A}_t|$ be the number of active arms in round t and $A = \max_{t\in[T]} A_t$ be the maximum value of A_t over T rounds. Let $\mathbb{G}_t = \cup_{s=1,\ldots,t}\mathbb{A}_s$ be the set of arms that are active at least once in the first t rounds. Let $G_t = |\mathbb{G}_t|$ be the size of $\mathbb{G}_t$. We write $\hat{\ell}_t = \ell_{i_t,t}$ for the learner's loss in round t. Let $\Delta_n = \{p \in \mathbb{R}^n \mid p_i \geq 0, \sum_{i=1}^n p_i = 1\}$ be the n-dimensional probability simplex.

Remark 4.2.1. The total number of underlying arms K is fixed before learning, and the adversary cannot change K. On the other hand, A_t and G_t are decided by the adversary and vary over time. As some arms may never be active, G_T can be strictly smaller than K.

4.3 Near-Optimal Regret Upper Bounds

In sleeping bandits, for any constant $A \in \{2, 3, \ldots, K\}$, there exists an $\Omega(\sqrt{TA})$ minimax pseudo-regret lower bound. The construction follows that of the minimax lower bound for standard bandits [Auer et al., 2002b] with A arms always active and $K - A$ arms always inactive over T rounds. In Section 4.3.1, we present SB-EXP3 (Algorithm 4.1) and its near-optimal $O(\sqrt{TA\ln G_T})$ pseudo-regret and high probability regret bounds. Note that SB-EXP3 does not require knowing K. In Section 4.3.2, we show that when K is known, an FTRL-based algorithm called FTARL (Algorithm 4.2) obtains an $O(\sqrt{TA\ln K})$ bound with negative Shannon entropy and an $O(\sqrt{T\sqrt{AK}})$ bound with Tsallis entropy as the regularization function.

4.3.1 Generalized EXP3 for Sleeping Bandits

In standard (non-sleeping) bandits, the EXP3 and EXP3-IX algorithms compute a distribution p_t over arms in round t based on the estimated regrets in previous rounds. Specifically, arms for which the estimated regrets are higher have larger probability of being sampled. In sleeping bandits, because arms can have different and even non-overlapping sets of rounds, it is unclear what kind of statistics about the arms should be maintained in each round. In particular, in any given round, the estimated regret for an arm might be higher than for

Algorithm 4.1 SB-EXP3 for sleeping bandits

Input: $\eta > 0, \gamma \geq 0$
Initialize $\tilde{q}_{i,1} = 1$ for $i = 1, 2, \ldots, K$
for each round $t = 1, \ldots, T$ **do**
  The adversary selects and reveals $\mathbb{A}_t$
  Compute $W_t = \sum_{i \in \mathbb{A}_t} I_{i,t} \tilde{q}_{i,t}$
  Compute $p_{i,t}$ by (4.6)
  Draw $i_t \sim p_t$ and observe $\hat{\ell}_t = \ell_{i_t,t}$
  Compute loss estimate $\tilde{\ell}_{i,t}$ by (4.4)
  Update $\tilde{q}_{i,t+1}$ by (4.5).

other arms due to it being active in more previous rounds, not because its losses are small in those rounds. Nevertheless, we will show that the estimated regret can be used effectively for selecting arms in each round.

Algorithm 4.1 illustrates Sleeping Bandits using EXP3 (SB-EXP3), an adaptation of EXP3 and EXP3-IX to sleeping bandits. In round t, SB-EXP3 computes a probability vector $p_t \in \Delta_{G_t}$ over $\mathbb{G}_t$. In the first round, p_1 is the uniform distribution. The learner samples $i_t \sim p_t$ and computes the loss estimates $\tilde{\ell}_t$ as follows:

$$\tilde{\ell}_{i,t} = \begin{cases} \frac{\ell_{i,t} \mathbb{1}\{i_t = i\}}{p_{i,t} + \gamma I_{i,t}} & \text{for } i \in \mathbb{A}_t, \\ 0 & \text{for } i \in \mathbb{G}_t \setminus \mathbb{A}_t, \end{cases} \tag{4.4}$$

where $\gamma \geq 0$ is a parameter of the loss estimator. When $\gamma = 0$, (4.4) is equivalent to the unbiased loss estimate in EXP3. When $\gamma > 0$, due to $I_{i,t} = 1$ for all $i \in \mathbb{A}_t$, (4.4) is the IX-loss estimator in EXP3-IX with exploration factor γ [Neu, 2015]. More generally, having different exploration factors for different arms may be beneficial. We explore such a case in Section 4.3.3.

The weight $\tilde{q}_{t+1}$ of arm i at the beginning of round $t+1$ is defined as follows:

$$\tilde{q}_{i,t+1} = \exp\left(\eta \sum_{s=1}^{t} I_{i,s} (\ell_{i_s,s} - \tilde{\ell}_{i,s} - \gamma \sum_{j \in \mathbb{A}_s} I_{j,s} \tilde{\ell}_{j,s})\right), \tag{4.5}$$

where $\eta > 0$ is the learning rate. The sampling probability of arm i is proportional to its $\tilde{q}_{i,t}$, i.e.,

$$p_{i,t} = \begin{cases} \frac{I_{i,t} \tilde{q}_{i,t}}{W_t} & \text{for } i \in \mathbb{A}_t, \\ 0 & \text{for } i \in \mathbb{G}_t \setminus \mathbb{A}_t, \end{cases} \tag{4.6}$$

where $W_t = \sum_{k \in \mathbb{A}_t} I_{k,t} \tilde{q}_{k,t}$ is the normalization factor.

Apart from setting zero sampling probability for inactive arms, similar to EXP3 and EXP3-IX, SB-EXP3 still follows the strategy of setting the sampling probability proportional to the exponential of the estimated per-action regret. The key difference is the added $-\gamma \sum_{j \in \mathbb{A}_s} I_{j,s} \tilde{\ell}_{j,s}$ in the exponent, which is crucial for obtaining near-optimal high-probability bounds (where $\gamma > 0$). Intuitively, this term significantly reduces the sampling probability of arms that were frequently active in previous rounds but not frequently chosen. Theorems 4.3.1 and 4.3.2 state the regret bounds of SB-EXP3.

Theorem 4.3.1. With $\gamma = 0$, for any $\eta > 0$, Algorithm 4.1 guarantees

$$\max_{a \in [K]} \mathbb{E}[R(a)] \leq \frac{\ln G_T}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} A_t.$$

Tuning η leads to an $O\left(\sqrt{\ln(G_T) \sum_{t=1}^{T} A_t}\right)$ bound.

Theorem 4.3.2. For any $\gamma \geq \frac{\eta}{2} > 0$, Algorithm 4.1 guarantees

$$\max_{a \in [K]} R(a) \leq \frac{\ln G_T}{\eta} + \frac{\ln(2G_T/\delta)}{\gamma} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} A_t$$

with probability at least $1 - \delta$. Tuning η and γ leads to an $O\left(\sqrt{\ln(G_T/\delta) \sum_{t=1}^{T} A_t}\right)$ bound.

Note that $\sum_{t=1}^{T} A_t \leq TA$. Since $A \leq G_T \leq K$, the bounds in Theorems 4.3.1 and 4.3.2 are generally smaller than the $O(\sqrt{TK \ln K})$ bounds of EXP3 and EXP3-IX. The difference is significant whenever $A \ll K$, which holds in many practical applications where the sets of active arms are sparse.

Analysis Sketch. The analyses of EXP3 and EXP3-IX [Auer et al., 2002b; Neu, 2015] treat the normalization factor W_t as a potential function and bound the growth of $\frac{W_{t+1}}{W_t}$. This was possible in standard bandits because all K arms are always active, hence $p_{i,t+1}$ can always be related to $p_{i,t}$. However, in sleeping bandits, the sets $\mathbb{A}_t$ and $\mathbb{A}_{t+1}$ might be non-overlapping, hence there might be no relationship between W_{t+1} and W_t. Instead, we use

$$\tilde{Q}_t = \sum_{i \in \mathbb{G}_T} \tilde{q}_{i,t} \tag{4.7}$$

as the potential function. The following key technical lemma bounds the growth of $\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t}$.

Lemma 4.3.3. For any $t \geq 0$,

$$\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \leq \sum_{i \in \mathbb{A}_t} p_{i,t} \exp\left(\eta(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t})\right).$$

The proof of Lemma 4.3.3 is based on decomposing $\tilde{Q}_{t+1}$ and $\tilde{Q}_t$ by $\mathbb{A}_t$ and $\bar{\mathbb{A}}_t = \mathbb{G}_T \setminus \mathbb{A}_t$. If arm $i \in \mathbb{A}_t$, its weight $\tilde{q}_{i,t+1}$ at time $t+1$ can be related to $\tilde{q}_{i,t}$ via the update in (4.5). If arm $i \notin \mathbb{A}_t$, by construction $\tilde{q}_{i,t+1} = \tilde{q}_{i,t}$. These two observations control the growth of $\tilde{Q}_{t+1}$ over $\tilde{Q}_t$.

Lemma 4.3.3 implies a bound on $\tilde{q}_{a,T+1}$ for any arm a as

$$\ln \tilde{q}_{a,T+1} \leq \ln \tilde{Q}_{T+1} = \sum_{t=1}^{T} \ln \frac{\tilde{Q}_{t+1}}{\tilde{Q}_t}.$$

Since $\tilde{q}_{a,T+1}$ grows with the estimated regret, this bound leads to an upper bound on the estimated regret, which in turns bounds the actual regret.

Observe that the dependency on K in Algorithm 4.1 can be removed completely: the initialization step of assigning $\tilde{q}$ to 1 can be done implicitly. All other explicit computations only use the sets $\mathbb{A}_t$ and $\mathbb{G}_t$. As a result, SB-EXP3 is independent of K. This property is similar to that of the AdaNormalHedge algorithm [Luo and Schapire, 2015], which obtains a low regret bound for sleeping experts when the total number of experts is unknown. Because the analysis of AdaNormalHedge relies on $[0, 1]$-bounded loss vectors, it does not apply to sleeping bandits with the loss estimates in (4.4).

Remark 4.3.4. Lemma 4.3.3 enables the proofs of both pseudo-regret and high-probability bounds without significant modifications to the algorithm. This is a major advantage over existing works in sleeping bandits, which provided only pseudo-regret bounds.

In both Theorems 4.3.1 and 4.3.2, optimally tuning the learning rate requires knowing G_T and $\sum_{t=1}^{T} A_t$, which may not be available a priori. In Appendix 4.F, we present Algorithm 4.4 which uses a two-level doubling trick to obtain a pseudo-regret bound of the same order without knowing these quantities beforehand.

Algorithm 4.2 FTARL for sleeping bandits

Input: $\eta > 0, \gamma > 0, K \geq 2$
Initialize $\tilde{L}_{i,0} = 0$ for $i = 1, 2, \ldots, K$
for each round $t = 1, \ldots, T$ **do**
  The adversary selects and reveals $\mathbb{A}_t$
  Compute $q_t = \arg\min_{q \in \Delta_K} \psi_t(q) + \langle q, \tilde{L}_{t-1} \rangle$
  Compute $W_t = \sum_{i=1}^{K} I_{i,t} q_{i,t}$
  Compute $p_{i,t}$ by (4.8)
  Draw $i_t \sim p_t$ and observe $\hat{\ell}_t = \ell_{i_t,t}$
  for each arm $i \in [K]$ **do**
    If $I_{i,t} = 1$, compute $\tilde{\ell}_{i,t}$ by (4.4)
    If $I_{i,t} = 0$, compute $\tilde{\ell}_{i,t}$ by (4.9)
    Update $\tilde{L}_{i,t} = \tilde{L}_{i,t-1} + \tilde{\ell}_{i,t}$;
  end
end

Theorem 4.3.5. For any $T \geq 2$, Algorithm 4.4 (in Appendix 4.F) guarantees that

$$\max_{a \in [K]} \mathbb{E}[R(a)] \leq \frac{4}{(\sqrt{2}-1)^2} \sqrt{\ln(G_T) \sum_{t=1}^{T} A_t}.$$

The same technique can be used to obtain a high-probability regret bound; however, the resulting bound is slightly larger (in the logarithmic term) due to the union bound.

4.3.2 Generalized FTRL for Sleeping Bandits

In standard non-sleeping bandits, FTRL with Tsallis entropy obtains an $O(\sqrt{TK})$ minimax pseudo-regret bound, an $O(\sqrt{TK \ln(1/\delta)})$ high probability bound against an *oblivious* adversary, and an $O(\sqrt{TK \ln(K/\delta)})$ high-probability bound against a *non-oblivious* adversary [Audibert and Bubeck, 2009; Luo, 2017]. In sleeping bandits, under the assumption that K is known, we will show that the Follow-the-Active-and-Regularized-Leader (FTARL) strategy in Algorithm 4.2 with $\frac{1}{2}$-Tsallis entropy obtains $O(\sqrt{T\sqrt{AK}})$ pseudo-regret. Against a *non-oblivious* adversary, we further show that FTARL obtains an $O(\sqrt{T\sqrt{AK}} + \sqrt{TA \ln(K/\delta)})$ high-probability bound. If the adversary is *oblivious*, this bound is reduced to $O(\sqrt{T\sqrt{AK}} + \sqrt{TA \ln(1/\delta)})$. Thus, when $A = K$, the bounds of FTARL recover those of FTRL on standard bandits.

In round t, Algorithm 4.2 computes the weight vectors $q_t \in \Delta_K$ using FTRL, i.e.,

$$q_t = \arg\min_{q \in \Delta_K} \psi_t(q) + \langle q, \tilde{L}_{t-1} \rangle,$$

where ψ_t is the regularization function and $\tilde{L}_t \in \mathbb{R}^K$ is the cumulative (estimated) loss vector of K arms. In particular, $\tilde{L}_{i,t} = \sum_{s=1}^{t} \tilde{\ell}_{i,t}$. Note that this step is possible because K and the simplex Δ_K are known.

While q_t is a valid probability vector over K arms, it cannot be used directly for sampling because inactive arms might have non-zero elements in q_t. The sampling probability vector p_t is computed by taking elements in $\mathbb{A}_t$ from q_t and normalizing:

$$p_{i,t} = \frac{I_{i,t} q_{i,t}}{\sum_{i \in \mathbb{A}_t} q_{i,t}}. \tag{4.8}$$

Similar to Algorithm 4.1, once an arm $i_t \sim p_t$ is sampled, the loss estimates of active arms are constructed by (4.4). The key difference in Algorithm 4.2 compared to Algorithm 4.1 is that the inactive arms have non-zero loss estimates. For an arm $i \notin \mathbb{A}_t$,

$$\tilde{\ell}_{i,t} = \hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}. \tag{4.9}$$

In Appendix 4.G, we show that by having non-zero loss estimates for inactive arms as in (4.9), Algorithm 4.2 with negative Shannon entropy regularizer is equivalent to Algorithm 4.1. In general, the motivation behind (4.9) is mostly technical: in every round, it ensures that the loss vector $\tilde{L}_t$ contains only non-negative values. This facilitates using a local norm analysis of standard FTRL [Orabona, 2023]. For an unbiased loss estimator where $\gamma = 0$, using (4.9) implies that the estimated regret is equivalent to

$$\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t}) = \sum_{t=1}^{T} \left(\hat{\ell}_t - \tilde{\ell}_{a,t} \right),$$

which resembles the regret of a standard non-sleeping experts problem with input loss vector $\tilde{\ell}$. A similar technique is used by [Chernov and Vovk, 2009].

Using the Tsallis entropy $\psi_t(x) = \frac{1}{\eta} \left(\frac{1 - \sum_{i=1}^{K} x_i^{\beta}}{1 - \beta} \right)$ with parameter β, we obtain the regret bounds of Algorithm 4.2 in Theorems 4.3.6 and 4.3.7.

Theorem 4.3.6. With $\gamma = 0$, for any $\beta \in (0, 1), \eta > 0$, Algorithm 4.2 with Tsallis entropy

guarantees

$$\max_{a\in[K]} \mathbb{E}[R(a)] \le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta} T A^{\beta}.$$

Setting $\beta = \frac{1}{2}$ and tuning η leads to an $O\left(\sqrt{2T\sqrt{AK}}\right)$ bound.

Theorem 4.3.7. For any $\beta, \gamma, \eta \in (0,1)$, Algorithm 4.2 with Tsallis entropy guarantees

$$R(a) \le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta A^{\beta} T}{\beta} + \gamma A T + \left(\frac{\eta+\beta}{2\beta\gamma} + \frac{1}{2}\right)\ln(3/\delta) + \frac{\ln(3K/\delta)}{2\gamma}$$

simultaneously for all $a \in [K]$ with probability at least $1-\delta$. Letting $\beta = \frac{1}{2}$ and tuning η and γ leads to an $O\left(\sqrt{T\sqrt{AK}} + \sqrt{TA\ln(K/\delta)}\right)$ bound.

Remark 4.3.8. If $G_T \le K$ is known, then we can replace K by G_T in Algorithm 4.2 and in Theorems 4.3.6 and 4.3.7. Moreover, if the adversary is oblivious then the second term in the bound can be improved to $O(\sqrt{TA\ln(1/\delta)})$ and the bound becomes $O\left(\sqrt{T\sqrt{AK}} + \sqrt{TA\ln(1/\delta)}\right)$.

Analysis Sketch. Similar to the analysis of Algorithm 4.1, the regret bound of Algorithm 4.2 is obtained by bounding the estimated regret $\sum_{t=1}^{T} \hat{\ell}_t - \tilde{\ell}_{a,t}$. By the local norm analysis of FTRL and properties of Tsallis entropy, we have

$$\sum_{t=1}^{T} \hat{\ell}_t - \sum_{t=1}^{T} \tilde{\ell}_{a,t} \le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{1}{2}\sum_{t=1}^{T}\frac{\eta}{\beta}\sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta}.$$

In both cases $\gamma = 0$ and $\gamma > 0$, because inactive arms have non-zero elements in both $\tilde{\ell}_t$ and $\tilde{q}_t$, they contribute a non-zero positive amount in the last term on the right-hand side. Furthermore, as the computation of $\tilde{\ell}_{i,t}$ uses $p_{i,t}$, we wish to replace $q_{i,t}$ by $p_{i,t}$. The following technical lemma achieves this.

Lemma 4.3.9. For any $t \ge 1$ and $\beta \in (0,1)$,

$$\sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta} \le \sum_{i\in\mathbb{A}_t} \tilde{\ell}_{i,t}^2 p_{i,t}^{2-\beta}.$$

Then, we bound $\sum_{i\in\mathbb{A}_t} \tilde{\ell}_{i,t}^2 p_{i,t}^{2-\beta} \le A_t^{\beta}$ using standard tools such as Holder's inequality. Finally, either taking the expectation (when $\gamma = 0$) or using concentration inequalities of the IX-loss estimator [Neu, 2015] leads to the $O(\sqrt{T\sqrt{AK}})$ bound.

Remark 4.3.10. The technique of assigning the observed loss to sleeping arms in (4.9) is similar to the reduction from sleeping experts to standard prediction with expert advice [Chernov and Vovk, 2009; Gaillard et al., 2014]. However, without Lemma 4.3.9, this reduction by itself does not immediately imply Theorems 4.3.6 and 4.3.7.

Remark 4.3.11. When K is large compared to A (i.e. sparse action sets), the $O\left(\sqrt{T\sqrt{AK}}\right)$ bound of FTARL can be much larger than the $O\left(\sqrt{TA\ln G_T}\right)$ bound of SB-EXP3. In contrast, when $A = \Theta(K)$ as in standard non-sleeping bandits, FTARL gives smaller bounds.

4.3.3 Bounds on Confidence Regret

To the adversary, selecting the set $\mathbb{A}_t$ in round t is equivalent to selecting K binary values $I_{i,t} \in \{0,1\}$. We generalize the setting further by having the adversary select real-valued $I_{i,t} \in [0,1]$. This new setting is the bandit feedback variant of the experts that report their confidences setting, in which $I_{i,t}$ is the confidence of expert i in round t [Blum and Mansour, 2007]. In this case, $R(a)$ is the *confidence regret* with respect to arm a [Gaillard et al., 2014].

More concretely, at the beginning of round t the adversary selects and reveals K real-valued $I_{i,t} \in [0,1]$. The set of active arms becomes $\mathbb{A}_t = \{i \in [K] : I_{i,t} > 0\}$. We apply Algorithm 4.1 to this problem without any modifications: the computations of $\tilde{\ell}_{i,t}, \tilde{q}_{i,t}$ and $p_{i,t}$ are the same as in Equations (4.4), (4.5) and (4.6). The full protocol and algorithm are given in Appendix 4.C. We state the following high probability regret bound.

Theorem 4.3.12. With optimally tuned η and γ, Algorithm 4.1 guarantees

$$R(a) \leq O\left(\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t}\ln(G_T/\delta)}\right)$$

simultaneously for all $a \in [K]$ with probability $1-\delta$.

Observe that in this setting, because $I_{i,t}$ can be different for different arms in $\mathbb{A}_t$, SB-EXP3 uses a different implicit exploration factor $\gamma_{i,t} = \gamma I_{i,t}$ in (4.4). We call this novel strategy based on the arm-dependent IX-loss estimator *confident implicit exploration*. The proof of Theorem 4.3.12 mostly follows that of Theorem 4.3.2 with one added key insight: the concentration inequalities of the original IX-loss estimator also hold for the arm-dependent IX-loss estimator.

Algorithm 4.3 SE-EXP4 for bandits with advice from sleeping experts

Input: $\eta > 0, \gamma > 0$
Initialize $\tilde{q}_{m,1} = 1$ for $m = 1, 2, \ldots, M$
for each round $t = 1, \ldots, T$ **do**
  An adversary selects and reveals $\mathbb{B}_t$ and E_t
  Compute $z_{m,t}$ by (4.12) and $p_t = z_t E_t$
  Draw $i_t \sim p_t$ and observe $\hat{\ell}_t = \ell_{i_t,t}$
  Compute $\tilde{\ell}_{k,t} = \frac{\ell_{k,t}\mathbb{1}\{i_t=k\}}{p_{k,t}+\gamma}$ for arms $k \in [K]$
  for each awake expert $m \in \mathbb{B}_t$ **do**
    Construct loss estimate $\tilde{x}_{m,t} = \langle E_{m,t}, \tilde{\ell}_t \rangle$
    Update $\tilde{q}_{m,t+1}$ by (4.11).
  end
end

Remark 4.3.13. An $O\left(\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t} \ln G_T}\right)$ pseudo-regret bound is obtained via a similar analysis. When $I_{i,t}$ are binary, since $\sum_{i=1}^{K} I_{i,t} = A_t \leq A$, these bounds recover the bounds in Theorems 4.3.1 and 4.3.2.

Remark 4.3.14. Similar to Theorem 4.3.5, the two-level doubling trick presented in Appendix 4.F can be used on $\sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t}$ and $\ln(G_T)$ to obtain pseudo-regret and high-probability bounds of the same order (up to a logarithmic factor).

4.4 Bandits with Advice from Sleeping Experts

The bandits with expert advice framework [Auer et al., 2002b] considers the non-sleeping MAB where M experts give advice to the learner in each round. We study a variant of this problem where a time-varying set $\mathbb{B}_t \subseteq [M]$ of awake experts gives advice to the learner in round t. The advice of expert $m \in \mathbb{B}_t$ is $E_{m,t} \in \Delta_K$. The learner aims to compete with each expert during their active rounds. More formally, let $I_{m,t} = 0$ and $I_{m,t} = 1$ denote whether expert m is sleeping or awake in round t, respectively. Let ℓ_t be the (hidden) loss vector of K arms in round t. The regret with respect to m after T rounds is

$$R(m) = \sum_{t=1}^{T} I_{m,t}(\ell_{i_t,t} - \langle E_{m,t}, \ell_t \rangle). \tag{4.10}$$

All theorems in this section are high probability bounds. The pseudo-regret bounds are in Appendix 4.D.

4.4.1 Generalized EXP4

In the standard setting where all M experts are always active, EXP4 and EXP4-IX [Auer et al., 2002b; Neu, 2015] obtain $O(\sqrt{TK \ln M})$ pseudo- and $O(\sqrt{TK \ln(M/\delta)})$ high-probability bounds, respectively. We will show that SE-EXP4 (Algorithm 4.3) obtains these two bounds as well in the bandits with advice from sleeping experts setting.

Let $B_t = |\mathbb{B}_t|$ be the number of awake experts at round t. Let E_t be the $B_t \times K$ matrix whose columns are $(E_{m,t})_{m \in \mathbb{B}_t}$. In round t, SE-EXP4 samples an expert m_t from a distribution $z_t \in \Delta_{B_t}$, then samples an arm $i_t \sim E_{m_t,t}$. This is equivalent to sampling $i_t \sim p_t$ directly, where $p_t = z_t E_t$ [Lattimore and Szepesvári, 2020]. The main idea of SE-EXP4 is to take the sleeping experts as "augmented" sleeping arms and apply SB-EXP3. In particular, the weight of expert m is

$$\tilde{q}_{m,t} = \exp\left(\sum_{s=1}^{t-1} I_{m,s}\left(\hat{\ell}_s - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,s} - \tilde{x}_{m,s}\right)\right), \tag{4.11}$$

where $\tilde{x}_{m,s}$ is the estimated loss of expert m and $\tilde{\ell}_{j,s}$ is the estimated loss of arm j in round s. Initially, $\tilde{q}_{m,1} = 1$. For an expert $m \in \mathbb{B}_t$, its sampling probability $z_{m,t}$ is proportional to $\tilde{q}_{m,t}$, i.e.,

$$z_{m,t} = \frac{\tilde{q}_{m,t}}{\sum_{j \in \mathbb{B}_t} \tilde{q}_{j,t}}. \tag{4.12}$$

Note that $z_{m,t} = 0$ for $m \notin \mathbb{B}_t$. After $i_t \sim p_t$ is sampled, the loss estimates $\tilde{\ell}_t$ of K arms are computed as in (4.4). Then, the losses of the awake experts are estimated by the inner product of their advice and $\tilde{\ell}_t$:

$$\tilde{x}_{m,t} = \langle E_{m,t}, \tilde{\ell}_t \rangle. \tag{4.13}$$

Using the same analysis of SB-EXP3 implies:

Theorem 4.4.1. For any $\gamma, \eta \in (0,1), \eta \leq 2\gamma$, SE-EXP4 guarantees

$$R(u) \leq \frac{\ln M}{\eta} + \frac{\ln(2M/\delta)}{2\gamma} + (\gamma + \frac{\eta}{2})TK + \ln(2/\delta) \tag{4.14}$$

simultaneously for all experts $u \in [M]$ with probability at least $1 - \delta$, where the probability is taken over the sequence of the learner's selected arms. Tuning η and γ leads to an $O(\sqrt{TK \ln(M/\delta)})$ bound.

4.4.2 Adaptive and Tracking Bounds for Standard Adversarial Bandits

Following [Hazan and Seshadhri, 2009], the adaptive regret of the learner on an interval $[t_1, t_2]$ with respect to arm k in standard non-sleeping bandits is

$$R_{[t_1,t_2]}(k) = \sum_{t=t_1}^{t_2} \left(\ell_{i_t,t} - \ell_{k,t}\right),$$

where i_t is the arm chosen by the learner in round t. To obtain an adaptive regret bound using SE-EXP4, we follow the "virtual experts" strategy similar to [Adamskiy et al., 2016]: for each interval $[t_1, t_2]$ and arm k, we create a virtual expert indexed by (k, t_1, t_2) that is awake from round t_1 to round t_2 with advice $E_{(k,t),s} = e_k$ for any $s \geq t$, where e_k is the k^{th} vector in the standard basis of $\mathbb{R}^K$. There are $K\binom{T}{2} = \frac{KT(T+1)}{2}$ such experts. For any interval $[t_1, t_2]$, the regrets of experts $(1, t_1, t_2), (2, t_1, t_2), \ldots, (K, t_1, t_2)$ are bounded by Theorem 4.4.1. This implies the following result.

Theorem 4.4.2. For any $\gamma, \eta \in (0, 1), \eta \leq 2\gamma$, SE-EXP4 with virtual experts guarantees that

$$R_{[t_1,t_2]}(k) \leq \frac{2\ln(KT)}{\eta} + \frac{\ln(KT/\delta)}{\gamma} + (\gamma + \frac{\eta}{2})TK + \ln(2/\delta)$$

simultaneously for all intervals $[t_1, t_2]$ and arms $k \in [K]$ with probability $1 - \delta$. Tuning η and γ leads to an $O(\sqrt{TK\ln(KT/\delta)})$ bound.

Next, we use Theorem 4.4.2 to bound the tracking regret. The tracking regret of algorithm $\mathcal{A}$ with respect to a sequence of arms $j_{1:T} = (j_1, j_2, \ldots, j_T)$ is

$$R(j_{1:T}) = \sum_{t=1}^{T} (\hat{\ell}_t - \ell_{j_t,t}).$$

Let $S = \sum_{t=2}^{T} \mathbb{1}\{j_t \neq j_{t-1}\}$ be the (unknown) number of switches in the sequence $j_{1:T}$. Since there are $S + 1$ intervals of different competing arms, Theorem 4.4.2 immediately implies an $O\left(S\sqrt{TK\ln(KT)}\right)$ tracking bound. Furthermore, the following corollary shows that if S is known, then a tracking bound of order $O(\sqrt{STK\ln\left(\frac{KT}{\delta S}\right)})$ is attainable as well. Thus, techniques based on sleeping bandits recover the tracking bounds in [Auer et al., 2002b] and [Neu, 2015].

Corollary 4.4.3. Restarting SE-EXP4 with virtual experts after every T/S rounds guarantees that for any $j_{1:T}$ where $H(j_{1:T}) \leq S$, with probability at least $1-\delta$

$$R(j_{1:T}) \leq O\left(\sqrt{STK \ln\left(\frac{KT}{\delta S}\right)}\right).$$

Remark 4.4.4. [Luo et al., 2018] obtained adaptive and tracking pseudo-regret bounds for contextual bandits which recover the same bounds (up to constants and logarithmic factors) in standard bandits. Our results hold for both pseudo-regret and high probability bounds.

4.5 A Per-Action Strongly Adaptive Lower Bound

Let $T_a = |\{t \in [T] : I_{a,t} = 1\}|$ be the number of rounds in which arm a is active. Prior works in sleeping experts established per-action regret bounds of order $O(\sqrt{T_a})$, independent of T [Chernov and Vovk, 2010; Gaillard et al., 2014]. In this work, we call T-independent bounds that guarantee $R(a) = o(T_a)$ for all action $a \in [K]$ *per-action strongly adaptive bounds*. The bounds obtained by SB-EXP3 and FTARL in Section 4.3 have $O(\sqrt{T})$ dependency for all arms a, thus they are not strongly adaptive. We study whether strongly adaptive bounds are achievable in sleeping bandits. The following theorem shows a linear per-action strongly adaptive lower bound for a large class of algorithms that contains SB-EXP3, FTARL in Section 4.3 as well as any minimax optimal algorithm.

Theorem 4.5.1. For any (possibly randomized) algorithm with guarantee

$$\sup_{a \in [K]} E[R(a)] \leq O(T^{\gamma} A^{\beta} (\ln(T))^{\mu}), \tag{4.15}$$

where $\gamma \in (0,1), \beta \geq 0, \mu \geq 0$ are constants, there exists a number of arms K and sequence of sets of active arms and their losses such that $A = 2$ and for at least one arm $a \in [K]$,

$$T_a \geq \Omega(T^{1-\gamma} (\ln(T))^{-\mu}) \text{ and } \mathbb{E}[R(a)] \geq \Omega(T_a).$$

Remark 4.5.2. By setting $\gamma = \beta = \frac{1}{2}, \mu = 0$, (4.15) corresponds to the $O(\sqrt{TA})$ bound of any minimax optimal algorithms. This implies that no algorithms can simultaneously have an optimal per-action regret bound and a sublinear per-action strongly adaptive bound.

Our proof extends that of [Daniely et al., 2015] to sleeping bandits. Specifically, the setup contains arm 1 that is always active with small losses, while the other $K-1$ arms are active in $K-1$ non-overlapping intervals with large losses, one interval for each arm. To guarantee small regret against arm 1, with high probability, the learner must pull only arm 1 in the interval of some arm $j > 1$. Consider a slightly different setup where the losses of arm j are smaller than that of arm 1. Because the sequence of active arms and their losses in the two setups are identical from the first round until the start of the interval of arm j, the learner is unable to distinguish between the two setups and incurs linear regret against arm j in the second setup.

The limitation of this construction is that K has to grow with T. In particular, we require K of order $T^{1-\gamma}2^{-\beta}(\ln(T))^{-\mu}$. As a result, algorithms with bounds that are sublinear in T but have large dependency on K, for example $O(T^{\gamma}K)$, do not satisfy (4.15). Note that SB-EXP3 and FTARL always satisfy (4.15), since $\sqrt{TA\ln K} \leq \sqrt{TA\ln T}$ and $\sqrt{T\sqrt{AK}} \leq T^{3/4}A^{1/4}$ for any $K \leq T$.

4.6 Conclusion

We derived algorithms for sleeping bandits and proved their near-optimal per-action regret bounds. These algorithms and their regret bounds strictly generalize existing approaches and results for standard non-sleeping bandits. We showed that sleeping bandits-based approaches both imply new bounds and recover a number of existing order-optimal $\tilde{O}(\sqrt{T})$ bounds in related settings with fundamentally different proofs. Furthermore, the analysis can be used to show both pseudo-regret and high probability bounds by using either the unbiased or IX-loss estimators.

A direction for future work is to either prove an $\Omega(\sqrt{TA\ln G_T})$ lower bound, or show that an $O(\sqrt{TA})$ upper bound is possible, thereby obtaining minimax optimal bounds. For the former, such a lower bound must hold only in restricted conditions such as $A\ln G_T < K$, so there is no contradiction to the optimal $O(\sqrt{TK})$ bound in non-sleeping bandits. The latter likely requires new analysis techniques other than the potential-based analysis and FTRL.

4.A Proofs for Section 4.3.1

Recall that $\tilde{Q}_t = \sum_{i\in\mathbb{G}_T} \tilde{q}_{i,t}$. For any set $S \subseteq \mathbb{G}_T$, let $\tilde{Q}_{S,t} = \sum_{i\in S} \tilde{q}_{i,t}$ be the projection of $\tilde{Q}_t$ onto the set S. Let $\bar{S} = \mathbb{G}_T \setminus S$ be the complement of S. Note that the normalization factor W_t is equal to $\sum_{i\in\mathbb{A}_t} \tilde{q}_{i,t} = \tilde{Q}_{\mathbb{A}_t,t}$.

First, we prove a technical lemma which holds for any $I_{i,t} \in [0,1]$ for all $i \in [K]$ and $t \in [T]$.

Lemma 4.A.1. For any $t \geq 1$,

$$\mathbb{E}_{i\sim p_t}[\tilde{\ell}_{i,t}] = \hat{\ell}_t - \gamma \sum_{j\in\mathbb{A}_t} I_{j,t}\tilde{\ell}_{j,t}.$$

Proof. Using $\tilde{\ell}_{j,t} = 0$ for $j \neq i_t$, we obtain

$$\begin{aligned}
\mathbb{E}_{i\sim p_t}[\tilde{\ell}_{i,t}] &= \sum_{i\in\mathbb{A}_t} p_{i,t}\tilde{\ell}_{i,t} \\
&= p_{i_t,t}\tilde{\ell}_{i_t,t} \\
&= p_{i_t,t}\frac{\hat{\ell}_t}{p_{i_t,t}+\gamma I_{i,t}} \\
&= \hat{\ell}_t - \gamma I_{i_t,t}\frac{\hat{\ell}_t}{p_{i_t,t}+\gamma I_{i_t,t}} \\
&= \hat{\ell}_t - \gamma I_{i_t,t}\tilde{\ell}_{i_t} \\
&= \hat{\ell}_t - \gamma \sum_{j\in\mathbb{A}_t} I_{j,t}\tilde{\ell}_{j,t}.
\end{aligned}$$

□

4.A.1 Proof of Lemma 4.3.3

We make use of the following two facts that can be proved easily:

- Fact 1: the function $f(x) = e^{-\eta x}$ is convex for any $\eta \in \mathbb{R}$.
- Fact 2: For any $a, b > 0, c \geq 0$, if $a \geq b$ then

$$\frac{a}{b} \geq \frac{a+c}{b+c}.$$

Lemma 4.3.3. For any $t \geq 0$,

$$\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \leq \sum_{i\in\mathbb{A}_t} p_{i,t} \exp\left(\eta(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j\in\mathbb{A}_t} \tilde{\ell}_{j,t})\right).$$

Proof. By Jensen's inequality and Fact 1, we have

$$\begin{aligned}\sum_{i\in\mathbb{A}_t} p_{i,t}\exp\Big(-\eta\tilde{\ell}_{i,t}\Big) &= \mathbb{E}_{i\sim p_t}[\exp\Big(-\eta\tilde{\ell}_{i,t}\Big)] \\ &\geq \exp\left(-\eta\mathbb{E}_{i\sim p_t}[\tilde{\ell}_{i,t}]\right) \\ &= \exp\left(-\eta\left(\hat{\ell}_t - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right),\end{aligned}$$

where the last equality is due to Lemma 4.A.1.

Multiplying by $\exp\left(\eta\left(\hat{\ell}_t - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right)\tilde{Q}_{A_t,t} > 0$ on both sides, we obtain

$$\sum_{i\in\mathbb{A}_t} p_{i,t}\tilde{Q}_{\mathbb{A}_t,t}\exp\left(\eta\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right) \geq \tilde{Q}_{\mathbb{A}_t,t}. \tag{4.16}$$

For all $i\in\mathbb{A}_t$, we have $I_{i,t}=1$ and thus $p_{i,t} = \frac{\tilde{q}_{i,t}}{\tilde{Q}_{\mathbb{A}_t,t}}$. Equation (4.16) is equivalent to

$$\sum_{i\in\mathbb{A}_t}\tilde{q}_{i,t}\exp\left(\eta\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right) \geq \tilde{Q}_{\mathbb{A}_t,t}.$$

By our update rule, $\tilde{q}_{i,t+1} = \tilde{q}_{i,t}$ for $i\notin\mathbb{A}_t$ and $\tilde{q}_{i,t+1} = \tilde{q}_{i,t}\exp\left(\eta\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right)$ for $i\in\mathbb{A}_t$. Hence,

$$\sum_{i\in\mathbb{A}_t}\tilde{q}_{i,t+1} = \sum_{i\in\mathbb{A}_t}\tilde{q}_{i,t}\exp\left(\eta\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right) \geq \tilde{Q}_{\mathbb{A}_t,t}.$$

Applying Fact 2 for $a = \sum_{i\in\mathbb{A}_t} \tilde{q}_{i,t+1}, b = \tilde{Q}_{\mathbb{A}_t,t}$ and $c = \tilde{Q}_{\bar{\mathbb{A}}_t,t}$, we obtain

$$\begin{aligned}\sum_{i\in\mathbb{A}_t} p_{i,t} \exp\left(\eta\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right) &= \frac{\sum_{i\in\mathbb{A}_t}\tilde{q}_{i,t}\exp\left(\eta\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\right)\right)}{\tilde{Q}_{\mathbb{A}_t,t}} \\ &= \frac{\sum_{i\in\mathbb{A}_t}\tilde{q}_{i,t+1}}{\tilde{Q}_{\mathbb{A}_t,t}} \\ &\geq \frac{\sum_{i\in\mathbb{A}_t}\tilde{q}_{i,t+1} + \tilde{Q}_{\bar{\mathbb{A}}_t,t}}{\tilde{Q}_{\mathbb{A}_t,t} + \tilde{Q}_{\bar{\mathbb{A}}_t,t}} \\ &= \frac{\tilde{Q}_{t+1}}{\tilde{Q}_t},\end{aligned} \tag{4.17}$$

where the last equality is due to the fact that $\sum_{i\in\bar{\mathbb{A}}_t}\tilde{q}_{i,t+1} = \sum_{i\in\bar{\mathbb{A}}_t}\tilde{q}_{i,t} = \tilde{Q}_{\bar{\mathbb{A}}_t,t}$. □

4.A.2 Bounding the Estimated Regret

Using Lemma 4.3.3, we bound the estimated regret as a function of the cumulative estimated losses of all active arms over T rounds.

Lemma 4.A.2. For any $\gamma \geq 0$, any arm $a \in [K]$,

$$\sum_{t=1}^{T} I_{a,t}\hat{\ell}_t \leq \frac{\ln G_T}{\eta} + \sum_{t=1}^{T} I_{a,t}\tilde{\ell}_{a,t} + \left(\frac{\eta}{2} + \gamma\right)\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t}\tilde{\ell}_{i,t}.$$

Proof. For any arm $a \in \mathbb{G}_T$, we have

$$\ln\tilde{Q}_{T+1} = \ln\sum_{i\in\mathbb{G}_T}\tilde{q}_{i,T+1} \geq \ln\tilde{q}_{a,T+1} = \eta\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t} - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}). \tag{4.18}$$

On the other hand, we have

$$\begin{aligned}
\ln \tilde{Q}_{T+1} &= \ln \tilde{Q}_1 + \sum_{t=1}^{T} \ln \frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \\
&\leq \ln G_T + \sum_{t=1}^{T} \ln \left(\sum_{i \in \mathbb{A}_t} p_{i,t} \exp\left(\eta \left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) \right) \right) \\
&= \ln G_T + \sum_{t=1}^{T} \ln \left(\exp\left(\eta \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) \right) \sum_{i \in \mathbb{A}_t} p_{i,t} \exp\left(-\eta \tilde{\ell}_{i,t} \right) \right) \\
&= \ln G_T + \sum_{t=1}^{T} \left(\eta \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) + \ln \left(\sum_{i \in \mathbb{A}_t} p_{i,t} \exp\left(-\eta \tilde{\ell}_{i,t} \right) \right) \right) \\
&\leq \ln G_T + \sum_{t=1}^{T} \left(\eta \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) + \ln \left(\sum_{i \in \mathbb{A}_t} p_{i,t} (1 + \frac{\eta^2 \tilde{\ell}_{i,t}^2}{2} - \eta \tilde{\ell}_{i,t}) \right) \right) \\
&= \ln G_T + \sum_{t=1}^{T} \left(\eta \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) + \ln \left(1 + \eta^2 \sum_{i \in \mathbb{A}_t} \frac{p_{i,t} \tilde{\ell}_{i,t}^2}{2} - \eta \sum_{i \in \mathbb{A}_t} p_{i,t} \tilde{\ell}_{i,t} \right) \right) \\
&\leq \ln G_T + \sum_{t=1}^{T} \left(\eta \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) + \eta^2 \sum_{i \in \mathbb{A}_t} \frac{p_{i,t} \tilde{\ell}_{i,t}^2}{2} - \eta \sum_{i \in \mathbb{A}_t} p_{i,t} \tilde{\ell}_{i,t} \right) \\
&= \ln G_T + \eta^2 \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \frac{p_{i,t} \tilde{\ell}_{i,t}^2}{2},
\end{aligned}$$

where

- the first inequality is due to Lemma 4.3.3,
- the second inequality is $\exp(-x) \leq 1 + \frac{x^2}{2} - x$ for all $x \geq 0$,
- the third inequality is $\ln(1 + x) \leq x$ for all $x \geq -1$,
- the last equality is due to Lemma 4.A.1.

Since the losses are bounded in $[0, 1]$ we also have $\sum_{i \in \mathbb{A}_t} p_{i,t} \tilde{\ell}_{i,t}^2 \leq \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t}$. This implies

$$\ln \tilde{Q}_{T+1} \leq \ln G_T + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t}. \tag{4.19}$$

From (4.18) and (4.19), we obtain

$$\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t} - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}) \leq \frac{\ln G_T}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t}.$$

Adding $\sum_{t=1}^{T} I_{a,t}(\tilde{\ell}_{a,t} + \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t})$ to both sides yields

$$\begin{aligned} \sum_{t=1}^{T} I_{a,t}\hat{\ell}_t &\leq \frac{\ln G_T}{\eta} + \sum_{t=1}^{T} I_{a,t}\tilde{\ell}_{a,t} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} + \gamma \sum_{t=1}^{T} I_{a,t} \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \\ &\leq \frac{\ln G_T}{\eta} + \sum_{t=1}^{T} I_{a,t}\tilde{\ell}_{a,t} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t}, \end{aligned} \tag{4.20}$$

where the second inequality is due to $I_{a,t} \leq 1$. □

4.A.3 Proof of Theorem 4.3.1

Theorem 4.3.1. With $\gamma = 0$, for any $\eta > 0$, Algorithm 4.1 guarantees

$$\max_{a \in [K]} \mathbb{E}[R(a)] \leq \frac{\ln G_T}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} A_t.$$

Tuning η leads to an $O\left(\sqrt{\ln(G_T) \sum_{t=1}^{T} A_t}\right)$ bound.

Proof. If an arm a is not in $\mathbb{G}_T$ then it has never been active in any round, thus $R(a) = 0$. For an arm $a \in \mathbb{G}_T$, taking the expectation of both sides of Lemma 4.A.2 and using

$$\begin{aligned} \mathbb{E}_{i_t \sim p_t}[\tilde{\ell}_{i,t}] &= p_{i,t} \frac{\ell_{i,t}}{p_{i,t}} \\ &= \ell_{i,t} \\ &\leq 1, \end{aligned}$$

we obtain

$$\mathbb{E}[R(a)] \leq \frac{\ln G_T}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} A_t.$$

□

4.A.4 Proof of Theorem 4.3.2

Before proving Theorem 4.3.2, we state the following lemma and its corollary which provide high-probability guarantees that the sum of the loss estimators is a lower confidence bound for the sum of the true losses of all active arms. This lemma is adapted from Lemma 1 of [Neu, 2015].

Lemma 4.A.3 (Lemma 1 of [Neu, 2015])**.** For all $i \in [K], t \in [T]$ and $I_{i,t} \in [0,1]$, let $\alpha_{i,t}$ satisfy $0 \leq \alpha_{i,t} \leq 2\gamma I_{i,t}$. With probability $1-\delta'$,

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \alpha_{i,t}\mathbb{1}\{I_{i,t}>0\}(\tilde{\ell}_{i,t}-\ell_{i,t}) \leq \ln(1/\delta').$$

The proof of Lemma 4.A.3 is identical to that of Lemma 1 of [Neu, 2015] (with one additional straightforward step of handling $I_{i,t}=0$) and thus is omitted here. For any fixed $j \in \mathbb{G}_T$, applying Lemma 4.A.3 with $\alpha_{i,t} = 2\gamma I_{i,t}\mathbb{1}\{i=j\}$ and taking a union bound over all $j \in \mathbb{G}_T$ leads to the following corollary.

Corollary 4.A.4. With probability at least $1-\delta'$,

$$\sum_{t=1}^{T} I_{j,t}(\tilde{\ell}_{j,t}-\ell_{j,t}) \leq \frac{\ln(G_T/\delta')}{2\gamma}$$

holds simultaneously for all $j \in \mathbb{G}_T$.

Theorem 4.3.2. For any $\gamma \geq \frac{\eta}{2} > 0$, Algorithm 4.1 guarantees

$$\max_{a\in[K]} R(a) \leq \frac{\ln G_T}{\eta} + \frac{\ln(2G_T/\delta)}{\gamma} + \left(\frac{\eta}{2}+\gamma\right)\sum_{t=1}^{T} A_t$$

with probability at least $1-\delta$. Tuning η and γ leads to an $O\left(\sqrt{\ln(G_T/\delta)\sum_{t=1}^{T} A_t}\right)$ bound.

Proof. Applying Lemma 4.A.3 with $\alpha_{i,t} = (\eta/2+\gamma)I_{i,t}, \delta' = \delta/2$ and Corollary 4.A.4 with $\delta' = \delta/2$ gives

$$\left(\frac{\eta}{2}+\gamma\right)\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t}\tilde{\ell}_{i,t} \leq \ln(2/\delta) + \left(\frac{\eta}{2}+\gamma\right)\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t}\ell_{i,t} \tag{4.21}$$

and

$$\sum_{t=1}^{T} I_{a,t}\tilde{\ell}_{a,t} \le \frac{\ln(2G_T/\delta)}{2\gamma} + \sum_{t=1}^{T} I_{a,t}\ell_{a,t} \quad \text{for any } a \in [K]. \tag{4.22}$$

Plugging (4.21) and (4.22) into the right-hand side of Lemma 4.A.2 and taking a union bound, we have with probability at least $1-\delta$, simultaneously for all $a \in [K]$,

$$\sum_{t=1}^{T} I_{a,t}\hat{\ell}_t \le \frac{\ln G_T}{\eta} + \frac{\ln(2G_T/\delta)}{2\gamma} + \sum_{t=1}^{T} I_{a,t}\ell_{a,t} + \ln(2/\delta) + \left(\frac{\eta}{2}+\gamma\right)\sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t}\ell_{i,t}.$$

Subtracting $\sum_{t=1}^{T} I_{a,t}\ell_{a,t}$ on both sides and using $I_{i,t}\ell_{i,t} \le 1$, we obtain

$$R(a) \le \frac{\ln G_T}{\eta} + \frac{\ln(2G_T/\delta)}{2\gamma} + \ln(2/\delta) + \left(\frac{\eta}{2}+\gamma\right)\sum_{t=1}^{T} A_t \tag{4.23}$$

with probability at least $1-\delta$. □

4.B Proofs for Section 4.3.2

Let $D_{\psi_t} : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}_+$ be the Bregman divergence generated by ψ_t. Let $\tilde{\ell}_t = \begin{bmatrix} \tilde{\ell}_{1,t} \\ \dots \\ \tilde{\ell}_{K,t} \end{bmatrix}$ be the estimated loss vector in round t. We write $Q_t = \sum_{i=1}^{K} q_{i,t}$ for the sum of the weights of all arms at the beginning round t. Note that $Q_t = 1$. For a set $S \subseteq [K]$, we write $Q_{S,t} = \sum_{i\in S} q_{i,t}$ for the projection of Q_t on S. The complement of S is $\bar{S} = [K] \setminus S$.

4.B.1 Proof of Lemma 4.3.9

Lemma 4.3.9. For any $t \ge 1$ and $\beta \in (0,1)$,

$$\sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta} \le \sum_{i\in\mathbb{A}_t} \tilde{\ell}_{i,t}^2 p_{i,t}^{2-\beta}.$$

Proof. Since $\tilde{\ell}_{i,t} = 0$ if $i \in \mathbb{A}_t$ and $i \ne i_t$, the right-hand side can be reduced to

$$\sum_{i\in\mathbb{A}_t} \tilde{\ell}_{i,t}^2 p_{i,t}^{2-\beta} = \tilde{\ell}_{i_t,t}^2 p_{i_t,t}^{2-\beta}.$$

For the left-hand side, the estimated losses of inactive arms are equal to

$$\begin{aligned}\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} &= \hat{\ell}_t - \gamma \tilde{\ell}_{i_t,t} \\ &= \hat{\ell}_t - \gamma \frac{\hat{\ell}_t}{p_{i_t,t} + \gamma} \\ &= \frac{p_{i_t,t} \hat{\ell}_t}{p_{i_t,t} + \gamma}.\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta} &= \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta} + \sum_{i \in \bar{\mathbb{A}}_t} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta} \\ &= \tilde{\ell}_{i_t,t}^2 q_{i_t,t}^{2-\beta} + \frac{p_{i_t,t}^2 \hat{\ell}_t^2}{(p_{i_t,t} + \gamma)^2} \sum_{i \in \bar{\mathbb{A}}_t} q_{i,t}^{2-\beta} \\ &= \frac{\hat{\ell}_t^2}{(p_{i_t,t} + \gamma)^2} q_{i_t,t}^{2-\beta} + \frac{p_{i_t,t}^2 \hat{\ell}_t^2}{(p_{i_t,t} + \gamma)^2} \sum_{i \in \bar{\mathbb{A}}_t} q_{i,t}^{2-\beta} \\ &= \frac{\hat{\ell}_t^2}{(p_{i_t,t} + \gamma)^2} p_{i_t,t}^{2-\beta} (Q_{\mathbb{A}_t}^{2-\beta} + p_{i_t,t}^{\beta} \sum_{i \in \bar{\mathbb{A}}_t} q_{i,t}^{2-\beta}) \qquad \text{since } p_{i_t,t} = \frac{q_{i_t,t}}{Q_{\mathbb{A}_t,t}} \\ &\leq \frac{\hat{\ell}_t^2}{(p_{i_t,t} + \gamma)^2} p_{i_t,t}^{2-\beta} (Q_{\mathbb{A}_t}^{2-\beta} + \sum_{i \in \bar{\mathbb{A}}_t} q_{i,t}^{2-\beta}) \qquad \text{since } p_{i_t,t}^{\beta} \leq 1 \\ &\leq \tilde{\ell}_{i_t,t}^2 p_{i_t,t}^{2-\beta} (Q_{\mathbb{A}_t} + \sum_{i \in \bar{\mathbb{A}}_t} q_{i,t}) \\ &= \sum_{i \in \mathbb{A}_t} \tilde{\ell}_t^2 p_{i,t}^{2-\beta},\end{aligned}$$

where the last two inequalities are due to applying $x^\alpha \leq x$ for all $x \in [0,1], \alpha > 1$ on $p_{i_t,t}, Q_{\mathbb{A}_t,t}$ and each $q_{i,t}$ for $i \in \bar{\mathbb{A}}_t$ as well as the fact that $q_t \in \Delta_K$. □

4.B.2 Bounding the Estimated Regret

Lemma 4.B.1. For any arm $a \in [K]$,

$$\sum_{t=1}^{T} I_{a,t} (\hat{\ell}_t - \tilde{\ell}_{a,t}) \leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \gamma \sum_{t=1}^{T} \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} + \frac{\eta}{2\beta} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} p_{i,t}^{1-\beta},$$

where e_a is the a-th standard basis vector of $\mathbb{R}^K$.

Proof. From the regret bound of FTRL using local norms [Orabona, 2023, Lemma 7.12] we have

$$\sum_{t=1}^{T}\langle\tilde{\ell}_t, q_t - e_a\rangle \le \psi_{T+1}(e_a) - \min_{x\in\Delta_K}\psi_1(x) + \frac{1}{2}\sum_{t=1}^{T}\left\|\tilde{\ell}_t\right\|^2_{(\nabla^2\psi_t(u_t))^{-1}}, \tag{4.24}$$

where u_t is a point between q_t and

$$\bar{q}_{t+1} = \arg\min_{x\in\mathbb{R}^K}\langle\tilde{\ell}_t, x\rangle + D_{\psi_t}(x; q_t). \tag{4.25}$$

First, we examine the left-hand side in (4.24). Obviously $\langle\tilde{\ell}_t, e_a\rangle = \tilde{\ell}_{a,t}$. Furthermore,

$$\begin{aligned}
\langle\tilde{\ell}_t, q_t\rangle &= \sum_{i=1}^{K}\tilde{\ell}_{i,t}q_{i,t} \\
&= \sum_{i\in\mathbb{A}_t}\tilde{\ell}_{i,t}q_{i,t} + \sum_{i\in\bar{\mathbb{A}}_t}\tilde{\ell}_{i,t}q_{i,t} \\
&= \hat{\ell}_t\frac{q_{i_t,t}}{p_{i_t,t}+\gamma} + \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t}+\gamma}\sum_{i\in\bar{\mathbb{A}}_t}q_{i,t} \qquad \left(\text{since } \tilde{\ell}_{i,t} = \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t}+\gamma} \text{ for } i\in\bar{\mathbb{A}}_t\right) \\
&= \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t}+\gamma}\left(\frac{q_{i_t,t}}{p_{i_t,t}} + \sum_{i\in\bar{\mathbb{A}}_t}q_{i,t}\right) \\
&= \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t}+\gamma}\left(\sum_{i\in\mathbb{A}_t}q_{i,t} + \sum_{i\in\bar{\mathbb{A}}_t}q_{i,t}\right) \qquad \left(\text{since } p_{i,t} = \frac{q_{i,t}}{\sum_{j\in\mathbb{A}_j}q_{j,t}}\right) \\
&= \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t}+\gamma} \qquad (\text{since } q_t\in\Delta_K) \\
&= \hat{\ell}_t - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}.
\end{aligned} \tag{4.26}$$

Therefore, $\langle\tilde{\ell}_t, q_t - e_a\rangle = \hat{\ell}_t - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t} - \tilde{\ell}_{a,t}$. By construction, if $I_{a,t} = 0$ then $\tilde{\ell}_{a,t} = \hat{\ell}_t - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}$. Hence,

$$\sum_{t=1}^{T}\langle\tilde{\ell}_t, q_t - e_a\rangle = \sum_{t=1}^{T}I_{a,t}\left(\hat{\ell}_t - \gamma\sum_{j\in\mathbb{A}_t}\tilde{\ell}_t - \tilde{\ell}_{a,t}\right).$$

Plugging this into (4.24) implies that

$$\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t}) \leq \psi_{T+1}(e_a) - \min_{x \in \Delta_K} \psi_1(x) + \gamma \sum_{t=1}^{T} I_{a,t} \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} + \frac{1}{2} \sum_{t=1}^{T} \left\| \tilde{\ell}_t \right\|^2_{(\nabla^2 \psi_t(u_t))^{-1}}. \tag{4.27}$$

Next, we bound $\left\| \tilde{\ell}_t \right\|^2_{(\nabla^2 \psi_t(u_t))^{-1}}$ on the right-hand side. It can be shown [Orabona, 2023, Section 10.1.2] that for the Tsallis entropy regularizer, the solution of the optimization problem (4.25) satisfies $\bar{q}_{i,t+1} \leq q_{i,t}$ whenever $\tilde{\ell}_{i,t} \geq 0$ for all $i \in [K]$. In our construction,

- $\tilde{\ell}_{i,t} = \frac{\ell_{i_t,t}}{p_{i_t,t}+\gamma} \geq 0$ if $i = i_t$,
- $\tilde{\ell}_{i,t} = 0$ if $i \in \mathbb{A}_t$ and $i \neq i_t$,
- $\tilde{\ell}_{i,t} = \ell_{i_t,t} - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} = \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t}+\gamma} \geq 0$ if $i \notin \mathbb{A}_t$.

Hence, the condition $\tilde{\ell}_{i,t} \geq 0$ holds for all $i \in [K]$. It follows that $u_{i,t} \leq q_{i,t}$ for all $i \in [K]$.

It is well-known [Abernethy et al., 2015] that the Hessian of Tsallis entropy is diagonal and equal to

$$(\nabla^2 \psi_t(x))_{ii} = \frac{\beta}{\eta x_i^{2-\beta}}.$$

It follows that its inverse is a diagonal matrix with entries $\left(\eta \frac{x_i^{2-\beta}}{\beta}\right)_{i=1,2,\ldots,K}$ on the main diagonal. Hence,

$$\begin{aligned}
\left\| \tilde{\ell}_t \right\|^2_{(\nabla^2 \psi_t(u_t))^{-1}} &= \frac{\eta}{\beta} \sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 u_{i,t}^{2-\beta} \\
&\leq \frac{\eta}{\beta} \sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\beta} \\
&\leq \frac{\eta}{\beta} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t}^2 p_{i,t}^{2-\beta} \\
&\leq \frac{\eta}{\beta} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} p_{i,t}^{1-\beta},
\end{aligned}$$

where the first inequality is due to $u_{i,t} \leq q_{i,t}$, the second inequality is due to Lemma 4.3.9 and the last inequality is due to $\tilde{\ell}_{i,t} p_{i,t} \leq 1$ for all $i \in \mathbb{A}_t$. Plugging this into (4.27) and using

$I_{a,t} \leq 1$ gives

$$\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t}) \leq \psi_{T+1}(e_a) - \min_{x \in \Delta_K} \psi_1(x) + \gamma \sum_{t=1}^{T} \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} + \frac{\eta}{2\beta} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} p_{i,t}^{1-\beta}$$
$$\leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \gamma \sum_{t=1}^{T} \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} + \frac{\eta}{2\beta} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} p_{i,t}^{1-\beta},$$

where, in the last inequality, where we used $\psi_{T+1}(e_a) - \min_{x \in \Delta_K} \psi_1(x) \leq \frac{K^{1-\beta}}{\eta(1-\beta)}$ as a standard property of the Tsallis entropy regularizer [Abernethy et al., 2015]. □

4.B.3 Proof of Theorem 4.3.6

Theorem 4.3.6. With $\gamma = 0$, for any $\beta \in (0,1), \eta > 0$, Algorithm 4.2 with Tsallis entropy guarantees

$$\max_{a \in [K]} \mathbb{E}[R(a)] \leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta} T A^{\beta}.$$

Setting $\beta = \frac{1}{2}$ and tuning η leads to an $O\left(\sqrt{2T\sqrt{AK}}\right)$ bound.

Proof. With $\gamma = 0$, Lemma 4.B.1 implies that

$$\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t}) \leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta} \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} p_{i,t}^{1-\beta}$$
$$= \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta} \sum_{t=1}^{T} \ell_{i_t,t} p_{i_t,t}^{-\beta},$$

where the equality is due to the fact that for $i \in \mathbb{A}_t$, $\ell_{i,t} = 0$ if $i \neq i_t$ and $\tilde{\ell}_{i_t,t} = \frac{\ell_{i_t,t}}{p_{i_t,t}}$. Taking the expectation over $i_t \sim p_t$ on both sides and using

$$\mathbb{E}_{i_t \sim p_t}[\ell_{i_t,t} p_{i_t,t}^{-\beta}] = \sum_{i \in \mathbb{A}_t} p_{i,t}^{1-\beta} \ell_{i,t}$$
$$\leq \sum_{i \in \mathbb{A}_t} p_{i,t}^{1-\beta}$$
$$\leq A_t^{\beta},$$

where the first inequality is due to $\ell_{i,t} \in [0,1]$ and the second inequality is Holder's inequality,

we obtain

$$\mathbb{E}[R(a)] \leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta} T A^{\beta}.$$

□

4.B.4 Proof of Theorem 4.3.7

Theorem 4.3.7. For any $\beta, \gamma, \eta \in (0,1)$, Algorithm 4.2 with Tsallis entropy guarantees

$$R(a) \leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta A^{\beta} T}{\beta} + \gamma A T + \left(\frac{\eta+\beta}{2\beta\gamma} + \frac{1}{2}\right) \ln(3/\delta) + \frac{\ln(3K/\delta)}{2\gamma}$$

simultaneously for all $a \in [K]$ with probability at least $1-\delta$. Letting $\beta = \frac{1}{2}$ and tuning η and γ leads to an $O\left(\sqrt{T\sqrt{AK}} + \sqrt{TA\ln(K/\delta)}\right)$ bound.

Proof. We apply Lemma 4.A.3 twice:

- the first time with $\delta' = \delta/3, \alpha_{i,t} = 2\gamma I_{i,t} p_{i,t}^{1-\beta}$ to obtain

$$\sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} p_{i,t}^{1-\beta} \tilde{\ell}_{i,t} \leq \frac{\ln(3/\delta)}{2\gamma} + \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} p_{i,t}^{1-\beta} \ell_{i,t} \tag{4.28}$$

  with probability at least $1-\delta/3$,

- the second time with $\delta' = \delta/3, \alpha_{i,t} = 2\gamma I_{i,t}$ to obtain

$$\sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} \leq \frac{\ln(3/\delta)}{2\gamma} + \sum_{t=1}^{T} \sum_{i \in \mathbb{A}_t} \ell_{i,t} \tag{4.29}$$

  with probability at least $1-\delta/3$.

We also apply Corollary 4.A.4 once with $\delta' = \frac{\delta}{3}$ to obtain

$$\sum_{t=1}^{T} I_{a,t} \tilde{\ell}_{a,t} \leq \frac{\ln(3G_T/\delta)}{2\gamma} + \sum_{t=1}^{T} I_{a,t} \ell_{a,t} \tag{4.30}$$

with probability at least $1-\delta/3$. Plugging (4.28), (4.29), (4.30) into Lemma 4.B.1 and taking

a union bound, we obtain with probability at least $1-\delta$,

$$\sum_{t=1}^{T} I_{a,t}\hat{\ell}_t \le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\ln(3G_T/\delta)}{2\gamma} + \sum_{t=1}^{T} I_{a,t}\ell_{a,t} + \frac{\ln(3/\delta)}{2} + \gamma\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t}\ell_{i,t} + \frac{\eta\ln(3/\delta)}{4\beta\gamma} + \frac{\eta}{2\beta}\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t} p_{i,t}^{1-\beta}\ell_{i,t}.$$

Subtracting $\sum_{t=1}^{T} I_{a,t}\ell_{a,t}$ on both sides, we obtain

$$\begin{aligned}
R(a) &\le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\ln(3G_T/\delta)}{2\gamma} + \frac{\ln(3/\delta)}{2} + \gamma\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t}\ell_{i,t} + \frac{\eta\ln(3/\delta)}{4\beta\gamma} + \frac{\eta}{2\beta}\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t} p_{i,t}^{1-\beta}\ell_{i,t} \\
&\le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\ln(3G_T/\delta)}{2\gamma} + \left(\frac{\eta}{4\beta\gamma}+\frac{1}{2}\right)\ln(3/\delta) + \gamma AT + \frac{\eta}{2\beta}\sum_{t=1}^{T}\sum_{i\in\mathbb{A}_t} p_{i,t}^{1-\beta} \\
&\le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta}TA^{\beta} + \gamma AT + \left(\frac{\eta}{4\beta\gamma}+\frac{1}{2}\right)\ln(3/\delta) + \frac{\ln(3G_T/\delta)}{2\gamma} \\
&\le \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{2\beta}TA^{\beta} + \gamma AT + \left(\frac{\eta}{4\beta\gamma}+\frac{1}{2}\right)\ln(3/\delta) + \frac{\ln(3K/\delta)}{2\gamma},
\end{aligned}$$

where

- the second inequality is due to $\ell_{i,t}\in[0,1]$ and $A_t \le A$,
- the third inequality is Holder's inequality and $A_t \le A$,
- the last inequality is due to $G_T \le K$.

□

4.C Proofs for Section 4.3.3

First, we state a more general definition of the active sets of arms. An arm i is active in round t if $I_{i,t} > 0$ i.e. $\mathbb{A}_t = \{i\in[K] : I_{i,t} > 0\}$. Let $\mathbb{G}_t = \cup_{s=1}^{t}\mathbb{A}_t$ and $G_t = |\mathbb{G}_t|$. Let the potential function $\tilde{Q}_t$ be

$$\tilde{Q}_t = \sum_{i\in\mathbb{G}_T}\tilde{q}_{i,t}.$$

Input: $\eta > 0, \gamma > 0, \eta \leq \gamma$
Initialize $\tilde{q}_{i,1} = 1$ for $i = 1, 2, \ldots, K$;
for each round $t = 1, \ldots, T$ **do**
  An adversary selects and reveals K values $I_{i,t} \in [0, 1]$;
  Compute $w_{i,t} = I_{i,t}\tilde{q}_{i,t}$ and $W_t = \sum_{i=1}^{K} w_{i,t}$;
  Compute $p_{i,t} = \frac{w_{i,t}}{W_t}$;
  Draw $i_t \sim p_t$ and observe $\hat{\ell}_t = \ell_{i_t,t}$;
  Construct loss estimate $\tilde{\ell}_{i,t}$ by Equation (4.31) ;
  Update $\tilde{q}_{i,t+1}$ by Equation (4.32).
end

Algorithm 1: SB-EXP3 for experts that report their confidences with bandit feedback

The protocol is given in Algorithm 1. In round t, the loss estimator is

$$\tilde{\ell}_{i,t} = \begin{cases} \frac{\ell_{i,t}\mathbb{1}\{i_t = i\}}{p_{i,t} + \gamma I_{i,t}} & \text{if } p_{i,t} > 0, \\ 0 \quad \text{otherwise.} \end{cases} \tag{4.31}$$

and the update rule for $\tilde{q}_{i,t+1}$ becomes

$$\tilde{q}_{i,t+1} = \tilde{q}_{i,t} \exp\left(\eta I_{i,t}\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}\right)\right). \tag{4.32}$$

Before proving Theorem 4.3.12, we present the following lemma which bounds the growth of $\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t}$.

Lemma 4.C.1. For any $t \geq 1$,

$$\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \leq 1 + \eta \frac{W_t}{\tilde{Q}_t} \sum_{i=1}^{K} p_{i,t}\left((\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}) + \eta I_{i,t}(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})^2\right)$$

Proof. By the definition of $\tilde{Q}_{t+1}$, we have

$$\begin{aligned}
\tilde{Q}_{t+1} - \tilde{Q}_t &= \sum_{i \in \mathbb{G}_T} \tilde{q}_{i,t} \exp\left(\eta I_{i,t}\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}\right)\right) - \sum_{i \in \mathbb{G}_T} \tilde{q}_{i,t} \\
&= \sum_{i \in \mathbb{G}_T} \tilde{q}_{i,t}\left(\exp\left(\eta I_{i,t}(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})\right) - 1\right) \\
&\leq \sum_{i \in \mathbb{G}_T} \tilde{q}_{i,t}\left(\eta I_{i,t}(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}) + \eta^2 I_{i,t}^2(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})^2\right) \\
&= \eta \sum_{i \in \mathbb{G}_T} I_{i,t}\tilde{q}_{i,t}\left((\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}) + \eta I_{i,t}(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})^2\right) \\
&= \eta W_t \sum_{i \in \mathbb{G}_T} p_{i,t}\left((\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}) + \eta I_{i,t}(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})^2\right),
\end{aligned}$$

where

- the inequality is obtained by applying $\exp(x) - 1 \leq x + x^2$ for any $x \leq 1$ on $x = \eta I_{i,t}(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})$ and multiplying both sides by $\tilde{q}_{i,t} \geq 0$,
- the last equality is due to the computation of $p_{i,t} = \frac{I_{i,t}\tilde{q}_{i,t}}{W_t}$.

Dividing both sides by $\tilde{Q}_t > 0$, we obtain the desired expression. □

Note that in each round t, Lemma 4.A.1 still holds for $I_{i,t} \in [0, 1]$. This implies the following corollary.

Corollary 4.C.2. For any $t \geq 1$,

$$\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \leq 1 + \eta^2 \sum_{i=1}^{K} p_{i,t} I_{i,t}\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}\right)^2.$$

Proof. We write the second term in the sum on the right-hand side of Lemma 4.C.1 as

follows:

$$\begin{aligned}\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} &\leq 1+\eta\frac{W_t}{\tilde{Q}_t}\sum_{i=1}^{K}p_{i,t}\left((\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t})+\eta I_{i,t}(\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t})^2\right)\\ &= 1+\underbrace{\frac{\eta W_t}{\tilde{Q}_t}\sum_{i=1}^{K}p_{i,t}(\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t})}_{(a)}+\underbrace{\frac{\eta^2 W_t}{\tilde{Q}_t}\sum_{i=1}^{K}p_{i,t}I_{i,t}(\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t})^2}_{(b)}\end{aligned}$$

We bound (a) and (b) separately. By Lemma 4.A.1 we have

$$\begin{aligned}\sum_{i=1}^{K}p_{i,t}(\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t}) &= \hat{\ell}_t-\sum_{i=1}^{K}p_{i,t}\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t}\\ &= 0,\end{aligned}$$

and thus the quantity (a) is equal to 0. To bound (b), we have

$$\begin{aligned}W_t &= \sum_{i=1}^{K}I_{i,t}\tilde{q}_{i,t}\\ &\leq \sum_{i=1}^{K}\tilde{q}_{i,t}\\ &= \tilde{Q}_t,\end{aligned}$$

where the inequality is due to $I_{i,t}\in[0,1]$ and $\tilde{q}_{i,t}>0$. This implies that $0<\frac{W_t}{\tilde{Q}_t}\leq 1$. Multiplying both sides by $\eta^2\sum_{i=1}^{K}p_{i,t}I_{i,t}(\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}\tilde{\ell}_{j,t})^2\geq 0$, we obtain

$$(b)\leq\eta^2\sum_{i=1}^{K}p_{i,t}I_{i,t}\left(\hat{\ell}_t-\tilde{\ell}_{i,t}-\gamma\sum_{j=1}^{K}I_{j,t}\tilde{\ell}_{j,t}\right)^2,$$

which implies the desired statement. □

4.C.1 Proof of Theorem 4.3.12

Theorem 4.3.12. With optimally tuned η and γ, Algorithm 4.1 guarantees

$$R(a) \le O\left(\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t}\ln(G_T/\delta)}\right)$$

simultaneously for all $a \in [K]$ with probability $1-\delta$.

Proof. We still employ the standard strategy of lower and upper bounding Q_{T+1}. We have

$$\ln \tilde{Q}_{T+1} \ge \ln \tilde{q}_{a,T+1} = \eta \sum_{t=1}^{T} I_{a,t}\left(\hat{\ell}_t - \tilde{\ell}_{a,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}\right). \tag{4.33}$$

On the other hand,

$$\begin{aligned}
\ln \tilde{Q}_{T+1} &= \ln \tilde{Q}_1 + \sum_{t=1}^{T} \ln \frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \\
&\le \ln G_T + \sum_{t=1}^{T} \ln\left(1 + \eta^2 \sum_{i=1}^{K} p_{i,t} I_{i,t} (\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t})^2\right) \\
&\le \ln G_T + \eta^2 \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t} I_{i,t}\left(\hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}\right)^2
\end{aligned} \tag{4.34}$$

where the first inequality is due to Corollary 4.C.2 and the second inequality is due to $\ln(1+x) \le x$ for all $x \ge -1$. Let $c_t = \gamma \sum_{j=1}^{K} I_{j,t}\tilde{\ell}_{j,t}$. For any $t \ge 1$, we have:

$$\begin{aligned}
\hat{\ell}_t - c_t &= \hat{\ell}_t - \gamma I_{i_t,t}\tilde{\ell}_{i_t,t} \qquad (\text{ since } \tilde{\ell}_{j,t} = 0 \text{ if } j \ne i_t) \\
&= \hat{\ell}_t - \frac{\gamma I_{i_t,t}\hat{\ell}_t}{p_{i_t,t} + \gamma I_{i_t,t}} \\
&= \frac{p_{i_t,t}\hat{\ell}_t}{p_{i_t,t} + \gamma I_{i_t,t}} \\
&= p_{i_t,t}\tilde{\ell}_{i_t,t}.
\end{aligned}$$

It follows that

$$\begin{aligned}\sum_{i=1}^{K} p_{i,t} I_{i,t} (\hat{\ell}_t - \tilde{\ell}_{i,t} - c_t)^2 &= \sum_{i=1}^{K} p_{i,t} I_{i,t} (\hat{\ell}_t - c_t)^2 + \sum_{i=1}^{K} p_{i,t} I_{i,t} \tilde{\ell}_{i,t}^2 - 2 \sum_{i=1}^{K} p_{i,t} I_{i,t} (\hat{\ell}_t - c_t) \tilde{\ell}_{i,t} \\ &\leq (\hat{\ell}_t - c_t)^2 \sum_{i=1}^{K} p_{i,t} I_{i,t} + \sum_{i=1}^{K} p_{i,t} I_{i,t} \tilde{\ell}_{i,t}^2 \\ &\leq \sum_{i=1}^{K} p_{i,t} I_{i,t} + \sum_{i=1}^{K} I_{i,t} \tilde{\ell}_{i,t},\end{aligned} \tag{4.35}$$

where

- the first inequality is due to $\hat{\ell}_t - c_t = p_{i_t,t} \tilde{\ell}_{i_t,t} \geq 0$,
- the second inequality is due to $\hat{\ell}_t - c_t = p_{i_t,t} \tilde{\ell}_{i_t,t} \leq 1$.

Combining (4.33), (4.34) and (4.35), we obtain

$$\sum_{t=1}^{T} I_{a,t} (\hat{\ell}_t - \tilde{\ell}_{a,t} - \gamma \sum_{j=1}^{K} I_{j,t} \tilde{\ell}_{j,t}) \leq \frac{\ln G_T}{\eta} + \eta \sum_{t=1}^{T} \left(\sum_{i=1}^{K} p_{i,t} I_{i,t} + \sum_{i=1}^{K} I_{i,t} \tilde{\ell}_{i,t} \right), \tag{4.36}$$

which implies that

$$\sum_{t=1}^{T} I_{a,t} \hat{\ell}_t \leq \sum_{t=1}^{T} I_{a,t} \tilde{\ell}_{a,t} + \frac{\ln G_T}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} I_{i,t} + \sum_{t=1}^{T} \sum_{i=1}^{K} (I_{a,t} \gamma + \eta) I_{i,t} \tilde{\ell}_{i,t} \tag{4.37}$$

$$\leq \sum_{t=1}^{T} I_{a,t} \tilde{\ell}_{a,t} + \frac{\ln G_T}{\gamma} + \gamma \sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} I_{i,t} + 2\gamma \sum_{t=1}^{T} \sum_{i=1}^{K} I_{i,t} \tilde{\ell}_{i,t}, \tag{4.38}$$

where the last inequality is due to $I_{a,t} \leq 1$ and picking $\eta = \gamma$.

Note that Lemma 4.A.3 holds for $I_{i,t} \in [0,1]$. Applying Lemma 4.A.3 with $\alpha_{i,t} = 2\gamma I_{i,t}, \delta' = \delta/2$, applying Corollary 4.A.4 with $\delta' = \delta/2$ and using $\sum_{i=1}^{K} p_{i,t} I_{i,t} \leq \sum_{i=1}^{K} I_{i,t}$, we obtain that with probability at least $1 - \delta$, simultaneously for all $a \in [K]$,

$$\sum_{t=1}^{T} I_{a,t} \hat{\ell}_t \leq \sum_{t=1}^{T} I_{a,t} \ell_{a,t} + \frac{\ln(2G_T/\delta)}{2\gamma} + \frac{\ln G_T}{\gamma} + \gamma \sum_{t=1}^{T} \sum_{i=1}^{K} I_{i,t} + \ln(2/\delta) + 2\gamma \sum_{t=1}^{T} \sum_{j=1}^{K} I_{j,t} \ell_{j,t}.$$

Subtracting $\sum_{t=1}^{T} I_{a,t}\ell_{a,t}$ on both sides and using $\ell_{j,t} \leq 1$ and $\gamma, \delta \in (0,1)$, we obtain

$$R(a) \leq \frac{\ln(2G_T/\delta)}{\gamma} + \frac{\ln G_T}{\gamma} + \ln(2/\delta) + 3\gamma \sum_{t=1}^{T}\sum_{j=1}^{K} I_{j,t} \tag{4.39}$$

$$\leq \frac{3\ln(G_T/\delta)}{\gamma} + 3\gamma \sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t}. \tag{4.40}$$

Setting $\gamma = \eta = \sqrt{\frac{\ln(G_T/\delta)}{\sum_{t=1}^{T}\sum_{i=1}^{K} I_{i,t}}}$ leads to the desired bound. □

4.D Proofs for Section 4.4

Recall that the sleeping experts are considered *sleeping augmented arms*. Note that $\mathbb{B}_t = \{m \in [M] : I_{m,t} = 1\}$ is the set of awake experts as defined in the main text. For an expert $u \in [M]$, the actual loss of expert u in round t is

$$x_{u,t} = \langle E_{u,t}, \ell_t \rangle.$$

First, we prove a technical lemma showing that the z_t-weighted average of these estimated losses of the augmented arms is equivalent to $\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}$. This lemma is the counterpart of Lemma 4.A.1.

Lemma 4.D.1. For any $t \in [K]$ and $m \in \mathbb{B}_t$,

$$\mathbb{E}_{m \sim z_t}[\tilde{x}_{m,t}] = \hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}.$$

Proof. Let $E_{m,t}^{(k)}$ be the value of the element at index k in $E_{m,t}$. We have

$$\begin{aligned}
\mathbb{E}_{m\sim z_t}[\tilde{x}_{m,t}] &= \sum_{m\in\mathbb{B}_t} z_{m,t}\tilde{x}_{m,t} \\
&= \sum_{m\in\mathbb{B}_t} z_{m,t} \sum_{k=1}^{K} E_{m,t}^{(k)}\tilde{\ell}_{k,t} \\
&= \sum_{m\in\mathbb{B}_t} z_{m,t} E_{m,t}^{(i_t)}\tilde{\ell}_{i_t,t} \\
&= \frac{\hat{\ell}_t}{p_{i_t,t}+\gamma} \sum_{m\in\mathbb{B}_t} z_{m,t} E_{m,t}^{(i_t)} \\
&= \frac{\hat{\ell}_t p_{i_t,t}}{p_{i_t,t}+\gamma} \\
&= \hat{\ell}_t - \gamma\frac{\hat{\ell}_t}{p_{i_t,t}+\gamma} \\
&= \hat{\ell}_t - \gamma\tilde{\ell}_{i_t,t},
\end{aligned}$$

where

- the second equality is by Equation (4.13)
- the third equality is due to $\tilde{\ell}_{k,t} = 0$ whenever $k \neq i_t$
- the fourth equality is due to $p_{k,t} = \sum_{m\in\mathbb{B}_t} z_{m,t}E_{m,t}^{(k)}$ for all $k \in [K]$.

□

Let $\tilde{Q}_t = \sum_{m=1}^{M} \tilde{q}_{m,t}$. For a set $S \in [M]$, let $\tilde{Q}_{S,t} = \sum_{m\in S} \tilde{q}_{m,t}$ be the projection of $\tilde{Q}_t$ on S. Let $\bar{S} = [M] \setminus S$ for any $S \subseteq [M]$. Lemma 4.D.1 leads to the following technical lemma that resembles Lemma 4.3.3.

Lemma 4.D.2. For any $t \geq 0$,

$$\frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \leq \sum_{m=1}^{M} z_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma\sum_{j=1}^{K}\tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right).$$

Proof. The proof makes use of the following two facts that can be proved easily:

- Fact 1: The function $f(x) = e^{-\eta x}$ is convex for any $\eta \in \mathbb{R}$.

- Fact 2: For any $a, b > 0, c \geq 0$, if $a \geq b$ then

$$\frac{a}{b} \geq \frac{a+c}{b+c}.$$

By Jensen's inequality and Fact 1, we have

$$\begin{aligned}\sum_{m=1}^{M} z_{m,t} \exp(-\eta \tilde{x}_{m,t}) &= \mathbb{E}_{m \sim z_t}[\exp(-\eta \tilde{x}_{m,t})] \\ &\geq \exp\left(-\eta \mathbb{E}_{m \sim z_t}[\tilde{x}_{m,t}]\right) \\ &= \exp\left(-\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t})\right),\end{aligned}$$

where the last equality is due to Lemma 4.D.1. Since $z_{m,t} = 0$ for $m \notin \mathbb{B}_t$, the expression above is equivalent to

$$\sum_{m \in \mathbb{B}_t} z_{m,t} \exp(-\eta \tilde{x}_{m,t}) \geq \exp\left(-\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t})\right).$$

Multiplying $\exp\Big(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t})\Big) \tilde{Q}_{\mathbb{B}_t,t} > 0$ on both sides, we obtain

$$\sum_{m \in \mathbb{B}_t} z_{m,t} \tilde{Q}_{\mathbb{B}_t,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right) \geq \tilde{Q}_{\mathbb{B}_t,t}. \tag{4.41}$$

By definition, $z_{m,t} = \frac{\tilde{q}_{m,t}}{\tilde{Q}_{\mathbb{B}_t,t}}$. Hence, Equation (4.41) is equivalent to

$$\sum_{m \in \mathbb{B}_t} \tilde{q}_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right) \geq \tilde{Q}_{\mathbb{B}_t,t}.$$

By our update rule, $\tilde{q}_{m,t+1} = \tilde{q}_{m,t} \exp\Big(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\Big)$ for $m \in \mathbb{B}_t$ and $\tilde{q}_{m,t+1} = \tilde{q}_{m,t}$ for $m \notin \mathbb{B}_t$. Hence,

$$\sum_{m \in \mathbb{B}_t} \tilde{q}_{m,t+1} = \sum_{m \in \mathbb{B}_t} \tilde{q}_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right) \geq \tilde{Q}_{\mathbb{B}_t,t}.$$

Applying Fact 2 for $a = \sum_{m\in\mathbb{B}_t} \tilde{q}_{m,t+1}, b = \tilde{Q}_{\mathbb{B}_t,t}$ and $c = \tilde{Q}_{\bar{\mathbb{B}}_t,t}$, we obtain

$$\begin{aligned}
\sum_{i=m}^{M} z_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right) &= \sum_{m\in\mathbb{B}_t} z_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right) \\
&= \frac{\sum_{m\in\mathbb{B}_t} \tilde{q}_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t})\right)}{\tilde{Q}_{\mathbb{B}_t,t}} \\
&= \frac{\sum_{m\in\mathbb{B}_t} \tilde{q}_{m,t+1}}{\tilde{Q}_{\mathbb{B}_t,t}} \\
&\geq \frac{\sum_{m\in\mathbb{B}_t} \tilde{q}_{m,t+1} + \tilde{Q}_{\bar{\mathbb{B}}_t,t}}{\tilde{Q}_{\mathbb{B}_t,t} + \tilde{Q}_{\bar{\mathbb{B}}_t,t}} \\
&= \frac{\tilde{Q}_{t+1}}{\tilde{Q}_t},
\end{aligned} \tag{4.42}$$

where the last equality is due to the fact that $\sum_{m\in\bar{\mathbb{B}}_t} \tilde{q}_{m,t+1} = \sum_{m\in\bar{\mathbb{B}}_t} \tilde{q}_{m,t} = \tilde{Q}_{\bar{\mathbb{B}}_t,t}$. □

4.D.1 Bounding the Estimated Regret

The following lemma bounds the estimated regret of each expert $u \in [M]$.

Lemma 4.D.3. For any $\gamma \geq 0$, for any $u \in [M]$, SE-EXP4 guarantees that

$$\sum_{t=1}^{T} I_{u,t}(\hat{\ell}_t - \tilde{x}_{u,t}) \leq \frac{\ln M}{\eta} + \left(\gamma + \frac{\eta}{2}\right) \sum_{t=1}^{T} \sum_{j=1}^{K} \tilde{\ell}_{j,t}. \tag{4.43}$$

Proof. We have

$$\begin{aligned}
\ln \tilde{Q}_{T+1} &= \ln \sum_{m=1}^{K} \tilde{q}_{m,T+1} \\
&\geq \ln \tilde{q}_{u,T+1} \\
&= \eta \sum_{t=1}^{T} I_{u,t}(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{u,t}).
\end{aligned} \tag{4.44}$$

On the other hand, we have

$$\begin{aligned}
\ln \tilde{Q}_{T+1} &= \ln \tilde{Q}_1 + \sum_{t=1}^{T} \ln \frac{\tilde{Q}_{t+1}}{\tilde{Q}_t} \\
&\leq \ln M + \sum_{t=1}^{T} \ln \left(\sum_{m \in \mathbb{B}_t} z_{m,t} \exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{m,t}) \right) \right) \\
&= \ln M + \sum_{t=1}^{T} \ln \left(\exp\left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}) \right) \sum_{m \in \mathbb{B}_t} z_{m,t} \exp(-\eta \tilde{x}_{m,t}) \right) \\
&= \ln M + \sum_{t=1}^{T} \left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}) + \ln \left(\sum_{m \in \mathbb{B}_t} z_{m,t} \exp(-\eta \tilde{x}_{m,t}) \right) \right) \\
&\leq \ln M + \sum_{t=1}^{T} \left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}) + \ln \left(\sum_{m \in \mathbb{B}_t} z_{m,t} (1 + \frac{\eta^2 \tilde{x}_{m,t}^2}{2} - \eta \tilde{x}_{m,t}) \right) \right) \\
&= \ln M + \sum_{t=1}^{T} \left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}) + \ln \left(1 + \eta^2 \sum_{m \in \mathbb{B}_t} \frac{z_{m,t} \tilde{x}_{m,t}^2}{2} - \eta \sum_{m \in \mathbb{B}_t} z_{m,t} \tilde{x}_{m,t} \right) \right) \\
&\leq \ln M + \sum_{t=1}^{T} \left(\eta(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t}) + \eta^2 \sum_{m \in \mathbb{B}_t} \frac{z_{m,t} \tilde{x}_{m,t}^2}{2} - \eta \sum_{m \in \mathbb{B}_t} z_{m,t} \tilde{x}_{m,t} \right) \\
&= \ln M + \eta^2 \sum_{t=1}^{T} \sum_{m \in \mathbb{B}_t} \frac{z_{m,t} \tilde{x}_{m,t}^2}{2},
\end{aligned}$$

where

- the first inequality is due to Lemma 4.D.2,
- the second inequality is $\exp(-x) \leq 1 + \frac{x^2}{2} - x$ for all $x \geq 0$,
- the third inequality is $\ln(1 + x) \leq x$ for all $x \geq -1$,
- the last equality is due to Lemma 4.D.1.

We obtain

$$\sum_{t=1}^{T} I_{u,t}(\hat{\ell}_t - \gamma \sum_{j=1}^{K} \tilde{\ell}_{j,t} - \tilde{x}_{u,t}) \leq \frac{\ln M}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{m \in \mathbb{B}_t} z_{m,t} \tilde{x}_{m,t}^2. \tag{4.45}$$

Next, we proceed in the same way as in [Neu, 2015]. We have

$$\sum_{m\in\mathbb{B}_t} z_{m,t}\tilde{x}_{m,t}^2 = \sum_{m\in\mathbb{B}_t} z_{m,t}\left(\sum_{k=1}^{K} E_{m,t}^{(k)}\tilde{\ell}_{k,t}\right)^2 \tag{4.46}$$

$$\leq \sum_{m\in\mathbb{B}_t} z_{m,t}\sum_{k=1}^{K} E_{m,t}^{(k)}(\tilde{\ell}_{k,t})^2 \tag{4.47}$$

$$= \sum_{k=1}^{K}(\tilde{\ell}_{k,t})^2\sum_{m\in\mathbb{B}_t} z_{m,t}E_{m,t}^{(k)} \tag{4.48}$$

$$= \sum_{k=1}^{K} p_{k,t}(\tilde{\ell}_{k,t})^2 \tag{4.49}$$

$$\leq \sum_{k=1}^{K}\tilde{\ell}_{k,t}, \tag{4.50}$$

where the first inequality is Jensen's inequality. This implies

$$\sum_{t=1}^{T} I_{u,t}(\hat{\ell}_t - \gamma\sum_{j=1}^{K}\tilde{\ell}_{j,t} - \tilde{x}_{u,t}) \leq \frac{\ln M}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{k=1}^{K}\tilde{\ell}_{k,t}. \tag{4.51}$$

Moving $\gamma\sum_{t=1}^{T} I_{u,t}\sum_{j=1}^{K}\tilde{\ell}_{j,t}$ to the right-hand side and using $I_{u,t}\in[0,1]$, we obtain the desired statement. □

4.D.2 Proof of Theorem 4.4.1

Theorem 4.4.1. For any $\gamma,\eta\in(0,1),\eta\leq 2\gamma$, SE-EXP4 guarantees

$$R(u) \leq \frac{\ln M}{\eta} + \frac{\ln(2M/\delta)}{2\gamma} + (\gamma+\frac{\eta}{2})TK + \ln(2/\delta) \tag{4.14}$$

simultaneously for all experts $u\in[M]$ with probability at least $1-\delta$, where the probability is taken over the sequence of the learner's selected arms. Tuning η and γ leads to an $O(\sqrt{TK\ln(M/\delta)})$ bound.

Proof. Lemma 4.D.3 implies that

$$\begin{aligned}\sum_{t=1}^{T} I_{u,t}\hat{\ell}_t &\le \frac{\ln M}{\eta} + \sum_{t=1}^{T} I_{u,t}\tilde{x}_{u,t} + \sum_{t=1}^{T}\sum_{j=1}^{K}\left(\gamma+\frac{\eta}{2}\right)\tilde{\ell}_{j,t} \\ &= \frac{\ln M}{\eta} + \sum_{t=1}^{T} I_{u,t}\sum_{j=1}^{K} E_{u,t}^{(j)}\tilde{\ell}_{j,t} + \sum_{t=1}^{T}\sum_{j=1}^{K}\left(\gamma+\frac{\eta}{2}\right)\tilde{\ell}_{j,t}.\end{aligned} \tag{4.52}$$

Next, we apply Lemma 4.A.3 twice, where

- the first time with $\alpha_{i,t} = 2\gamma I_{u,t}E_{u,t}^{(i)}, \delta' = \frac{\delta}{2M}$ and a union bound over $[M]$ implies

$$\sum_{t=1}^{T}\sum_{j=1}^{K} I_{u,t}E_{u,t}^{(j)}(\tilde{\ell}_{j,t} - \ell_{j,t}) \le \frac{\ln(2M/\delta)}{2\gamma} \tag{4.53}$$

  with probability at least $1-\delta/2$;

- the second time with $\alpha_{i,t} = \gamma + \eta/2, \delta' = \delta/2$ implies

$$\left(\gamma+\frac{\eta}{2}\right)\sum_{t=1}^{T}\sum_{j=1}^{K}(\tilde{\ell}_{j,t} - \ell_{j,t}) \le \ln(2/\delta) \tag{4.54}$$

  with probability at least $1-\delta/2$.

Plugging (4.53) and (4.54) into (4.52) yields

$$\begin{aligned}\sum_{t=1}^{T} I_{u,t}\hat{\ell}_t &\le \frac{\ln M}{\eta} + \sum_{t=1}^{T} I_{u,t}\sum_{j=1}^{K} E_{u,t}^{(j)}\ell_{j,t} + \frac{\ln(2M/\delta)}{2\gamma} + \left(\gamma+\frac{\eta}{2}\right)\sum_{t=1}^{T}\sum_{j=1}^{K}\ell_{j,t} + \ln(2/\delta) \\ &\le \frac{\ln M}{\eta} + \sum_{t=1}^{T} I_{u,t}x_{u,t} + \frac{\ln(2M/\delta)}{2\gamma} + \left(\gamma+\frac{\eta}{2}\right)TK + \ln(2/\delta).\end{aligned}$$

Moving $\sum_{t=1}^{T} I_{u,t}x_{u,t}$ to the left-hand side, we obtain

$$R(u) \le \frac{\ln M}{\eta} + \frac{\ln(2M/\delta)}{2\gamma} + \left(\gamma+\frac{\eta}{2}\right)TK + \ln(2/\delta). \tag{4.55}$$

Letting $\eta = 2\gamma$ and tuning η implies the $O(\sqrt{TK\ln(M/\delta)})$ bound. □

4.D.3 A Pseudo-Regret Bound of SE-EXP4

We bound the pseudo-regret $\mathbb{E}[R(u)]$ for any expert $u \in [M]$. Note that $\gamma = 0$. Taking the expectation on both side of Lemma 4.D.3 and using

$$\mathbb{E}_{i_t \sim p_t}\left[\tilde{\ell}_{j,t}\right] = \ell_{j,t} \leq 1,$$

we obtain

$$\mathbb{E}[\sum_{t=1}^{T} I_{u,t}(\hat{\ell}_t - \tilde{x}_{u,t})] \leq \frac{\ln M}{\eta} + \frac{\eta}{2} TK. \tag{4.56}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}_{i_t \sim p_t}[\tilde{x}_{u,t}] &= \sum_{k=1}^{K} p_{k,t} \mathbb{E}[\tilde{x}_{u,t} \mid i_t = k] \\
&= \sum_{k=1}^{K} p_{k,t} \mathbb{E}[\sum_{j=1}^{K} E_{u,t}^{(j)} \tilde{\ell}_{j,t} \mid i_t = k] \\
&= \sum_{k=1}^{K} p_{k,t} E_{u,t}^{(k)} \frac{\ell_{k,t}}{p_{k,t}} \\
&= \sum_{k=1}^{K} E_{u,t}^{(k)} \ell_{k,t} \\
&= \langle E_{u,t}, \ell_t \rangle.
\end{aligned}$$

We conclude that

$$\mathbb{E}[R(u)] \leq \frac{\ln M}{\eta} + \frac{\eta TK}{2}. \tag{4.57}$$

By setting $\eta = \sqrt{\frac{2 \ln M}{TK}}$ we obtain the bound

$$\mathbb{E}[R(u)] \leq \sqrt{2TK \ln M}. \tag{4.58}$$

4.D.4 Proof of Theorem 4.4.2

Theorem 4.4.2. For any $\gamma, \eta \in (0,1), \eta \leq 2\gamma$, SE-EXP4 with virtual experts guarantees that

$$R_{[t_1,t_2]}(k) \leq \frac{2\ln(KT)}{\eta} + \frac{\ln(KT/\delta)}{\gamma} + (\gamma + \frac{\eta}{2})TK + \ln(2/\delta)$$

simultaneously for all intervals $[t_1, t_2]$ and arms $k \in [K]$ with probability $1-\delta$. Tuning η and γ leads to an $O(\sqrt{TK\ln(KT/\delta)})$ bound.

Proof. For any triple (k, t_1, t_2) we create a virtual expert that is active from round t_1 to round t_2 and give advice e_k. There are $M = K\binom{T}{2} = \frac{KT(T+1)}{2}$ such experts. Because $M \leq KT^2 \leq (KT)^2$, we have

$$\ln M \leq \ln\big((KT)^2\big) = 2\ln(KT).$$

Furthermore, for all $\delta \in (0,1)$,

$$\ln(2M/\delta) \leq \ln\big(4M/\delta^2\big) \leq \ln\big((2KT/\delta)^2\big) = 2\ln(2KT/\delta).$$

Let u_{k,t_1,t_2} denote the expert indexed by (k, t_1, t_2). By Theorem 4.4.1, with probability at least $1-\delta$,

$$\begin{aligned} R_{[t_1,t_2]}(k) &= R(u_{k,t_1,t_2}) \\ &\leq \frac{\ln M}{\eta} + \frac{\ln(2M/\delta)}{2\gamma} + (\gamma + \frac{\eta}{2})TK + \ln(2/\delta) \\ &\leq \frac{2\ln(KT)}{\eta} + \frac{\ln(2KT/\delta)}{\gamma} + (\gamma + \frac{\eta}{2})TK + \ln(2/\delta) \end{aligned} \tag{4.59}$$

holds simultaneously for all u_{k,t_1,t_2}. □

4.D.5 Proof of Corollary 4.4.3

Corollary 4.4.3. Restarting SE-EXP4 with virtual experts after every T/S rounds guarantees that for any $j_{1:T}$ where $H(j_{1:T}) \leq S$, with probability at least $1-\delta$

$$R(j_{1:T}) \leq O\left(\sqrt{STK\ln\left(\frac{KT}{\delta S}\right)}\right).$$

Proof. Without loss of generality, assume that T is divisible by S. We use the following

algorithm called SE-EXP4-Restart, which runs in S episodes. In each episode, a new instance of SE-EXP4 with virtual experts is run for T/S rounds. Let $b = 1, 2, \ldots, S$ be an index for the episodes, and $t_{(b)} = \frac{bT}{S}$ be the ending round of episode b. Note that each episode b starts from round $\frac{(b-1)T}{S} + 1$ to round $\frac{bT}{S}$.

We examine the regret of this SE-EXP4-Restart with respect to the competing arms $j_{1:T}$ in each episode b. Divide T rounds into $Z = H(j_{1:T}) + 1$ non-overlapping segments, where the competing arms are the same within each segment $z = 1, \ldots, Z$. Let S_z, E_z be the first and last rounds of segment z. For every pair (b, z) of episode b and segment z, let

$$F_{b,z} = [S_z, E_z] \cap [\frac{(b-1)T}{S}, \frac{bT}{S}] \tag{4.60}$$

be the intersection between the rounds of episode b and segment z. Since the episodes are non-overlapping and the segments are non-overlapping, the intervals $F_{b,z}$'s are non-overlapping. In addition, their union is $\cup_{b,z} F_{b,z} = [T]$.

Fix an episode b and a segment z. There are two cases:

- $F_{b,z} = \emptyset$: Obviously, the regret of SE-EXP4-Restart on this empty interval is zero.
- $F_{b,z} \neq \emptyset$: in this case, because $F_{b,z}$ is an interval within episode b, the tracking regret of SE-EXP4-Restart on $F_{b,z}$ cannot exceed the adaptive regret of running (a new instance of) SE-EXP4 with virtual experts under horizon T/S. By Theorem 4.4.2, this is bounded by

$$R_{F_{b,s}} \leq \frac{2}{\eta} \ln \left(\frac{KT}{S\delta} \right) + \eta TK/S + \ln(2S/\delta) \tag{4.61}$$

  with probability at least $1 - \delta/S$.

Taking a union bound over all S episodes and setting $\eta = 2\gamma$ implies that with probability at least $1 - \delta$,

$$R_{F_{b,s}} \leq \frac{2}{\eta} \ln \left(\frac{KT}{S\delta} \right) + \eta TK/S + \ln(2S/\delta)$$

simultaneously for all intervals $F_{b,s}$. Because $F_{b,z}$'s are non-overlapping and their union is

$[T]$, the tracking regret of SE-EXP4-Restart is bounded by

$$R(j_{1:T}) = \sum_{b,z} R_{F_{b,z}} \tag{4.62}$$

$$= \left(\sum_{b,z} \mathbb{1}\{F_{b,z} \neq \emptyset\}\right)\left(\frac{2}{\eta}\ln\left(\frac{KT}{S\delta}\right) + \eta TK/S + \ln(2S/\delta)\right). \tag{4.63}$$

Next, we show that the count $C = \sum_{b,z} \mathbb{1}\{F_{b,z} \neq \emptyset\}$ is smaller than $2S$. Observe that the S episodes split the sequence of T rounds into S intervals with $S-1$ splitting points (not counting the two ends at 0 and T). Similarly, the $S+1$ segments of $j_{1:T}$ have S splitting points. In total, there are at most $2S-1$ splitting points from the episodes and the segments of $j_{1:T}$. Each non-empty interval $F_{b,z}$ has an ending point that is either T or one of the $2S-1$ splitting points. Therefore, there can be at most $2S$ such $F_{b,z}$. We conclude that $C \leq 2S$. As a result, with probability at least $1-\delta$,

$$R(j_{1:T}) \leq \frac{4S}{\eta}\ln\left(\frac{KT}{S\delta}\right) + 2\eta TK + 2S\ln(2S/\delta).$$

Letting $\eta = \sqrt{\frac{2S\ln\left(\frac{KT}{S\delta}\right)}{TK}}$ leads to the desired bound. □

4.E Proofs for Section 4.5

Theorem 4.5.1. For any (possibly randomized) algorithm with guarantee

$$\sup_{a\in[K]} E[R(a)] \leq O(T^{\gamma}A^{\beta}(\ln(T))^{\mu}), \tag{4.15}$$

where $\gamma \in (0,1), \beta \geq 0, \mu \geq 0$ are constants, there exists a number of arms K and sequence of sets of active arms and their losses such that $A = 2$ and for at least one arm $a \in [K]$,

$$T_a \geq \Omega(T^{1-\gamma}(\ln(T))^{-\mu}) \text{ and } \mathbb{E}[R(a)] \geq \Omega(T_a).$$

Proof. Let $\mathcal{A}$ be any algorithm with the stated worst-case guarantee and $f(T,A) = O(T^{\gamma}A^{\beta}(\ln(T))^{\mu})$ represent the worst-case regret bound of $\mathcal{A}$. We show a construction with $A_t = 2$ for all $t = 1,\ldots,T$. Our construction is adapted from the lower bound construction of [Daniely et al., 2015] for strongly adaptive regret in the standard adversarial MAB setting. Without loss of generality, assume $4f(T,A)$ divides T. Let $L = \frac{T}{4f(T,A)}$ and $K = 1 + 4f(T,A)$.

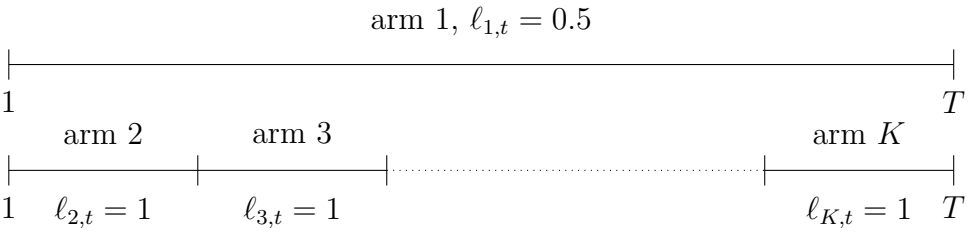

Figure 4.1: Environment V_0. All arms have loss equal 1 when they are active

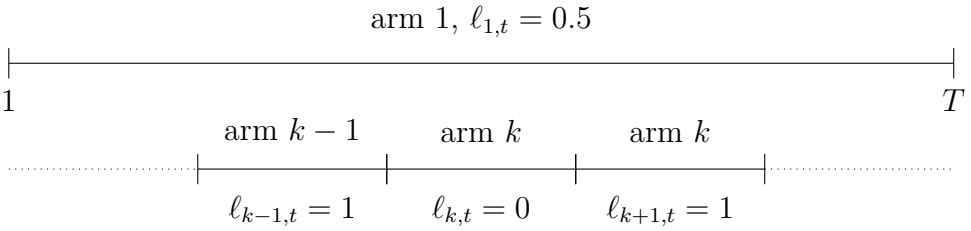

Figure 4.2: Environment V_k. Except for arm k which has losses equal to 0, all arms have losses equal to 1 when they are active.

Obviously $L \geq \Omega(T^{1-\gamma}A^{-\beta}(\ln(T))^{-\mu})$. Consider an environment V_0 defined as follows:

- Arm 1 is always active. Its losses are $\ell_{1,t} = 0.5$ for all $t \in [T]$.
- Arm $k = 2, 3, \ldots, K$ are active for the rounds in $\mathcal{I}_k = \left[\frac{(k-2)T}{4f(T,A)} + 1, \frac{(k-1)T}{4f(T,A)}\right]$. The length of each interval is L. Their losses are $\ell_{k,t} = 1$ for the rounds t in which they are active.

Figure 4.1 illustrates this environment V_0. We also define $K-1$ competing environments V_k for $k = 2, 3, \ldots, K$, defined as follows:

- The number of arms and their active rounds are identical to that of V_0. That is, arm 1 is always active for all rounds while each of the arms $k = 2, 3, \ldots, K$ are active for rounds within $\mathcal{I}_k$, respectively.
- The losses of every arm are the same as in V_0, except for that of arm k: all losses of arm k are 0. Figure 4.2 illustrates environment $V_{k,j}$.

Following the standard strategy of comparing the behavior of the algorithm on V_0 and competing environments [Auer et al., 2002b; Daniely et al., 2015], we first consider the neutral environment V_0. Let $\mathbb{E}_0$ and $\Pr_0$ indicate the expectation and probability taken in this environment over the randomness of the algorithm $\mathcal{A}$, respectively. Let $U = \{t : i_t \neq 1\}$ be the rounds in which the arm chosen by $\mathcal{A}$ is not arm 1. On V_0, since arm 1 is the best arm

and the gaps between the losses of arm 1 and that of every other arm is 0.5, the inequality $\sup_a \mathbb{E}_0[R(a)] \leq f(T, A)$ implies

$$\mathbb{E}_0 [|U|] \leq 2f(T, A). \tag{4.64}$$

For any $k \in \{2, 3, \ldots, K\}$, let

$$\mathcal{E}_k = \{U \cap \mathcal{I}_k = \emptyset\}$$

be the event that only arm 1 is chosen by $\mathcal{A}$ on $\mathcal{I}_k$. Because the $\mathcal{I}_k$ are non-overlapping and $\cup_{k=2,\ldots,K} \mathcal{I}_k = [T]$, we can write

$$U = \cup_{k=2,\ldots,K} (U \cap \mathcal{I}_k),$$

and

$$|U| = \sum_{k=2}^{K} |U \cap \mathcal{I}_k|.$$

Next, we show that for some segment $\mathcal{I}_{k^*}$ of size $L = \frac{T}{4f(T,A)}$, we have $\mathbb{E}_0 [|U \cap \mathcal{I}_{k^*}|] \leq \frac{1}{2}$. Assume on the contrary that $\mathbb{E}_0 [|U \cap \mathcal{I}_k|] > \frac{1}{2}$ for all $k = 2, \ldots, K$. Then,

$$\begin{aligned} \mathbb{E}_0[U] &= \sum_{k=2}^{K} \mathbb{E}_0 [U \cap \mathcal{I}_k] \\ &> \frac{K-1}{2} \\ &= 2f(T, A), \end{aligned}$$

which contradicts (4.64). Since $|U \cap \mathcal{I}_{k^*}|$ is a non-negative integer, the inequality $\mathbb{E}_0 [|U \cap \mathcal{I}_{k^*}|] \leq \frac{1}{2}$ implies $\Pr_0 [\mathcal{E}_{k^*}] \geq \frac{1}{2}$. Next, we consider the environment V_{k^*}. From round 1 up to (and including) round $t^* = \frac{(k^*-2)T}{4f(T,A)}$, the set of active arms and their losses on V_0 and V_{k^*} are identical. As a result, the distribution over the past chosen arms and observed losses induced by $\mathcal{A}$ up to round t^* is the same on both environment. Moreover, once the algorithm enters $\mathcal{I}_{k^*}$ at round $t^* + 1$ and chooses only arm 1 subsequently, it also observes the same sequence of chosen arms and losses on both V_0 and V_{k^*} until the end of $\mathcal{I}_{k^*}$. Hence,

$$\Pr_{k^*} [\mathcal{E}_{k^*}] = \Pr_0 [\mathcal{E}_{k^*}] \geq \frac{1}{2}, \tag{4.65}$$

where the subscript k^* indicates a probability measured in environment V_{k^*}. In other words, on V_{k^*}, with probability at least 0.5, arm 1 is always chosen on I_{k^*} and arm k^* is never chosen. Under this event $\mathcal{E}_{k^*}$, the regret of $\mathcal{A}$ with respect to arm k^* is $(0.5-0)L = 0.5L$. When $\mathcal{E}_{k^*}$ does not hold, the regret with respect to k^* is non-negative because $\ell_{k^*,t} = 0$. Overall, on V_{k^*} the expected regret is at least

$$\mathbb{E}_{k^*}\left[R(k^*)\right] \geq \frac{L}{4}.$$

□

4.F Proof of Theorem 4.3.5: Doubling-Trick for Adapting to $\sum_{t=1}^{T} A_t$ AND G_T

For SB-EXP3, computing the optimal learning rates require the fraction $\sqrt{\frac{\ln G_T}{\sum_{t=1}^{T} A_t}}$ of $\sqrt{\ln G_T}$ and the sum $\sqrt{\sum_{t=1}^{T} A_t}$. Both of these quantities are monotonically non-decreasing. Therefore, we can apply the doubling trick on these two quantities. First, we prove a simple lemma justifying doing this.

Lemma 4.F.1. Let $a, b, c, d > 0$ be constants such that $a \leq c$ and $b \leq d$. Let

$$f(x) = \frac{a}{x} + \frac{bx}{2}$$

be a function on $\mathbb{R}_+$. Then,

$$f\left(\sqrt{\frac{2c}{d}}\right) \leq \sqrt{2cd}.$$

Proof. Due to $a \leq c$ and $b \leq d$, for any $x \geq 0$ we have

$$f(x) \leq \frac{c}{x} + \frac{dx}{2}.$$

Plugging $x = \sqrt{\frac{2c}{d}}$ into the right-hand side gives

$$f\left(\sqrt{\frac{2c}{d}}\right) \leq c\sqrt{\frac{d}{2c}} + d\sqrt{\frac{c}{2d}}$$
$$= \sqrt{2cd}.$$

□

Lemma 4.F.1 implies that for any horizon T, if we set $\eta_t = \sqrt{\frac{2c}{d}}$ for c and c such that $c \geq \ln G_T$ and $d \geq \sum_{t=1}^{T} A_t$ then we obtain a regret of bound $\sqrt{2cd}$.

We proceed to perform the doubling trick on $\ln G_T$ and $\sum_{t=1}^{T} A_t$. The full procedure is given in Algorithm 4.4. The main idea is a two-level doubling trick which divides the learning process into episodes as follows:

- The first level: throughout the learning process, we maintain a set $\mathbb{V}$ for the arms that have been active at least once in each episode and an upper bound 2^C for $\ln(|\mathbb{V}|)$. Initially, $\mathbb{V} = \emptyset$ and $C = 1$. At the beginning of round t, we check if $\ln(|\mathbb{V} \cup \mathbb{A}_t|)$ exceeds 2^C. If $\ln(|\mathbb{V} \cup \mathbb{A}_t|) \leq 2^C$ then we continue the learning process and update $\mathbb{V} = \mathbb{V} \cup \mathbb{A}_t$. Otherwise, we reset $\mathbb{V}$ to $\emptyset$, increment C by at least one until $2^C \geq \ln(A_t)$ and start a new episode from round t.

- The second level: throughout the rounds of each episode, we maintain a cumulative sum U for the sum of A_t and an upper bound 2^b for U. Note that C is fixed within these rounds. Before the first round of an episode, we initialize $U = 0, b = 1$. As long as $U + A_t \leq 2^b$, we run SB-EXP3 with $\eta = \sqrt{\frac{2^{C+1}}{2^b}}$ and update $U = U + A_t$. Once U exceeds 2^b at some round t, we increment b by at least one until $A_t \leq 2^b$, reset $U = 0$ and run a new instance of SB-EXP3 onwards.

The pseudo-regret of Algorithm 4.4 is shown in the following theorem.

Theorem 4.3.5. For any $T \geq 2$, Algorithm 4.4 (in Appendix 4.F) guarantees that

$$\max_{a \in [K]} \mathbb{E}[R(a)] \leq \frac{4}{(\sqrt{2}-1)^2}\sqrt{\ln(G_T)\sum_{t=1}^{T} A_t}.$$

Proof. Let C_T be the last value of C after T rounds. Note that C_T is also the number of episodes. Let $c = 1, 2, \ldots, C_T$ be the index of the episodes. We first bound the regret within each episode c, and then sum up this bound over C_T episodes to get the total regret bound.

Bounding The Regret Within Each Episode

Fix an episode c. Let $\mathbb{T}_c$ be the rounds in this episode, and $T_c = |\mathbb{T}_c|$. Let $\mathbb{V}_c$ be the set of arms that are active at least once during this episode, and $V_c = |\mathbb{V}_c|$. By construction, $b = 1$ at the beginning of this episode. Let B be the last value of b after T_c rounds starting from the first round of episode c. Divide the rounds in $\mathbb{T}_c$ into B time intervals, where b does not change in each interval. For $b = 1, 2, \ldots, B$, let F_b be the time interval of b. Let $U_b = \sum_{t \in F_b} A_t$ be the sum of A_t within F_b. Since $U_b \leq 2^b$ and $\ln(V_c) \leq 2^c$, by Lemma 4.F.1 and Theorem 4.3.1, the regret of the learner in this interval is bounded by

$$\max_{a \in [K]} \mathbb{E}[R_{F_b}(a)] \leq \sqrt{2^b 2^{c+1}}.$$

Let $R_c(a)$ be the regret incurred during episode c with respect to arm a. We have

$$\begin{aligned}
\max_{a \in [K]} \mathbb{E}[R_c(a)] &\leq \sum_{b=1}^{B} \max_{a \in [K]} \mathbb{E}[R_{F_b}(a)] \\
&\leq \sum_{b=1}^{B} \sqrt{2^b 2^{c+1}} \\
&\leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{2^B 2^{c+1}}.
\end{aligned}$$

For $b = 1, \ldots, B$, let t_b be the first round of F_b. Since b is increased from $B-1$ to B at the beginning of round t_B, we have $U_{B-1} + A_{t_B} > 2^{B-1}$. It follows that $2^B \leq 2(U_{B-1} + A_{t_B}) \leq 2 \sum_{t \in \mathbb{T}_c} A_t$. Hence,

$$\max_{a \in [K]} \mathbb{E}[R_c(a)] \leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{2^{c+1} \sum_{t \in \mathbb{T}_c} A_t}. \tag{4.66}$$

Bounding The Total Regret

Summing up (4.66) for $c = 1, \ldots, C_T$, we obtain

$$\begin{aligned}
\max_{a\in[K]} \mathbb{E}[R(a)] &\le \sum_{c=1}^{C_T} \max_{a\in[K]} \mathbb{E}[R_c(a)] \\
&\le \frac{\sqrt{2}}{\sqrt{2}-1} \sum_{c=1}^{C_T} \sqrt{2^{c+1} \sum_{t\in\mathbb{T}_c} A_t} \\
&\le \frac{\sqrt{2}}{\sqrt{2}-1} \left(\sqrt{\sum_{t=1}^{T} A_t} \right) \sum_{c=1}^{C_T} \sqrt{2^{c+1}} \\
&\le \frac{2\sqrt{2}}{(\sqrt{2}-1)^2} \left(\sqrt{\sum_{t=1}^{T} A_t} \right) \sqrt{2^{C_T}},
\end{aligned}$$

where the third inequality is due to $\sum_{t\in\mathbb{T}_c} A_t \le \sum_{t=1}^{T} A_t$ and the last inequality is due to $\sum_{c=1}^{C_T} \sqrt{2^c} = \frac{\sqrt{2}}{\sqrt{2}-1}(\sqrt{2^{C_T}} - 1)$.

Assume C was increased at least once i.e. $C_T > 1$, otherwise we immediately have an $O\left(\sqrt{\sum_{t=1}^{T} A_t}\right)$ total regret bound. Let τ be the first round of the last episode. Since C was increased from $C_T - 1$ to C_T at round τ, we have $2^{C_T-1} < \ln(|\mathbb{V}_{C_T-1} \cup \mathbb{A}_\tau|)$. On the other hand, $(\mathbb{V}_{C_T-1} \cup \mathbb{A}_\tau) \subseteq \mathbb{G}_T$ implies $\ln(|\mathbb{V}_{C_T-1} \cup \mathbb{A}_\tau|) \le \ln(G_T)$. This implies that $2^{C_T} \le 2\ln(G_T)$, hence the total regret is bounded by

$$\max_{a\in[K]} \mathbb{E}[R(a)] \le \frac{4}{(\sqrt{2}-1)^2} \sqrt{\ln(G_T) \sum_{t=1}^{T} A_t}.$$

□

Remark 4.F.2. While the two-level doubling trick works, for practical purposes we can set a small constant (e.g. 16) to be an upper bound for $\ln G_T$ and perform a one-level doubling trick only on $\sum_{t=1}^{T} A_t$. This is because $\ln G_T$ increases exponentially slowly: if the upper bound for $\ln G_T$ is doubled in each increment starting from $2^0 = 1$, then to have k such increments G_T must be as large as $G_T \ge e^{2^k}$. For $k = 5$, this is approximately 8×10^{13}, which is exceedingly large for the number of arms. Overall, this implies that the doubling level on $\ln G_T$ would change at most 4 times in any practical scenario, and setting $\ln(G_T) \le 16$ would contribute at most a multiplicative factor of $\sqrt{16} = 4$ on the total regret bound.

4.G FTARL with Negative Entropy is Equivalent to SB-EXP3

In Algorithm 4.2, the loss estimate of non-active arms $a \notin \mathbb{A}_t$ is $\tilde{\ell}_{a,t} = \hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}$. Because $I_{a,t} = 1$ for $a \in \mathbb{A}_t$ and $I_{a,t} = 0$ for $a \notin \mathbb{A}_t$, it follows that in Algorithm 4.2, for all $a \in [K]$,

$$\sum_{t=1}^{T} \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} - \tilde{\ell}_{a,t} \right) = \sum_{t=1}^{T} I_{a,t} \left(\hat{\ell}_t - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} - \tilde{\ell}_{a,t} \right) \tag{4.67}$$

Next, we show that in each round, the sampling probability p_t of FTARL with negative entropy is the same as that of SB-EXP3. With $\psi_t(x) = \frac{1}{\eta} \sum_{i=1}^{K} x_i \ln x_i$ the negative Shannon entropy, in Algorithm 4.2 the weight vector q_t of FTARL is the solution of the optimization problem

$$q_t = \min_{x \in \Delta_K} \frac{1}{\eta} \sum_{i=1}^{K} x_i \ln x_i + \sum_{i=1}^{K} x_i \tilde{L}_{i,t-1}.$$

Solving for q_t, we obtain for any $i \in [K]$,

$$q_{i,t} = \frac{\exp\left(-\eta \tilde{L}_{i,t-1}\right)}{\sum_{k=1}^{K} \exp\left(-\eta \tilde{L}_{k,t-1}\right)}.$$

It follows that for an active arm $i \in \mathbb{A}_t$,

$$\begin{aligned} p_{i,t} &= \frac{q_{i,t}}{\sum_{k \in \mathbb{A}_t} q_{k,t}} \\ &= \frac{\exp\left(-\eta \tilde{L}_{i,t-1}\right)}{\sum_{k \in \mathbb{A}_t} \exp\left(-\eta \tilde{L}_{k,t-1}\right)} \\ &= \frac{\exp\left(\eta \sum_{s=1}^{t-1} \hat{\ell}_s - \gamma \sum_{j \in \mathbb{A}_s} \tilde{\ell}_{j,s} - \tilde{\ell}_{i,s}\right)}{\sum_{k \in \mathbb{A}_t} \exp\left(\eta \sum_{s=1}^{t-1} \hat{\ell}_s - \gamma \sum_{j \in \mathbb{A}_s} \tilde{\ell}_{j,s} - \tilde{\ell}_{k,s}\right)} \\ &= \frac{\exp\left(\eta \sum_{s=1}^{t-1} I_{i,s} (\hat{\ell}_s - \gamma \sum_{j \in \mathbb{A}_s} \tilde{\ell}_{j,s} - \tilde{\ell}_{i,s})\right)}{\sum_{k \in \mathbb{A}_t} \exp\left(\eta \sum_{s=1}^{t-1} I_{k,s} (\hat{\ell}_s - \gamma \sum_{j \in \mathbb{A}_s} \tilde{\ell}_{j,s} - \tilde{\ell}_{k,s})\right)}, \end{aligned} \tag{4.68}$$

where

- the second-to-last equality is by multiplying $\exp\left(\eta \sum_{s=1}^{t-1} \hat{\ell}_s - \gamma \sum_{j \in \mathbb{A}_s} \tilde{\ell}_{j,s}\right)$ to both the denominator and numerator.
- the last equality is due to (4.67).

Observe that (4.68) is equal to the sampling probability of arm $i \in \mathbb{A}_t$ computed in (4.6) of Algorithm 4.3.1. Moreover, $p_{i,t} = 0$ for $i \notin \mathbb{A}_t$ in both Algorithms 4.1 and 4.2. This implies that FTARL with negative Shannon entropy is equivalent to SB-EXP3.

Algorithm 4.4 SB-EXP3-ATGT adapted to G_T and $\sum_{t=1}^{T} A_t$

Initialize $U = 0, C = 1, b = 1, \mathbb{V} = \emptyset$
Initialize $L_i = 0$ for $i = 1, 2, \ldots, K$
for each round $t = 1, \ldots,$ **do**
  An adversary selects and reveals $\mathbb{A}_t$
  if $\ln(|\mathbb{V} \cup \mathbb{A}_t|) > 2^C$ **then**
    $C = C + 1$
    while $\ln(A_t) > 2^C$ **do**
      $C = C + 1$
    end
    Set $\mathbb{V} = \emptyset$
    Set $U = 0, b = 1$
    for arm $i \in \mathbb{G}_t$ **do**
      $L_i = 0$
    end
  end
  if $U + A_t > 2^b$ **then**
    $b = b + 1$
    while $A_t > 2^b$ **do**
      $b = b + 1$
    end
    Set U $= 0$
    for arm $i \in \mathbb{G}_t$ **do**
      $L_i = 0$
    end
  end
  Update $\mathbb{V} = \mathbb{V} \cup \mathbb{A}_t$
  Update $U = U + A_t$
  Compute $\eta = \sqrt{\frac{2^{C+1}}{2^b}}$
  for arm $i \in \mathbb{A}_t$ **do**
    $\tilde{q}_{i,t} = \exp(\eta L_i)$
  end
  Compute p_t by (4.6)
  Sample $i_t \sim p_t$
  Compute $\tilde{\ell}_{i,t}$ by (4.4)
  for arm $i \in \mathbb{A}_t$ **do**
    $L_i = L_i + \hat{\ell}_t - \tilde{\ell}_{i,t} - \gamma \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}$
  end
end

Chapter 5

Beyond minimax rates in group distributionally robust optimization via a novel notion of sparsity

5.1 Introduction

Performing well across different data subpopulations and being robust to distribution-shift in testing are two of the most important goals in building machine learning models [Ben-Tal et al., 2013; Williamson and Menon, 2019; Sagawa et al., 2020]. These goals are especially important for models making decisions that could have societal and safety impacts. A recently proposed framework for achieving these goals is the group distributionally robust optimization (GDRO) framework, in which a learner aims to find a single hypothesis that minimizes the maximum risk over a finite number of data distributions. This minimax objective is often considered in the context of fairness [Rawls, 1971; Williamson and Menon, 2019; Abernethy et al., 2022] when the distributions represent different demographic groups, or as a means to promote robustness when they represent possible shifts in the data distribution [Mohri et al., 2019; Sagawa et al., 2020].

More formally, given an n-dimensional hypothesis set Θ and a group of K distributions $\mathcal{P}_i$, the learner aims to solve the optimization $\min_{\theta\in\Theta} \max_{i\in\{1,\dots,K\}} R_i(\theta)$, where $R_i(\theta)$ is the risk of the learner with respect to $\mathcal{P}_i$. Intuitively, this objective encourages the learner to find a model with good balance in performance with respect to a finite number of distributions of data, and avoid models that might perform extremely well on one distribution but have significantly worse performance on others. The GDRO framework assumes that the

learner has access to a sampling oracle, which returns an i.i.d sample from $\mathcal{P}_i$ upon receiving a request $i \in [K]$. The sample complexity of the learner is the number of samples needed to find an ϵ-optimal hypothesis $\bar{\theta}$ such that the optimality gap $\max_i R_i(\bar{\theta}) - \max_i R_i(\theta^*)$ is smaller than a target value ϵ, where θ^* is an optimal hypothesis.

Throughout the paper, the $\tilde{O}$ notation hides logarithmic factors. Existing works [Soma et al., 2022; Zhang et al., 2023a] have shown a sample complexity lower bound of order $\Omega(\frac{G^2D^2+K}{\epsilon^2})$ and a near-matching $\tilde{O}\left(\frac{G^2D^2+K}{\epsilon^2}\right)$ worst-case upper bound, where D is the ℓ_2 diameter of Θ and G is the Lipschitz constant of the loss function. While these existing results are useful for understanding worst-case scenarios, practical problems may have additional structure that allows for significantly lower sample complexity. In particular, the $\Omega(\frac{G^2D^2+K}{\epsilon^2})$ lower bound construction in [Soma et al., 2022] relies on having arbitrarily small gaps (i.e., difference in risks) between groups for all $\theta \in \Theta$. This property rarely holds in practice, where most hypotheses can have significant gaps between groups. For example, in car manufacturing, each car model often has noticeably different effects on different surfaces and road conditions.

5.1.1 Contributions and Techniques

We transcend the established minimax bounds by considering problem instances with additional structure. We formally define such a structure called (λ, β)-sparsity in Section 5.2.1. The main idea of (λ, β)-sparsity is that for all θ, the groups can be divided into two sets: one contains groups with large risks and the other contains groups with small risks. The parameter λ specifies the risk-difference between these two sets of groups, while β specifies the number of groups with large risks. Let β_λ denote the smallest β for which (λ, β)-sparsity holds. For problem with (λ, β)-sparsity, we show that the dependence on K in the leading term (here and throughout, the term for which $1/\epsilon$ is of the highest order) can be reduced from $O(K \ln K)$ to $O(\beta_\lambda \ln K)$. Table 5.1 summarizes our main results, which consist of three high-probability upper bounds and a lower bound. The leading terms in the upper bounds grow with $\tilde{O}\left(\frac{D^2G^2+\beta_\lambda}{\epsilon^2}\right)$ instead of $\tilde{O}\left(\frac{D^2G^2+K}{\epsilon^2}\right)$, where β_λ could be much smaller than K.[1] To the best of our knowledge, these are the first bounds that go beyond the established minimax bound in [Soma et al., 2022; Zhang et al., 2023a]. The near-matching lower bound is of order $\Omega\left(\frac{D^2G^2+\beta}{\epsilon^2}\right)$, generalizing the minimax lower bound in [Soma et al., 2022].

Technically, our results are based on improving the sample complexity of the two-player zero-sum game framework for GDRO [Nemirovski et al., 2009; Zhang et al., 2023a]. In this

[1] See Table 5.1 for full results.

Table 5.1: Summary of main results. λ-adapt indicates if the bound is adaptive to the best λ^* possible. n-free indicates whether the bound depends on the dimension of Θ. δ is the failure probability.

Upper and Lower Bounds	λ-adapt?	n-free?
$O\left(\frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{\lambda^2}+\frac{(G^2D^2+\beta_\lambda)\ln\left(\frac{K}{\delta}\right)}{\epsilon^2}\right)$(Thm 5.3.4)	×	×
$O\left(\left(\frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{(\lambda^*)^2}+\frac{(G^2D^2+\beta_{\lambda^*})\ln\left(\frac{K}{\delta}\right)}{\epsilon^2}\right)\ln\left(\frac{1}{\epsilon}\right)\right)$(Thm 5.4.1)	✓	×
$O\left(\frac{DKG\sqrt{(D^2G^2+\beta)\ln(K/\delta)}\ln\left(\frac{KDG}{\epsilon\lambda\delta}\right)}{\lambda^3\epsilon}+\frac{(D^2G^2+\beta)\ln(K/\delta)}{\epsilon^2}\right)$ (Thm 5.C.1)	×	✓
$O\left(\frac{(D^2G^2+\max(\ln(K),\beta_{\lambda^*}))\ln(K/\delta)}{\epsilon^2}\right)$ (Thm 5.4.2)	✓	✓
$\Omega\left(\frac{D^2G^2+\beta}{\epsilon^2}\right)$(Thm 5.3.5)	-	-

framework, a game is played repeatedly as follows: in round t, the first player (the min-player) plays a hypothesis $\theta_t \in \Theta$ and the second player (the max-player) plays a group index $i_t \in [K]$ and draws a sample from distribution $\mathcal{P}_{i_t}$. For the max-player, choosing one of K groups and getting an i.i.d sample from that group is similar to pulling one of K arms and getting feedback from that arm in a multi-armed bandit problem. While existing works [Soma et al., 2022; Zhang et al., 2023a] use a fixed set of K arms in every round for the max-player to choose from, the (λ, β)-sparsity condition allows us to use a smaller, time-varying subset of active arms of size at most β. To handle this time-varying action set, we use the sleeping bandits framework [Kleinberg et al., 2010] to model the learning process of the max-player. Critically, recent progress [Nguyen and Mehta, 2024] in bounding the per-action regret in sleeping bandits (details in Section 5.3.2) enables us to reduce the max player's regret bound and improve the dependency on the number of groups from $K \ln K$ to $\beta \ln K$ in the leading term of the sample complexity.

For the two dimension-dependent bounds, the computation of the time-varying subsets of arms for the max-player is based on a uniform convergence bound for Θ that uses $\tilde{O}\left(\frac{Kn}{\lambda^2}\right)$ samples. The first bound is obtained by an algorithm called `SB-GDRO` that takes λ as input and outputs an ϵ-optimal hypothesis using $\tilde{O}\left(\frac{Kn}{\lambda^2}+\frac{G^2D^2+\beta_\lambda}{\epsilon^2}\right)$ samples. Letting λ^* be the λ that minimizes the sample complexity bound of `SB-GDRO`, a natural question is whether it is possible to nearly obtain this minimum sample complexity *without* knowing λ^*. Surprisingly, in Section 5.4 we show that such adaptivity is possible. A disadvantage of the fully-adaptive approach is that it is computationally expensive due to the explicit computation of covers

of the potentially high-dimensional set Θ. In Section 5.4.2, we partially resolve this by proposing a computationally efficient semi-adaptive algorithm with a dimension-independent $\tilde{O}(\frac{D^2G^2+\max(\ln(K),\beta_{\lambda^*})}{\epsilon^2})$ bound in high-precision settings where $\epsilon \ll \lambda^*$.

In Section 5.5, we present experimental results showing that not only this (λ, β)-sparsity condition holds for high-dimensional practical setting *around* the optimal hypothesis θ^*, but also our algorithms can efficiently (in both sample and computational complexity) compute an estimate of λ^* and leverage it to find ϵ-optimal hypotheses with significantly fewer samples compared to baseline methods.

5.1.2 Related Works

We consider the GDRO problem where the loss function is real-valued in $[0, 1]$ and the hypothesis space Θ is convex and compact. In this setting, [Nemirovski et al., 2009][2] converts GDRO to a stochastic saddle point problem and uses stochastic mirror descent methods with $O(\frac{K(G^2D^2+\ln(K))}{\epsilon^2})$ sample complexity guarantee. [Sagawa et al., 2020] adopts the two-player convex-concave game framework from the deterministic min-max optimization literature [Freund and Schapire, 1999; Cesa-Bianchi and Lugosi, 2006] to obtain $O\left(\frac{K^2(G^2D^2+\ln(K))}{\epsilon^2}\right)$ sample complexity bound, which was improved to $\tilde{O}(\frac{D^2G^2+K\ln(K)}{\epsilon^2})$ by [Zhang et al., 2023a] by refining the approach. An $\Omega\left(\frac{G^2D^2+K}{\epsilon^2}\right)$ information-theoretic lower bound was shown in [Soma et al., 2022].

A related, more constrained setting is the class of multi-distribution binary classification problems in which Θ has finite VC-dimension d [Haghtalab et al., 2022; Awasthi et al., 2023]. Recent works have established tight minimax sample complexity bounds of order $\frac{d+K}{\epsilon^2}$ for this setting [Zhang et al., 2023b; Peng, 2023]. Multi-distribution learning with multi-label prediction with offline data was recently explored in [Jang et al., 2024]. We refer interested readers to [Haghtalab et al., 2022; Zhang et al., 2023b] for a more comprehensive discussion of related works in min-max fairness and federated learning settings.

5.2 Problem Setup

Let $\Theta \subset \mathbb{R}^n$ be a compact convex set of hypotheses, $\mathcal{Z}$ be a sample space and $\ell : \Theta \times \mathcal{Z} \to [0, 1]$ be a loss function measuring the performance of a hypothesis on a data point. Similar to

[2] [Nemirovski et al., 2009] do not impose the bounded loss assumption, although [Zhang et al., 2023a] do adopt this assumption.

previous works in GDRO [Sagawa et al., 2020; Haghtalab et al., 2022], we use the following assumption.

Assumption 5.2.1. The diameter of Θ is bounded as $\|\theta\|_2 \leq D$ for all $\theta \in \Theta$. The loss function ℓ is convex and G-Lipschitz in the first argument, i.e., $|\ell(\theta, \cdot) - \ell(\theta', \cdot)| \leq G\|\theta - \theta'\|_2$ for all $\theta, \theta' \in \Theta$.

There are K groups, each associated with a distribution $(\mathcal{P}_i)_{i,i=1,\ldots,K}$ over $\mathcal{Z}$. Let $[K] = \{1, 2, \ldots, K\}$. Let $R_i(\theta) = \mathbb{E}_{z\sim\mathcal{P}_i}[\ell(\theta, z)]$ be the risk of θ with respect to group i. The worst-case risk of a hypothesis θ is measured by its maximum risk over these distributions:

$$\mathcal{L}(\theta) = \max_{i\in[K]} R_i(\theta).$$

The objective is find a hypothesis θ^* with minimum $\mathcal{L}(\theta)$:

$$\theta^* = \arg\min_{\theta\in\Theta} \mathcal{L}(\theta) = \arg\min_{\theta\in\Theta} \max_{i\in[K]} R_i(\theta). \tag{5.1}$$

The optimality gap of $\bar{\theta} \in \Theta$ is $\text{err}(\bar{\theta}) = \mathcal{L}(\bar{\theta}) - \mathcal{L}(\theta^*)$. Similar to previous works, we assume that the learner has access to a sampling oracle that, for every query $i \in [K]$, returns an i.i.d sample $z \sim \mathcal{P}_i$. Given a target optimality ϵ, the sample complexity of a learner is the number of samples to find an ϵ-optimal hypothesis $\bar{\theta}$ such that $\text{err}(\bar{\theta}) \leq \epsilon$.

5.2.1 (λ, β)-Sparsity Structure

First, we formally define the notion of a λ-dominant set.

Definition 5.2.1. For any $\lambda \in [0, 1]$ and $\theta \in \Theta$, a non-empty set of groups $S \subseteq [K]$ is λ-dominant at θ if for all $j \notin S$,

$$\min_{i\in S} R_i(\theta) \geq R_j(\theta) + \lambda. \tag{5.2}$$

Note that $S = [K]$ is a dominant set, as there is no j in the empty set $[K] \setminus S$ such that $R_j(\theta) + \lambda > \min_{i\in[K]} R_i(\theta)$. Next, we introduce (λ, β)-sparsity, our novel condition for GDRO problems.

Definition 5.2.2. For $\lambda \geq 0$ and $\beta \in [1, K]$, a GDRO problem is (λ, β)-sparse if for all $\theta \in \Theta$, there exists a λ-dominant set whose size is at most β. If $\lambda > 0$ and $\beta < K$, we say that (λ, β) is nontrivial.

Algorithm 5.1 SB-GDRO with a known λ

Input: Constants $K, D, G, \lambda, \epsilon$, hypothesis set $\Theta \subset \mathbb{R}^n$
Draw m (defined in Lemma 5.3.1) samples from each group into set V
Initialize $\theta_1 = \arg\min_{\theta \in \Theta} \|\theta\|_2$
for each round $t = 1, \ldots, T$ **do**
  $\hat{S}_{\theta_t} = \texttt{DominantSet}(\theta_t, V, 0.7\lambda)$ // 0.4λ-dominant set at θ_t
  $q_t = \texttt{MaxP}(t, \hat{S}_{\theta_t})$ // Action of max-player
  Draw $i_t \sim q_t$ and $z_{i_t,t} \sim \mathcal{P}_{i_t}$
  $\theta_{t+1} = \texttt{MinP}(\theta_t, z_{i_t,t})$ // Action of min-player
Return: $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$

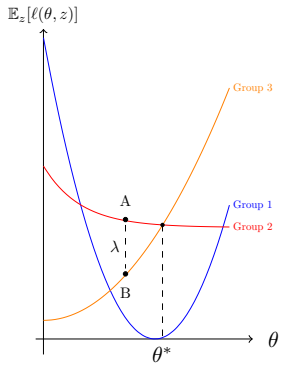

Figure 5.1: (Left) SB-GDRO with known λ. (Right) A (λ, β)-sparse example with $K = 3, \beta = 2$.

By definition, a GDRO instance can be (λ, β)-sparse for multiple (λ, β). For example, a $(0.2, 10)$-sparse problem with $K = 20$ is also $(0.2, 11)$ and $(0.1, 10)$-sparse. Similarly, there can be multiple λ-dominant sets at each θ. Let $\mathbb{S}_{\lambda,\theta}$ be the collection of all λ-dominant sets at θ. Since $[K]$ is always a λ-dominant set, this collection always contains $[K]$. Let $\beta_{\lambda,\theta} = \min_{S \in \mathbb{S}_{\lambda,\theta}} |S|$ be the size of the smallest λ-dominant set at $\theta \in \Theta$. Then, we have $\beta_\lambda = \max_{\theta \in \Theta} \beta_{\lambda,\theta}$ is the smallest value of β such that (λ, β)-sparsity holds. Moreover, all GDRO instances are trivially $(0, 1)$-sparse, in which case the 0-dominant set contains one of the groups with maximum expected loss. If (λ, β)-sparsity holds for nontrivial (λ, β), then for every model, there is a prominent gap in the outcome (i.e., risks) of applying that model across different groups. Figure 5.1 (Right) illustrates the mathematical plausibility of nontrivial (λ, β)-sparsity in the continuous domain via a simple example with $\Theta = [0, 1]$.

In Section 5.3, we begin by presenting an algorithm which, for any input $\lambda \in (0, 1]$, returns an ϵ-optimal hypothesis with sample complexity $\tilde{O}\left(\frac{Kn}{\lambda^2} + \frac{D^2G^2+\beta_\lambda}{\epsilon^2}\right)$. For *any* such λ, including trivial choices for which $\beta_\lambda = K$, this algorithm (with high probability) provides a valid sample complexity guarantee, but the guarantee is most useful for the unknown, optimal λ — call it λ^* — that minimizes the sample complexity. The focus of Section 5.4 is adaptive algorithms that obtain, without any knowledge of λ^*, sample complexity whose order is only larger than that of our previous algorithm (were it given λ^*) by a logarithmic factor.

5.3 Two-Player Zero-Sum Game Approach

In this section, we present a new algorithm SB-GDRO that, for a given input $\lambda \in (0, 1]$, obtains an $O\left(\frac{Kn\ln(GDK/\delta)}{\lambda^2} + \frac{(G^2D^2+\beta_\lambda)\ln(K/\delta)}{\epsilon^2}\right)$ sample complexity. Let Δ_K be the K-dimensional probability simplex. For any $q \in \Delta_K$, let $\phi(\theta, q) = \sum_{i=1}^K q_i R_i(\theta)$ be the weighted sum of the risks of θ over K groups. Following [Nemirovski et al., 2009], we write the objective function in (5.1) as

$$\min_{\theta\in\Theta} \mathcal{L}(\theta) = \min_{\theta\in\Theta} \max_{q\in\Delta_K} \phi(\theta, q).$$

The duality gap of $\bar{\theta} \in \Theta$ and $\bar{q} \in \Delta_K$ is defined as

$$\mathrm{err}(\bar{\theta}, \bar{q}) = \max_{q\in\Delta_K} \phi(\bar{\theta}, q) - \min_{\theta\in\Theta} \phi(\theta, \bar{q}).$$

Since $\mathcal{L}(\theta) \geq \phi(\theta, \bar{q})$ for all θ, we have $\mathrm{err}(\bar{\theta}) \leq \mathrm{err}(\bar{\theta}, \bar{q})$. To minimize $\mathrm{err}(\bar{\theta}, \bar{q})$, similar to existing works [Nemirovski et al., 2009; Soma et al., 2022], we employ the following two-player zero-sum game approach: a game is run in T rounds, where in each round, there are two players $\mathcal{A}_\theta$ and $\mathcal{A}_q$ corresponding to the min and max operators in the objective function (5.1). In round t, the min-player $\mathcal{A}_\theta$ first plays a hypothesis θ_t, and then the max-player $\mathcal{A}_q$ plays a vector $q_t \in \Delta_K$. Then, a random group $i_t \sim q_t$ is drawn, and the sampling oracle returns a sample $z_{i_t,t} \sim \mathcal{P}_{i_t}$. The two players compute θ_{t+1} and q_{t+1} for the next round based on i_t and $z_{i_t,t}$. The min-player's goal is to minimize its regret with respect to the best hypothesis in hindsight:

$$R_{\mathcal{A}_\theta} = \sum_{t=1}^T \phi(\theta_t, q_t) - \min_{\theta\in\Theta} \sum_{t=1}^T \phi(\theta, q_t). \tag{5.3}$$

The max-player's goal is to minimize its regret with respect to the best weight vector in hindsight:

$$R_{\mathcal{A}_q} = \max_{q\in\Delta_K} \sum_{t=1}^T \phi(\theta_t, q) - \sum_{t=1}^T \phi(\theta_t, q_t). \tag{5.4}$$

The SB-GDRO algorithm is illustrated in Algorithm 5.1. Before the game starts, SB-GDRO draws a set V_i of m samples from each group $i \in [K]$, where m is defined in Lemma 5.3.1. Let $V = \{V_1, \ldots, V_K\}$ be the collection of these sets. The strategies of the two players are as follows:

Algorithm 5.2 `MaxP`: the sleeping bandits max-player $\mathbb{A}_q$

Input: Time step $t > 0$, a dominant set $\hat{S}_{\theta_t}$
if $t = 1$ **then** Initialize $\tilde{q}_{i,t} = 1$ for $i \in [K]$
else
Let $h_{i,s} = 1 - \ell(\theta_s, z_{i,s})$ for $s = 1, 2, \ldots, t-1$
Compute $\tilde{q}_{i,t}$ by Equation (5.7)
Return: q_t where $q_{i,t} = \frac{\mathbb{1}\{i \in \hat{S}_{\theta_t}\}\tilde{q}_{i,t}}{\sum_{j=1}^{K} \mathbb{1}\{j \in \hat{S}_{\theta_t}\}\tilde{q}_{j,t}}$

- The min-player $\mathcal{A}_\theta$ follows the stochastic mirror descent framework similar to [Zhang et al., 2023a]. Specifically, given a sample $z_{i_t,t} \sim \mathcal{P}_{i_t}$ and an existing θ_t, $\mathcal{A}_\theta$ computes θ_{t+1} by

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \eta_{w,t} \langle \tilde{g}_t, \theta - \theta_t \rangle + \frac{1}{2} \|\theta - \theta_t\|_2^2 \right\} \tag{5.5}$$

  where $\eta_{w,t} = \frac{D}{G\sqrt{t}}$ is a time-varying learning rate and $\tilde{g}_t = \nabla \ell(\theta_t, z_{i_t,t})$ is a stochastic gradient of $R_{i_t}(\theta_t)$. Note that $\theta_1 = \arg\min_{\theta \in \Theta} \|\theta\|_2$. We refer to the strategy of the min-player as `MinP`, whose formal procedure is given in Algorithm 5.4 in Appendix 5.A.

- The max-player $\mathcal{A}_q$ uses θ_t and V to compute a set of "active" groups $\hat{S}_{\theta_t}$. A group i is *active* if the empirical risk of θ with respect to V_i is sufficiently large. Then, a sleeping bandits algorithm called SB-EXP3 is used to compute a group-sampling probability vector $q_t \in \Delta_K$, where $q_{i,t} > 0$ for $i \in \hat{S}_{\theta_t}$ and $q_{i,t} = 0$ for $i \notin \hat{S}_{\theta_t}$. We refer to the strategy of the max-player as `MaxP`, whose details are given in Algorithm 5.2.

Compared to existing works [Soma et al., 2022; Zhang et al., 2023a; Haghtalab et al., 2022], our two-player zero-sum game procedure has two additional steps: the construction of the collection V and the computation of the set $\hat{S}_{\theta_t}$. At the end of round T, the hypothesis $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$ is returned. As shown in [Soma et al., 2022], $\mathrm{err}(\bar{\theta}, \bar{q})$ is bounded by $\frac{1}{T}(R_{\mathcal{A}_\theta} + R_{\mathcal{A}_q})$. The min-player $\mathcal{A}_\theta$ uses a variant of the stochastic online mirror descent algorithm in [Zhang et al., 2023a] that uses time-varying learning rates instead of fixed learning rates, and obtains the same high-probability $O(DG\sqrt{T \ln(1/\delta)})$ regret bound. Our focus is to obtain an improved bound for the max-player $\mathcal{A}_q$ with a modified strategy.

Next, in Section 5.3.1, we compute the size of V needed to construct the λ-dominant sets in each round. Section 5.3.2 presents the strategy of using V to improve the regret of $\mathcal{A}_q$.

Algorithm 5.3 `DominantSet`: compute a dominant set $\hat{S}_{\theta_t}$

Input: $\theta_t \in \Theta$, collection of samples V, threshold $\tau > 0$
Compute $\hat{R}_i(\theta_t) = \frac{1}{m}\sum_{j=1}^{m} \ell(\theta_t, V_{i,j})$ for $i \in [K]$
Sort $\hat{R}_i(\theta_t)$ in *decreasing* order and let $\mathrm{ord}(i)$ be the sorted order of group i
Compute $\mathrm{nxt}(i)$ by $\mathrm{ord}(\mathrm{nxt}(i)) = \mathrm{ord}(i) + 1$
Let $\hat{i}$ be the first group in ord such that $\hat{R}_i(\theta_t) \geq \hat{R}_{\mathrm{nxt}(i)}(\theta_t) + \tau$, or -1 if no such groups exist.
Return: $\hat{S}_{\theta_t} = \{i \in [K] : \mathrm{ord}(i) \leq \mathrm{ord}(\hat{i})\}$ if $\hat{i} \neq -1$, otherwise $\hat{S}_{\theta_t} = [K]$.

5.3.1 Computing the Dominant Sets

Before the game starts, a set of m samples is drawn from each of K groups. Let $V_{i,j} \in V_i$ be the j-th sample collected from group i. Let $\hat{R}_i(\theta) = \frac{1}{m}\sum_{j=1}^{m} \ell(\theta, V_{i,j})$ be the empirical risk of θ with respect to V_i. To compute a 0.4λ-dominant set at θ_t, we use the algorithm `DominantSet` (Algorithm 5.3) which traverses the groups in order of decreasing $\hat{R}_i(\theta_t)$ and returns a set $\hat{S}_{\theta_t}$ of groups up to (and including) the first group whose empirical risk exceeds the next group's empirical risk by at least $\tau = 0.7\lambda$. The following lemma shows that if m is sufficiently large, then the set $\hat{S}_{\theta_t}$ returned by Algorithm 5.3 is a 0.4λ-dominant set at θ_t whose size does not exceed β_λ. This implies that the max-player only needs to sample the groups in $\hat{S}_{\theta_t}$ in order to maximize the cumulative risks over T rounds.

Lemma 5.3.1. Let $m = \frac{384n\ln\left(\frac{741GDK}{\delta}\right)}{0.01\lambda^2}$. With probability at least $1 - \delta/2$, for any $t \in [T]$, `DominantSet` returns a 0.4λ-dominant set $\hat{S}_{\theta_t}$ at θ_t satisfying $\left|\hat{S}_{\theta_t}\right| \leq \beta_\lambda$.

5.3.2 Non-Oblivious Sleeping Bandits

In this section, we discuss the sleeping bandits problem [Kleinberg et al., 2010]. Sleeping bandits is a variant of the adversarial multi-armed bandit problem with K arms, where arms can be non-active in each round. Formally, in round $t = 1, 2, \ldots, T$, an adaptive adversary gives the learner a set $\mathbb{A}_t \subseteq [K]$ of active arms. For each arm $i \in \mathbb{A}_t$, the adversary also selects a (hidden) loss value $h_{i,t} \in [0, 1]$. The learner pulls one active arm $i_t \in \mathbb{A}_t$, observes and incurs the loss $h_{i_t,t}$. Let $I_{i,t} = \mathbb{1}\{i \in \mathbb{A}_t\}$. For any $a \in [K]$, the per-action regret of the learner with respect to arm a is the difference in the cumulative loss of the learner and that of arm a over the rounds in which a is active:

$$\mathrm{Regret}(a) = \sum_{t=1}^{T} I_{a,t}(h_{i_t,t} - h_{a,t}). \tag{5.6}$$

Modified EXP3-IX for sleeping bandits. We use an algorithm called SB-EXP3 [Nguyen and Mehta, 2024] for sleeping bandits. SB-EXP3 uses the standard IX-loss estimate [Neu, 2015] as the loss estimate $\tilde{h}_{i,t}$ in round t, i.e., $\tilde{h}_{i,t} = \frac{h_{i,t}\mathbb{1}\{i_t=i\}}{q_{i,t}+\gamma_t}$, where $\gamma_t > 0$ is the exploration factor in round t. For each arm i, over T rounds SB-EXP3 maintains a weight vector $\tilde{q}_t \in \mathbb{R}_+^K$ defined as

$$\tilde{q}_{i,t} = \exp\left(\eta_{q,t} \sum_{s=1}^{t-1} I_{i,s}(h_{i_s,s} - \tilde{h}_{i,s} - \gamma_s \sum_{j\in\hat{S}_{\theta_s}} \tilde{h}_{j,s})\right), \tag{5.7}$$

where $\eta_{q,s} > 0$ is the learning rate and $\tilde{h}_{i,s}$ is the loss estimate of arm i in round s. Initially $\tilde{q}_{i,1} = 1$ for $i \in [K]$. The sampling probability q_t is computed by a filtering step, where inactive arms have $q_{i,t} = 0$ and the weights of active arms are normalized as $q_{i,t} = \frac{I_{i,t}\tilde{q}_{i,t}}{\sum_{j=1}^K I_{j,t}\tilde{q}_{j,t}}$. The following theorem bounds the per-action regret of SB-EXP3.

Theorem 5.3.2. With $\eta_{q,t} = 2\gamma_t = \sqrt{\frac{\ln(3K/\delta)}{\sum_{s=1}^t |\mathbb{A}_s|}}$, SB-EXP3 guarantees that with probability $1-\delta$,

$$\max_{a\in[K]} \text{Regret}(a) \leq O\left(\sqrt{\ln(K/\delta)\sum_{t=1}^T |\mathbb{A}_t|}\right).$$

Our Theorem 5.3.2 is a relatively straightforward but important extension of Nguyen and Mehta [2024, Theorem 3]. While the latter requires knowing $\max(|\mathbb{A}_t|)_t$ for tuning $\eta_{q,t}$ and γ_t, we obtain the same bound using adaptive learning rates without knowing anything about future active sets.

5.3.3 Sample Complexity of `SB-GDRO`

In `SB-GDRO`, the max-player uses SB-EXP3 to compute the group-sampling probability q_t. For the max-player, the set $\hat{S}_{\theta_t}$ in Algorithm 5.2 is similar to the set $\mathbb{A}_t$ in sleeping bandits as the set of "active groups" in round t depends on θ_t, which is decided by a non-oblivious adversary (i.e., the min-player). Furthermore, choosing a group $i_t \sim q_t$ and then drawing $z_{i_t,t} \sim \mathcal{P}_{i_t}$ is mathematically equivalent to having K samples $\{z_{i,t} \sim \mathcal{P}_i \mid i \in [K]\}$ (one from each group) but observing only $z_{i_t,t}$. The hidden stochastic loss of group i in round t is $\ell(\theta_t, z_{i,t})$. Note that SB-EXP3 is formulated in terms of minimizing losses rather than maximizing gains, so similar to [Zhang et al., 2023a], we set $h_{i,t} = 1 - \ell(\theta_t, z_{i,t})$ to be the

(hidden) stochastic losses of arms i for SB-EXP3. A fundamental connection between the two-player zero-sum game approach in GDRO and sleeping bandits is shown in the following lemma, which states that the regret of the max-player $R_{\mathcal{A}_q}$ is bounded by the per-action regret with $\hat{S}_{\theta_t}$ being the set of active groups at round t.

Lemma 5.3.3. With probability at least $1 - \delta/2$, the regret of the max-player is bounded by

$$R_{\mathcal{A}_q} \leq \max_{i \in [K]} \sum_{t=1}^{T} \mathbb{1}\{i \in \hat{S}_{\theta_t}\} \left(R_i(\theta_t) - \phi(\theta_t, q_t)\right).$$

Theorem 5.3.2 and Lemma 5.3.3 imply the following sample complexity bound for `SB-GDRO`.

Theorem 5.3.4. For any $\epsilon > 0, \delta \in (0,1)$, with probability $1 - \delta$, Algorithm 5.1 has sample complexity

$$O\left(\frac{Kn \ln(GDK/\delta)}{\lambda^2} + \frac{(D^2G^2 + \beta_\lambda)\ln(K/\delta)}{\epsilon^2}\right). \tag{5.8}$$

In Theorem 5.3.4, because λ is a fixed problem-dependent quantity while the required optimality gap ϵ can be arbitrarily small, the dependency on K in Theorem 5.3.4 is dominated by $O\left(\frac{\beta_\lambda \ln(K/\delta)}{\epsilon^2}\right)$. The following lower bound shows that the upper bound in Theorem 5.3.4 is essentially near-optimal.

Theorem 5.3.5. For any algorithm $\mathcal{A}$ and any $\lambda \geq 0.5, \beta \geq 3$, there exists a (λ, β)-sparse GDRO instance with $\beta_\lambda = \beta$ so that the sample complexity of $\mathcal{A}$ is at least $\Omega\left(\frac{G^2D^2+\beta}{\epsilon^2}\right)$.

5.4 λ^*-Adaptive Sample Complexity

Theorem 5.3.4 suggests that a desirable λ must be significantly larger than ϵ (so that $\frac{K}{\lambda^2} \ll \frac{K}{\epsilon^2}$) but also small enough that $\beta_\lambda \ll K$. In this section, we define the notion of an optimal λ^* and present a sample-efficient approach for adapting to this unknown λ^*. First, we write the sample complexity in Theorem 5.3.4 in the form

$$\ln(K/\delta)\left(\frac{C}{\lambda^2} + \frac{\beta_\lambda}{\epsilon^2}\right) + \frac{D^2G^2\ln(K/\delta)}{\epsilon^2}, \tag{5.9}$$

where $C = \frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{\ln(K/\delta)}$. By definition, $\beta'_\lambda \le \beta_\lambda$ for any $\lambda' \le \lambda$, and thus $\lambda \mapsto \beta_\lambda$ is non-decreasing. Let λ^* be the λ that minimizes (5.9). Our goal is to develop a sample-efficient method to find λ^*.

To describe our approach for finding λ^*, it will be useful to frame the idea of an optimal λ more generically. Consider any $C > K \ge 1$ (not necessarily taking the value above), $\epsilon \in (0,1)$, and δ in $(0,1)$. Let $g\colon [0,1] \to [1,K]$ be a nondecreasing function which is unknown. Now, let $\lambda^*_{C,g}$ be the minimizer, among all $\lambda \in [0,1]$, of

$$\mathrm{Cost}^{(\mathrm{GDRO})}_{C,g}(\lambda) := \frac{C}{\lambda^2} + \frac{g(\lambda)}{\epsilon^2}. \tag{5.10}$$

Clearly, if $C = \frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{\ln(K/\delta)}$ and $g(\lambda) = \beta_\lambda$, then $\lambda^*_{C,g} = \lambda^*$. In general, g (e.g., $\lambda \mapsto \beta_\lambda$) is unknown. However, g can be evaluated at any $\lambda \in (0,1]$ at a cost of $\mathrm{Cost}^{(\mathrm{Query})}_{C,g}(\lambda) = O(C\ln(K/\delta)/\lambda^2)$ samples. The problem $\mathrm{OPT}(C,g)$ is to find $\lambda^*_{C,g}$ using as few samples as possible.

Now, at a high level (our actual approach in Section 5.4.1 slightly differs), by solving $\mathrm{OPT}(C,g)$ for C as above and $g(\lambda) = \beta_\lambda$, we obtain a fully adaptive algorithm for GDRO that adapts to λ^* and has total (including the cost of finding λ^*) sample complexity whose rate (in big-O) is equal to the product of $\ln(1/\epsilon)$ and (5.9) with λ replaced by λ^*; here, $\ln(1/\epsilon)$ is the price paid for adaptivity. We present this algorithm in Section 5.4.1. However, this algorithm is computationally intractable for large n, and so Section 5.4.2 introduces a computationally efficient semi-adaptive algorithm with total sample complexity that, in high-precision settings where $\epsilon \ll \lambda^*$, swaps the β_{λ^*} in the fully adaptive algorithm's sample complexity with $\max\{\ln K, \beta_{\lambda^*}\}$; moreover it entirely avoids the dimension-dependent term $\frac{C}{(\lambda^*)^2}$, making it dimension-free.

5.4.1 λ^*-Adaptive Sample Complexity for GDRO

We present an algorithm called `SB-GDRO-A`, shown in full in Algorithm 5.8 in Appendix 5.B.2. The idea of this algorithm is to *(i)* construct a non-decreasing function $\hat{g}$ so that $\mathrm{Cost}^{(\mathrm{GDRO})}_{C,\hat{g}}(\lambda^*_{C,\hat{g}})$ is sufficiently close to $\mathrm{Cost}^{(\mathrm{GDRO})}_{C,\beta(\cdot)}(\lambda^*_{C,\beta})$ with high probability; *(ii)* solve $\mathrm{OPT}(C,\hat{g})$ to get $\lambda^*_{C,\hat{g}}$; *(iii)* input $\lambda^*_{C,\hat{g}}$ into `SB-GDRO`. Our approach uses at most $O(\mathrm{Cost}^{(\mathrm{GDRO})}_{C,\beta(\cdot)}(\lambda^*_{C,\beta})\ln\left(\frac{1}{\epsilon}\right))$ samples for steps *(i)* and *(ii)*, which, together with Theorem 5.3.4, gives us the following theorem (proved in Appendix 5.B.2).

Theorem 5.4.1. For any $\epsilon > 0, \delta \in (0,1)$, with probability at least $1-\delta$, `SB-GDRO-A` (Algorithm 5.8) with $\eta_{w,t}, \eta_{q,t}$ and γ_t defined in Theorem 5.3.4 has sample complexity

$$O\left(\left(\frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{(\lambda^*)^2} + \frac{(D^2G^2 + \beta_{\lambda^*})\ln\left(\frac{K}{\delta}\right)}{\epsilon^2}\right)\ln\left(\frac{1}{\epsilon}\right)\right).$$

Compared to Theorem 5.3.4, the sample complexity bound in Theorem 5.4.1 contains an additional multiplicative factor of $O(\ln(1/\epsilon))$, which we consider a small price for not knowing λ^* beforehand. Next, we briefly describe the two main steps above, with the full details in Appendix 5.B.

First Step: Constructing $\hat{g}$ We first describe a method that, given $\lambda \in [0,1]$, returns an estimate $\hat{\beta}_\lambda$ for β_λ using at most $O(\frac{C\ln(K/\delta)}{\lambda^2})$ samples. This method constructs a $\frac{0.1\lambda}{G}$-cover for Θ, uses Algorithm 5.3 to compute a 0.4λ-dominant set at each element of the cover, and then returns as its estimate $\hat{\beta}_\lambda$ the maximum cardinality among these dominant sets. Now, the function $\hat{g}$ is defined by setting $\hat{g}(\lambda)$ equal to 1 for $\lambda \le \frac{\epsilon}{2}$, setting it to $\hat{\beta}_\lambda$ for λ in the geometric sequence $(1, \frac{1}{5}, \frac{1}{5^2}, \ldots)$, and then interpolating at other λ to form a non-decreasing step function. In Appendix 5.B.2, we prove that with high probability, $\beta_{0.2\lambda} \le \hat{\beta}_\lambda \le \beta_\lambda$ and $\hat{g}$ is non-decreasing, leading to $\lambda^*_{C,\hat{g}}$ being close to $\lambda^*_{C,\beta}$.

Second Step: Solving for $\lambda^*_{C,\hat{g}}$ Our method for solving $\mathrm{OPT}(C,g)$ is called `SolveOpt`. It outputs $\hat{\lambda}$ such that $\mathrm{Cost}^{(\mathrm{GDRO})}(\hat{\lambda}) = O(\mathrm{Cost}^{(\mathrm{GDRO})}(\lambda^*))$ while using $O(\mathrm{Cost}^{(\mathrm{GDRO})}(\lambda^*)\ln(1/\epsilon))$ samples; note that we drop the subscripts C and g. The main idea of `SolveOpt` is to maintain two variables U and L which specify an interval $[L,U]$ that always contains a good estimate of λ^*. We iteratively evaluate $g(\lambda)$ for $\lambda \in [L,U]$ and shrink this interval, i.e., U monotonically decreases while L monotonically increases. The shrinking process is based on comparing $\mathrm{Cost}^{(\mathrm{GDRO})}(\lambda)$ and $\mathrm{Cost}^{(\mathrm{GDRO})}(U)$: if $\mathrm{Cost}^{(\mathrm{GDRO})}(\lambda) < \mathrm{Cost}^{(\mathrm{GDRO})}(U)$, then U is set to λ and L is increased accordingly. The process stops when $\lambda < L$, at which point the algorithm return the last value of U as its estimate of λ^*. The value of λ is taken from a geometric sequence; this ensures that at most $\ln(1/\epsilon)$ values of $g(\lambda)$ are evaluated, leading to the $\ln(1/\epsilon)$ multiplicative factor in the final bound.

5.4.2 A Semi-Adaptive Bound in High-Precision Settings

While `SB-GDRO-A` is fully adaptive to λ^*, it relies on building covers for Θ, which is computationally intensive when n is large. We now propose a semi-adaptive, computationally efficient algorithm called `SB-GDRO-SA` that avoids covers. The main idea is to merge the λ^*-estimation process into the two-player zero-sum game: starting with $\lambda = 1$, if the dominant sets S_{λ,θ_t} computed in round t of the game is bigger than a threshold (e.g. $\ln(K)$), then similar to `SolveOpt`, we decrease λ exponentially (e.g. $\lambda \leftarrow \lambda/2$). To avoid a too small λ, we also set a lower threshold L so that λ stops decreasing once $\lambda \leq L$. These two thresholds, one for $|S_{\lambda,\theta_t}|$ and one for λ, determine the trade-off between adaptivity and sample complexity. In `SB-GDRO-SA`, we use $\ln(K)$ and $L = \tilde{O}(\epsilon\sqrt{Kn})$ as the two thresholds. Let $\lambda_{\ln(K)}$ be the largest λ such that $\beta_\lambda = \ln(K)$. In high-precision settings where $\epsilon \ll \lambda^*$, the following theorem states that Algorithm 5.9 is adaptive to $\max(\lambda_{\ln(K)}, \lambda^*)$.

Theorem 5.4.2. If $\epsilon\sqrt{\frac{C}{\ln(K)}} < \lambda^*$, then with probability at least $1-\delta$, `SB-GDRO-SA` (Algorithm 5.9 in Appendix 5.B.3) has sample complexity

$$O\left(\frac{(D^2G^2 + \max(\ln(K), \beta_{\lambda^*}))}{\epsilon^2} \ln(K/\delta) \ln\frac{1}{\epsilon}\right)$$

We emphasize that Theorem 5.4.2 holds without knowing λ^*. This bound guarantees that in high-precision settings, Algorithm 5.9 enjoys (on average) dominant sets of small sizes that never exceed $\max(\beta_{\lambda^*}, \ln(K))$. Remarkably, this bound is also dominantly dimension-free although the algorithm still uses the dimension n. In Appendix 5.C, we present a completely dimension-free approach that, if a (λ, β)-sparsity condition is known, obtains an $\tilde{O}(\frac{DKG\sqrt{D^2G^2+\beta}}{\lambda^3\epsilon} + \frac{D^2G^2+\beta}{\epsilon^2})$ sample complexity based on the stability property of the regularized update (5.5) and the Lipschitzness of the loss function ℓ.

5.5 Experimental Results

We support our theoretical findings with empirical results in two different GDRO instances: one with the lower bound environment constructed in Theorem 5.3.5, and another with the Adult dataset [Becker and Kohavi, 1996]. On the lower bound environment, we set $\epsilon = 0.005, K = 10, \lambda^* = 0.2$ and $\beta_{\lambda^*} = 2$ so that the maximum risks can only be attained by the first two groups for any θ. On the Adult dataset, we use the same setup as [Soma et al., 2022] and divide 48 842 samples into groups based on `race` $\times$ `gender` with the goal of

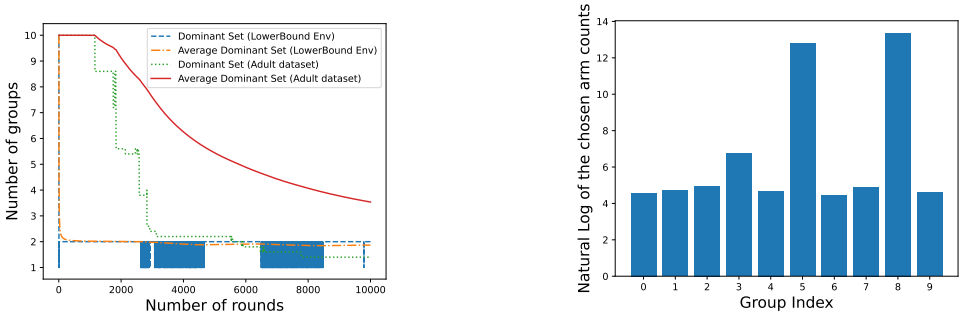

Figure 5.2: (Left) Sizes of the dominant sets in the first 10000 rounds computed by `SB-GDRO-SA`. (Right) The number of times a group is selected by the max-player, displayed in natural log. The highest group (group 8) is female Amer-Indian-Eskimo people.

finding a linear classifier that determines whether the annual outcome of a person exceeds USD 50 000 based on $n = 5$ features: age, years of education, capital gain, capital loss, and number of working hours. Similar to [Soma et al., 2022], $\mathcal{P}_i$ is the empirical distribution over samples in group i. One difference from [Soma et al., 2022] is we have $K = 10$ groups from 5 races and 2 genders instead of 6 groups, so that the difference between $\ln(K)$ and K is amplified. With $\epsilon = 0.001$, we use hinge loss and normalize the features so that the losses are in $[0, 1]$. We set $T = 10^6$ and $\delta = 0.01$ on both GDRO instances. The results are aggregated from five independent runs with random seeds $\{0, 1, 2, 3, 4\}$. To compute θ^*, we run the two-player zero-sum game with *ideal players* who have access to the underlying distributions $\mathcal{P}_i$. More experimental details are in Appendix 5.F.

On both GDRO instances, we compare `SB-GDRO-SA` (Algorithm 5.9) to the Stochastic Mirror Descent for GDRO algorithm (`SMD-GDRO`) proposed by [Zhang et al., 2023a]. To the best of our knowledge, `SMD-GDRO` is the only suitable baseline with a near-optimal high-probability guarantee in the minimax regime.

5.5.1 Discovering non-trivial (λ, β)-sparsity

Figure 5.2 (Left) shows the sizes $\left|\hat{S}_{\theta_t}\right|$ and the average $\frac{1}{t}\sum_{h=1}^{t}\left|\hat{S}_{\theta_h}\right|$ computed by `SB-GDRO-SA` in the first 10 000 rounds. On GDRO with Adult dataset, it indicates that `SB-GDRO-SA` quickly discovers dominant sets of sizes smaller than $\lceil \ln(K) \rceil$ within the first 3000 rounds. This shows that a non-trivial $(\lambda, \ln(K))$-sparsity condition indeed holds for hypotheses *around* θ^* in practical settings. Further inspection reveals this $(\lambda, \ln(K))$-sparsity is discovered early in the game without using too many samples: on the lower bound environment the final λ is $0.125 \approx 0.5\lambda^*$ using roughly 3000 samples, while on the Adult dataset the

final λ is $\frac{1}{2^9} \approx \epsilon\sqrt{\frac{C}{\ln(K)}}$ using roughly 36 000 samples. Both of these values are much smaller than T, and as T is scaled with $\frac{1}{\epsilon^2}$, this empirically supports the insight in Theorem 5.4.2 that the sample complexity is dominated by the number of rounds needed in the two-player zero-sum game.

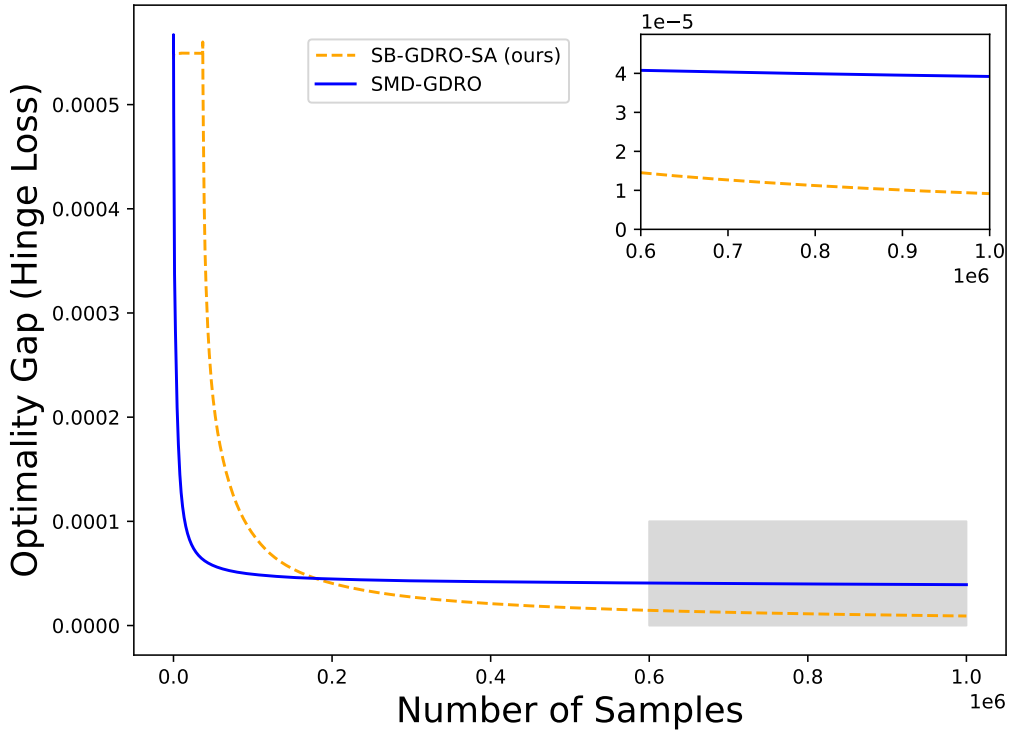

Figure 5.3: The optimality gap of `SB-GDRO-SA` and `SMD-GDRO` on GDRO with the Adult dataset. Lower is better.

5.5.2 Convergence Properties of `SB-GDRO-SA`

Next, we show results indicating that `SB-GDRO-SA` finds a ϵ-optimal hypothesis using fewer samples than `SMD-GDRO`. Figure 5.3 shows the optimality gap err of $\bar{\theta}_t$ of `SB-GDRO-SA` and `SMD-GDRO` as a function of the number of drawn samples on the Adult dataset. Initially, `SB-GDRO-SA` uses more samples than `SMD-GDRO` because `SB-GDRO-SA` needs to estimate λ^*. However, as θ_t gets closer to θ^*, the optimality gap of `SB-GDRO-SA` decreases much quicker since it only collects samples from the two groups with the largest risks. While `SMD-GDRO` struggles to get an optimality gap under 4×10^{-5} even after nearly $T = 10^6$ samples, `SB-GDRO-SA` manages to do so well below 4×10^5 samples. Figure 5.2 (Right) shows an interesting ob-

servation that more than 60% of the samples drawn by `SB-GDRO-SA` are from the *female Amer-Indian-Eskimo* group. This is in stark contrast to the fact that this group constitutes only 0.3% of the dataset (186 out of 48 842 samples). This underlines the *robustness* aspect of GDRO, which is different compared to the traditional empirical risk minimization regime where samples from the largest groups contribute more to the optimization process.

5.6 Conclusion and Future Work

We introduced a new structure called (λ, β)-sparsity into the GDRO problem. We showed a fundamental connection between the per-action regret in sleeping bandits and the optimality gap of the two-player zero-sum game approach for the GDRO problem, and then improved the dependency from $O(K \ln(K))$ to $O(\beta \ln(K))$ in the leading term of the sample complexity of (λ, β)-sparse problems, even when the optimal λ is unknown. We also showed a near-matching lower bound, which both extends and generalizes the lower bound construction in minimax settings to the (λ, β)-sparse settings. One interesting future direction is relax the (λ, β)-sparsity to hold only within some neighborhood of θ^*. This seems to require last iterate convergence of the sequence of θ_t's in stochastic games, which is still an open problem.

5.A Proofs for Section 5.3

For a pseudo-metric space $(\mathcal{F}, \|.\|)$, for any $\nu > 0$, let $\mathcal{N}(\mathcal{F}, \nu, \|.\|)$ be the ν-covering number of $\mathcal{F}$; that is $\mathcal{N}(\mathcal{F}, \nu, \|.\|)$ is the minimal number of balls of radius ν needed to cover $\mathcal{F}$.

First, we prove the following lemma on a uniform convergence bound that holds for a sufficiently large value of m.

Lemma 5.A.1. Let $m = \frac{384n \ln\left(\frac{741GDK}{\delta}\right)}{0.01\lambda^2}$. With probability at least $1 - \delta/2$, the event

$$\mathcal{E}_{i,\theta} = \{|\hat{R}_i(\theta) - R_i(\theta)| \leq 0.15\lambda\} \tag{5.11}$$

holds simultaneously for all $i \in [K]$ and $\theta \in \Theta$.

5.A.1 Proof of Lemma 5.A.1

Our proof for the uniform convergence bound in Lemma 5.A.1 is based on the Rademacher complexity bound of the class of functions L_Θ defined as follows:

$$L_\Theta = \{\ell(\theta, .) : \mathcal{Z} \to [0, 1], \theta \in \Theta\}, \tag{5.12}$$

which is the set of all possible functions $\ell(\theta, .)$ for $\theta \in \Theta$. First, we state the following bound for the empirical Rademacher complexity based on the chaining argument [Dudley, 1967; Liao, 2020].

Lemma 5.A.2. (Dudley's Entropy Integral Bound [Dudley, 1967; Liao, 2020]) Let $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$ be a class of real-valued functions, $S = \{z_1, z_2, \ldots, z_m\}$ be a set of m random i.i.d samples. For a function $f \in \mathcal{F}$, let

$$\|f\|_{2,S} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(f(z_j))^2} \tag{5.13}$$

be an S-dependent seminorm of f. Assuming

$$\sup_{f\in\mathcal{F}} \|f\|_{2,S} \leq c,$$

where c is a positive constant, we have

$$\text{Rad}(\mathcal{F}, S) \leq \inf_{\epsilon \in [0, \frac{c}{2}]} \left(4\epsilon + \frac{12}{\sqrt{m}} \int_{\epsilon}^{\frac{c}{2}} \sqrt{\ln\Big(\mathcal{N}(\mathcal{F}, \nu, \|.\|_{2,S})\Big)} d\nu \right), \tag{5.14}$$

where $\text{Rad}(\mathcal{F}, S) = \frac{1}{m}\mathbb{E}_{\sigma \in \{\pm 1\}^m} \left[\sup_{f \in \mathcal{F}} \sum_{j=1}^{m} \sigma_j f(z_j) \right]$ is the empirical Rademacher complexity of $\mathcal{F}$ and $\mathcal{N}(\mathcal{F}, \nu, \|.\|_{2,S})$ is the size of a ν-cover of $\mathcal{F}$.

A proof of this lemma can be found in [Liao, 2020]. We now prove Lemma 5.A.1.

Proof (of Lemma 5.A.1). For $i \in [K]$, Theorem 26.5 in [Shalev-Shwartz and Ben-David, 2014] states that with probability at least $1 - \frac{\delta}{4K}$ over the set V_i of size m, for all $\theta \in \Theta$,

$$\left| \frac{1}{m} \sum_{j=1}^{m} \ell(\theta, V_{i,j}) - R_i(\theta) \right| \leq 2\text{Rad}(L_\Theta, V_i) + \sqrt{\frac{32 \ln(4K/\delta)}{m}}.$$

Our proof is based on the fact that the covering number of the compact set $\Theta \subset \mathbb{R}^n$ is finite, and hence the empirical Rademacher complexity $\text{Rad}(L_\Theta, V_i)$ is bounded for all $i \in [K]$. Because the values of the loss function ℓ is in $[0, 1]$, we have $\|f\|_{2,V_i} \leq 1$ for all $f \in L_\Theta$. Moreover, the diameter of L_Θ measured in $\|.\|_{2,V_i}$ is

$$\begin{aligned}
\max_{\theta, \theta' \in \Theta} \sqrt{\frac{1}{m} \sum_{j=1}^{m} \left(\ell(\theta, V_{i,j}) - \ell(\theta', V_{i,j})\right)^2} &\leq \max_{\theta, \theta' \in \Theta} \sqrt{\frac{1}{m} \sum_{j=1}^{m} G^2 \|\theta - \theta'\|_2^2} \\
&\leq \sqrt{\frac{1}{m} \sum_{j=1}^{m} G^2 D^2} \\
&= GD,
\end{aligned}$$

where the first inequality is due to the Lipschitzness of the loss function ℓ and the second inequality is due to D being the diameter of Θ measured in ℓ_2-norm. Applying Lemma 5.A.2

with $c = 1$ and $\epsilon = 0$, we have

$$\begin{aligned}
\text{Rad}(L_\Theta, V_i) &\leq \frac{12}{\sqrt{m}} \int_0^{\frac{1}{2}} \sqrt{\ln\Big(\mathcal{N}(L_\Theta, \nu, \|.\|_{2,V_i})\Big)} d\nu \\
&\leq \frac{12G}{\sqrt{m}} \int_0^{\frac{1}{2}} \sqrt{\ln\left(\frac{4GD}{\nu}\right)^n} d\nu \\
&= \frac{12\sqrt{n}}{\sqrt{m}} \int_0^{\frac{1}{2}} \sqrt{\ln\left(\frac{4GD}{\nu}\right)} d\nu
\end{aligned}$$

where the second inequality is due to a result that the size of the smallest ν-cover on a set $\mathcal{F}$ with diameter d is bounded by $(\frac{4d}{\nu})^n$ [see e.g. Carl and Stephani, 1990, Equation 1.1.10]. To compute this integral, let $u = \sqrt{\ln(4GD/\nu)}$. We then have $\nu = 4GDe^{-u^2}$, and $d\nu = 4GDd(e^{-u^2})$. As $\nu \to 0, u \to \infty$. As $\nu \to \frac{1}{2}, u \to \sqrt{\ln(8GD)}$. Hence,

$$\begin{aligned}
\int_0^{\frac{1}{2}} \sqrt{\ln\left(\frac{4GD}{\nu}\right)} d\nu &= 4GD \int_\infty^{\sqrt{\ln(8GD)}} u d(e^{-u^2}) \\
&= 4GD \left(ue^{-u^2} \,|_\infty^{\sqrt{\ln(8GD)}} - \int_\infty^{\sqrt{\ln(8GD)}} e^{-u^2} du \right) \\
&= 4GD \left(\frac{\sqrt{\ln(8GD)}}{8GD} + \int_{\sqrt{\ln(8GD)}}^\infty e^{-u^2} du \right) \\
&\leq 4GD \left(\frac{\sqrt{\ln(8GD)}}{8GD} + \frac{\sqrt{\pi}}{2} e^{-\ln(8GD)} \right) \\
&= \frac{2\sqrt{\ln(8GD)} + \sqrt{\pi}}{4},
\end{aligned}$$

where the second equality is integration by parts and the inequality is by a Chernoff-type bound on the Gaussian error function $\frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \leq e^{-x^2}$ [Chang et al., 2011]. Overall, we have

$$\text{Rad}(L_\Theta, V_i) \leq \frac{3\sqrt{n}}{\sqrt{m}} \left(2\sqrt{\ln(8GD)} + \sqrt{\pi} \right).$$

We conclude that the uniform convergence bound is

$$\left| \frac{1}{m} \sum_{j=1}^m \ell(\theta, V_{i,j}) - R_i(\theta) \right| \leq \frac{3\sqrt{n}}{\sqrt{m}} \left(2\sqrt{\ln(8GD)} + \sqrt{\pi} \right) + \sqrt{\frac{32 \ln(4K/\delta)}{m}}.$$

Algorithm 5.4 `MinP`: the stochastic-OMD min-player $\mathbb{A}_\theta$

Input: $\theta_t \in \Theta$, sample $z_{i_t,t}$
Compute $\tilde{g}_t = \nabla \ell(\theta_t, z_{i_t,t})$
Compute $\theta_{t+1} = \arg\min_{\theta\in\Theta}\{\eta_{w,t}\langle \tilde{g}_t, \theta - \theta_t\rangle + \frac{1}{2}\|\theta - \theta_t\|_2^2\}$ by Equation (5.5)
Return: θ_{t+1}

By setting the right-hand side to 0.15λ, solving for m and simplifying, we obtain the following sufficient condition on m:

$$m \geq \frac{384n\ln\left(\frac{741GDK}{\delta}\right)}{0.01\lambda^2} \tag{5.15}$$

so that with probability at least $1 - \frac{\delta}{4K}$, we have $\left|\frac{1}{m}\sum_{j=1}^{m}\ell(\theta, V_{i,j}) - R_i(\theta)\right| \leq 0.15\lambda$ for all $\theta \in \Theta$. Taking a union bound over all K groups leads to the desired statement. □

5.A.2 Proof of Lemma 5.3.1

Proof. From Lemma 5.A.1, we immediately have the event $\mathcal{E}_{i,\theta}$ holds simultaneously for all $i \in [K]$ and $\theta \in \Theta$ with probability at least $1 - \frac{\delta}{2}$. Thus, it suffices to prove the desired statement assuming that all $\mathcal{E}_{i,\theta}$ hold. Let $R_{i,t} = R_i(\theta_t)$ and $\hat{R}_{i,t} = \hat{R}_i(\theta_t)$ be the risk and empirical risk of θ_t with respect to group i, respectively. We consider two cases: $\beta_\lambda < K$ and $\beta_\lambda = K$.

When $\beta_\lambda < K$:

In this case, there exists a non-empty set λ-dominant set S_{λ,θ_t} whose size is smaller than $\beta_\lambda < K$. This implies that the set $[K] \setminus S_{\lambda,\theta_t}$ is also non-empty. For any $i \in S_{\lambda,\theta_t}$ and $k \in [K] \setminus S_{\lambda,\theta_t}$, due to $\mathcal{E}_{i,\theta_t}, \mathcal{E}_{k,\theta_t}$ and by Definition 5.2.1, we have

$$\begin{aligned}
\hat{R}_{i,t} - \hat{R}_{k,t} &\geq (R_{i,t} - 0.15\lambda) - (R_{k,t} + 0.15\lambda) \\
&= R_{i,t} - R_{k,t} - 0.3\lambda \\
&\geq \lambda - 0.3\lambda \\
&= \tau > 0.
\end{aligned}$$

Thus, at any time t, the sorted sequence of groups can be divided into two non-empty parts: the first contains all groups in S_{λ,θ_t} and the second contains the rest. Since $|S_{\lambda,\theta_t}| \leq \beta_\lambda$,

the size of the first part is at most β_λ. Let $i^* = \arg\max_{j \in S_{\lambda,\theta_t}} \{\text{ord}(j)\}$ be the last group in the first part. Since $\text{nxt}(i^*) \in [K] \setminus S_{\lambda,\theta_t}$, we have $\hat{R}_{i^*,t} \geq \hat{R}_{\text{nxt}(i^*),t} + \tau$. This satisfies the condition in Algorithm 5.3, therefore the resulting set $\hat{S}_{\theta_t}$ is non-empty and its size does not exceed β_λ. To show that $\hat{S}_{\theta_t}$ is a 0.4λ-dominant set, for any $i' \in \hat{S}_{\theta_t}$ and $k' \in [K] \setminus \hat{S}_{\theta_t}$, we have

$$\begin{aligned} R_{i',t} - R_{k',t} &\geq (\hat{R}_{i',t} - 0.15\lambda) - (\hat{R}_{k',t} + 0.15\lambda) \\ &= \hat{R}_{i',t} - \hat{R}_{k',t} - 0.3\lambda \\ &\geq \hat{R}_{\hat{i},t} - \hat{R}_{\text{nxt}(\hat{i}),t} - 0.3\lambda \\ &\geq \tau - 0.3\lambda = 0.4\lambda, \end{aligned} \tag{5.16}$$

where the second inequality is from the definition of $\hat{i}$ and $\hat{S}_{\theta_t} = \{i \in [K] : \text{ord}(i) \leq \text{ord}(\hat{i})\}$, we have $\text{ord}(i') \leq \text{ord}(\hat{i}), \text{ord}(\text{nxt}(\hat{i})) \geq \text{ord}(k')$ and the empirical risks are sorted in decreasing order.

When $\beta_\lambda = K$:

In this case, the inequality $\left|\hat{S}_{\theta_t}\right| \leq \beta_\lambda$ holds trivially. To show that $\hat{S}_{\theta_t}$ is a 0.4λ-dominant set, we further consider two sub-cases: $\hat{i} \neq -1$ and $\hat{i} = -1$.

- $\hat{i} \neq -1$: in this case, the set $\hat{S}_{\theta_t} = \{i \in [K] : \text{ord}(i) \leq \text{ord}(\hat{i})\}$ has size at most $K - 1$ because the group with the largest empirical risk is excluded. Therefore, by the same argument as in (5.16), the set $\hat{S}_{\theta_t}$ is a 0.4λ-dominant set.
- $\hat{i} = -1$: in this case, we have $\hat{S}_{\theta_t} = [K]$ is trivially a 0.4λ-dominant set by Definition 5.2.1.

We conclude that the set $\hat{S}_{\theta_t}$ is a 0.4λ-dominant set at θ_t and $\left|\hat{S}_{\theta_t}\right| \leq \beta_\lambda$. □

5.A.3 Proof of Theorem 5.3.2

Let $A_t = |\mathbb{A}_t|$ be the number of active arms in round t. Throughout this section, we write $\eta_t = \eta_{q,t}$ for the learning rate of the SB-EXP3 algorithm used by the max-player.

The $O\left(\sqrt{\ln(K/\delta)\sum_{t=1}^{T} A_t}\right)$ high-probability per-action regret bound of the SB-EXP3 algorithm in [Nguyen and Mehta, 2024] was established for a fixed learning rate $\eta_t = \eta$ and a fixed exploration factor $\gamma_t = \gamma$. In this section, we generalize their result to algorithms

Algorithm 5.5 `FTARLShannon`: Follow the regularized and active leader with Shannon entropy regularizer and time-varying learning rates for sleeping bandits

Input: $K \geq 2$
Initialize $\tilde{L}_{i,0} = 0$ for all arms $i \in [K]$.
for each round $t = 1, \ldots,$ **do**
  The non-oblivious adversary selects and reveals $\mathbb{A}_t$
  Compute $q_{i,t} = \frac{\exp\left(-\eta_t \tilde{L}_{i,t}\right)}{\sum_{j=1}^{K} \exp\left(-\eta_t \tilde{L}_{j,t}\right)}$
  Compute $p_{i,t} = \frac{I_{i,t} q_{i,t}}{\sum_{j=1}^{K} I_{j,t} q_{j,t}}$ by Equation (5.18)
  Draw arm $i_t \sim p_t$ and observe $\hat{\ell}_t = \ell_{i_t,t}$
  for each arm $i \in [K]$ **do**
    If $I_{i,t} = 1$, compute $\tilde{\ell}_{i,t} = \frac{\mathbb{1}\{i_t = i\}\hat{\ell}_t}{p_{i,t} + \gamma_t}$ by Equation (5.19)
    If $I_{i,t} = 0$, compute $\tilde{\ell}_{i,t} = \hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}$ by Equation (5.20)
    Update $\tilde{L}_{i,t} = \tilde{L}_{i,t-1} + \tilde{\ell}_{i,t}$

with time-varying learning rates and exploration factors defined as follows:

$$\eta_t = 2\gamma_t = \sqrt{\frac{\ln(3K/\delta)}{\sum_{s=1}^{t} A_s}}. \tag{5.17}$$

Note that η_t and γ_t are chosen *after* the set of active arms $\mathbb{A}_t$ is revealed. As pointed out in Nguyen and Mehta [2024, Appendix G], the SB-EXP3 algorithm is equivalent to their Follow-the-Regularized-and-Active-Leader (FTARL) algorithm with the Shannon entropy regularizer. Therefore, a high-probability regret bound of FTARL with Shannon entropy regularizer and η_t and γ_t defined in (5.17) would imply Theorem 5.3.2. For completeness, we provide the full procedure of FTARL with Shannon entropy regularizer in Algorithm 5.5. For each arm $i \in [K]$ and round $t \in [T]$, this algorithm maintains an estimated cumulative loss $\tilde{L}_{i,t}$ defined as

$$\tilde{L}_{i,t} = \sum_{s=1}^{t} \tilde{\ell}_{i,t},$$

and computes the weight of arm i in round t by

$$q_{i,t} = \frac{\exp\left(-\eta_t \tilde{L}_{i,t-1}\right)}{\sum_{j=1}^{K} \exp\left(-\eta_t \tilde{L}_{j,t-1}\right)},$$

where η_t is the learning rate in round t. Initially, $\tilde{L}_{i,0} = 0$ for all arms $i \in [K]$. Upon receiving the set $\mathbb{A}_t$ of active arms, the sampling probability p_t is computed by normalizing $I_{i,t}q_{i,t}$ as follows:

$$p_{i,t} = \frac{I_{i,t}q_{i,t}}{\sum_{j=1}^{K} I_{j,t}q_{j,t}}. \tag{5.18}$$

Note that $I_{i,t} = \mathbb{1}\{i \in \mathbb{A}_t\}$, hence $p_{i,t}$ is non-zero only for active arms. An arm $i_t \sim p_t$ is drawn according to p_t and its loss $\hat{\ell}_t = \ell_{i_t,t}$ is observed. For an active arm $i \in \mathbb{A}_t$, its loss estimate is the IX-loss estimator [Neu, 2015]:

$$\tilde{\ell}_{i,t} = \frac{\mathbb{1}\{i_t = i\}\hat{\ell}_t}{p_{i,t} + \gamma_t}, \tag{5.19}$$

where γ_t is the exploration factor in round t. For a non-active arm $i \notin \mathbb{A}_t$, its loss estimate is defined as the difference between the observed loss $\hat{\ell}_t$ and the weighted sum of estimated losses of active arms [Nguyen and Mehta, 2024]:

$$\tilde{\ell}_{i,t} = \hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}. \tag{5.20}$$

The following theorem states the per-action regret bound of Algorithm 5.5.

Theorem 5.A.3. Let $(\eta_t)_{t=1,\dots}$ and $(\gamma_t)_{t=1,\dots}$ be two sequences of non-increasing learning rates and exploration factors such that $\eta_t \leq 2\gamma_t$. With probability at least $1-\delta$, `FTARLShannon` (Algorithm 5.5) guarantees that

$$\max_{a \in [K]} \text{Regret}(a) \leq \frac{\ln(K)}{\eta_T} + \frac{\ln(3K/\delta)}{2\gamma_T} + \ln(3/\delta) + \sum_{t=1}^{T} \left(\frac{\eta_t}{2} + \gamma_t\right) A_t. \tag{5.21}$$

The proof of this theorem is in Appendix 5.D. We are now ready to prove Theorem 5.3.2

Proof (of Theorem 5.3.2). By plugging (5.17) into the bound in Theorem 5.A.3, we obtain

$$\begin{aligned}
\max_{a\in[K]} \text{Regret}(a) &\leq \frac{\ln(K)}{\eta_T} + \frac{\ln(3K/\delta)}{\eta_T} + \ln\left(\frac{3}{\delta}\right) + \sum_{t=1}^{T} \eta_t A_t \\
&\leq \frac{2\ln(3K/\delta)}{\eta_T} + \ln\left(\frac{3}{\delta}\right) + \sum_{t=1}^{T} \eta_t A_t \\
&= \frac{2\ln(3K/\delta)}{\eta_T} + \ln\left(\frac{3}{\delta}\right) + \sqrt{\ln(3K/\delta)} \sum_{t=1}^{T} \frac{A_t}{\sqrt{\sum_{s=1}^{t} A_s}} \\
&= 2\sqrt{\ln(3K/\delta) \sum_{t=1}^{T} A_t} + \ln\left(\frac{3}{\delta}\right) + \sqrt{\ln(3K/\delta)} \sum_{t=1}^{T} \frac{A_t}{\sqrt{\sum_{s=1}^{t} A_s}}.
\end{aligned}$$

We bound $\sum_{t=1}^{T} \frac{A_t}{\sqrt{\sum_{s=1}^{t} A_s}}$ as follows: let $C_t = \sum_{s=1}^{t} A_t$ and $C_0 = 0$. Then,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{A_t}{\sqrt{\sum_{s=1}^{t} A_s}} &= \sum_{t=1}^{T} \frac{C_t - C_{t-1}}{\sqrt{C_t}} \\
&= \sum_{t=1}^{T} \int_{C_{t-1}}^{C_t} \frac{dx}{\sqrt{C_t}} \\
&\leq \sum_{t=1}^{T} \int_{C_{t-1}}^{C_t} \frac{dx}{\sqrt{x}} \\
&= \int_{C_0}^{C_T} \frac{dx}{\sqrt{x}} \\
&= 2\sqrt{C_T},
\end{aligned}$$

where the inequality holds because $\frac{1}{\sqrt{x}} \geq \frac{1}{\sqrt{C_t}}$ for all $C_{t-1} \leq x \leq C_t$. This implies that

$$\begin{aligned}
\max_{a\in[K]} \text{Regret}(a) &\leq 2\sqrt{\ln(3K/\delta) \sum_{t=1}^{T} A_t} + \ln\left(\frac{2}{\delta}\right) + 2\sqrt{\ln(3K/\delta) \sum_{t=1}^{T} A_t} \\
&= O\left(\sqrt{\ln(K/\delta) \sum_{t=1}^{T} A_t}\right).
\end{aligned}$$

□

5.A.4 Proof of Lemma 5.3.3

Proof. Since Δ_K is convex, we can write

$$\begin{aligned}\max_{q\in\Delta_K}\sum_{t=1}^{T}\phi(\theta_t,q) &= \max_{q\in\Delta_K}\sum_{t=1}^{T}\sum_{i=1}^{K}q_i R_i(\theta_t)\\ &= \max_{q\in\Delta_K}\sum_{i=1}^{K}q_i\sum_{t=1}^{T}R_{i,t}\\ &= \max_{i\in[K]}\sum_{t=1}^{T}R_{i,t}.\end{aligned}$$

Thus,

$$R_{\mathcal{A}_q} = \max_{i\in[K]}\sum_{t=1}^{T}R_{i,t} - \sum_{t=1}^{T}\phi(\theta_t,q_t).$$

If a group i is not included in $\hat{S}_{\theta_t}$ at time t, then by Lemma 5.3.1, for any $k\in\hat{S}_{\theta_t}$ we have

$$R_{i,t} < R_{i,t} + 0.4\lambda \le R_{k,t}.$$

By construction, the probability vector q_t contains non-zero elements only for groups in $\hat{S}_{\theta_t}$, hence for any $i\notin\hat{S}_{\theta_t}$, we have

$$R_{i,t} - \phi(\theta_t,q_t) = \sum_{k\in\hat{S}_{\theta_t}} q_{k,t}(R_{i,t}-R_{k,t}) \le 0.$$

We conclude that for any $i\in[K]$,

$$\sum_{t=1}^{T}R_{i,t} - \phi(\theta_t,q_t) \le \sum_{t=1}^{T}\mathbb{1}\{i\in\hat{S}_{\theta_t}\}(R_{i,t}-\phi(\theta_t,q_t)),$$

hence

$$\begin{aligned}R_{\mathcal{A}_q} &= \max_{i\in[K]}\sum_{t=1}^{T}R_{i,t} - \sum_{t=1}^{T}\phi(\theta_t,q_t)\\ &\le \max_{i\in[K]}\sum_{t=1}^{T}\mathbb{1}\{i\in\hat{S}_{\theta_t}\}\left(R_i(\theta_t)-\phi(\theta_t,q_t)\right).\end{aligned}$$

□

5.A.5 Proof of Theorem 5.3.4

Let $\beta_t = \left|\hat{S}_{\theta_t}\right|$ be the size of $\hat{S}_{\theta_t}$. Let $\bar{\beta}_T = \frac{1}{T}\sum_{t=1}^{T}\beta_t$ be the average number of active groups over T rounds. We first state the following bound for the regret of the max-player as a function of β_t, which is obtained directly by combining Theorem 5.3.2 and Lemma 5.3.3.

Lemma 5.A.4. With probability at least $1-\delta/4$, the regret of the max-player in `SB-GDRO-SA` (Algorithm 5.9) is bounded by

$$R_{\mathcal{A}_q} \leq O\left(\sqrt{\sum_{t=1}^{T}\beta_t \ln(K/\delta)}\right).$$

Proof. The max-player in Algorithm 5.9 uses the sleeping bandits algorithm SB-EXP3 with the stochastic loss of arm i at round t is

$$h_{i,t} = 1 - \ell(\theta_t, z_{i,t}).$$

Let $H_{i,t} = \mathbb{E}_{z_{i,t}\sim\mathcal{P}_i}[h_{i,t}]$ be the expected value of $h_{i,t}$. We have $H_{i,t} = 1 - R_i(\theta_t)$. Note that both $h_{i,t}$ and $H_{i,t}$ are in $[0,1]$. Fix a group $a \in [K]$ and let $I_{a,t} = \mathbb{1}\{a \in \hat{S}_{\theta_t}\}$. The per-action regret of group a is

$$\begin{aligned}
\text{GroupRegret}(a) &= \sum_{t=1}^{T} I_{a,t}(R_a(\theta_t) - \phi(\theta_t, q_t)) \\
&= \sum_{t=1}^{T} I_{a,t}\left(R_a(\theta_t) - \sum_{i=1}^{K} q_{i,t}R_i(\theta_t)\right) \\
&= \sum_{t=1}^{T} I_{a,t}\left(\sum_{i=1}^{K} q_{i,t}H_{i,t} - H_{a,t}\right) \\
&= \sum_{t=1}^{T} I_{a,t}\left(\sum_{i=1}^{K} q_{i,t}H_{i,t} - h_{i_t,t} + h_{i_t,t} - h_{a,t} + h_{a,t} - H_{a,t}\right) \\
&= \underbrace{\sum_{t=1}^{T} I_{a,t}\left(\sum_{i=1}^{K} q_{i,t}H_{i,t} - h_{i_t,t}\right)}_{(A)} + \underbrace{\sum_{t=1}^{T} I_{a,t}(h_{a,t} - H_{a,t})}_{(B)} + \underbrace{\sum_{t=1}^{T} I_{a,t}(h_{i_t,t} - h_{a,t})}_{(C)}.
\end{aligned}$$

The term C is exactly the per-action regret of arm a in SB-GDRO-SA defined in Equation (5.6) which, by Theorem 5.3.2, is bounded by $O\left(\sqrt{\ln(K/\delta)\sum_{t=1}^{T}\beta_t}\right)$ with probability at least $1-\frac{\delta}{12}$ simultaneously for all $a \in [K]$. Next, we bound the terms A and B. Since

$$\begin{aligned}\mathbb{E}_{i_t\sim q_t}[\mathbb{E}_{z_{i_t,t}\sim\mathcal{P}_{i_t}}[h_{i_t,t}]] &= \mathbb{E}_{i_t\sim q_t}[H_{i_t,t}] \\ &= \sum_{i=1}^{K} q_{i,t}H_{i,t}\end{aligned}$$

and

$$\begin{aligned}\left|\sum_{i=1}^{K} q_{i,t}H_{i,t} - h_{i_t,t}\right| &= \left|\sum_{i=1}^{K} q_{i,t}(H_{i,t} - h_{i_t,t})\right| \\ &\leq \sum_{i=1}^{K} q_{i,t}|H_{i,t} - h_{i_t,t}| \\ &\leq 1,\end{aligned}$$

A is a sum of a martingale difference sequence in which the absolute values of its elements are bounded by 1. By Azuma-Hoeffding inequality, with probability at least $1-\frac{\delta}{12K}$, we have

$$A \leq \sqrt{2T\ln(12K/\delta)}. \tag{5.22}$$

For term B, we also have $\mathbb{E}_{z_{a,t}\sim\mathcal{P}_a}[h_{a,t}] = H_{a,t}$, therefore B is also a sum of a martingale difference sequence with elements' absolute values bounded by 1. We then have $B \leq \sqrt{2T\ln(12K/\delta)}$ with probability at least $1-\frac{\delta}{12K}$. By taking a union bound twice: once over A and B for each action a and once all K actions, we obtain with probability at least $1-\frac{\delta}{6}$,

$$A + B \leq 2\sqrt{2T\ln(12K/\delta)}$$

simultaneously for all $a \in [K]$. Furthermore, since

$$T = \sum_{t=1}^{T} 1 \leq \sum_{t=1}^{T} \beta_t$$

due to $1 \leq \beta_t$, we obtain that with probability at least $1 - \frac{\delta}{4}$,

$$\max_{a \in [K]} \text{GroupRegret(a)} \leq O\left(\sqrt{\ln(K/\delta) \sum_{t=1}^{T} \beta_t}\right). \tag{5.23}$$

By Lemma 5.3.3, when $\mathcal{E}_{i,\theta}$ holds simultaneously for all $i \in [K]$ and $\theta \in \Theta$, we have

$$\begin{aligned} R_{A_q} &\leq \max_{a \in [K]} \text{GroupRegret(a)} \\ &\leq O\left(\sqrt{\ln(K/\delta) \sum_{t=1}^{T} \beta_t}\right). \end{aligned}$$

□

Corollary 5.A.5. For any $T \geq 1$, `SB-GDRO-SA` (Algorithm 5.9) guarantees that with probability at least $1 - \frac{\delta}{2}$,

$$\text{err}(\bar{\theta}, \bar{q}) \leq O\left(\frac{(DG + \sqrt{\bar{\beta}_t})\sqrt{\ln(K/\delta)}}{\sqrt{T}}\right) \tag{5.24}$$

Proof. By [Zhang et al., 2023a], the duality gap is bounded by the average regret of the two players:

$$\text{err}(\bar{\theta}, \bar{q}) \leq \frac{1}{T}\left(R_{\mathcal{A}_\theta} + R_{\mathcal{A}_q}\right). \tag{5.25}$$

In Appendix 5.E, we prove that with probability $1 - \delta/4$, the regret of the min-player is bounded by

$$R_{\mathcal{A}_\theta} \leq O\left(DG\sqrt{T \ln(1/\delta)}\right). \tag{5.26}$$

For the max-player, Lemma 5.A.4 implies that with probability $1 - \delta/4$,

$$\begin{aligned} R_{\mathcal{A}_q} &\leq O\left(\sqrt{\sum_{t=1}^{T} \beta_t \ln(K/\delta)}\right) \\ &= O\left(\sqrt{T\bar{\beta}_T \ln(K/\delta)}\right) \end{aligned} \tag{5.27}$$

where the equality is from the definition of $\bar{\beta}_T = \frac{\sum_{t=1}^{T} \beta_t}{T}$. Plugging (5.26) and (5.27) into (5.25) and taking a union bound, we obtain that with probability at least $1 - \delta/2$

$$\begin{aligned}\mathrm{err}(\bar{\theta}, \bar{q}) &\leq O\left(\frac{DG\sqrt{\ln(1/\delta)} + \sqrt{\bar{\beta}_T \ln(K/\delta)}}{\sqrt{T}}\right) \\ &\leq O\left(\frac{(DG + \sqrt{\bar{\beta}_T})\sqrt{\ln(K/\delta)}}{\sqrt{T}}\right).\end{aligned}$$

□

Corollary 5.A.5 implies $T = O\left(\frac{(D^2G^2 + \bar{\beta}_T)\ln(K/\delta)}{\epsilon^2}\right)$ is sufficient for a target optimality gap ϵ. This is a self-bounding condition on T since the quantity $\bar{\beta}_T$ is dependent on (and changes with) T. Nevertheless, it represents a valid stopping condition because $\bar{\beta}_T$ is fully observable and bounded above by a constant K. We are now ready to prove Theorem 5.3.4.

Proof (of Theorem 5.3.4). In Corollary 5.A.5, by setting the right-hand side to ϵ and solving for T, we obtain that with probability at least $1 - \frac{\delta}{2}$, the number of samples collected during the game for having $\mathrm{err}(\bar{\theta}, \bar{q}) \leq \epsilon$ is

$$O\left(\frac{(D^2G^2 + \bar{\beta}_T)\ln(K/\delta)}{\epsilon^2}\right).$$

By Lemma 5.A.1, we collect

$$O\left(\frac{n\ln(GDK/\delta)}{\lambda^2}\right)$$

samples from each group before the game starts so that with probability at least $1 - \frac{\delta}{2}$, we have $\beta_t \leq \beta_\lambda$ simultaneously for all $t \in [T]$. This implies that $\bar{\beta}_T \leq \beta_\lambda$ and thus the bound can be written as

$$O\left(\frac{(D^2G^2 + \beta_\lambda)\ln(K/\delta)}{\epsilon^2}\right).$$

By taking a union bound, we obtain that with probability at least $1 - \delta$, the total sample complexity is

$$O\left(\frac{Kn(\ln(GDK/\delta))}{\lambda^2} + \frac{(D^2G^2 + \beta_\lambda)\ln(K/\delta)}{\epsilon^2}\right).$$

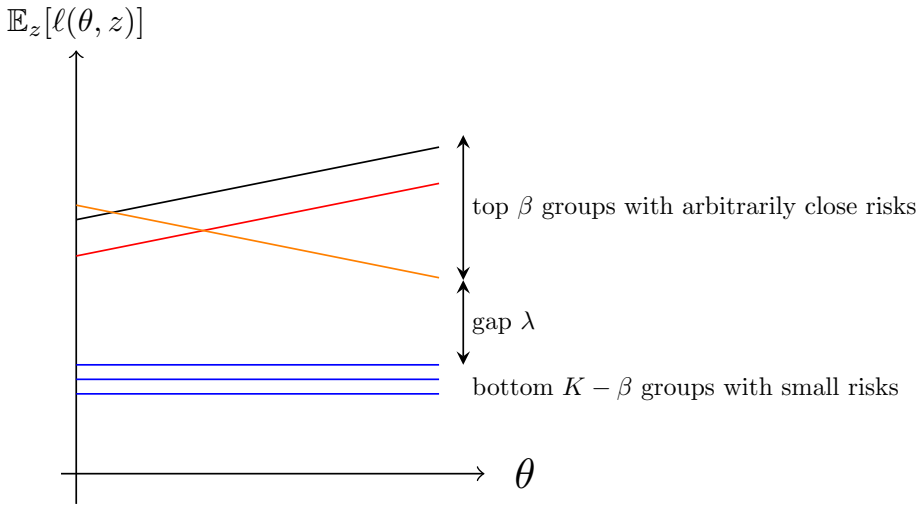

Figure 5.4: The construction for the $\Omega\left(\frac{G^2D^2+\beta}{\epsilon^2}\right)$ lower bound.

□

5.A.6 Proof of Theorem 5.3.5

Proof. Our lower bound construction directly extends that of [Soma et al., 2022]. In particular, let $\mathcal{Z} = [0, 1]^3$ be the set of samples and $\Theta = [0, 1]$ be the hypothesis set. The loss of a hypothesis θ on a sample $z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$ is

$$\ell(\theta, z) = \delta(z_1\theta + z_2(1 - \theta)) + z_3,$$

where $\delta \in (0, 1)$ is a constant defined later.

The distributions of the first β groups are similar to that of [Soma et al., 2022], where

- The first $\beta - 1$ distributions are

$$P_i = \begin{cases} z_1 = 0 & \text{almost surely} \\ z_2 = 1 & \text{almost surely} \\ z_3 \sim \text{Bernoulli}(\mu_i), \end{cases}$$

  where $\mu_i = \frac{1}{2}$ for $i = 1, 2, \ldots, \beta - 1$.

- The β^{th} distribution is

$$P_\beta = \begin{cases} z_1 = 1 & \text{almost surely} \\ z_2 = 0 & \text{almost surely} \\ z_3 \sim \text{Bernoulli}(\mu_\beta), \end{cases}$$

  where $\mu_\beta = \frac{1}{2}$.

The last $K - \beta$ distributions are

$$P_i = \begin{cases} z_1 = 0 & \text{almost surely} \\ z_2 = 0 & \text{almost surely} \\ z_3 = \frac{1}{2} - \lambda & \text{almost surely} \end{cases}$$

for $i = \beta + 1, \beta + 2, \ldots, K$. Figure 5.4 illustrates this construction. The risks of the groups are

$$R_i(\theta) = \mathbb{E}_{z \sim P_i}[\ell(\theta, z)] = \begin{cases} \Delta(1 - \theta) + \mu_i & (i = 1, 2, \ldots, \beta - 1) \\ \Delta\theta + \mu_\beta & (i = \beta) \\ \frac{1}{2} - \lambda & (i = \beta + 1, \beta + 2, \ldots, K). \end{cases}$$

Since $\Delta \geq 0, \theta \in (0, 1)$ and $\mu_i = \frac{1}{2}$ for $i = 1, \ldots, \beta$, we have $R_i(\theta) - R_j(\theta) \geq \lambda$ for any $1 \leq i \leq \beta$ and $\beta + 1 \leq j \leq K$. It follows that the set $[\beta] = \{1, 2, \ldots, \beta\}$ is a λ-dominant set, and this GDRO instance is (λ, β)-sparse. Because the risk differences between the top β groups are upper bounded by

$$|R_1(\theta) - R_\beta(\theta)| = |\Delta(1 - 2\theta)|,$$

which is arbitrarily smaller than λ, there can be no λ-dominant sets of size smaller than β. Thus, we have $\beta_\lambda = \beta$. Moreover, for any θ, its maximal risk is attained on a group within the set $[\beta]$ only. Therefore, the sample complexity of algorithm $\mathcal{A}$ is lower bounded by the total samples drawn from the first β groups.

On the other hand, by setting

$$\Delta = O\left(\sqrt{\frac{\beta}{T}}\right),$$

where T is the expected total number of samples drawn by $\mathcal{A}$, the first β groups are identical to the groups that give rise to the minimax lower bound in [Soma et al., 2022]. It follows that for any algorithm $\mathcal{A}$, there exists a GDRO instance which requires at least

$$\Omega\left(\frac{G^2D^2+\beta}{\epsilon^2}\right)$$

samples to find a ϵ-optimal hypothesis. □

5.B Proofs for Section 5.4

Remark 5.B.1. For notational simplicity, we use a short-hand notation $f_{C,g}$ for $\text{Cost}_{C,g}^{(GDRO)}$. In other words, we will write

$$f_{C,g}(\lambda) = \frac{C}{\lambda^2} + \frac{g(\lambda)}{\epsilon^2}. \tag{5.28}$$

We will also drop C, g when it is clear from the context and simply write $f(\lambda)$.

Remark 5.B.2. Throughout the proofs for Section 5.4, some of our bounds contain a $\ln\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ factor. While we will always present this term explicitly the first time they appear in the bounds, for ease of exposition we generally are not pedantic about this term and will treat it as a constant. For example, we will write

$$\begin{aligned}\ln\left(\frac{K\ln\left(\frac{1}{\epsilon}\right)}{\delta}\right) &= \ln\left(\frac{K}{\delta}\right) + \ln\left(\ln\left(\frac{1}{\epsilon}\right)\right)\\ &= O\left(\ln\left(\frac{K}{\delta}\right)\right),\end{aligned}$$

assuming that in practice, the number of arms $K > 1$ is not too small and the failure probability $\delta < 1$ is not too large so that $\frac{K}{\delta} > \ln\left(\frac{1}{\epsilon}\right)$.

5.B.1 A Sample-Efficient Approach for Estimating $\lambda^*_{C,g}$

We present an algorithm called `SolveOpt` for solving $\texttt{OPT}(C, \epsilon, g)$. `SolveOpt` outputs a $\hat{\lambda}$ such that $f(\hat{\lambda}) = O(f(\lambda^*))$ while using at most $O(f(\lambda^*)\ln(K/\delta)\ln(1/\epsilon))$ samples. The significance of this result in the context of GDRO is as follows: by using $\tilde{O}(f(\lambda^*))$ samples to obtain an estimate $\hat{\lambda}$ and then using $\hat{\lambda}$ for GDRO, we guarantee that the total sample complexity is of order $\tilde{O}(f(\lambda^*))$. This implies that without knowing λ^*, we can achieve a

Algorithm 5.6 `SolveOpt`: algorithm for solving $\texttt{OPT}(C, g)$

Input: $\epsilon \in (0,1), K \geq 3, C > K$, function g
Evaluate $g(1)$
Initialize $U = 1, L = \sqrt{\frac{C}{C + \frac{g(1)-1}{\epsilon^2}}}, \lambda = 1$;
while $\lambda \geq L$ **do**
  Evaluate $g(\lambda)$
  if $f(\lambda) < f(U)$ **then**
    Assign $U \leftarrow \lambda$, $L \leftarrow \sqrt{\frac{C}{\frac{C}{\lambda^2} + \frac{g(\lambda)-1}{\epsilon^2}}}$
  Update $\lambda = \lambda/5$
Return: $\hat{\lambda} = U$.

bound with only a logarithmic factor overhead than the bound obtained when λ^* is known. Our results and techniques are applicable to other trade-off problems similar to (5.10), and thus they could be of independent interest.

As mentioned in the main text, `SolveOpt` maintains two variables U and L which specify an interval $[L, U]$ that always contains a good estimate for λ^*. This $[L, U]$ shrinks over time based on how large $f(\lambda)$ is in comparison to $f(U)$: U is set to λ and L is increased accordingly if $f(\lambda) < f(U)$ holds. A crucial element of this process is choosing the geometric sequence $(1, \frac{1}{5}, \frac{1}{25}, \dots)$ of common ratio $\frac{1}{5}$ as the sequence of λ at which $g(\lambda)$ is evaluated. The process stops when $\lambda < L$, at which point the algorithm return the last value of U as an estimate for λ^*. The first key technical insight of this process is that L and U can be computed using only readily known quantities such as C, K and evaluated $g(\lambda)$. The second key technical insight is after some finite number of steps, it is guaranteed that *any* value in the interval $[L, U]$ is a good estimate for λ^*. The full procedure is given in Algorithm 5.6. The following lemma states the sample complexity of this approach.

Theorem 5.B.3. For any $\texttt{OPT}(C, g)$ problem defined in (5.10), `SolveOpt` (Algorithm 5.6) returns a $\hat{\lambda}$ such that $f(\hat{\lambda}) \leq 50 f(\lambda^*)$ while using at most $O\left(f(\lambda^*) \ln(K/\delta) \ln(1/\epsilon)\right)$ samples.

Before proving Theorem 5.B.3, we note that `SolveOpt` (Algorithm 5.6) maintains a range of values $[L, U]$ that always contains at least one good estimate for λ^*, and evaluates $g(\lambda)$ at elements of the geometric series $(U, \frac{U}{5}, \frac{U}{25}, \dots, \frac{U}{5^{\lfloor \log_5 (\frac{U}{L}) \rfloor}})$ to compute this estimate. Note that all elements of this series are in $[L, U]$. Whenever $f(\lambda)$ is strictly smaller than $f(U)$ for some λ, we shrink the range $[L, U]$ by setting $U = \lambda$ and $L = \sqrt{\frac{C}{\frac{C}{\lambda^2} + \frac{g(\lambda)-1}{\epsilon^2}}}$. We first prove

the following lemma which shows that L is always smaller than or equal to λ^*, thus at least one $g(\lambda)$ for $\lambda \leq \lambda^*$ will be evaluated while running `SolveOpt`.

Lemma 5.B.4. For any $U \in (0, 1]$, let

$$L = \sqrt{\frac{C}{\frac{C}{U^2} + \frac{g(U)-1}{\epsilon^2}}}.$$

Then, $L \leq \min\{\lambda^*, U\}$.

Proof. Since $\beta_U \geq 1$, we have $L \leq U$. By definition of λ^*, we have

$$f(\lambda^*) = \frac{C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2} \leq f(U) = \frac{C}{U^2} + \frac{g(U)}{\epsilon^2}.$$

Since $g(\lambda^*) \geq 1$, this implies

$$\frac{C}{(\lambda^*)^2} + \frac{1}{\epsilon^2} \leq \frac{C}{U^2} + \frac{g(U)}{\epsilon^2}.$$

Subtracting $\frac{1}{\epsilon^2}$ and dividing C on both sides, we obtain

$$\begin{aligned}(\lambda^*)^2 &\geq \frac{C}{\frac{C}{U^2} + \frac{g(U)-1}{\epsilon^2}} \\ &= L^2.\end{aligned}$$

We conclude that $L \leq \min\{\lambda^*, U\}$. □

The next lemma shows that if λ falls into the range $[\frac{\lambda^*}{5}, \lambda^*]$ when $f(U)$ is much larger than $f(\lambda^*)$, then the inequality $f(\lambda) < f(U)$ holds.

Lemma 5.B.5. For any $U \in (0, 1]$, if

$$f(U) > \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2},$$

then for any $\lambda \in [\frac{\lambda^*}{5}, \lambda^*]$, we have

$$f(\lambda) < f(U).$$

Proof. For any $\lambda \in [\frac{\lambda^*}{5}, \lambda^*]$, we have $\frac{C}{\lambda^2} \leq \frac{25C}{(\lambda^*)^2}$ and $g(\lambda) \leq g(\lambda^*)$. Hence,

$$\begin{aligned} f(\lambda) &= \frac{C}{\lambda^2} + \frac{g(\lambda)}{\epsilon^2} \\ &\leq \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2} \\ &< f(U). \end{aligned}$$

□

We need one last lemma, showing that once U is sufficiently close to λ^* such that $f(U) = O(f(\lambda^*))$, then any values between $[L, U]$ can be used as an estimate for λ^*.

Lemma 5.B.6. For any $U \in (0, 1]$, if

$$f(U) \leq \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2},$$

then with $L = \sqrt{\frac{C}{\frac{C}{U^2} + \frac{g(U)-1}{\epsilon^2}}}$, we have for any $\lambda \in [L, U]$,

$$f(\lambda) \leq 50 f(\lambda^*).$$

Proof. For any $\lambda \in [L, U]$, we have $\frac{C}{\lambda^2} \leq \frac{C}{L^2}$ and $g(\lambda) \leq g(U)$. Hence,

$$\begin{aligned} f(\lambda) &= \frac{C}{\lambda^2} + \frac{g(\lambda)}{\epsilon^2} \\ &\leq \frac{C}{L^2} + \frac{g(U)}{\epsilon^2} \\ &= \frac{C}{U^2} + \frac{g(U)-1}{\epsilon^2} + \frac{g(U)}{\epsilon^2} \qquad \text{since } L = \sqrt{\frac{C}{\frac{C}{U^2} + \frac{g(U)-1}{\epsilon^2}}} \\ &\leq \frac{C}{U^2} + \frac{2g(U)}{\epsilon^2} \\ &\leq 2f(U) \\ &\leq 50 \left(\frac{C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2} \right) \\ &= 50 f(\lambda^*). \end{aligned}$$

□

Proof (of Theorem 5.B.3). First, we prove that `SolveOpt` (Algorithm 5.6) always terminates after a finite number of steps. Observe that during the **while** loop, the sequence of values of λ is $(1, \frac{1}{5}, \frac{1}{25}, \dots)$, which is monotonically decreasing. On the other hand, L is non-decreasing from the initial value of $\sqrt{\frac{C}{C+\frac{g(1)-1}{\epsilon^2}}}$. This is because whenever $f(\lambda) \geq f(U)$, the value of L is

$$L = \sqrt{\frac{C}{\frac{C}{U^2} + \frac{g(U)-1}{\epsilon^2}}} = \sqrt{\frac{C}{f(U) - \frac{1}{\epsilon^2}}}.$$

Once the inequality $f(\lambda) < f(U)$ holds, U is assigned to λ and L is assigned to a new value L', where

$$L' = \sqrt{\frac{C}{f(\lambda) - \frac{1}{\epsilon^2}}} > \sqrt{\frac{C}{f(U) - \frac{1}{\epsilon^2}}} = L.$$

It follows that the condition $\lambda \geq L$ of the **while** loop must be false after a finite number of steps.

Next, we consider two cases: $f(1) \leq \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2}$ and $f(1) > \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2}$.

Case 1: $f(1) \leq \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2}$

In this case, by Lemma 5.B.6, any value in the range $[L, 1]$ where $L = \sqrt{\frac{C}{C+\frac{g(1)-1}{\epsilon^2}}}$ is a good estimate for λ^*. Because these values are at least L and there are at most $\log_5\left(\frac{1}{L}\right)$ evaluations, the maximum number of samples needed for testing all values of λ in the sequence

$(1, \frac{1}{5}, \frac{1}{5^2}, \ldots, \frac{1}{5^{\lfloor \log_5(1/L) \rfloor}})$ is bounded by

$$\begin{aligned} O\left(\frac{C\ln(K/\delta)\log_5\left(\frac{1}{L}\right)}{L^2}\right) &= O\left(\ln(K/\delta)\left(C+\frac{g(1)-1}{\epsilon^2}\right)\log_5\left(\frac{1}{L}\right)\right) \\ &= O\left(\ln(K/\delta)\left(C+\frac{g(1)-1}{\epsilon^2}\right)\log_5\left(\sqrt{1+\frac{g(1)-1}{C\epsilon^2}}\right)\right) \\ &\leq O\left(\ln(K/\delta)\left(C+\frac{g(1)}{\epsilon^2}\right)\ln\left(1+\frac{g(1)-1}{C\epsilon^2}\right)\right) \\ &\leq O\left(\ln(K/\delta)\left(C+\frac{g(1)}{\epsilon^2}\right)\ln\left(1+\frac{1}{\epsilon^2}\right)\right) \\ &\leq O\left(\ln(K/\delta)f(\lambda^*)\ln(1/\epsilon)\right), \end{aligned}$$

where the first inequality is from $\log_5(\sqrt{x}) = \frac{\ln(x)}{2\ln(5)} \leq \ln(x)$, the second inequality is due to $g(1) - 1 < K < C$, and the third inequality is from $C + \frac{g(1)}{\epsilon^2} = f(1) \leq 25f(\lambda^*)$ and $\ln\left(1+\frac{1}{\epsilon^2}\right) \leq \ln\left(\frac{2}{\epsilon^2}\right) = 2\ln\left(\frac{\sqrt{2}}{\epsilon}\right)$.

Case 2: $f(1) > \frac{25C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2}$

In this case, since $f(1) > \frac{24C}{(\lambda^*)^2} + f(\lambda^*) > f(\lambda^*)$, initially, we have $U = 1 > \lambda^*$. By Lemma 5.B.4, we have λ^* belongs to the range $[L, 1]$, where $L = \sqrt{\frac{C}{C+\frac{g(1)-1}{\epsilon^2}}}$. As λ repeatedly shrinks by $\frac{1}{5}$ from 1, after at most $\log_5(\frac{5}{\lambda^*})$ iterations of the **while** loop, it must fall into the range $[\frac{\lambda^*}{5}, \lambda^*]$. Each of these iterations makes one evaluation $g(\lambda)$ for some $\lambda \geq \lambda^*/5$. Hence, the total number of samples to evaluate $g(\lambda)$ for λ from 1 to $\frac{1}{5^{\lfloor \log_5(\frac{5}{\lambda^*}) \rfloor}}$ (i.e., the first element of the geometric series that lies inside the range $[\lambda^*/5, \lambda^*]$), is at most

$$O\left(\frac{C\ln(K/\delta)\log_5(\frac{5}{\lambda^*})}{(\lambda^*)^2}\right). \tag{5.29}$$

Let U_* be the largest value of λ being tested for which $f(\lambda) \leq \frac{25C}{\lambda^*} + \frac{g(\lambda^*)}{\epsilon^2}$. By Lemma 5.B.5, $U_* \geq \lambda^*/5$. Note that U_* might be larger than λ^*. Because the algorithm starts with $f(1) > \frac{25C}{\lambda^*} + \frac{g(\lambda^*)}{\epsilon^2}$, the inequality $f(\lambda) < f(U)$ must be true at $\lambda = U_*$. It follows that U is set to a U_*, and $f(U_*) \leq 50f(\lambda^*)$ by Lemma 5.B.6.

Let $L_* = \sqrt{\frac{C}{\frac{C}{U_*^2}+\frac{g(U_*)-1}{\epsilon^2}}}$ be the corresponding value of L after U is assigned to U^*. In each of the subsequent iterations, since L is non-decreasing and U is non-increasing, the returned value $\hat{\lambda}$ must be in this range $[L_*, U_*]$. The number of iterations needed until termination

starting from U_* is at most

$$\begin{aligned}
\log_5\left(\frac{U_*}{L_*}\right) &= \log_5\left(\frac{U_*\sqrt{\frac{C}{U_*^2}+\frac{g(U_*)-1}{\epsilon^2}}}{\sqrt{C}}\right) \\
&\le \log_5\left(\frac{U_*\sqrt{\frac{C}{U_*^2}+\frac{g(U_*)}{\epsilon^2}}}{\sqrt{C}}\right) \\
&= \log_5\left(\sqrt{1+\frac{g(U_*)U_*^2}{C\epsilon^2}}\right) \\
&\le \log_5\left(\sqrt{1+\frac{1}{\epsilon^2}}\right) \\
&\le \ln\left(1+\frac{1}{\epsilon^2}\right) \\
&\le \ln\left(\frac{2}{\epsilon^2}\right),
\end{aligned} \tag{5.30}$$

where the second inequality is from $g(U_*) \le K < C$ and $U_* \le 1$, the third inequality is due to $\log_5(\sqrt{x}) = \frac{1}{2}\log_5(x) = \frac{1}{2}\frac{\ln(x)}{\ln(5)} \le \ln(x)$ for $x > 1$ and the last inequality is $1 + \frac{1}{\epsilon^2} \le \frac{2}{\epsilon^2}$ for $\epsilon \le 1$. In each of these iterations, `SolveOpt` evaluates $g(\lambda)$ once for $\lambda \ge L_*$. In total, the number of samples in these iterations is at most

$$\begin{aligned}
O\left(\frac{C\ln(K/\delta)\log_5\left(\frac{U_*}{L_*}\right)}{L_*^2}\right) &= O\left(\ln(K/\delta)\left(\frac{C}{U_*^2}+\frac{g(U_*)-1}{\epsilon^2}\right)\log_5\left(\frac{U_*}{L_*}\right)\right) \\
&\le O\left(\ln(K/\delta)f(U_*)\ln\left(1/\epsilon^2\right)\right) \\
&= O\left(\ln(K/\delta)f(\lambda^*)\ln(1/\epsilon)\right)
\end{aligned} \tag{5.31}$$

where the first inequality is due to (5.30) and the second inequality is due to $f(\lambda^*) \le f(U_*) \le 50f(\lambda^*)$ and $\ln(1/\epsilon^2) = 2\ln(1/\epsilon)$. The total number of samples used by Algorithm 5.6 is the bounded by the sum of the number of samples for testing λ from 1 to U_*, and then from U_* to L_*. Combining (5.29) and (5.31), we have the total number of samples needed until

Algorithm 5.7 `EstG`: estimating $g(\lambda)$ for λ in the geometric sequence of common ratio $\frac{1}{5}$

Input: $\lambda \in (0,1)$

Compute a $\frac{0.1\lambda}{G}-$cover $\widehat{\Theta}$ of Θ with centers $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(\left|\widehat{\Theta}\right|)}$

Let $N = \ln\left(\frac{2}{\epsilon}\right)$

Draw $m_N = \frac{384n\ln\left(\frac{741GDKN}{\delta}\right)}{0.01\lambda^2}$ samples from each of the K groups into a set V_λ

Compute $S^{(i)} = \texttt{DominantSet}(\hat{\theta}^{(i)}, V_\lambda, 0.7\lambda)$ for $i = 1, 2, \ldots, \left|\widehat{\Theta}\right|$ by Algorithm 5.3

Return: $\hat{\beta}_\lambda = \max_{i=1,2,\ldots,\left|\widehat{\Theta}\right|} \left|S^{(i)}\right|$

Algorithm 5.6 terminates is at most

$$O\left(\frac{C\ln(K/\delta)\log_5(\frac{5}{\lambda^*})}{(\lambda^*)^2} + \frac{C\ln(K/\delta)\log_5\left(\frac{U_*}{L_*}\right)}{L_*^2}\right)$$
$$\leq O\left(f(\lambda^*)\ln(K/\delta)\ln(1/\epsilon)\right),$$

where the inequality is from $f(\lambda^*) = \frac{C}{(\lambda^*)^2} + \frac{g(\lambda^*)}{\epsilon^2} \geq \frac{C}{(\lambda^*)^2}$. □

5.B.2 Proofs for Section 5.4.1

Let $\mathcal{B}(\theta, r) = \{\theta' \in \Theta : \|\theta - \theta'\|_2 \leq r\}$ be a ℓ_2-ball of radius r centered at $\theta \in \Theta$. In this section, we prove Theorem 5.4.1 which specifies the sample complexity of `SB-GDRO-A` (Algorithm 5.8) for the setting where no λ is known beforehand. The most important component of `SB-GDRO-A` is computing an estimate $\hat{\lambda}$ for the optimal λ^* using the algorithm `SolveOpt` (Algorithm 5.6). This computation uses the algorithm `EstG` (Algorithm 5.7) to compute an estimate of β_λ for λ in the geometric sequence $(1, \frac{1}{5}, \frac{1}{25}, \ldots)$ of common ratio $\frac{1}{5}$.

First, we prove the following lemma which bounds the number of λ tested in Algorithm 5.6.

Lemma 5.B.7. For any $\texttt{OPT}(C, g)$ problem, in `SolveOpt` (Algorithm 5.6), the number of values λ whose $g(\lambda)$ need to be evaluated is at most

$$N = \ln\left(\frac{2}{\epsilon}\right). \tag{5.32}$$

Proof. In `SolveOpt`, the values of λ belong to a geometric sequence of common ratio $\frac{1}{5}$

Algorithm 5.8 `SB-GRDRO-A`: SB-GDRO without knowing any λ

Input: Constants $T, K, D, G > 0, \delta > 0, \epsilon > 0$
Compute $\hat{\lambda} = \texttt{SolveOpt}(\epsilon, K, \hat{C}, \hat{g})$ by Algorithm 5.6 where $\hat{g}$ is defined in Equation (5.37).

Let $\hat{\Theta} = \{\hat{\theta}^{(i)}\}_{i=1,\dots,\left|\hat{\Theta}\right|}$ be the $\frac{0.1\hat{\lambda}}{G}$-cover of Θ constructed when querying $\hat{\lambda}$ in `EstG`
Initialize $\theta_1 = \arg\min_{\theta\in\Theta} \|\theta\|_2$
for each round $t = 1, \dots, T$ **do**
  Let $c_t = \arg\min_{i\in\left|\hat{\Theta}\right|} \left\|\hat{\theta}^{(i)} - \theta_t\right\|$ be the index of the center in $\widehat{\Theta}$ closest to θ_t
  Let $\hat{S}_{\theta_t}$ be the pre-computed $0.4\hat{\lambda}$-dominant set at $\hat{\theta}^{(c_t)}$
  Compute $q_t = \texttt{MaxP}(t, \hat{S}_{\theta_t})$ by Algorithm 5.2
  Draw a group $i_t \sim q_t$ and a sample $z_{i_t,t} \sim \mathcal{P}_{i_t}$
  Compute $\theta_{t+1} = \texttt{MinP}(\theta_t, z_{i_t,t})$ by Algorithm 5.4
Return: $\bar{\theta}$

starting at 1 and terminating at a value no smaller than $\sqrt{\frac{C}{C+\frac{g(1)-1}{\epsilon^2}}}$, where $C > K \geq g(1)$. Therefore, the number of values in this sequence is at most

$$\begin{aligned}
\log_5\left(\frac{1}{\sqrt{\frac{C}{C+\frac{g(1)-1}{\epsilon^2}}}}\right) &= \log_5\left(\sqrt{\frac{C+\frac{g(1)-1}{\epsilon^2}}{C}}\right) \\
&= \log_5\left(\sqrt{1-\frac{1}{C\epsilon^2}+\frac{g(1)}{C\epsilon^2}}\right) \\
&< \log_5\left(\sqrt{1+\frac{g(1)}{C\epsilon^2}}\right) \\
&\leq \log_5\left(\sqrt{1+\frac{1}{\epsilon^2}}\right) \qquad \text{since } g(1) < C \\
&\leq \frac{1}{2}\log_5\left(\frac{4}{\epsilon^2}\right) \qquad \text{since } 1+\frac{1}{\epsilon^2} \leq \frac{4}{\epsilon^2} \\
&\leq \ln\left(\frac{2}{\epsilon}\right)
\end{aligned}$$

where the last inequality is due to $\frac{\log_5(x^2)}{2} = \frac{\ln(x)}{\ln(5)} \leq \ln(x)$ for any $x > 0$. □

Next, we show that any 0.4λ-dominant set $S_{0.4\lambda,\theta}$ at a $\theta \in \Theta$ is also a 0.2λ-dominant set at any θ' within the Euclidean ball $\mathcal{B}\left(\theta, \frac{0.1\lambda}{G}\right)$.

Lemma 5.B.8. Let $\theta \in \Theta$ and $\lambda \in [0,1]$. For any $\theta' \in \mathcal{B}\left(\theta, \frac{0.1\lambda}{G}\right)$, any 0.4λ-dominant set $S_{0.4\lambda,\theta}$ at θ is also a 0.2λ-dominant set at θ'.

Proof. The statement holds trivially if $S_{0.4\lambda,\theta} = [K]$. If $S_{0.4\lambda,\theta} \neq [K]$, for any $\theta' \in \mathcal{B}(\theta, \frac{0.1\lambda}{G})$ and any group $k \in [K]$, we have

$$\begin{aligned} |R_k(\theta') - R_k(\theta)| &\leq G\|\theta' - \theta\|_2 \\ &\leq 0.1\lambda, \end{aligned}$$

where the first inequality is due to the Lipschitzness of the loss function, and the second inequality is due to $\|\theta' - \theta\|_2 \leq \frac{0.1\lambda}{G}$. It follows that for any $k \in S_{0.4\lambda,\theta}$ and $k' \in [K] \setminus S_{0.4\lambda,\theta}$, we have

$$\begin{aligned} R_k(\theta') - R_{k'}(\theta') &\geq R_k(\theta) - 0.1\lambda - (R_{k'}(\theta) + 0.1\lambda) \\ &\geq 0.2\lambda, \end{aligned}$$

where the second inequality is due to $R_k(\theta) - R_{k'}(\theta) \geq 0.4\lambda$. This implies that $S_{0.4\lambda,\theta}$ is also a 0.2λ-dominant set at θ'. □

Using Lemma 5.B.8, we prove the following guarantee of `EstG`, which is obtained directly from Lemma 5.A.1 and Lemma 5.3.1 by re-scaling δ to δ/N.

Lemma 5.B.9. For any input $\lambda \in [0,1]$, `EstG` (Algorithm 5.7) outputs a $\hat{\beta}_\lambda$ such that with probability at least $1 - \frac{\delta}{4N}$, the following condition hold:

$$\beta_{0.2\lambda} \leq \hat{\beta}_\lambda \leq \beta_\lambda. \tag{5.33}$$

Moreover, the number of samples needed to compute $\hat{\beta}_\lambda$ is

$$\begin{aligned} Km_N &= \frac{384Kn\ln\left(\frac{741GDK\ln\left(\frac{2}{\epsilon}\right)}{\delta}\right)}{0.01\lambda^2} \\ &= O\left(\frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{\lambda^2}\right) \end{aligned} \tag{5.34}$$

Proof. In `EstG`, for each $g(\lambda)$ being evaluated, the number of samples drawn from each of

the K groups is

$$m_N = \frac{384n \ln\left(\frac{741GDK\ln(\frac{2}{\epsilon})}{\delta}\right)}{0.01\lambda^2}. \tag{5.35}$$

By Lemma 5.A.1 and Lemma 5.3.1, this value of m_N is sufficiently large so that with probability at least $1 - \frac{\delta}{4N}$, for all $i = 1, 2, \ldots, \left|\widehat{\Theta}\right|$, the set $S^{(i)}$ is a 0.4λ-dominant set at $\hat{\theta}^{(i)}$. Since $\left|S^{(i)}\right| \leq \beta_\lambda$ by Lemma 5.3.1, we have

$$\hat{\beta}_\lambda = \max_{i=1,2,\ldots,\left|\widehat{\Theta}\right|} \left|S^{(i)}\right| \leq \beta_\lambda.$$

Moreover, by Lemma 5.B.8, at any $\theta \in \mathcal{B}(\hat{\theta}^{(i)}, \frac{0.1\lambda}{G})$, $S^{(i)}$ is also a 0.2λ-dominant set at θ. It follows that

$$\left|S^{(i)}\right| \geq \beta_{0.2\lambda,\theta} \tag{5.36}$$

where we recall the definition of $\beta_{0.2\lambda,\theta}$ being the size of the smallest 0.2λ-dominant set at θ. Taking the maximum over i on both sides, we obtain

$$\begin{aligned}
\hat{\beta}_\lambda &= \max_{i\in\{1,2,\ldots,\left|\hat{\Theta}\right|\}} \left|S^{(i)}\right| \\
&\geq \max_{i\in\{1,2,\ldots,\left|\hat{\Theta}\right|\}} \max\left\{\beta_{0.2\lambda,\theta} : \theta \in \mathcal{B}\left(\hat{\theta}^{(i)}, \frac{0.1\lambda}{G}\right)\right\} \\
&= \max_{\theta\in\Theta} \beta_{0.2\lambda,\theta} \\
&= \beta_{0.2\lambda},
\end{aligned}$$

where the second equality (third line) is due to $\hat{\Theta}$ being a cover of Θ and the last equality is due to the definition of $\beta_{0.2\lambda}$. We conclude that $\beta_{0.2\lambda} \leq \hat{\beta}_\lambda \leq \beta_\lambda$.

Finally, since m_N samples are drawn from each of K groups, the total number of samples needed to compute $\hat{\beta}_\lambda$ is Km_N. □

Next, we define a function $\hat{g} : [0, 1] \to [K]$ as follows.

$$\hat{g}(\lambda) = \begin{cases} 1 & \text{if } \lambda < \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}} \\ \texttt{EstG}(\lambda) & \text{if } \lambda \in (1, \frac{1}{5}, \frac{1}{25}, \ldots, \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}}) \\ \hat{g}(x) \text{ for } x = \arg\max\{t < \lambda : t \in (1, \frac{1}{5}, \frac{1}{25}, \ldots, \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}})\} & \text{otherwise} \end{cases} \tag{5.37}$$

In other words, we define $\hat{g}(\lambda) = 1$ for any sufficiently small λ that will never be called during `SolveOpt`, which consists of values smaller than $\frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}}$. For any $\lambda \geq \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}}$, if λ that belongs to the geometric sequence $\left(1, \frac{1}{5}, \frac{1}{25}, \ldots, \right)$, then $\hat{g}(\lambda)$ is the output of `EstG` with input λ. Otherwise, $\hat{g}(\lambda)$ is equal to the output $\hat{\beta}_x$ of `EstG` with input $x = \frac{1}{5^{\lceil \log_5(\frac{1}{\lambda}) \rceil}}$, which is the first value in the geometric sequence that is smaller than λ. Let

$$\hat{C} = \frac{Kn \ln\left(\frac{GDK \ln(\frac{1}{\epsilon})}{\delta}\right)}{\ln(K/\delta)}, \tag{5.38}$$

and

$$\hat{f}(\lambda) = \frac{\hat{C}}{\lambda^2} + \frac{\hat{g}(\lambda)}{\epsilon^2}, \tag{5.39}$$

We have $\hat{g}(\lambda) \in [1, K]$ due to the fact that `DominantSet` always returns a non-empty subset of $[K]$. Moreover, $\hat{C} > K$. The following lemma shows that with high probability, this function $\hat{g}(.)$ is non-decreasing.

Lemma 5.B.10. With probability at least $1 - \frac{\delta}{4}$, the function $\hat{g}$ defined in (5.37) is non-decreasing.

Proof. Since $\frac{\epsilon}{2} \leq \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}}$, within the range $[0, \frac{\epsilon}{2}]$ we have $\hat{g}(\lambda) = 1$ which is never larger than any possible returned value by `EstG`. Therefore, we only need show that $\hat{g}(\lambda)$ is non-decreasing for $\lambda > \frac{\epsilon}{2}$. To this end, we will prove that $\hat{g}(\frac{\lambda}{5}) \leq \hat{g}(\lambda)$ for any value λ in the truncated geometric sequence $(1, \frac{1}{5}, \frac{1}{25}, \ldots, \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}})$. This trivially holds for the last value $\lambda_{\text{last}} = \frac{1}{5^{\lfloor \log_5(\frac{2}{\epsilon}) \rfloor}}$ in this sequence, since by definition $\hat{g}(\frac{\lambda_{\text{last}}}{5}) = 1$ and the returned value of `EstG` is always greater than or equal to 1. For other λ in this sequence, let $\lambda' = \frac{\lambda}{5} = 0.2\lambda$. Observe that the number of values in this truncated geometric sequence is at most

$$\log_5\left(\frac{2}{\epsilon}\right) \leq \ln\left(\frac{2}{\epsilon}\right) = N,$$

hence we can apply Lemma 5.B.9 and take a union bound (over at most N values of the

truncated geometric sequence) to obtain that with probability at least $1-\frac{\delta}{4}$, we have $\hat{g}(\lambda') = \hat{\beta}_{\lambda'} \le \beta_{\lambda'} = \beta_{0.2\lambda}$ and $\beta_{0.2\lambda} \le \hat{\beta}_{\lambda}$ simultaneously for any $\lambda > \lambda_{\text{last}}$. We conclude that $\hat{g}(\lambda') \le \beta_{0.2\lambda} \le \hat{\beta}_{\lambda} = \hat{g}(\lambda)$ for any $\lambda' = \lambda/5$ and $\lambda \le 1$ in the geometric sequence $(1, \frac{1}{5}, \frac{1}{25}, \dots)$. Furthermore, this implies that for any pair (λ', λ) where $\lambda' \le \lambda$ from this geometric sequence, we have $\hat{g}(\lambda') \le \hat{g}(\lambda)$.

More generally, for any $0 \le x < y \le 1$, we have three possibilities:

- if $y \le \frac{\epsilon}{2}$, then $g(x) = g(y) = 1$
- if $x \le \frac{\epsilon}{2} < y$, then $g(x) = 1 \le g(y)$
- if $\frac{\epsilon}{2} < x$, then

$$\begin{aligned} g(x) &= g\left(\frac{1}{5^{\lceil \log_5(\frac{1}{x}) \rceil}}\right) \\ &\le g\left(\frac{1}{5^{\lceil \log_5(\frac{1}{y}) \rceil}}\right) \\ &= g(y), \end{aligned}$$

In all cases, $g(x) \le g(y)$. We conclude that the function g is piecewise-constant and non-decreasing. □

Lemma 5.B.10 indicates that with high probability, the function $\hat{g}$ defined in (5.37) satisfies the conditions of the optimization problem (5.10), thus enabling the use of `SolveOpt` (Algorithm 5.6) and Theorem 5.B.3. From Lemma 5.B.9, taking a union bound over all queried λ throughout `SolveOpt` and note that there are at most N such λ by Lemma 5.B.7, we immediately obtain the following result.

Corollary 5.B.11. Running `SolveOpt` (Algorithm 5.6) with $\hat{g}(\lambda)$ defined in (5.37) guarantees that with probability at least $1-\delta/2$, simultaneously for all λ queried in `SolveOpt`, `EstG` (Algorithm 5.7) returns a value $\hat{\beta}_{\lambda}$ such that $\beta_{0.2\lambda} \le \hat{\beta}_{\lambda} \le \beta_{\lambda}$.

We are now ready to prove Theorem 5.4.1.

Proof (of Theorem 5.4.1). Let

$$\lambda^* = \arg\min_{\lambda \in [0,1]} \left(\frac{Kn \ln\left(\frac{GDK \ln(1/\epsilon)}{\delta}\right)}{\lambda^2} + \frac{(D^2G^2 + \beta_{\lambda}) \ln(K/\delta)}{\epsilon^2} \right). \tag{5.40}$$

We run `SolveOpt` for solving $\texttt{OPT}(\hat{C}, \hat{g})$ and obtain $\hat{\lambda}$ as an estimate for $\lambda^*_{\hat{C},\hat{g}}$. Theorem 5.B.3 and Corollary 5.B.11 implies that with probability at least $1 - \frac{\delta}{2}$, the returned value $\hat{\lambda}$ is an element of the geometric sequence $(1, \frac{1}{5}, \frac{1}{25}, \dots)$ and satisfies $\hat{f}(\hat{\lambda}) \leq 50\hat{f}(\lambda^*_{\hat{C},\hat{g}})$, which is equivalent to

$$\begin{aligned}
\frac{\hat{C}}{\hat{\lambda}^2} + \frac{\hat{\beta}_{\hat{\lambda}}}{\epsilon^2} &= \frac{\hat{C}}{\hat{\lambda}^2} + \frac{\hat{g}(\hat{\lambda})}{\epsilon^2} \\
&\leq 50\left(\frac{\hat{C}}{(\lambda^*_{\hat{C},\hat{g}})^2} + \frac{\hat{g}(\lambda^*_{\hat{C},\hat{g}})}{\epsilon^2}\right) \\
&\leq 50\left(\frac{\hat{C}}{(\lambda^*)^2} + \frac{\hat{g}(\lambda^*)}{\epsilon^2}\right) \\
&\leq 50\left(\frac{\hat{C}}{(\lambda^*)^2} + \frac{\beta_{\lambda^*}}{\epsilon^2}\right),
\end{aligned} \tag{5.41}$$

where the second inequality is from the definition of $\lambda^*_{\hat{C},\hat{g}}$, and the last inequality is due to $\hat{g}(\lambda^*) = \hat{g}(x_*) \leq \beta_{x_*} \leq \beta_{\lambda^*}$, where $x_* = \frac{1}{5^{\lceil \log_5(\frac{1}{\lambda^*}) \rceil}} \leq \lambda^*$. Moreover, the number of samples needed for running `SolveOpt` is at most

$$\begin{aligned}
O\left(\hat{f}(\lambda_{\hat{C},\hat{g},*})\ln(K/\delta)\ln(1/\epsilon)\right) &\leq O\left(\hat{f}(\lambda^*)\ln(K/\delta)\ln(1/\epsilon)\right) \\
&= O\left(\left(\frac{\hat{C}}{(\lambda^*)^2} + \frac{\hat{g}(\lambda^*)}{\epsilon^2}\right)\ln(K/\delta)\ln(1/\epsilon)\right) \\
&\leq O\left(\left(\frac{\hat{C}}{(\lambda^*)^2} + \frac{\beta_{\lambda^*}}{\epsilon^2}\right)\ln(K/\delta)\ln(1/\epsilon)\right).
\end{aligned} \tag{5.42}$$

In each round t of the two-player zero-sum game in `SB-GDRO-A`, the dominant set used by the max-player is taken to be the pre-computed $0.4\hat{\lambda}$-dominant set of the center c_t closest to θ_t, where $c_t \in \{1, 2, \dots, \left|\widehat{\Theta}\right|\}$:

$$c_t = \operatorname*{arg\,min}_{c=1,2,\dots,\left|\widehat{\Theta}\right|} \left\|\theta_t - \hat{\theta}^{(c)}\right\|.$$

As a result, the sizes of the dominant sets used by the max-player never exceeds $\hat{\beta}_{\hat{\lambda}}$. Together with Corollary 5.A.5, this implies that with probability at least $1-\delta/2$, the number of samples

used by the two-player zero-sum game in SB-GDRO-A is

$$O\left(\frac{(D^2G^2+\hat{\beta}_{\hat{\lambda}})\ln(2K/\delta)}{\epsilon^2}\right) \tag{5.43}$$

Finally, combining (5.42) and (5.43) and taking a union bound, we obtain that with probability at least $1-\delta$, SB-GDRO-A returns an ϵ-optimal hypothesis $\bar{\theta}$ with sample complexity

$$O\left(\left(\frac{Kn\ln\left(GDK\ln\left(\frac{1}{\epsilon}\right)/\delta\right)}{(\lambda^*)^2}+\frac{(D^2G^2+\beta_{\lambda^*})\ln(K/\delta)}{\epsilon^2}\right)\ln(1/\epsilon)\right) = \tag{5.44}$$

$$O\left(\left(\frac{C}{(\lambda^*)^2}+\frac{D^2G^2+\beta_{\lambda^*}}{\epsilon^2}\right)\ln(K/\delta)\ln(1/\epsilon)\right), \tag{5.45}$$

where $C=\frac{Kn\ln\left(\frac{GDK}{\delta}\right)}{\ln(K/\delta)}$ and we dropped the $\ln(\ln(1/\epsilon))$ term in the final bound for ease of exposition. □

5.B.3 Proofs for Section 5.4.2

The detailed procedure of the computationally efficient approach SB-GDRO-SA is given in Algorithm 5.9. Similar to SB-GDRO (Algorithm 5.1), SB-GDRO-SA uses the two-player zero-sum game framework. The main difference is that SB-GDRO-SA does not assume any input λ. Instead, it uses λ from the geometric sequence $\left(1,\frac{1}{2},\frac{1}{4},\dots\right)$. A new value of λ_{t+1} in this sequence is used for computing the dominant set in round $t+1$ if *both* of the following conditions hold:

- The size $|S_t|$ of the dominant set in round t is larger than $\ln(K)$
- The value of λ_t used in round t is not smaller than $L=\epsilon\sqrt{\frac{C}{\ln(K)}}$.

If at least one of the two conditions does not hold, we set $\lambda_{t+1}=\lambda_t$.

Whenever a new value of λ_t is used, i.e., either $t=1$ or $\lambda_t\neq\lambda_{t-1}$, a new set of samples of size m is drawn from each of K groups. The value of m is set by Lemma 5.A.1 and Lemma 5.3.1, that is $m_t=\frac{384n\ln\left(\frac{741GDK}{\delta_t}\right)}{0.01\lambda_t^2}$. Here, the failure probability δ_t is set by a geometric sequence of the form (recall that δ is the global failure probability of the algorithm)

$$\delta_t=\frac{3\delta}{\pi^2(\sum_{s=2}^{t}\mathbb{1}\{\lambda_s\neq\lambda_{s-1}\})^2}, \tag{5.46}$$

Algorithm 5.9 `SB-GDRO-SA`: adaptive and computationally efficient approach without knowing any λ

Input: Constants $K \geq 2, D, G > 0, \delta > 0, \epsilon > 0$
Compute constant $C = \frac{Kn \ln\left(\frac{GDK}{\delta}\right)}{\ln(K/\delta)}$
Initialize $\lambda_1 = 1, L = \epsilon\sqrt{\frac{C}{\ln(K)}}$
Initialize $\theta_1 = \arg\min_{\theta \in \Theta}(\|\theta\|_2)$
Initialize $\delta_1 = \frac{\delta}{2}$, counter $c_1 = 1$
Draw a new set of samples V_1 of size Km_1, where $m_1 = \frac{384n \ln\left(\frac{741GDK}{\delta_t}\right)}{0.01\lambda_1^2}$.
for each round $t = 1, \ldots,$ **do**
  Min-player plays θ_t
  Compute a $0.4\lambda_t$-dominant set $S_t = \text{DominantSet}(\theta_t, V_t, 0.7\lambda_t)$ at θ_t using Algorithm 5.3
  if $|S_t| > \ln(K)$ and $\lambda_t \geq L$ **then**
    Increase counter $c_{t+1} = c_t + 1$
    Reduce $\lambda_{t+1} \leftarrow \frac{\lambda_t}{2}$
    Reduce $\delta_{t+1} \leftarrow \frac{6\delta_t}{\pi^2 c_{t+1}^2}$
    Draw a new set of samples V_{t+1} of size Km_t, where $m_t = \frac{384n \ln\left(\frac{741GDK}{\delta_{t+1}}\right)}{0.01\lambda_{t+1}^2}$
  else
    Set $\lambda_{t+1} \leftarrow \lambda_t, V_{t+1} \leftarrow V_t$ and $\delta_{t+1} \leftarrow \delta_t, c_{t+1} \leftarrow c_t$
  Compute $q_t = \text{MaxP}(t, S_t)$
  Draw $i_t \sim q_t$ and $z_{i_t,t} \sim \mathcal{P}_{i_t}$
  Compute $\theta_{t+1} = \text{MinP}(\theta_t, z_{i_t,t})$
Return: $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$.

so that the total failure probability of computing the dominant sets is bounded by

$$\sum_{s=1}^{\infty} \delta_s \mathbb{1}\{\lambda_s \neq \lambda_{s-1}\} \leq \frac{3\delta}{\pi^2} \sum_{s=1}^{\infty} \frac{1}{s^2} \leq \frac{\delta}{2}. \tag{5.47}$$

Note that we define $\lambda_0 = -1$ by convention, so that $\lambda_s \neq \lambda_{s-1}$ holds for $s = 1$.

We will prove the following theorem, which is more general than Theorem 5.4.2

Theorem 5.B.12. Let $C = \frac{Kn \ln\left(\frac{GDK}{\delta}\right)}{\ln(K/\delta)}$ and $L = \epsilon\sqrt{\frac{C}{\ln(K)}}$. For any $\epsilon > 0, \delta \in (0,1)$, with probability at least $1 - \delta$, `SB-GDRO-SA` (Algorithm 5.9) returns an ϵ-optimal hypothesis with sample complexity

$$O\left(\frac{(D^2G^2 + \max(\ln(K), \beta_L))}{\epsilon^2} \ln(K/\delta) \ln\frac{1}{\epsilon}\right) \tag{5.48}$$

Obviously, Theorem 5.B.12 immediately implies Theorem 5.4.2 since $\beta_L \leq \beta_{\lambda^*}$ for all $L < \lambda^*$.

Before proving Theorem 5.B.12, we first prove a lemma showing that the sets S_t in all $t = 1, 2, \ldots, T$ rounds are indeed $0.4\lambda_t$-dominant sets with probability $1 - \delta/2$.

Lemma 5.B.13. With probability at least $1 - \frac{\delta}{2}$, `SB-GDRO-SA` (Algorithm 5.9) guarantees that for all $t \geq 1$, the set S_t is a $0.4\lambda_t$-dominant set at θ_t.

Proof. Fix a λ in the geometric sequence $(1, \frac{1}{2}, \frac{1}{4}, \ldots)$. Let t and t' be the first and last rounds in which λ is used for computing the dominant sets, respectively. By Lemma 5.A.1 and Lemma 5.3.1, m_t is sufficiently large so that with probability at least $1 - \frac{\delta_t}{2}$, for all the rounds from t to t', the set S_h for $h = t, t+1, \ldots, t'$ is a 0.4λ-dominant set of θ_h. By construction, $\delta_t = \frac{3\delta}{\pi^2(\sum_{s=2}^{t} \mathbb{1}\{\lambda_s \neq \lambda_{s-1}\})^2}$. Taking a union bound over all λ in the geometric sequence $(1, \frac{1}{2}, \frac{1}{4}, \ldots)$ and using

$$\sum_{s=1}^{\infty} \frac{1}{s^2} = \frac{\pi^2}{6},$$

we obtain with probability at least $1 - \sum_{s=1}^{\infty} \delta_s \mathbb{1}\{\lambda_s \neq \lambda_{s-1}\} \geq 1 - \frac{\delta}{2}$, the set S_t is a $0.4\lambda_t$-dominant set at θ_t for all $t \geq 1$. □

The next technical lemma helps bounding the sum $\sum_{s=1}^{T} m_s \mathbb{1}\{\lambda_s \neq \lambda_{s-1}\}$.

Lemma 5.B.14. Let $\delta > 0, G \geq 1, D \geq 1, K \geq 1$ and $C = \frac{Kn \ln\left(\frac{GDK}{\delta}\right)}{\ln(K/\delta)}$. For any $x \in (0, 1)$, we have

$$Kn \sum_{s=1}^{\lceil -\log_2(x) \rceil} \ln\left(\pi^2 \frac{s^2 GDK}{3\delta}\right) \leq 2C \ln(K/\delta) \ln\left(\frac{1}{x}\right) + O\left(Kn \ln\left(\frac{1}{x}\right) \ln\left(\ln\left(\frac{1}{x}\right)\right)\right).$$

Proof. Without loss of generality, assume $1/x$ is a power of e. We have

$$\begin{aligned}\sum_{s=1}^{\lceil -\log_2(x)\rceil} \ln\left(\frac{\pi^2 s^2 GDK}{3\delta}\right) &\le \sum_{s=1}^{-\ln(x)} \left(\ln\left(\frac{GDK}{\delta}\right) + \ln\left(\frac{\pi^2}{3}\right) + \ln\left(s^2\right)\right) \\ &\le 2\ln\left(\frac{GDK}{\delta}\right)\ln\left(\frac{1}{x}\right) + \ln\left(\prod_{s=1}^{-\ln(x)} s^2\right) \\ &= 2C\ln(K/\delta)\left(\frac{1}{x}\right) + O\left(\ln\left(\frac{1}{x}\right)\ln\left(\ln\left(\frac{1}{x}\right)\right)\right),\end{aligned} \tag{5.49}$$

where the inequalities are from $\log_2(1/x) \le \ln(1/x)$ and $\ln(n!) = O(n\ln(n))$. Multiplying Kn to both sides leads to the desired statement. □

We are now ready to prove Theorem 5.B.12.

Proof (of Theorem 5.B.12). Let $\lambda_{\ln K}$ be the largest λ such that $\beta_\lambda = \ln(K)$. If no such λ exists, we set $\lambda_{\ln K} = 0$. Let $\bar{\lambda} = \max(L, \lambda_{\ln(K)})$. Note that $\bar{\lambda} \ge L = \epsilon\sqrt{\frac{C}{\ln(K)}} > \epsilon$ since $C > K > \ln(K)$.

In the worst case, Algorithm 5.9 draw a new set of samples until $\frac{\bar{\lambda}}{2} \le \lambda_t \le \bar{\lambda}$. Without loss of generality, we can assume $\bar{\lambda} < \frac{1}{4}$. Otherwise, Algorithm 5.9 draws only three sets of samples and stops doing so immediately after some $\lambda_t \ge \frac{1}{4}$, which trivially leads to a sample complexity of $O\left(\frac{G^2D^2+\ln(K)}{\epsilon^2}\right)$.

With $\bar{\lambda} < \frac{1}{4}$, the total number of samples of used for computing the dominant sets in Algorithm 5.9 are

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{1}\{\lambda_t \ne \lambda_{t-1}\}\frac{384Kn\ln\left(741GDK/\delta_t\right)}{0.01\lambda_t^2} &\le O\left(\frac{Kn}{\bar{\lambda}^2}\sum_{s=1}^{-\log_2(\bar{\lambda})}\ln\left(\frac{2^s GDK}{\delta_1}\right)\right) \\ &\le O\left(\frac{C}{\bar{\lambda}^2}\ln(K/\delta)\ln\left(1/\bar{\lambda}\right)\right) \\ &\le O\left(\frac{C}{\bar{\lambda}^2}\ln(K/\delta)\ln(1/\epsilon)\right),\end{aligned} \tag{5.50}$$

where the second inequality is from Lemma 5.B.14 and the last inequality is from $\bar{\lambda} > \epsilon$. Note that we dropped the $\ln\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ for ease of exposition.

Next, we bound the regret bound of the max-player. Let $\hat{\beta} = \max(\ln(K), \beta_L)$. We show the average $\bar{\beta}_T = \frac{1}{T}\sum_{t=1}^{T}|S_t|$ is not much larger than $\hat{\beta}$.

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}|S_t| &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{1}\{|S_t|>\hat{\beta}\}+\mathbb{1}\{|S_t|\leq\hat{\beta}\}\right)|S_t| \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\{|S_t|>\hat{\beta}\}K+\mathbb{1}\{|S_t|<\hat{\beta}\}\hat{\beta} \\ &\leq \frac{1}{T}\left(K\log_2\left(\frac{1}{\bar{\lambda}}\right)+\hat{\beta}T\right) \\ &\leq \frac{1}{T}\left(2\hat{\beta}T\right)\log_2(1/\bar{\lambda}) \\ &\leq 2\hat{\beta}\ln(1/\epsilon),\end{aligned} \tag{5.51}$$

where the first inequality is from $|S_t| < K$ for all t, the second inequality is because there are at most $\log_2(\frac{1}{\bar{\lambda}})$ rounds where $|S_t| > \hat{\beta}$, the third inequality is from $K < \hat{\beta}T$ as $\hat{\beta} \geq 1$, and the last inequality is from $\log_2(1/\bar{\lambda}) \leq \ln\left(1/\bar{\lambda}\right)$ as $\bar{\lambda} > \epsilon$. Combining this with (5.27), the regret of the max-player is bounded by

$$\begin{aligned}R_{\mathcal{A}_q} &\leq O\left(\sqrt{T\bar{\beta}_T\ln(K/\delta)}\right) \\ &= O\left(\sqrt{T\hat{\beta}\ln(K/\delta)\ln(1/\epsilon)}\right).\end{aligned} \tag{5.52}$$

Plugging (5.52) into (5.25) and combining with (5.50), we have the total amount of samples to get an ϵ-optimal hypothesis is

$$O\left(\frac{C}{(\bar{\lambda})}\ln(1/\epsilon)^2\right)+O\left(\frac{(G^2D^2+\hat{\beta})\ln(K/\delta)}{\epsilon^2}\ln(1/\epsilon)\right)\leq \tag{5.53}$$

$$O\left(\left(\min\left\{\frac{\ln(K)}{\epsilon^2},\frac{C}{\lambda_{\ln(K)}^2}\right\}+\frac{(D^2G^2+\max(\ln(K),\beta_L))}{\epsilon^2}\right)\ln(K/\delta)\left(\ln\frac{1}{\epsilon}\right)\right) \tag{5.54}$$

where the inequality is from $\bar{\lambda} = \max(\lambda_{\ln(K)}, L)$ and $\frac{C}{L^2} = \frac{\ln(K)}{\epsilon^2}$, thus

$$\frac{C}{(\bar{\lambda})^2} \leq \min\left\{\frac{C}{L^2},\frac{C}{\lambda_{\ln(K)}^2}\right\} = \min\left\{\frac{\ln(K)}{\epsilon^2},\frac{C}{\lambda_{\ln(K)}^2}\right\}.$$

Since $\min\left\{\frac{\ln(K)}{\epsilon^2},\frac{C}{\lambda_{\ln(K)}^2}\right\} \leq \frac{\ln(K)}{\epsilon^2} \leq \frac{\max(\ln(K),\beta_L)}{\epsilon^2}$, the final bound can be simplified to $O\left(\frac{(D^2G^2+\max(\ln(K),\beta_L))}{\epsilon^2}\right.$ l

□

5.C A Completely Dimension-Independent Approach

In this section, we present SB-GDRO-DF, a modified version of Algorithm 5.1 that uses $O\left(\frac{KDG\sqrt{(D^2G^2+\beta)\ln(K/\delta)}}{\lambda^3\epsilon}\right)$ samples for computing the dominant sets over T rounds of the two-player zero-sum game. This bound avoids the dependency on n, the dimension of Θ, which might be preferable in high-dimensional settings. The trade-off for getting rid of n is an additional $\frac{1}{\lambda\epsilon}$ multiplicative factor in the non-leading term of the sample complexity bound.

We assume that a pair (λ, β) is known such that the problem instance is (λ, β)-sparse. Unlike SB-GDRO, SB-GDRO-DF does not use a fixed set of samples V for computing the dominant sets of all $\theta \in \Theta$. Instead, SB-GDRO-DF computes the dominant sets only for the hypotheses θ_t that the learner encounters during the game. In particular, the T rounds are divided into $\frac{T}{\sigma}$ episodes, in which each episode has σ consecutive rounds that use the same dominant set. By the stability property of the regularized update (5.5) and the Lipschitzness of the loss function ℓ, if σ is sufficiently small then the differences between the risks of the hypotheses within each episode is small. This implies that a dominant set for θ_t will remain a dominant set (possibly with smaller gaps) and therefore can be reused for the hypotheses $\theta_{t+1}, \theta_{t+2}, \ldots, \theta_{t+\sigma}$. The full procedure is given in Algorithm 5.10 in Appendix 5.C, and its sample complexity is stated in the following theorem.

Theorem 5.C.1. For any $\epsilon > 0, \delta \in (0,1)$, with probability $1-\delta$, SB-GDRO-DF with $\eta_{w,t} = \frac{2D}{G\sqrt{T}}$, $\eta_{q,t}$ and γ_t defined in Theorem 5.3.4 returns an ϵ-optimal hypothesis with sample complexity

$$O\left(\frac{DKG\sqrt{(D^2G^2+\beta)\ln(K/\delta)}\ln\left(\frac{KDG}{\epsilon\lambda\delta}\right)}{\lambda^3\epsilon} + \frac{(D^2G^2+\beta)\ln(K/\delta)}{\epsilon^2}\right).$$

Next, we give a detailed description of SB-GDRO-DF (Algorithm 5.10) and prove its sample complexity bound in Theorem 5.C.1. Essentially, SB-GDRO-DF also uses the two-player zero-sum game framework similar to SB-GDRO. Note that since β is known, we can compute the number of rounds $T = O(\frac{(G^2D^2+\beta)\ln(K/\delta)}{\epsilon^2})$ before the game starts. Unlike the previous algorithms, knowing T before the game starts allows us to use a fixed learning rate

$$\eta_{w,t} = \eta_t = \frac{2D}{G\sqrt{T}} \tag{5.55}$$

for the min-player in Algorithm 5.10. Another difference is that SB-GDRO-DF proceeds in

Algorithm 5.10 `SB-GDRO-DF`: Dimension-free `SB-GDRO` Algorithm with known (λ, β)

Input: Constants $K, D, G, \eta_w, \lambda, \beta, \epsilon > 0$, hypothesis set $\Theta \subset \mathbb{R}^n$
Compute $T = O(\frac{(D^2G^2+\beta)\ln(K/\delta)}{\epsilon^2})$
Compute the maximum length of each episode $\sigma = \left\lfloor \frac{0.1\lambda}{\eta_w G^2} \right\rfloor$
Initialize an episode counter $\rho = 1$
Compute $m' = \frac{24\ln\left(\frac{4KT}{\sigma\delta}\right)}{\lambda^2}$
Draw m' samples from each K groups into set V^1
Initialize $\theta_1 = \arg\min_{\theta\in\Theta} \|\theta\|_2$
Compute a dominant set $S^1 = \texttt{DominantSet}(\theta_1, V^1, \lambda)$ at θ_1 by Algorithm 5.3
Let $\hat{S}_{\theta_1} = S^1$
Compute $q_1 = \texttt{MaxP}(\theta_1, \hat{S}_{\theta_1})$ by Algorithm 5.2
for each round $t = 1, \ldots, T$ **do**
  Draw a group $i_t \sim q_t$ and a sample $z_{i_t,t} \sim \mathcal{P}_{i_t}$
  Compute $\theta_{t+1} = \texttt{MinP}(\theta_t, z_{i_t,t})$ by Algorithm 5.4
  if t is divisible by σ **then**
    Increase episode counter $\rho \leftarrow \rho + 1$
    Draw new m' samples from each of K groups into V^ρ.
    Compute a dominant set $S^\rho = \texttt{DominantSet}(\theta_{t+1}, V^\rho, \lambda)$ at θ_{t+1} by Algorithm 5.3
  Let $\hat{S}_{\theta_{t+1}} = S^\rho$
  Compute $q_{t+1} = \texttt{MaxP}(\theta_{t+1}, \hat{S}_{\theta_{t+1}})$ by Algorithm 5.2 using the last computed S^ρ
Return: $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ and $\bar{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$

episodes, each consists of multiple consecutive rounds, and the max-player uses the same dominant set for the rounds within each episode. More concretely, in `SB-GDRO-DF`, the T rounds of the game are divided into $\lceil \frac{T}{\sigma} \rceil$ episodes, each is of length σ, except for the last episode which may have fewer than σ rounds if T is not divisible by σ. The value σ is defined as follows:

$$\sigma = \left\lfloor \frac{0.1\lambda}{\eta_w G^2} \right\rfloor. \tag{5.56}$$

By this construction, the first episode contains rounds $(1, 2, \ldots, \sigma)$, the second episode contains rounds $(\sigma + 1, \ldots, 2\sigma)$ and so on, until the last episode which contains rounds $(\lfloor \frac{T}{\sigma} \rfloor \sigma + 1, \ldots, T)$. Let $\rho = 1, 2, \ldots, \lceil \frac{T}{\sigma} \rceil$ be the running index of the episodes. Within an episode ρ,

- Before the first round of this episode, a set V^ρ of Km' samples are drawn from the K groups, where m' i.i.d samples are drawn from each group. The value of m' is $\frac{24\ln\left(\frac{4KT}{\sigma\delta}\right)}{\lambda^2}$.

- Let t^ρ be the index of the first round in episode ρ and $\theta^\rho = \theta_{t^\rho}$ be either the initial hypothesis (if $\rho = 1$) or the hypothesis played by the min-player using the algorithm `MinP` (Algorithm 5.4) (if $\rho > 1$) in round t^ρ. A 0.4λ-dominant set S^ρ is computed using `DominantSet` (Algorithm 5.3) with input θ^ρ and V^ρ.
- In rounds $t \in (t^\rho, t^\rho + 1, \ldots, \min\{t^\rho + \sigma, T\})$ of this episode, the max-player plays q_t using the algorithm `MaxP` (Algorithm 5.2) with the same input S^ρ. Then, a group $i_t \sim q_t$ is drawn and a sample $z_{i_t,t} \sim \mathcal{P}_{i_t}$ is drawn from group i_t. The min-player then follows the `MinP` strategy (Algorithm 5.4) with input θ_t and z_{i_t} to compute θ_{t+1}.

The algorithm returns $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ after T rounds. The following lemma shows that for any episode ρ, with high probability, S^ρ is a 0.4λ-dominant set at θ^ρ.

Lemma 5.C.2. At the beginning of episode ρ in `SB-GDRO-DF`, with probability at least $1 - \frac{\sigma\delta}{2T}$, the set S^ρ is a 0.4λ-dominant set at θ^ρ.

Proof. In each episode ρ, we draw $m' = \frac{24\ln\left(\frac{4KT}{\sigma\delta}\right)}{\lambda^2}$ samples from each group. By Hoeffding's inequality, for each group k, we have

$$\begin{aligned}
\Pr_{V_k^\rho}\left(\frac{1}{m}\left|\sum_{j=1}^{m'}\ell(\theta^\rho, V_{k,j}^\rho) - R_k(\theta^\rho)\right| \geq 0.15\lambda\right) &\leq 2\exp\left(-0.045\lambda^2 m'\right) \\
&= 2\exp\left(-1.08\ln\left(\frac{4KT}{\sigma\delta}\right)\right) \\
&\leq 2\exp\left(-\ln\left(\frac{4KT}{\sigma\delta}\right)\right) \\
&= \frac{\sigma\delta}{2KT}.
\end{aligned}$$

By taking a union bound over K groups, we have

$$\left|\frac{1}{m}\sum_{j=1}^{m'}\ell(\theta^\rho, V_{k,j}^\rho) - R_k(\theta^\rho)\right| \leq 0.15\lambda \tag{5.57}$$

holds simultaneously for all $k \in [K]$ with probability at least $1 - \frac{\sigma\delta}{2T}$. The condition (5.57) of V^ρ is the same as the event $\mathcal{E}_{k,\theta^\rho}$ in (5.11) of V in Lemma 5.A.1. Hence, we can apply Lemma 5.3.1 and conclude that with probability at least $1 - \frac{\sigma\delta}{2T}$, the set S^ρ is a 0.4λ-dominant set at θ^ρ. □

The next lemma shows that the set S^ρ is a dominant set not only at θ^ρ but also at the hypotheses within the episode ρ.

Lemma 5.C.3. SB-GDRO-DF guarantees that if S is a 0.4λ-dominant set at θ_t for some $t \in [T]$, then for any non-negative integer $\sigma' \leq \min\left\{\left\lfloor \frac{0.1\lambda}{\eta_w G^2} \right\rfloor, T-t\right\}$, S is also a 0.2λ-dominant set at $\theta_{t+\sigma'}$.

Proof. SB-GDRO-DF uses the update rule (5.5) to compute θ_{t+1}. This update rule can be written as follows:

$$\begin{aligned}\theta_{t+1} &= \arg\min_{\theta\in\Theta}\{2\langle \eta_w \tilde{g}_t, \theta - \theta_t\rangle + \|\theta - \theta_t\|^2 + \eta_w^2 \|\tilde{g}_t\|^2\} \\ &= \arg\min_{\theta\in\Theta}\{\|\theta_t - \eta_w \tilde{g}_t - \theta\|^2\}\end{aligned}$$

which is equivalent to projecting $\theta_t - \eta_w \tilde{g}_t$ onto the convex set Θ. By properties of projection onto convex sets [see e.g. Orabona, 2023, Proposition 2.11], for any $1 \leq t < T$, we have

$$\begin{aligned}\|\theta_{t+1} - \theta_t\| &\leq \|(\theta_t - \eta_w \tilde{g}_t) - \theta_t\| \\ &= \eta_w \|\tilde{g}_t\| \\ &\leq \eta_w G,\end{aligned} \tag{5.58}$$

where the last inequality is $\|\tilde{g}_t\| \leq G$ by the Lipschitzness of the loss function ℓ. Combining (5.58) and triangle inequality, we obtain

$$\begin{aligned}\|\theta_{t+\sigma'} - \theta_t\| &\leq \|\theta_{t+\sigma'} - \theta_{t+\sigma'-1}\| + \|\theta_{t+\sigma'-1} - \theta_t\| \\ &\leq \underbrace{\|\theta_{t+\sigma'} - \theta_{t+\sigma'-1}\| + \|\theta_{t+\sigma'-1} - \theta_{t+\sigma'-2}\| + \cdots + \|\theta_{t+1} - \theta_t\|}_{\sigma' \text{ elements}} \\ &\leq \sigma' \eta_w G \\ &\leq \frac{0.1\lambda}{G},\end{aligned}$$

where the last inequality is due to $\sigma' \leq \lfloor \frac{0.1\lambda}{\eta_w G^2} \rfloor \leq \frac{0.1\lambda}{\eta_w G^2}$. This implies that $\theta_{t+\sigma'} \in \mathcal{B}(\theta_t, \frac{0.1\lambda}{G})$. By Lemma 5.B.8, it follows that if a set is 0.4λ-dominant at θ_t, then it is also a 0.2λ-dominant set at the hypotheses $\theta_{t+1}, \theta_{t+2}, \ldots, \theta_{t+\sigma}$ played in σ subsequent rounds of the game. □

Finally, we show the proof of Theorem 5.C.1.

Proof (of Theorem 5.C.1). Since the maximum number of rounds in each episode is $\sigma \leq \frac{0.1\lambda}{\eta_w G^2}$, there are at most $\frac{T}{\sigma}$ episodes. Combining Lemma 5.C.2, Lemma 5.C.3 and taking a

union bound over $\frac{T}{\sigma}$ episodes, in total we draw

$$O\left(\frac{\eta_w KTG^2 \ln\left(\frac{KT}{\sigma\delta}\right)}{\lambda^3}\right) \tag{5.59}$$

samples over $\frac{T}{\sigma}$ episodes to guarantee that with probability at least $1-\delta/2$, all the computed sets over $\frac{T}{\sigma}$ episodes are dominant sets at $(\theta_t)_{t=1,2,\ldots,T}$ with sizes no larger than $\beta_{0.4\lambda}$. Plugging $\eta_w = \frac{2D}{G\sqrt{T}}$ and $\sigma = \frac{0.1\lambda}{\eta_w G^2} = \frac{0.1\lambda\sqrt{T}}{2DG}$ into (5.59), we obtain a sample complexity of order

$$O\left(\frac{\eta_w KTG^2 \ln\left(\frac{KT}{\sigma\delta}\right)}{\lambda^3}\right) = O\left(\frac{DKG\sqrt{T}\ln\left(\frac{KDG\sqrt{T}}{\lambda\delta}\right)}{\lambda^3}\right). \tag{5.60}$$

From Corollary 5.A.5, we have $T = O(\frac{(D^2G^2+\beta)\ln(K/\delta)}{\epsilon^2})$ is sufficient for obtaining an ϵ-optimal hypothesis with probability at least $1-\frac{\delta}{2}$. By plugging $T = O(\frac{(D^2G^2+\beta)\ln(K/\delta)}{\epsilon^2})$ into (5.60), we obtain the number of samples collected for computing the dominant sets over $\frac{T}{\sigma}$ episodes is

$$O\left(\frac{DKG\sqrt{(D^2G^2+\beta)\ln(K/\delta)}\ln\left(\frac{KDG}{\epsilon\lambda\delta}\right)}{\lambda^3\epsilon}\right). \tag{5.61}$$

In addition, each of the T rounds uses exactly one sample to compute the outputs of the two players in the next round. Hence, with probability at least $1-\delta$, the total sample complexity of the two-player zero-sum game needed to return an ϵ-optimal hypothesis is of order

$$O\left(\frac{DKG\sqrt{(D^2G^2+\beta)\ln(K/\delta)}\ln\left(\frac{KDG}{\epsilon\lambda\delta}\right)}{\lambda^3\epsilon} + \frac{(D^2G^2+\beta)\ln(K/\delta)}{\epsilon^2}\right).$$

□

5.D FTARL with Time-Varying Learning Rates

We consider a variant of the FTARL algorithm in [Nguyen and Mehta, 2024] with time-varying learning rates. The procedure is given in Algorithm 5.11. The only difference between this algorithm and the `FTARLShannon` algorithm (Algorithm 5.5) is that Algorithm 5.11 uses

Algorithm 5.11 `FTARL`: Follow the regularized and active leader with α-Tsallis entropy regularizer and time-varying learning rates for sleeping bandits

Input: $K \geq 2$, α-Tsallis entropy function $\psi(x) = \frac{1-\sum_{i=1}^{K} x_i^\alpha}{1-\alpha}$
Initialize $\tilde{L}_{i,0} = 0$ for all arms $i \in [K]$.
for each round $t = 1, \ldots,$ **do**
  The non-oblivious adversary selects and reveals $\mathbb{A}_t$
  Compute $q_t = \arg\min_{q \in \Delta_K} \psi_t(q) + \langle q, \tilde{L}_{t-1} \rangle$
  Compute $p_{i,t} = \frac{I_{i,t} q_{i,t}}{\sum_{j=1}^{K} I_{j,t} q_{j,t}}$ by Equation (5.18)
  Draw arm $i_t \sim p_t$ and observe $\hat{\ell}_t = \ell_{i_t,t}$
  for each arm $i \in [K]$ **do**
    If $I_{i,t} = 1$, compute $\tilde{\ell}_{i,t} = \frac{\mathbb{1}\{i_t = i\}\hat{\ell}_t}{p_{i,t} + \gamma_t}$ by Equation (5.19)
    If $I_{i,t} = 0$, compute $\tilde{\ell}_{i,t} = \hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}$ by Equation (5.20)
    Update $\tilde{L}_{i,t} = \tilde{L}_{i,t-1} + \tilde{\ell}_{i,t}$

the α-Tsallis entropy regularizer to compute the weight q_t as follows:

$$q_t = \arg\min_{q \in \Delta_K} \psi_t(q) + \langle q, \tilde{L}_{t-1} \rangle, \tag{5.62}$$

where $\psi_t(q) = \frac{\psi(q) - \min_{v \in \Delta_K} \psi(v)}{\eta_t}$ for $\psi(q) = \frac{1-\sum_{i=1}^{K} q_i^\alpha}{1-\alpha}$ and $\alpha \in (0, 1)$ is a constant. The computation of the sampling probability p_t and the loss estimates of active and non-active arms $\tilde{\ell}_{i,t}$ are identical to that of `FTARLShannon`. Since the α-Tsallis entropy tends to Shannon entropy when $\alpha \to 1$ [see e.g. Nielsen and Nock, 2011], we will prove the following high-probability per-action regret bound of Algorithm 5.11 and then take the limit $\alpha \to 1$ to obtain Theorem 5.A.3.

Theorem 5.D.1. Let $(\eta_t)_{t=1,\ldots}$ and $(\gamma_t)_{t=1,\ldots}$ be two sequences of non-increasing learning rates and exploration factors such that $\eta_t \leq 2\gamma_t$. With probability at least $1 - \delta$, `FTARL` (Algorithm 5.11) guarantees that

$$\max_{a \in [K]} \text{Regret}(a) \leq \frac{K^{1-\alpha} - 1}{\eta_T (1 - \alpha)} + \frac{\ln(3K/\delta)}{2\gamma_T} + \left(\frac{1}{2\alpha} + \frac{1}{2}\right) \ln(3/\delta) + \sum_{t=1}^{T} \left(\frac{\eta_t}{2\alpha} + \gamma_t\right) A_t$$

Before proving Theorem 5.D.1, similar to [Nguyen and Mehta, 2024], we state the following results on the concentration bound of the IX-loss estimator. These results are adapted from Neu [2015, Lemma 1 and Corollary 1] in the non-sleeping bandits setting to the sleeping

bandits setting with nearly identical proofs.

Lemma 5.D.2 (Lemma 1 of [Neu, 2015])**.** Let $(\nu_{i,t})$ be non-negative random variables satisfying $\nu_{i,t} \leq 2\gamma_t$ for all $i \in [K]$ and $t \geq 1$. With probability at least $1 - \delta'$,

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \nu_{i,t} \mathbb{1}\{I_{i,t} > 0\}(\tilde{\ell}_{i,t} - \ell_{i,t}) \leq \ln(1/\delta').$$

Since the sequence $(\gamma_t)_{t=1,\dots}$ is non-increasing, we have $\gamma_T \leq \gamma_t$ for all $t \leq T$. Hence, for any fixed arm $a \in [K]$, we can apply Lemma 5.D.2 with $\nu_{i,t} = 2\gamma_T \mathbb{1}\{i = a\} \leq 2\gamma_t$ and take a union bound over K arms to obtain the following corollary.

Corollary 5.D.3. With probability at least $1 - \delta'$, simultaneously for all $a \in [K]$,

$$\sum_{t=1}^{T} I_{a,t}(\tilde{\ell}_{a,t} - \ell_{a,t}) \leq \frac{\ln(K/\delta')}{2\gamma_T}$$

We turn to the proof of Theorem 5.D.1.

Proof (of Theorem 5.D.1). Fix an arm $a \in [K]$ and let e_a be the a-th standard basis vector of $\mathbb{R}^K$. Let $\tilde{\ell}_t = \begin{bmatrix} \tilde{\ell}_{1,t} \\ \tilde{\ell}_{2,t} \\ \dots \\ \tilde{\ell}_{K,t} \end{bmatrix}$ be the vector of estimated losses of K arms in round t. Since the sequence of learning rates is non-increasing and positive, we have $\psi_t(x) \geq 0$ and $\psi_{t+1}(x) \geq \psi_t(x)$ for all $x \in \Delta_K$. Hence, we can invoke the standard local-norm analysis of FTRL with Tsallis entropy regularizer [e.g. Orabona, 2023, Lemma 7.14] on non-negative loss estimates $(\tilde{L}_t)_{t=1,\dots,}$ to obtain

$$\sum_{t=1}^{T} \langle \tilde{\ell}_t, q_t - e_a \rangle \leq \psi_{T+1}(e_a) - \min_{x \in \Delta_K} \psi_1(x) + \frac{1}{2\alpha} \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\alpha} \tag{5.63}$$

Following the proof of Nguyen and Mehta [2024, Lemma 26] and by definition of $\tilde{\ell}_t$, we obtain

$$\begin{aligned}
\langle \tilde{\ell}_t, q_t \rangle &= \sum_{i \in \mathbb{A}_t} \tilde{\ell}_{i,t} q_{i,t} + \sum_{i \notin \mathbb{A}_t} \tilde{\ell}_{i,t} q_{i,t} \\
&= \tilde{\ell}_{i_t,t} q_{i_t,t} + \sum_{i \notin \mathbb{A}_t} \tilde{\ell}_{i,t} q_{i,t} \\
&= \tilde{\ell}_{i_t,t} q_{i_t,t} + \left(\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) \sum_{i \notin \mathbb{A}_t} q_{i,t} \\
&= \tilde{\ell}_{i_t,t} p_{i_t,t} \sum_{i \in \mathbb{A}_t} q_{i,t} + \left(\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) \sum_{i \notin \mathbb{A}_t} q_{i,t} \\
&= \left(\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} \right) \sum_{i=1}^{K} q_{i,t} \\
&= \hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t},
\end{aligned}$$

where the second equality is due to $\tilde{\ell}_{i,t} = 0$ for $i \in \mathbb{A}_t, i \neq i_t$, the third equality is due to $\tilde{\ell}_{i,t} = \hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t}$ for all non-active arms $i \notin \mathbb{A}_t$, the second-to-last equality is due to

$$\tilde{\ell}_{i_t,t} p_{i_t,t} = \frac{p_{i_t,t} \hat{\ell}_t}{p_{i_t,t} + \gamma_t} = \hat{\ell}_t - \frac{\gamma_t \hat{\ell}_t}{p_{i_t,t} + \gamma_t} = \hat{\ell}_t - \gamma_t \tilde{\ell}_{i_t,t} = \hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t},$$

and the last equality is due to $q \in \Delta_K$. Plugging this into (5.63) and using $\langle \tilde{\ell}_t, e_a \rangle = \tilde{\ell}_{a,t}$ implies that

$$\sum_{t=1}^{T} \left(\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} - \tilde{\ell}_{a,t} \right) \leq \psi_{T+1}(e_a) - \min_{x \in \Delta_K} \psi_1(x) + \frac{1}{2\alpha} \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\alpha}.$$

By the definition of the loss estimate for non-active arms in (5.20), in the rounds where $I_{a,t} = 0$, we have $\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} - \tilde{\ell}_{a,t} = 0$. It follows that

$$\begin{aligned}
\sum_{t=1}^{T} I_{a,t} \left(\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} - \tilde{\ell}_{a,t} \right) &= \sum_{t=1}^{T} \left(\hat{\ell}_t - \gamma_t \sum_{j \in \mathbb{A}_t} \tilde{\ell}_{j,t} - \tilde{\ell}_{a,t} \right) \\
&\leq \psi_{T+1}(e_a) - \min_{x \in \Delta_K} \psi_1(x) + \frac{1}{2\alpha} \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\alpha}.
\end{aligned} \tag{5.64}$$

By the non-negativity of the regularizer function, we have

$$\begin{aligned}\psi_{T+1}(e_a) - \min_{x\in\Delta_K} \psi_1(x) &= \psi_{T+1}(e_a) \\ &= \frac{1}{\eta_{T+1}}\left(\psi(e_a) - \min_{v\in\Delta_K}\psi(v)\right) \\ &= \frac{K^{1-\alpha}-1}{\eta_{T+1}(1-\alpha)},\end{aligned}$$

where the third equality is from $\psi(e_a) = 0$ and $\min_{v\in\Delta_K}\psi(v) = \frac{1-K^{1-\alpha}}{1-\alpha}$ by properties of Tsallis entropy function [Abernethy et al., 2015]. Since the round $T+1$ does not contribute to the total regret, we can set $\eta_{T+1} = \eta_T$ and obtain $\psi_{T+1}(e_a) - \min_{x\in\Delta_K}\psi_1(x) \le \frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)}$. Furthermore, by Lemma 10 in [Nguyen and Mehta, 2024], for all $t \ge 1$,

$$\sum_{i=1}^{K} \tilde{\ell}_{i,t}^2 q_{i,t}^{2-\alpha} \le \sum_{j\in\mathbb{A}_t} \tilde{\ell}_{j,t}^2 p_{j,t}^{2-\alpha}.$$

It follows that the right-hand side in (5.64) can be further bounded by

$$\begin{aligned}\sum_{t=1}^{T} I_{a,t}\left(\hat{\ell}_t - \gamma_t\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t} - \tilde{\ell}_{a,t}\right) &\le \frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)} + \frac{1}{2\alpha}\sum_{t=1}^{T}\eta_t\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}^2 p_{j,t}^{2-\alpha} \\ &\le \frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)} + \frac{1}{2\alpha}\sum_{t=1}^{T}\eta_t\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t} p_{j,t}^{1-\alpha} \\ &\le \frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)} + \frac{1}{2\alpha}\sum_{t=1}^{T}\eta_t\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\end{aligned}$$

where the second inequality is due to $\tilde{\ell}_{j,t}p_{j,t} = \frac{\hat{\ell}_t \mathbb{1}\{i_t=j\}p_{j,t}}{p_{j,t}+\gamma_t} \le 1$ for all $j \in \mathbb{A}_t$ and the last inequality is due to $p_{j,t}^{1-\alpha} \le 1$ for $p_{j,t} \in [0,1]$ and $\alpha \in (0,1)$. Moving $\sum_{t=1}^{T} I_{a,t}\gamma_t \sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}$ to the right-hand side and using $I_{a,t} \le 1$, we obtain

$$\sum_{t=1}^{T} I_{a,t}(\hat{\ell}_t - \tilde{\ell}_{a,t}) \le \frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)} + \sum_{t=1}^{T}\left(\frac{\eta_t}{2\alpha}+\gamma_t\right)\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}. \tag{5.65}$$

We then apply Lemma 5.D.2 twice and Corollary 5.D.3 once, each of them with $\delta' = \frac{\delta}{3}$. The first application of Lemma 5.D.2 uses $\nu_{i,t} = \eta_t \le 2\gamma_t$ and obtains with probability at least

$1-\delta/3$,

$$\sum_{t=1}^{T}\eta_t\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\le\ln\left(\frac{3}{\delta}\right)+\sum_{t=1}^{T}\eta_t\sum_{j\in\mathbb{A}_t}\ell_{j,t}. \tag{5.66}$$

The second application of Lemma 5.D.2 uses $\nu_{i,t}=2\gamma_t$ and obtains with probability at least $1-\delta/3$,

$$\sum_{t=1}^{T}2\gamma_t\sum_{j\in\mathbb{A}_t}\tilde{\ell}_{j,t}\le\ln\left(\frac{3}{\delta}\right)+\sum_{t=1}^{T}2\gamma_t\sum_{j\in\mathbb{A}_t}\ell_{j,t}. \tag{5.67}$$

An application of Corollary 5.D.3 leads to

$$\sum_{t=1}^{T}I_a\tilde{\ell}_{a,t}\le\frac{\ln(3K/\delta)}{2\gamma_T}+\sum_{t=1}^{T}I_a\ell_{a,t} \tag{5.68}$$

with probability at least $1-\delta/3$. Plugging (5.66), (5.67) and (5.68) into (5.65) and taking a union bound, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T}I_{a,t}(\hat{\ell}_t-\ell_{a,t})&\le\frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)}+\frac{\ln(3K/\delta)}{2\gamma_T}+\left(\frac{1}{2\alpha}+\frac{1}{2}\right)\ln(3/\delta)+\sum_{t=1}^{T}\left(\frac{\eta_t}{2\alpha}+\gamma_t\right)\sum_{j\in\mathbb{A}_t}\ell_{j,t}\\&\le\frac{K^{1-\alpha}-1}{\eta_T(1-\alpha)}+\frac{\ln(3K/\delta)}{2\gamma_T}+\left(\frac{1}{2\alpha}+\frac{1}{2}\right)\ln(3/\delta)+\sum_{t=1}^{T}\left(\frac{\eta_t}{2\alpha}+\gamma_t\right)A_t,\end{aligned}$$

holds for simultaneously for all $a\in[K]$, where the last inequality is $\sum_{j\in\mathbb{A}_t}\ell_{j,t}\le\sum_{j\in\mathbb{A}_t}1=A_t$. □

Finally, we prove Theorem 5.A.3.

Proof (of Theorem 5.A.3). Since Theorem 5.D.1 holds for any α arbitrarily close to 1, we can take the limit of α to 1 on the right-hand side of its bound and obtain the desired bound in Theorem 5.A.3:

$$\max_{a\in[K]}\mathrm{Regret}(a)\le\frac{\ln(K)}{\eta_T}+\frac{\ln(3K/\delta)}{2\gamma_T}+\ln(3/\delta)+\sum_{t=1}^{T}\left(\frac{\eta_t}{2}+\gamma_t\right)A_t. \tag{5.69}$$

□

5.E Stochastic OMD with non-increasing, time-varying learning rate

[Zhang et al., 2023a] lamented that they could not find an analysis of stochastic mirror descent for non-oblivious online convex optimization with stochastic gradients, and they therefore proved their own high probability result. Their result uses a fixed learning rate, whereas we would like to avoid needing knowledge of the time horizon T and therefore will describe how one can trivially (in light of known results) extend their derivation to the case of a non-increasing learning rate. All that is needed is to extend their upper bounds in equations (40) and (44) in their work to the case of a non-increasing learning rate $\eta_{w,t}$ (so that $\eta_{w,t+1} \leq \eta_{w,t}$ for $t \in [T]$). Such an extension is for free using, e.g., Theorem 6.10 of [Orabona, 2023]. All other steps of the proof of Theorem 2 of [Zhang et al., 2023a] can proceed without any important modifications, including the application of the Hoeffding-Azuma inequality. Here, we just highlight a few keyframes of the proof.

Using our notation and with non-increasing learning rate sequence $(\eta_{w,t})_{t\geq 1}$ and applying Theorem 6.10 of [Orabona, 2023], the bound in equation (40) of [Zhang et al., 2023a] becomes, for any $\theta \in \Theta$,

$$\sum_{t=1}^{T} \langle \tilde{g}_t, \theta_t - \theta \rangle \leq \frac{D^2}{\eta_{w,T}} + \frac{G^2}{2} \sum_{t=1}^{T} \eta_{w,t}.$$

Fastforwarding to our analogue of equation (42) of [Zhang et al., 2023a], we now get

$$\max_{\theta\in\Theta} \sum_{t=1}^{T} \big[\phi(\theta_t, q_t) - \phi(\theta, q_t)\big] \leq \frac{D^2}{\eta_{w,T}} + \frac{G^2}{2} \sum_{t=1}^{T} \eta_{w,t} + \max_{\theta\in\Theta} \left\{ \sum_{t=1}^{T} \big\langle \nabla_\theta \phi(\theta_t, q_t) - \tilde{g}_t, \theta_t - \theta \big\rangle \right\}.$$

Setting

$$\tilde{\theta}_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \eta_{w,t} \big\langle \nabla_\theta \phi(\theta_t, q_t) - \tilde{g}_t, \theta - \tilde{\theta}_t \big\rangle + \frac{1}{2} \left\| \theta - \tilde{\theta}_t \right\|^2 \right\},$$

we now again apply Theorem 6.10 of [Orabona, 2023] to get the following analogue of equation (44) of [Zhang et al., 2023a]:

$$\sum_{t=1}^{T} \big\langle \nabla_\theta \phi(\theta_t, q_t) - \tilde{g}_t, \tilde{\theta}_t - \theta \big\rangle \leq \frac{D^2}{\eta_{w,T}} + 2G^2 \sum_{t=1}^{T} \eta_{w,t}.$$

All of the remaining steps of the analysis of [Zhang et al., 2023a] go through without any interesting modification, giving the result that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\phi(\theta_t,q_t)-\min_{\theta\in\Theta}\sum_{t-1}^{T}\phi(\theta,q_t)\leq\frac{D^2}{\eta_{w,T}}+\frac{G^2}{2}\sum_{t=1}^{T}\eta_{w,t}+\frac{D^2}{\eta_{w,T}}+2G^2\sum_{t=1}^{T}\eta_{w,t}+8DG\sqrt{T\ln\frac{1}{\delta}},$$

where the last term is from applying the Hoeffding-Azuma inequality in precisely the same way as in [Zhang et al., 2023a]. Using a learning rate schedule of $\eta_{w,t}=\eta_0\cdot\frac{D}{G\sqrt{t}}$ gives the upper bound

$$DG\sqrt{T}\left(\frac{1}{\eta_0}+\eta_0+\frac{1}{\eta_0}+4\eta_0+\sqrt{\ln\frac{1}{\delta}}\right)=DG\sqrt{T}\left(\frac{2}{\eta_0}+5\eta_0+\sqrt{\ln\frac{1}{\delta}}\right),$$

which, letting $\eta_0=1$, gives

$$DG\sqrt{T}\left(7+\sqrt{\ln\frac{1}{\delta}}\right)=O\left(DG\sqrt{T\ln\frac{1}{\delta}}\right),$$

as desired.

5.F Details of the Experiments

In this section, we provide the full setup details of the experiments presented in Section 5.5.

5.F.1 The Lower Bound Environment

For the GDRO problem instance constructed based on the lower bound construction in the proof of Theorem 5.3.5, we scale the loss by $\frac{1}{2}$ to ensure that the losses are in $[0,1]$. This implies that for a hypothesis $\theta\in[0,1]$, its maximum risk over K groups is

$$\mathcal{L}(\theta)=\frac{1}{2}\max\left(\Delta\theta+\frac{1}{2},\Delta(1-\theta)+\frac{1}{2}\right).$$

We set $\Delta=0.1$. The optimal hypothesis is $\theta^*=0.5$ with $\mathcal{L}(\theta^*)=\frac{1.1}{4}=0.275$. The optimality gap of a hypothesis θ is

$$\mathrm{err}(\theta)=\mathcal{L}(\theta)-\mathcal{L}(\theta^*)=\frac{1}{2}\Delta\left|\frac{1}{2}-\theta\right|=0.05\left|\frac{1}{2}-\theta\right|.$$

With the desired optimality gap of $\epsilon = 0.005$, the acceptable range of the risk of $\bar{\theta}_T$ is $[0.27, 0.28]$. The set of ϵ-optimal hypotheses is obtained by solving $0.05\left|\frac{1}{2} - \theta\right| \leq 0.005$, which implies that $\theta \in [0.4, 0.6]$ is the set of ϵ-optimal hypotheses.

5.F.2 The Adult Dataset

Loss function and data normalization. Similar to [Soma et al., 2022], we train a linear classifier with hinge loss

$$\ell(\theta, z, y) = \max(0, 1 - y\langle\theta, z\rangle),$$

where $z \in \mathbb{R}^5$ is a feature vector of a sample and $y \in \{-1, 1\}$ is the label.

On the Adult dataset, the default value of the features could be much larger than 1, leading to loss values larger than 1. To avoid exceedingly large losses, we compute the maximum norm of all feature vectors in the dataset and then divide all features by this maximum norm. Note that the same maximum norm value is used for all 10 groups.

UCI Adult Dataset As mentioned in the main text, we construction $K = 10$ groups from five races `White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other` and two genders `male, female`. The dataset of 48 842 samples is heavily imbalanced. The largest group is `(White, male)` having 28 736 samples while the smallest group is `(Other, female)` having 156 samples.

No batch processing. Our results in Section 5.5 are generated by the exact algorithms described in Sections 5.3 and 5.4 without adding any batch processing. This is different from [Soma et al., 2022], who used a batch of 10 samples to stabilize the gradients. We find that as the dominant sets quickly converge to just one or two groups, especially the groups with small amount of samples such as `(Amer-Indian-Eskimo, female)`, the gradients computed from just one random sample are sufficiently stable with the long horizon of $T = 10^6$.

Computing (an estimate of) θ^*. In order to obtain the optimality gap of `SB-GDRO-SA` and `SMD-GDRO`, we compute an estimate of $\mathcal{L}(\theta^*)$ using the following algorithm: we run a deterministic two-player zero-sum game in which both players have full knowledge of $(\mathcal{P}_i)_{i=1,2,\ldots,K}$. In each round t, the max-player is able to compute a dominant set consisting of just one group – that is, the group with maximum risk on θ_t. Similarly, the min-player is given the expected value of the gradient $\mathbb{E}[\tilde{g}_t]$ instead of the stochastic gradients. We run the game for $T = 10^7$ rounds and record the final maximum risk of $\bar{\theta}_T$ to be $\mathcal{L}(\theta^*) \approx 0.49945$. This final maximum risk of is observed to be on group 8 (i.e., female Amer-Indian-Eskimo).

5.G Discussion of the Competing Approach in Stochastically Constrained Adversarial Regime

Our approach to going beyond minimax bounds in GDRO is based on the (λ, β)-sparsity condition and, algorithmically, based on the sleeping bandits framework. The expert bandit reader may wonder about the viability of the following competing approach: suppose that after some unknown time horizon τ, all θ_t's fall within a radius-ρ ball of θ^*, and within such a ball, further suppose for simplicity that a unique group obtains the maximum risk in all subsequent rounds by a margin of at least λ. This setup generalizes the previously studied stochastically constrained adversarial (SCA) regime [Wei and Luo, 2018a; Zimmert and Seldin, 2021b] wherein the best arm's mean is separated with a gap from the other arms' means *for all rounds*. In this generalized SCA regime, one might hope for better regret bounds for the max player than we achieve using our sleeping bandits-based approach. However, there are at least three major challenges: first, to our knowledge, it is not known how to get high probability regret bounds in the SCA regime even when $\tau = 1$; second, we are not aware of results that provide last iterate convergence so that, eventually, all iterates θ_t are within distance ρ of θ^* (SCA requires such convergence); third, there could well be multiple best arms or multiple nearly best arms, which recently has been addressed in some different regimes but adds another layer of complexity for the generalized SCA regime.

As mentioned in Section 5.6, if we had last iterate convergence, then our (λ, β)-sparsity condition could be relaxed to hold only within some proximity of θ^*. However, our condition is more flexible as compared to SCA since it is not known how the latter can be analyzed when the best arm (or set of best arms) changes throughout the game, whereas such a changing set of approximate (within gap λ) maximizers fits naturally with sleeping bandits.

Chapter 6

Data-dependent bounds with T-optimal best-of-both-worlds guarantees in multi-armed bandits using stability-penalty matching

6.1 Introduction

The multi-armed bandits problem [Lai and Robbins, 1985; Auer et al., 2002a] is one of the most fundamental frameworks for modeling sequential decision making problems under limited feedback. In this problem, a learner sequentially interacts with the environment in T rounds. In round $t = 1, 2, \ldots,$ the learner chooses an action I_t from a set of K available actions and observes a numerical feedback $\ell_{t,I_t} \in \mathbb{R}$. This ℓ_{t,I_t} is an element of a hidden vector $\ell_t \in \mathbb{R}^K$ chosen at the beginning of round t by an oblivious adversary. The performance of the learner is its *pseudo-regret*

$$R_T = \max_{a \in [K]} R_{T,a} = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,a}\right], \tag{6.1}$$

where $\mathbb{E}$ denote the expectation taken over all randomness from all T rounds. Existing works have constructed algorithms with *worst-case* regret bounds that hold under the assumption on whether the adversary is adversarial (i.e., $(\ell_t)_t$ are arbitrary) or stochastic (i.e., $(\ell_t)_t$ are drawn i.i.d. from some distribution) [Lai and Robbins, 1985; Auer et al., 2002a,b], *best-of-both-worlds* (BOBW) bounds that have worst-case guarantees simultaneously for adversarial

and stochastic adversaries [e.g. Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021a; Dann et al., 2023; Ito et al., 2024], or *data-dependent* bounds that are adaptive to the sequence $(\ell_t)_t$ [e.g. Wei and Luo, 2018b; Bubeck et al., 2018; Ito, 2021; Ito et al., 2022; Tsuchiya et al., 2023]. Despite this vast amount of literature on different types of worst-case and adaptive bounds for multi-armed bandits, we are not aware of any works that establish bounds that are *simultaneously* data-dependent, best-of-both-worlds *and* have optimal dependency on T. In particular, existing works suffer from at least one of three limitations: being data-dependent but not BOBW [Hazan and Kale, 2011; Bubeck et al., 2018; Wei and Luo, 2018b], being BOBW but not data-dependent [Zimmert and Seldin, 2021a; Dann et al., 2023] or having sub-optimal dependency on T [Hazan and Kale, 2011; Wei and Luo, 2018b; Tsuchiya et al., 2023; Ito et al., 2024]. In this work, we close this gap in the literature by introducing novel algorithms with regret bounds that are simultaneously BOBW, data-dependent and T-optimal.

All of our algorithms are established in the Follow-the-Regularized-Leader (FTRL) framework [see e.g. Lattimore and Szepesvári, 2020], in which the time-varying learning rates are tuned by the Stability-Penalty Matching (SPM) method. SPM was originally proposed by [Ito et al., 2024] as a principled method for tuning learning rates in FTRL using both the *penalty* and *stability* terms. More specifically, in round t, our algorithms compute a probability vector

$$q_t = \arg\min_{x \in \Delta_K} \langle L_{t-1}, x \rangle + \phi_t(x), \tag{6.2}$$

where $\Delta_K = \{x \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1\}$ denotes the K-dimensional simplex, $L_{t-1} \in \mathbb{R}^K$ is the estimated cumulative loss vector up to round $t-1$ and $\phi_t(x) : \Delta_K \to \mathbb{R}$ is the regularization function. We use the following specific form for ϕ_t:

$$\phi_t(x) = \beta_t f(x) + \gamma u(x), \tag{6.3}$$

where $f(x) : \Delta_K \to \mathbb{R}_-, u(x) : \Delta_K \to \mathbb{R}_+$ are convex, $\beta_t > 0$ is the learning rate in round t and γ is a constant. Then, the learner draws an arm $I_t \sim q_t$ according to q_t (or some $p_t \in \Delta_K$ derived from q_t) and computes an estimated loss vector $\hat{\ell}_t$. Let $D_t(x, y) = \phi_t(x) - \phi_t(y) - \langle \nabla \phi_t(y), x - y \rangle$ denote the Bregman divergence associated with ϕ_t. The standard analysis of FTRL [e.g. Lattimore and Szepesvári, 2020, Exercise 28.12] implies

that

$$R_{T,a} \lesssim \phi_{T+1}(e_a) - \phi_1(q_1) + \mathbb{E}\left[\sum_{t=1}^{T}(\phi_t(q_{t+1}) - \phi_{t+1}(q_{t+1})) + \sum_{t=1}^{T}(\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t))\right]$$
$$\lesssim \gamma u(e_a) - \beta_1 f(q_1) + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(\beta_{t+1} - \beta_t)h_{t+1}\right]}_{\text{penalty term}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right]}_{\text{stability term}}$$

where e_a is the a-th vector in the standard basis of $\mathbb{R}^K$, h_{t+1} satisfies $(-f(q_{t+1})) \lesssim h_{t+1}$ and z_t satisfies $\beta_t \mathbb{E}[\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)] \lesssim \mathbb{E}[z_t]$. SPM carefully chooses β_1, z_t and h_t so that $h_{t+1} \leq O(h_t)$ and sets the learning rate of the next round to be

$$\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}. \tag{6.4}$$

This makes $(\beta_{t+1} - \beta_t)h_{t+1}$ match with $\frac{z_t}{\beta_t}$ and implies $R_{T,a} \lesssim \gamma u(e_a) - \beta_1 f(q_1) + \mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right]$. An important insight in SPM is that by picking $f(x)$ and $u(x)$ appropriately, $\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right]$ is naturally adaptive to the adversarial or stochastic nature of the environment [Ito et al., 2024]. In our work, we will show that SPM can be made adaptive not only to the nature of the environment but also to the underlying structure of the sequence of losses such as sparsity and total variation.

6.1.1 Main Contributions and Techniques

Throughout the paper, we will write $O(\square \ln(T), \square\sqrt{T})$ to denote a BOBW bound that holds for stochastic and adversarial regimes, respectively, where $\square$ contains problem-dependent terms. The original SPM method [Ito et al., 2024] used $z_t = \Omega(\beta_t \mathbb{E}_{I_t}[\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)])$, where $\mathbb{E}_{I_t}$ denotes an expectation taken over I_t. Because only one out of K arms is observable in each round t, this in-expectation form of z_t inevitably requires taking the trivial bounds (e.g. 1) of the losses into its computation, thus limiting its adaptivity to $(\ell_t)_t$. Our work overcomes this limitation by setting

$$z_t = \Omega(\beta_t(\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t))). \tag{6.5}$$

We call this *real-time SPM*, since z_t depends on the observed arm I_t. The main technical challenge is now z_t can be very large since it grows with $\text{poly}(\frac{1}{p_{t,I_t}})$. At the same time, we need to limit the amount of explicit exploration to obtain a BOBW bound for stochastic

bandits. Table 6.1 summarizes our main results, showing that real-time SPM can be controlled effectively to give BOBW *and* data-dependent bounds with optimal dependency on T. Our results also hold for the more general adversarial regime with self-bounding constraint setting [Zimmert and Seldin, 2021a]. Appendix 6.A gives a more detailed discussion on related works. Our paper is organized as follows:

- Section 6.2 introduces the real-time SPM method and states Lemma 6.2.1, a key technical lemma for bounding $\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right]$. While the original analysis of SPM [Ito et al., 2024] relies on having a small $\max_{t\in[T]} z_t$ and thus cannot be applied to real-time SPM, our Lemma 6.2.1 instead shows that real-time SPM incurs an additional regret of at most $O\left(\max_{t\in[T]} \frac{z_t}{\beta_t} \ln \sum_{t=1}^{T} \frac{z_t}{h_t}\right)$. Moreover, both $\frac{z_t}{\beta_t}$ and $\frac{z_t}{h_t}$ can be effectively controlled by appropriate choices of $\phi_t(x)$.

- Section 6.3 considers the bandits problems with signed sparse losses, where $\ell_{t,i} \in [-1, 1]$ and $\|\ell_t\|_0 \leq S$. We show that using α-Tsallis entropy and log-barrier functions in place of f and g in (6.3) leads to an $O\left(\frac{(K^{1-\alpha}-1)S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}, \left(\sqrt{\frac{(K^{1-\alpha}-1)S^\alpha T}{\alpha(1-\alpha)}}\right)\right)$. This bound is T-optimal and improves upon the best known bound for this setting established by Tsuchiya et al. [2023]. When S is known, we show that the adversarial bound is improved to $O(\sqrt{ST\ln(K/S)})$, resolving an open question in Kwon and Perchet [2016]. Furthermore, we prove a near-matching lower bound for problems in which the sparsity constraint holds in expectation.

- Section 6.4 considers problems with small total variation Q (defined in Section 6.1.2) and presents a new algorithm obtaining a $O\left(\frac{(K-1)^{1-\alpha}K^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}, \sqrt{Q\ln(K)}\right)$ BOBW bound. In the adversarial regime, the $O(\sqrt{Q\ln(K)})$ bound matches the best known bound in Bubeck et al. [2018] while having the advantage of not requiring knowledge of Q.

- Section 6.5 introduces a new SPM method called coordinate-wise SPM (CoWSPM), which maintain arm-dependent learning rates $\beta_{t,i}$ and performs real-time SPM on each arm separately. We show that CoWSPM achieves a BOBW bound with order $O\left(\frac{1}{\alpha(1-\alpha)} \sum_{i\neq i^*} \frac{\ln(T)}{\Delta_i}\right)$ in stochastic bandits. In adversarial bandits, CoWSPM achieves $O\left(\min\left\{\sqrt{K\ln(T)\min(Q_\infty, L^*, T-L^*)}, K^{\frac{\alpha}{2}}\sqrt{KT}\right\}\right)$ regret bound. where Q_∞ and L^* are ℓ_∞-norm total variation and total loss of the best arm, respectively (see Section 6.1.2 for their formal definitions).

Table 6.1: Summary of data-dependent results in existing and ours works. The three blocks of rows show bounds dependent on sparsity S, total variation Q and a combination of Q_∞ and L^*, respectively (formal definitions are in Section 6.1.2). We use $H^*_\infty = \min(Q_\infty, L^*, T - L^*)$. "$T$-opt BOBW" denote whether a bound is BOBW and T-optimal. "Param-free" denote whether a bound requires knowledge of the data-dependent quantities.

Algorithms	Stochastic	Adversarial	T-opt BOBW?	Param-free?
Bubeck et al. [2018]	–	$\sqrt{ST \ln K}$	×	×
Tsuchiya et al. [2023]	$\frac{S \ln(T) \ln(KT)}{\Delta_{\min}}$	$\sqrt{ST \ln T \ln K}$	×	✓
Theorem 6.3.1	$\frac{S \ln T \ln K}{\Delta_{\min}}$	$\sqrt{ST \ln K}$	✓	✓
Hazan and Kale [2011]	–	$\sqrt{Q \ln T \ln K}$	×	✓
Bubeck et al. [2018]	–	$\sqrt{Q \ln K}$	×	×
Theorem 6.4.1	$\frac{K \ln T}{\Delta_{\min}}$	$\sqrt{Q \ln K}$	✓	✓
Wei and Luo [2018b]	$\frac{K \ln T}{\Delta_{\min}}$	$\sqrt{KL^* \ln T}$	×	✓
Ito [2021]	$\sum_{i \neq i^*} \frac{\ln T}{\Delta_i}$	$\sqrt{K \min(Q_\infty, L^*) \ln T}$	×	✓
Ito et al. [2022]	$\sum_{i \neq i^*} (\frac{\sigma_i^2}{\Delta_i} + 1) \ln T$	$\sqrt{KH^*_\infty \ln T}$	×	✓
Theorem 6.5.1	$\sum_{i \neq i^*} \frac{\ln T}{\Delta_i}$	$\min(\sqrt{KH^*_\infty \ln T}, \sqrt{K^{1+\alpha} T})$	✓	✓

6.1.2 Problem Setup

For an integer N, let $[N] = \{1, 2, \ldots, N\}$ denote the set of integers from 1 to N. We study the multi-armed bandits problem [Lai and Robbins, 1985; Auer et al., 2002a] in which a learner is given K arms and interacts with the environment in T rounds. In each round t, an adversary selects a hidden vector $\ell_t = (\ell_{t,1}, \ell_{t,2}, \ldots, \ell_{t,K})^\top$. The learner chooses one arm $I_t \in [K]$ and observes its loss ℓ_{t,I_t}. We assume $|\ell_{t,i}| \leq 1$ for all $t \in [T], i \in [K]$. The learner aims to minimize its regret R_T over T rounds, defined by Equation (6.1).

We are interested in developing learning algorithms with provable upper bounds on R_T that hold simultaneously for two regimes: adversarial [Auer et al., 2002a] and adversarial with a (Δ, C, T) self-bounding constraint [Zimmert and Seldin, 2021a]. In the *adversarial regime*, no assumption is made on how the adversary generates $(\ell_t)_{t\in[T]}$. The adversarial regime with a (Δ, C, T) self-bounding constraint [Zimmert and Seldin, 2021a] is given below.

Definition 6.1.1. (Adversarial regime with a self-bounding constraint) For $T \geq 1, \Delta \in [0, 1]^K$ and $C \geq 0$, the problem is in adversarial regime with a (Δ, C, T) self-bounding constraint if the regret of any algorithm at time T satisfies $R_T \geq \sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i \Pr(I_t = i) - C$.

As noted in Zimmert and Seldin [2021a], the stochastic bandits setting [Lai and Robbins, 1985] satisfies Definition 6.1.1. We also use the common assumption that there exists an optimal arm i^* such that $\Delta_i > 0$ for all $i \neq i^*$, that is, the optimal arm is unique. Let $\Delta_{\min} = \min_{i\in[K]}\{\Delta_i : \Delta_i > 0\}$.

We focus on obtaining bounds that are adaptive not only to the adversary's regime but also to the data-dependent properties of the loss sequence $(\ell_t)_{t\in[T]}$. The following data-dependent quantities are considered in our work.

- **Sparsity of losses** [Kwon and Perchet, 2016]. All loss vectors have at most $1 \leq S \leq K$ non-zero elements, i.e., $\|\ell_t\|_0 \leq S$, where S is unknown.
- **Variation of losses** [Hazan and Kale, 2011; Ito et al., 2022] The total variation of the sequence $(\ell_t)_t$ is $Q = \sum_{t=1}^{T} \left\|\ell_t - \frac{1}{T}\sum_{s=1}^{T} \ell_s\right\|_2^2$. The ℓ_∞-norm total variation is $Q_\infty = \min_{\bar{\ell}\in[0,1]^K} \sum_{t=1}^{T} \left\|\ell_t - \bar{\ell}\right\|_\infty^2$.
- **Best-arm loss**. For non-negative losses, we consider the cumulative loss of the best arm $L_* = \min_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}$.

6.2 Stability-Penalty Matching with Real-Time Stability Term

Let $\tilde{p} = \min(1-p, p)$ for $p \in [0,1]$. We use the notation $f \lesssim g$ to denote $f = O(g)$. To obtain data-dependent bounds using SPM, we use SPM where the stability term is a function of the *observed* loss, i.e., z_t satisfies Equation (6.5). Note that z_t grows with ℓ_{t,I_t}^2 and $\frac{1}{p_{t,I_t}}$. The benefit of this real-time z_t is that data-dependent quantities such as sparsity and total variation naturally come out of $\mathbb{E}[z_t]$. For example, in Algorithm 6.1 for bandits with sparse losses $\|\ell_t\|_0 \le S$, we use $z_t = O(\tilde{p}_{t,I_t}^{2-\alpha} \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2})$ for some $\alpha \in (0,1)$. It follows that $\mathbb{E}[z_t] = O(\sum_{i:\ell_{t,i}\neq 0} \tilde{p}_{t,i}^{1-\alpha} \ell_{t,i}^2) \le O(S^\alpha)$, leading to the $O(\sqrt{ST \ln K})$ bound. The main challenge in using the real-time z_t is the value of z_t can be unbounded whenever p_{t,I_t} is very small. It follows that $z_{\max} = \max_{t\in[T]} z_t$ can be unbounded, which makes it difficult to apply existing techniques in Ito et al. [2024, Lemma 10] that bounds $\mathbb{E}[\sum_{t=1}^T \frac{z_t}{\beta_t}]$ by a quantity that grows with $\mathbb{E}[z_{\max}]$. We resolve this challenge by using the following technical lemma.

Lemma 6.2.1. For any $T \ge 1, z_{1:T} \ge 0, h_{1:T} > 0$ and a sequence $\beta_{1:T}$ defined by Equation (6.4), let $F(z_{1:T}, h_{1:T}) = \sum_{t=1}^T \frac{z_t}{\beta_t}$ and $G(z_{1:T}, h_{1:T}) = \sum_{t=1}^T \frac{z_t}{\sqrt{\sum_{s=1}^t \frac{z_s}{h_s}}}$. We have

$$F(z_{1:T}, h_{1:T}) \lesssim G(z_{1:T}, h_{1:T}) + \left(\max_{t\in[T]} \frac{z_t}{\beta_t}\right) \ln\left(\sum_{t=1}^T \frac{z_t}{h_t}\right). \tag{6.6}$$

Proof. (Sketch) Our proof extends from the proof of Ito et al. [2024, Lemma 10]. Similar to the proof of Ito et al. [2024, Lemma 10], we define a new sequence $\beta'_t = \sqrt{\beta_1^2 + 2\sum_{s=1}^{t-1} \frac{z_s}{h_s}}$ and consider the set of rounds $E = \{t \in [T] : \beta'_{t+1} \ge \sqrt{2}\beta'_t\}$. The complement of E is $E^c = [T] \setminus E$. We have

$$F(z_{1:T}, h_{1:T}) = \underbrace{\sum_{t\in E^c} \frac{z_t}{\beta_t}}_{(a)} + \underbrace{\sum_{t\in E} \frac{z_t}{\beta_t}}_{(b)},$$

where (a) is bounded by $G(z_{1:T}, h_{1:T})$ as in Ito et al. [2024, Lemma 10], and (b) is bounded by

$$(b) \le \left(\max_{t\in[T]} \frac{z_t}{\beta_t}\right) |E| \le \left(\max_{t\in[T]} \frac{z_t}{\beta_t}\right) \log_{\sqrt{2}}\left(\frac{\beta'_{T+1}}{\beta'_1}\right) \lesssim \left(\max_{t\in[T]} \frac{z_t}{\beta_t}\right) \ln\left(\sum_{t=1}^T \frac{z_t}{h_t}\right),$$

Algorithm 6.1 Real-time SPM with hybrid regularization for losses in $[-1, 1]$.

Input: $K \geq 3, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}, \gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2$.
Initialize $L_{0,i} = 0$ for $i \in [K]$
for each round $t = 1, \ldots, T$ **do**
  Compute $q_t = \arg\min_{p \in \Delta_K} \langle L_{t-1}, p \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(p_i)$
  Compute $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$
  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
  Compute loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\{I_t = i\}}{p_{t,i}}$ and update $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
  Compute $z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min(p_{t,I_t}, 1 - p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{\beta_t 18 d^2}{\gamma} \ell_{t,I_t}^2 \right)$
  Compute $h_t = \left(\frac{1}{\alpha}(\sum_{i=1}^K p_{t,i}^\alpha - 1) \right)$
  Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$
end

where the second inequality is from the fact that β'_t is multiplied by at least $\sqrt{2}$ after every round in E; thus there can be at most $\log_{\sqrt{2}} \frac{\beta'_{T+1}}{\beta'_1}$ such multiplications. □

Lemma 6.2.1 implies that if (I) the sum $\mathbb{E}[\sum_{t=1}^T \frac{z_t}{h_t}]$ grows with $\text{poly}(T)$ and (II) $\max_t \frac{z_t}{\beta_t}$ is small, then $\mathbb{E}[F(z_{1:T}, h_{1:T})]$ grows dominantly with $\mathbb{E}[G(z_{1:T}, h_{1:T})]$ plus an $O(\ln(T))$ term. Hence, we can safely ignore other terms and focus only on bounding $G(z_{1:T}, h_{1:T})$. The proof of Ito et al. [2024, Lemma 10] already showed that

$$G(z_{1:T}, h_{1:T}) \lesssim \min \left\{ \sqrt{\ln(T) \sum_{t=1}^T h_t z_t} + \sqrt{\frac{1}{T} h_{\max} \sum_{t=1}^T z_t}, \sqrt{h_{\max} \sum_{t=1}^T z_t} \right\}. \tag{6.7}$$

In the rest of the paper, we will show that different choices of the (hybrid) regularization function lead to specific forms of z_t and h_t such that not only do both conditions (I) and (II) hold but also they imply BOBW data-dependent bounds with optimal dependency on T from (6.7).

6.3 Application I: BOBW Bounds for Bandits with Sparse Losses

We consider the multi-armed bandits setting with sparse losses [Kwon and Perchet, 2016], in which the loss vector $\ell_t \in [-1, 1]^K$ has at most S non-zero elements, i.e., $\max_{t \in [T]} \|\ell_t\|_0 \leq S$. Note that S is unknown to the learner. Let $\psi_{TE}(p) = \frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)$ be the α-Tsallis

entropy with some $\alpha \in (0,1)$ and $\psi_{LB}(p) = -\sum_{i=1}^{K} \ln(p_i)$ be the log-barrier function. Our approach for this setting is in Algorithm 6.1, in which we use the hybrid regularizer $\phi_t(p) = \beta_t \psi_{TE}(p) + \gamma \psi_{LB}(p)$ to obtain

$$q_t = \arg\min_{p \in \Delta_K} \{\langle L_{t-1}, p \rangle + \beta_t \psi_{TE}(p) + \gamma \psi_{LB}(p)\},$$

Then, we mix q_t with $\frac{1}{T}$-uniform exploration to obtain the sampling probability p_t, i.e., $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$. The learning rates $(\beta_t)_t$ are set by the SPM rule by [Ito et al., 2024] as

$$\beta_1 = \frac{8K}{1-\alpha}, \ \beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}, \tag{6.8}$$

where

$$z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min(p_{t,I_t}, 1-p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \beta_t \frac{18d^2}{\gamma} \ell_{t,I_t}^2\right), \ h_t = (-\psi_{TE}(p_t)), \tag{6.9}$$

and $\gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2$. Note that $\beta_1 \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ for $\omega = 2$. The following theorem states the BOBW bounds of Algorithm 6.1.

Theorem 6.3.1. For any $K \geq 4, T \geq 4k$, Algorithm 6.1 guarantees the following bounds simultaneously

- In the adversarial regime:

$$R_T \leq O\left(\sqrt{\frac{(K^{1-\alpha}-1)S^\alpha T}{\alpha(1-\alpha)}}\right) \tag{6.10}$$

- In the adversarial regime with a self-bounding constraint:

$$R_T \leq O\left(\frac{(K-1)^{1-\alpha} S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha} S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha} S^\alpha}{\alpha(1-\alpha)}}\right) \tag{6.11}$$

In Appendix 6.G, we show that by setting $\alpha = 1 - \frac{1}{2\ln(K)}$, we obtain $\frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)} \lesssim S^\alpha \ln(K)$ and $\frac{(K-1)^{1-\alpha}S^\alpha}{\alpha(1-\alpha)} \lesssim S^\alpha \ln(K)$. In the adversarial regime, Theorem 6.3.1 recovers the $O(\sqrt{ST\ln(K)})$ bound in [Bubeck et al., 2018] and [Tsuchiya et al., 2023] while still being

S-agnostic. In the adversarial regime with a self-bounding constraint, the bound becomes $O(\frac{S\ln(K)\ln(T)}{\Delta_{\min}})$ which has an optimal dependency on T. To the best of our knowledge, Theorem 6.3.1 is the first result for bandits with sparse signed losses that is simultaneously S-agnostic, T-optimal and BOBW. Also, our approach is more computationally efficient than that of [Tsuchiya et al., 2023] as we do not need to solve any additional optimization problems to compute the learning rates $(\beta_t)_t$.

Remark 6.3.2. When S is known, then $\frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)}$ can be further bounded by $6S\ln\left(\frac{K}{S}\right)$. Consider only the case where S is sufficiently small so that $e^2S \leq K$ (the other direction trivially leads to $O(\frac{S}{\alpha(1-\alpha)})$. Letting $\alpha = 1 - \frac{1}{\ln(K/S)}$, then $(\frac{K-1}{S})^{1-\alpha} \leq (\frac{K}{S})^{1-\alpha} = e$. Since $\ln(K/S) \geq 2$, we have $\alpha \geq \frac{1}{2}$. Therefore,

$$\frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)} \leq \frac{K^{1-\alpha}S^\alpha}{\alpha(1-\alpha)} = S\left(\frac{K}{S}\right)^{1-\alpha}\frac{\ln(K/S)}{\alpha} = \frac{eS\ln(K/S)}{\alpha} \leq 6S\ln(K/S).$$

This result shows that an $O(\sqrt{ST\ln(K/S)})$ upper bound is attainable even for signed losses, which resolves an open question posed in [Kwon and Perchet, 2016, Remark 12].

Remark 6.3.3. In Appendix 6.F, we also show that Algorithm 6.1 can be applied in the related setting of adversarial sleeping bandits and obtain a regret bound that matches the best known bound in Nguyen and Mehta [2024] despite using fewer assumptions.

6.3.1 Proof Sketch for Theorem 6.3.1

As mentioned in Section 6.2, we first show that z_t and h_t in (6.9) satisfy the two conditions (I) $\mathbb{E}[\sum_{t=1}^T \frac{z_t}{h_t}] = O(\text{poly}(T))$ and (II) $\max_t \frac{z_t}{\beta_t}$ is small. The second condition is straightforward from the definition of z_t, γ and d, since $\frac{z_t}{\beta_t} \leq \frac{18d^2}{\gamma} \leq 6d^2 = 24$ which is a constant. To see that (I) is true, note that h_t is fixed with respect to I_t. Hence,

$$\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{h_t}\right] = \mathbb{E}\left[\sum_{t=1}^T \frac{\mathbb{E}_{I_t}[z_t]}{h_t}\right] \leq T\mathbb{E}\left[\frac{\max_{t\in[T]}\mathbb{E}_{I_t}[z_t]}{\min_{t\in[T]} h_t}\right].$$

Then, the condition (I) follows from Lemma 6.B.1, which shows that $h_t \geq \frac{1-\alpha}{4\alpha}T^{-\alpha}$ and $E_{I_t}[z_t] \leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)}S^\alpha$. Jensen's inequality implies that $\mathbb{E}\left[\ln\left(\sum_{t=1}^T \frac{z_t}{h_t}\right)\right] \leq \ln\left(\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{h_t}\right]\right)$. Combining this with Lemma 6.B.1, we conclude that $\mathbb{E}[F(z_{1:T}, h_{1:T})]$ grows dominantly with $\mathbb{E}[G(z_{1:T}, h_{1:T})]$. The last part of the proof is showing that plugging z_t and h_t from (6.9) into (6.7) yields the desired bounds. In the adversarial regime, the bound (6.10) follows directly from (6.7), Lemma 6.B.1, $h_t \leq \frac{K^{1-\alpha}-1}{\alpha}$ and Jensen's inequality $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$. In

the adversarial regime with a self-bounding constraint, we can prove (6.11) by first showing that

$$\mathbb{E}[h_t z_t] \leq \frac{(6d)^{2-\alpha}}{\alpha(1-\alpha)\Delta_{\min}}(K-1)^{1-\alpha}S^{\alpha}\mathbb{E}\left[\sum_{i=1}^{K} p_{t,i}\Delta_i\right].$$

and then following the same argument as in Ito et al. [2024].

6.3.2 A Lower Bound for Problems with Soft Sparsity Constraint

It remains an open question whether the BOBW bounds in Theorem 6.3.1 are tight under the hard constraint $\|\ell_t\|_0 \leq S$. This hard-constraint problem belongs to a broader class of settings with a more relaxed constraint, in which there exists an $\alpha \in (0,1)$ and $1 \leq U \leq K^{\alpha}$ such that for all $t \in [T]$,

$$\mathbb{E}\left[\left(\sum_{i=1}^{K} |\ell_{t,i}|^{2/\alpha}\right)^{\alpha}\right] \leq U. \tag{6.12}$$

In other words, the sparsity constraint holds in expectation. Obviously, the hard-constraint setting with $\|\ell_t\|_0 \leq S$ satisfies (6.12) for any $\alpha \in (0,1)$ and $U = S^{\alpha}$. Moreover, by using the same Algorithm 6.1 and straightforward modifications in its proof, we can obtain the corresponding $O(\frac{K^{1-\alpha}U}{\alpha(1-\alpha)\Delta_{\min}}\ln T)$ and $O(\sqrt{\frac{K^{1-\alpha}}{\alpha(1-\alpha)}UT})$ BOBW bounds for stochastic and adversarial regimes, respectively. The following theorem, whose proof is in Section 6.C, shows near-matching lower bounds for problems with soft sparsity constraint defined in (6.12).

Theorem 6.3.4. (Instance-Dependent Lower Bound) For any consistent algorithm, for any $\Delta \in (0,1), K \geq 4, \alpha \in (0,1)$ and $1 \leq U \leq \frac{K^{\alpha}}{4}$, there exists a K-armed stochastic bandit instance with $\Delta_{\min} = \Delta$ and loss distribution satisfying (6.12) such that

$$\lim_{T\to\infty} \frac{R_T}{\ln(T)} = \Omega\left(\frac{K^{1-\alpha}U}{\Delta}\right).$$

(Minimax Lower Bound) For any algorithm, for any $K \geq 4, \alpha \in (0,1)$ and $U \leq K^{\alpha}$, there exists an adversarial bandit instance with K arms and loss distribution satisfying (6.12) such that

$$R_T = \Omega(\sqrt{K^{1-\alpha}UT}).$$

6.4 Application II: $\sqrt{Q \ln(K)}$ Upper Bound with Unknown Q using Optimistic FTRL

In this section, we propose a new approach for obtaining a BOBW $O(\frac{K \ln T}{\Delta_{\min}}, \sqrt{Q \ln K})$-bound with unknown Q. For ease of exposition, we assume losses are in $[0, 1]$ and note that the analysis can be easily extended for losses in $[-1, 1]$. The new approach is based on applying real-time SPM on the Optimistic FTRL framework [Rakhlin and Sridharan, 2013], and then combining with the Reservoir Sampling algorithm [Hazan and Kale, 2011]. In principle, our algorithm follows the same framework as Hazan and Kale [2011]; Bubeck et al. [2018] where the learner maintains a reservoir $\mathbb{S}_i$ of observed losses for each arm $i \in [K]$ and then uses the estimated mean $m_{t,i} = \tilde{\mu}_{t,i}$ of these reservoirs as the optimistic vector m_t in Optimistic FTRL. In each round t, the learner chooses to perform either a reservoir sampling step for updating the reservoir $\mathbb{S}_i$, or a FTRL learning step for minimizing the regret. When the FTRL learning step is performed in round t, the vector q_t is computed by

$$q_t = \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} x_i^\alpha) \right) - \gamma \sum_{i=1}^{K} \ln(x_i). \tag{6.13}$$

Similar to Algorithm 6.1, the sampling probability vector p_t is obtained by mixing with $\frac{1}{T}$, i.e, $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$. After an arm $I_t \sim p_t$ is drawn, the loss estimates are $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{I_t = i\}}{p_{t,i}}$. The learning rates $(\beta_t)_t$ are computed by real-time SPM, with z_t and h_t defined as

$$z_t = \min \left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2 \right), \quad h_t = \frac{1}{\alpha} \left(\sum_{i=1}^{K} p_{t,i}^\alpha - 1 \right).$$

The full procedure is given in Algorithm 6.4 in Appendix 6.D. The following theorem states the BOBW bound for this approach.

Theorem 6.4.1. Algorithm 6.4 (in Appendix 6.D) guarantees the following bounds simultaneously

- In the adversarial regime:

$$R_T \leq O \left(\sqrt{\frac{(K^{1-\alpha} - 1)Q}{\alpha(1-\alpha)}} \right). \tag{6.14}$$

- In the adversarial regime with a self-bounding constraint:

$$R_T \leq O\left(\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}K^{\alpha}}{\alpha(1-\alpha)}}\right) \tag{6.15}$$

Proof. (Sketch) Our analysis follows from the analysis of Algorithm 6.1 and the observation by Hazan and Kale [2011] that the reservoir sampling steps only add an $O(\ln(T)^2)$ amount to the regret bound. As a result, the bound (6.15) for adversarial regime with a self-bounding constraint follows almost identically to that of Algorithm 6.1. For the bound (6.14) in the adversarial regime, the total variation Q naturally comes out of $\sum_{t=1}^{T} z_t$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} z_t\right] &\lesssim \frac{1}{1-\alpha}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\ell_{t,i}-m_{t,i})^2\right] = \frac{1}{1-\alpha}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t-\tilde{\mu}_t\|_2^2\right] \qquad (\text{since } m_t=\tilde{\mu}_t)\\
&\leq \frac{1}{1-\alpha}\left(\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t-\mu_T\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\|\tilde{\mu}_t-\mu_t\|_2^2\right]\right)\\
&\leq \frac{1}{1-\alpha}\left(Q + \sum_{t=1}^{T}\frac{Q}{t\ln(T)}\right) \leq O\left(\frac{Q}{1-\alpha}\right),
\end{aligned}$$

where the second inequality follows from triangle inequality and Lemma 10 in [Hazan and Kale, 2011], the third inequality is by Lemma 11 in [Hazan and Kale, 2011], and the last inequality is due to $\sum_{t=1}^{T}\frac{1}{t\ln T} \leq O(1)$. Together with (6.7) and $h_{\max} \leq \frac{K^{1-\alpha}-1}{\alpha}$, this implies (6.14). □

Remark 6.4.2. While existing works [Hazan and Kale, 2011; Bubeck et al., 2018] require either the knowledge of Q or sophisticated doubling tricks to estimate Q, our Algorithm 6.4 does not require such knowledge or any tricks. When $\alpha \to 1$, the bound in (6.14) becomes $O(\sqrt{Q\ln(K)})$. This bound matches the best known upper bound in Bubeck et al. [2018] and never exceeds $O(\sqrt{TK\ln(K)})$ in the worst case, all while simultaneously having a T-optimal best-of-both-worlds guarantee.

6.5 Coordinate-Wise Stability-Penalty Matching

We further generalize the SPM framework by introducing a new technique called coordinate-wise SPM (`CoWSPM`). As the name suggests, `CoWSPM` maintains separate learning rate $\beta_{t,i}$,

Algorithm 6.2 Coordinate-wise SPM with hybrid regularization for losses in $[0,1]$.

Input: $K \geq 3, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}\mathbf{1}, \gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $L_{0,i} = 0$ for $i \in [K]$
for each round $t = 1, \ldots, T$ **do**
  Compute $m_t \in [0,1]^K$ where $m_{t,i} = \frac{1}{1+\sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\}} \left(\frac{1}{2} + \sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\}\ell_{t,i}\right)$
  Compute q_t by Equation (6.13)
  Compute $p_t = (1 - \frac{K}{T})q_t + \frac{1}{T}\mathbf{1}$
  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
  Compute loss estimate $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{I_t = i\}}{p_{t,i}}$ and update $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
  Compute $z_{t,i} = \mathbb{1}\{i = I_t\}(\ell_{t,I_t} - m_{t,I_t})^2 \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{p_{t,I_t}^{-\alpha}, \frac{1-p_{t,I_t}}{p_{t,I_t}^2}\right\}, \frac{\beta_{t,i} 18 d^2}{\gamma}\right)$
  Compute $h_{t,i} = \frac{1}{\alpha} p_{t,i}^{\alpha}$
  Compute $\beta_{t+1,i} = \beta_{t,i} + \frac{z_{t,i}}{\beta_{t,i} h_{t,i}}$
end

stability term $z_{t,i}$ and penalty term $h_{t,i}$ for each arm $i \in [K]$. In each round t, `CoWSPM` updates the learning rates for each arm using the SPM update formula (6.4), i.e.,

$$\beta_{t+1,i} = \beta_{t,i} + \frac{z_{t,i}}{\beta_{t,i} h_{t,i}}. \tag{6.16}$$

Obviously, if $(z_{t,i})_{i\in[K]}$ and $(h_{t,i})_{i\in[K]}$ take the same values across all arms, then this approach recovers Algorithm 6.1. Instead, we adopt a different approach where $z_{t,i} = 0$ for all $i \neq I_t$ so that only the learning rate β_{t,I_t} of the observed arm I_t is updated in round t. The full procedure of `CoWSPM` is given in Algorithm 6.2, which uses the Optimistic FTRL framework with

$$\phi_t(x) = \sum_{i=1}^{K} \beta_{t,i} \left(\frac{-x_i^{\alpha}}{\alpha} + (1 - x_i)\ln(1 - x_i) + x_i\right) - \gamma \sum_{i=1}^{K} \ln(x_i). \tag{6.17}$$

This regularization function contains not only the α-Tsallis entropy, but also a part of the Shannon entropy and a linear term. The addition of these terms into the regularizer has been done in Ito et al. [2022] in order to have a regret bound containing the quantity $\tilde{p}_{t,i} = \min(p_{t,i}, 1 - p_{t,i})$ for the stochastic setting. This technique has a similar impact in our work, where it allows us to bound $z_{t,i}$ by a quantity containing $\tilde{p}_{t,i}$. However, while we use the same technique to introduce $\tilde{p}_{t,i}$ into our bounds, our analysis develops fundamentally different technical lemmas from that of [Ito et al., 2022] in order to use this new regularizer

in the real-time SPM framework. Next, to ensure that only β_{t,I_t} is updated, we set $z_{t,i}$ by

$$z_{t,i} = \mathbb{1}\{i = I_t\}(\ell_{t,I_t} - m_{t,I_t})^2 \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{p_{t,I_t}^{-\alpha}, \frac{1-p_{t,I_t}}{p_{t,I_t}^2}\right\}, \frac{\beta_{t,i} 18 d^2}{\gamma}\right), \tag{6.18}$$

so that $z_{t,I_t} \geq 0$ and $z_{t,i} = 0$ for $i \neq I_t$. The following theorem states the BOBW data-dependent bound of Algorithm 6.2, whose full proof is given in Appendix 6.E. The proof sketch outlines the main technical challenges in the analysis of Algorithm 6.2.

Theorem 6.5.1. For any $K \geq 4, T \geq 4k$, `CoWSPM` (Algorithm 6.2) with $\alpha \in (0,1)$ guarantees the following bounds simultaneously

- In the adversarial regime:

$$R_T \lesssim \min\left\{\sqrt{K \ln(T) \min(Q_\infty, L^*, T - L^*)}, K^{\frac{\alpha}{2}}\sqrt{KT}\right\}.$$

- In the stochastic regime:

$$R_T \lesssim \frac{1}{\alpha(1-\alpha)} \sum_{i \neq i^*} \frac{\ln(T)}{\Delta_i}.$$

Proof. (Sketch) Intuitively, coordinate-wise SPM consists of K separate real-time SPM processes, one for each arm. Similar to Ito et al. [2022], we find that this more refined approach enables deriving a bound (for the adversarial regime) that is adaptive to simultaneously different data-dependent quantities such as Q_∞ and L^*. However, having separate learning rates introduces several new technical challenges. First, the analysis developed for Algorithm 6.1 that bounds $q_{t+1,i} = O(q_{t,i})$ for all $i \in [K]$ no longer applies because in each round t, the learning rates $(\beta_{t,i})_{i\in[K]}$ can be arbitrarily different from each other. The `CoWSPM` algorithm resolves this by using $\beta_{t+1,i} = \beta_{t,i}$ for $i \neq I_T$ so that it only need $q_{t+1,i} = O(q_{t,i})$ to hold for $i = I_t$ since

$$\phi_t(q_{t+1}) - \phi_{t+1}(q_{t+1}) = \sum_{i=1}^{K} (\beta_{t+1,i} - \beta_{t,i})(-f(q_{t+1,i})) = (\beta_{t+1,I_t} - \beta_{t,I_t}) f(q_{t+1,I_t}).$$

The second and also more important challenge is that even if $q_{t+1,i} = O(q_{t,i})$, the naive decomposition of the α-Tsallis entropy into its coordinate-wise form $-\psi_{TE}(x) = \frac{1}{\alpha}\sum_{i=1}^{K}(x_i^\alpha - x_i)$ and then assigning $h_{t,i} = \frac{1}{\alpha}(x_i^\alpha - x_i)$ does *not* guarantee that $h_{t+1,i} = O(h_{t,i})$. This is

because the function $x \mapsto x^\alpha - x$ gets arbitrarily close to 0 when x gets close to 1. This prompts a different choice for $h_{t,i}$ rather than $-f(p_{t,i})$. Algorithm 6.2 uses $h_{t,i} = \frac{1}{\alpha} p_{t,i}^\alpha$, which is monotonically increasing and ensures that $h_{t+1,I_t} = O(h_{t,I_t})$ for $q_{t+1,i} = O(q_{t,i})$. This choice of $h_{t,i}$ is justified by the technical Lemma 6.E.5, which states that $(x-1)\ln(1-x) \le x^\alpha$ for any $x, \alpha \in [0,1]$.

Finally, we prove that with $z_{t,i}$ defined in (6.18), the product $\mathbb{E}[h_{t,i} z_{t,i}]$ is upper bounded by a quantity containing $\tilde{p}_{t,i}$ and thus an $O(\sum_{i \neq i^*} \frac{\ln T}{\Delta_i})$ regret bound holds for stochastic bandits. This is handled by Lemma 6.E.9, which shows that $\mathbb{E}_{I_t}[z_{t,i}] \le 2\min(p_{t,i}, 1-p_{t,i})$. □

Remark 6.5.2. Theorem 6.5.1 holds for all $\alpha \in (0,1)$. In particular, for $\alpha \neq \frac{1}{2}$, we do not require any additional assumptions such as the Δ_i being known in order to get the T-optimal BOBW bound. This is a major difference compared to the Tsallis-INF algorithm [Zimmert and Seldin, 2021a]. On the other hand, the adversarial bound in Theorem 6.5.1 has an extra factor $\sqrt{K^\alpha}$. It is unclear to us whether this extra factor is a fundamental limitation of `CoWSPM` or an artifact of our analysis.

6.6 Conclusion and Future Works

We developed real-time SPM, an extension of the SPM method originally developed for obtaining best-of-both-worlds bounds in bandits problems. We showed that real-time SPM algorithms achieve novel bounds that are simultaneously best-of-both-worlds, data-dependent and have optimal dependency on T in both stochastic and adversarial regimes. Our bounds also have optimal dependency on the data-dependent quantities such as sparsity or total variation of the loss sequence without knowing them nor using sophisticated estimation tricks. Future work includes applying real-time SPM on other bandits problems, such as contextual linear bandits, and making real-time SPM adaptive towards other challenging data-dependent quantities like ℓ_1 and ℓ_2-norm path-length bounds.

6.A Related Works

Due to the vast literature on BOBW and data-dependent bounds in various bandits learning settings, this sections presents only the most relevant works in multi-armed bandits. A more comprehensive list of related works can be found in Ito et al. [2024]; Tsuchiya et al. [2023] and references therein.

Best-of-both-worlds bounds. The BOBW bounds in our paper are derived using the SPM method for tuning learning rates in the FTRL framework, originally proposed in Ito et al. [2024]. For stochastic bandits, our $O(\frac{K \ln T}{\Delta_{\min}})$bound in Sections 6.3 and 6.4 matches that of Wei and Luo [2018b]; Ito et al. [2024], and our $O(\sum_{i \neq i^*} \frac{\ln T}{\Delta_i})$ bound in Section 6.5 matches that of Zimmert and Seldin [2021a]; Ito [2021]. Both of these bounds are looser than the $O(\sum_{i \neq i^*} \frac{\sigma_i^2 \ln T}{\Delta_i})$ in Ito et al. [2022] obtained by a more specialized approach, where σ_i^2 is the variance of the losses of a sub-optimal arm i. However, except for Ito et al. [2024], these existing works have an $O(\sqrt{T \ln T})$ worst-case bound for adversarial bandits, which contains an extra $\ln T$ factor compared to our work. Our BOBW bound also have data-dependent guarantees, which is an advantage over Ito et al. [2024]. For bandits with sparse losses, Tsuchiya et al. [2023] similarly obtained bounds that are both BOBW and dependent on the sparsity constraint; however their bounds contain extra factors of $\ln(KT)$ in stochastic bandits and $\sqrt{\ln T}$ in adversarial bandits compared to our results.

Data-dependent bounds. We study the following data-dependent quantities: sparsity of losses, total variations and small losses. For bandits with sparse negative losses where $\|\ell_t\| \leq S$ and S is unknown, our $O(\frac{S \ln T}{\Delta_{\min}}, \sqrt{ST \ln(K)})$ BOBW bound is the first S-agnostic and T-optimal BOBW bound for this setting, which improves upon on the bound of Tsuchiya et al. [2023] and matches the best known bound for adversarial bandits in Bubeck et al. [2018]. When the total variations Q, Q_∞ and/or the loss of the best arm L^* (defined in Section 6.1.2) are small, our algorithms are based on the optimistic FTRL (OFTRL) framework similar to Hazan and Kale [2011]; Bubeck et al. [2018]; Ito et al. [2022]. The dependency on Q, Q_∞ and L^* in our results match the best known bounds in these works, while our BOBW bounds have an optimal $O(\square \ln T, \square \sqrt{T})$ dependency on T. Particularly, our coordinate-wise real-time SPM algorithm in Section 6.5 can be seen as a T-optimal variant of the algorithm in Ito et al. [2022], which share the idea of using separate learning rates for each arm.

6.B Proofs for Section 6.3

6.B.1 Proof for Theorem 6.3.1

Let $D_{TE}(p, q)$ and $D_{LB}(p, q)$ denote the Bregman divergences induced by the α-Tsallis entropy and the log-barrier function, respectively. Let $D_t(p, q) = \beta_t D_{TE}(p, q) + \gamma D_{LB}(p, q)$ denote the Bregman divergence induced by the hybrid regularizer $\phi_t(p) = \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) -$

$\gamma \sum_{i=1}^{K} \ln(p_i)$. Let $\hat{\ell}_t = \begin{bmatrix} \hat{\ell}_{t,1} \\ \hat{\ell}_{t,2} \\ \cdots \\ \hat{\ell}_{t,K} \end{bmatrix}$ be the estimated loss vector at time t. We state the following three stability lemmas, whose proofs are in Section 6.B.2 and Section 6.B.3.

Lemma 6.B.1. For any $t \in [T]$, Algorithm 6.1 guarantees

$$h_t \geq \frac{1-\alpha}{4\alpha} T^{-\alpha} \quad \text{and} \quad E_{I_t}[z_t] \leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} S^{\alpha}.$$

Lemma 6.B.2. For any $t \in [T]$, Algorithm 6.1 guarantees

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} \min(p_{t,I_t}, 1-p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right). \tag{6.19}$$

Note that in Lemma 6.B.2, the right-hand side is exactly $\frac{z_t}{\beta_t}$.

Lemma 6.B.3. For any $t \in [T]$, Algorithm 6.1 guarantees that for all $i \in [K]$,

$$q_{t+1,i} \leq 3d q_{t,i} \leq 6d p_{t,i}. \tag{6.20}$$

Moreover, this implies that $(-\psi_{TE}(q_{t+1})) \leq 3d(-\psi_{TE}(q_t)) \leq 6d(-\psi_{TE}(p_t))$.

Proof. (Of Theorem 6.3.1) Next, let

$$\Phi_t(p) = \beta_t \psi_{TE}(p) + \gamma \psi_{LB}(p) \tag{6.21}$$

be the time-varying regularizer in Algorithm 6.1. For any $a \in [K]$, define

$$u_a = \left(1 - \frac{K}{T}\right) e_a + \frac{1}{T}\mathbf{1}.$$

The pseudo-regret with respect to arm a is

$$\begin{aligned}
R_{T,a} &= \mathbb{E}[\sum_{t=1}^{T}\langle \ell_t, p_t - e_a\rangle] \\
&= \mathbb{E}[\sum_{t=1}^{T}\langle \ell_t, q_t - u_a\rangle] + \mathbb{E}[\sum_{t=1}^{T}\langle \ell_t, p_t - q_t\rangle] + \mathbb{E}[\sum_{t=1}^{T}\langle \ell_t, u_a - e_a\rangle] \\
&\le \mathbb{E}[\sum_{t=1}^{T}\langle \ell_t, q_t - u_a\rangle] + 4K \\
&= \mathbb{E}[\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u_a\rangle] + 4K,
\end{aligned}$$

where the inequality is from $\langle \ell_t, p_t - q_t\rangle = \frac{1}{T}\sum_{i=1}^{K} \ell_{t,i}(1 - Kq_{t,i}) \le \frac{2K}{T}$ and $\langle \ell_t, u_a - e_a\rangle \le \frac{2K}{T}$, and the last equality is from $\mathbb{E}[\hat{\ell}_t] = \ell_t$. By the standard analysis of FTRL with time-varying regularizer [Lattimore and Szepesvári, 2020], we have

$$\begin{aligned}
\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u_a\rangle &\le \Phi_{T+1}(u_a) - \min_{p\in\Delta_K}\Phi_1(p) + \sum_{t=1}^{T}\Phi_t(q_{t+1}) - \Phi_{t+1}(q_{t+1}) \\
&\quad + \sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \\
&= \Phi_{T+1}(u_a) - \min_{p\in\Delta_K}\Phi_1(p) + \sum_{t=1}^{T}(\beta_{t+1} - \beta_t)(-\psi_{TE}(q_{t+1})) \\
&\quad + \sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \\
&\le \Phi_{T+1}(u_a) - \min_{p\in\Delta_K}\Phi_1(p) + 6d\left(\sum_{t=1}^{T}(\beta_{t+1} - \beta_t)h_t + \sum_{t=1}^{T}\frac{z_t}{\beta_t}\right) \\
&= \Phi_{T+1}(u_a) - \min_{p\in\Delta_K}\Phi_1(p) + 24\sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\le \gamma\psi_{LB}(u_a) - \beta_1 \min_{p\in\Delta_K}\psi_{TE}(p) + 24\sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\le \gamma K\ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + 24\sum_{t=1}^{T}\frac{z_t}{\beta_t},
\end{aligned}$$

where the second inequality is from Lemma 6.B.2 and Lemma 6.B.3, the second equality

is from the update rule $(\beta_{t+1} - \beta_t)h_t = \frac{z_t}{\beta_t}$, the third inequality is due to $\psi_{TE}(p) \leq 0$ and $\psi_{LB}(p) > 0$ for all $p \in \Delta_K$, and the last inequality is due to $(u_a)_i \geq \frac{1}{T}$.

Bounding $\mathbb{E}[\sum_{t=1}^T \frac{z_t}{\beta_t}]$

Note that we should not directly apply the SPM bound based on $z_{\max} = \max_{t\in[T]} z_t$ in Lemma 3 of [Ito et al., 2024], because $z_{\max}$ is of order $\max_t p_{t,I_t}^{-\alpha}$, which can be very large. Instead, let

$$G = \sum_{t=1}^T \frac{z_t}{\sqrt{\sum_{s=1}^t \frac{z_s}{h_s}}}$$
$$h_{\max} = \max_{t\in[T]} h_t,$$
$$z_{\mathbb{E},\max} = \max_{t\in[T]} \mathbb{E}_{I_t}[z_t].$$

By definitions of h_t and z_t, we have $h_t \leq \frac{K^{1-\alpha}-1}{\alpha}$ and

$$\mathbb{E}_{I_t}[z_t] \leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} S^\alpha,$$

from Lemma 6.B.1. It follows that

$$h_{\max} \leq \frac{K^{1-\alpha}-1}{\alpha},$$
$$z_{\mathbb{E},\max} \leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} S^\alpha.$$

Next, let

$$\beta'_t = \sqrt{\beta_1^2 + 2\sum_{s=1}^{t-1} \frac{z_s}{h_s}}. \tag{6.22}$$

Let $E = \{t \in [T] : \beta'_{t+1} \geq \sqrt{2}\beta'_t\}$ and $E^c = [T] \setminus E$. Also, let $N = |E|$ and $j = 1, 2, \ldots, N$ be the index running over the rounds in E. Similar to the proof of Lemma 2 in Ito et al. [2024], squaring both sides of $\beta_t = \beta_{t-1} + \frac{z_{t-1}}{\beta_{t-1}h_{t-1}}$ implies that

$$\beta_t^2 = \beta_{t-1}^2 + \frac{2z_{t-1}}{h_{t-1}} + \frac{z_{t-1}^2}{\beta_{t-1}^2 h_{t-1}^2} \geq \beta_{t-1}^2 + \frac{2z_{t-1}}{h_{t-1}} \geq \beta_1^2 + 2\sum_{s=1}^{t-1} \frac{z_s}{h_s},$$

which shows that $\beta'_t \leq \beta_t$. Furthermore, $\sum_{t\in E^c} \frac{z_t}{\beta_t} \leq G$ from the proof of Lemma 2 in Ito et al. [2024]. Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{z_t}{\beta_t} &= \sum_{t\in E^c} \frac{z_t}{\beta_t} + \sum_{t\in E} \frac{z_t}{\beta_t} \\
&\leq G + \sum_{t\in E} \frac{z_t}{\beta_t} \\
&\leq G + \frac{18d^2}{\gamma} N \\
&\leq G + \frac{18d^2}{\gamma} \log_{\sqrt{2}} \left(\frac{\beta'_{T+1}}{\beta'_1} \right) \\
&\leq G + \frac{26d^2}{\gamma} \ln \left(1 + 2 \sum_{t=1}^{T} \frac{z_t}{h_t} \right),
\end{aligned}$$

where the second inequality is from the definition of z_t and the third inequality is from the fact that β'_t is multiplied by at least $\sqrt{2}$ after every round in E, thus there can be at most $N \leq \log_{\sqrt{2}} \frac{\beta'_{T+1}}{\beta'_1}$ such multiplications.

Taking the expectation over $I_{1:T}$ on both sides and using $E[\ln(X)] \leq \ln(E[X])$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t} \right] &\leq \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(1 + 2\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{h_t} \right] \right) \\
&= \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(\mathbb{E}\left[1 + 2 \sum_{t=1}^{T} \frac{\mathbb{E}_{I_t}[z_t]}{h_t} \right] \right) \\
&\leq \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(1 + \frac{(6d)^{2-\alpha} S^\alpha}{(1-\alpha)} \mathbb{E}\left[\frac{4T}{\min_{t\in[T]} h_t} \right] \right) \\
&\leq \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(1 + \frac{(6d)^{2-\alpha} S^\alpha}{(1-\alpha)} \frac{4\alpha T^{\alpha+1}}{1-\alpha} \right) \\
&= \mathbb{E}[G] + O\left(\frac{1}{\gamma} \ln \left(\frac{\alpha S^\alpha T}{(1-\alpha)^2} \right) \right)
\end{aligned}$$

where the first equality is because h_t is $\mathbb{F}_{t-1}$-measurable and the last inequality is due to Lemma 6.B.1.

Next, Equation 45 in [Ito et al., 2024] shows that $G \leq 2\sqrt{h_{\max} \sum_{t=1}^{T} z_t}$. Moreover,

Equation 46 in [Ito et al., 2024] shows that for any fixed $J \geq 1$,

$$G \leq \sqrt{8J \sum_{t=1}^{T} h_t z_t} + 2\sqrt{2^{-J} h_{\max} \sum_{t=1}^{T} z_t}.$$

As a result, we obtain the following bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right] \leq \min\left\{\inf_{J\in\mathbb{N}} \mathbb{E}\left[\left\{\sqrt{8J \sum_{t=1}^{T} h_t z_t} + 2\sqrt{2^{-J} h_{\max} \sum_{t=1}^{T} z_t}\right\}\right], 2\mathbb{E}\left[\sqrt{h_{\max} \sum_{t=1}^{T} z_t}\right]\right\} + O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^{\alpha} T}{(1-\alpha)^2}\right)\right). \tag{6.23}$$

Adversarial Regime:

Using Jensen's inequality $E[\sqrt{X}] \leq \sqrt{E[X]}$ and Equation (6.23), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right] &\leq 2\sqrt{\frac{((K-1)^{1-\alpha}-1)}{\alpha} \sum_{t=1}^{T} \mathbb{E}[z_t]} + O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^{\alpha} T}{(1-\alpha)^2}\right)\right) \\
&\leq 2\sqrt{\frac{((K-1)^{1-\alpha}-1)}{\alpha} T \mathbb{E}[z_{\mathbb{E},\max}]} + O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^{\alpha} T}{(1-\alpha)^2}\right)\right) \\
&= O\left(\sqrt{\frac{(K^{1-\alpha}-1) S^{\alpha} T}{\alpha(1-\alpha)}}\right).
\end{aligned}$$

Adversarial Regime with a Self-Bounding Constraint:

Let $R_T = \max_{a\in[K]} R_{T,a}$. In this regime, we have $R_T + C \geq \mathbb{E}[\sum_{t=1}^{T} \sum_{i=1}^{K} q_{t,i} \Delta_i]$. Let $i^* \in [K]$ be the unique optimal arm.

Observe that given the sequence of randomly drawn arms until the beginning of round

t, the quantity h_t is fixed. Therefore, we can write $\mathbb{E}[h_t z_t] = \mathbb{E}[h_t \mathbb{E}_{I_t}[z_t]]$ and obtain

$$\begin{aligned}
\mathbb{E}[h_t z_t] &= \mathbb{E}[h_t \mathbb{E}_{I_t}[z_t]] \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}\left[h_t \left(\sum_{i=1}^{K} (\tilde{p}_{t,i}^{1-\alpha}) \ell_{t,i}^2 \right) \right] \\
&= \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}\left[\frac{1}{\alpha} \left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1 \right) \left(\sum_{i=1}^{K} (\tilde{p}_{t,i}^{1-\alpha}) \ell_{t,i}^2 \right) \right] \\
&\leq \frac{(6d)^{2-\alpha}}{2\alpha(1-\alpha)} \mathbb{E}\left[\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1 \right) \left(\sum_{\ell_{t,i} \neq 0} \tilde{p}_{t,i}^{1-\alpha} \right) \right],
\end{aligned} \tag{6.24}$$

where the second inequality is from $\ell_{t,i}^2 \leq 1$.

Using $p_{t,i^*}^{\alpha} - 1 \leq 0$ and $\sum_{i \neq i^*} p_{t,i}^{\alpha} \leq (K-1)^{1-\alpha} (\sum_{i \neq i^*} p_{t,i})^{\alpha}$ by Holder's inequality, we obtain

$$\begin{aligned}
\sum_{i \in [K]} p_{t,i}^{\alpha} - 1 &\leq (K-1)^{1-\alpha} (\sum_{i \neq i^*} p_{t,i})^{\alpha} \\
&\leq \frac{(K-1)^{1-\alpha}}{\Delta_{\min}^{\alpha}} (\sum_{i \in [K]} p_{t,i} \Delta_i)^{\alpha}.
\end{aligned}$$

Next, from $\tilde{p}_{t,i^*} \leq \sum_{i \neq i^*} p_{t,i}$ we have

$$\begin{aligned}
\tilde{p}_{t,i*}^{1-\alpha} &\leq \left(\sum_{i \neq i^*} \tilde{p}_{t,i} \right)^{1-\alpha} \\
&\leq \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha}.
\end{aligned}$$

Therefore, by Holder's inequality,

$$\begin{aligned}
\sum_{\ell_{t,i}\neq 0} \tilde{p}_{t,i}^{1-\alpha} &\leq \left(\sum_{\ell_{t,i}\neq 0, i\neq i^*} p_{t,i}^{1-\alpha}\right) + \tilde{p}_{t,i*}^{1-\alpha} \\
&\leq \sum_{\ell_{t,i}\neq 0, i\neq i^*} \left(\Delta_i^{-\frac{1-\alpha}{\alpha}}\right)^{\alpha} (p_{t,i}\Delta_i)^{1-\alpha} + \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i\neq i^*} p_{t,i}\Delta_i\right)^{1-\alpha} \\
&\leq \left(\sum_{\ell_{t,i}\neq 0, i\neq i^*} \Delta_i^{-\frac{1-\alpha}{\alpha}}\right)^{\alpha} \left(\sum_{\ell_{t,i}\neq 0, i\neq i^*} p_{t,i}\Delta_i\right)^{1-\alpha} + \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i\neq i^*} p_{t,i}\Delta_i\right)^{1-\alpha} \\
&\leq \frac{S^{\alpha}}{\Delta_{\min}^{1-\alpha}} \left(\sum_{\ell_{t,i}\neq 0, i\neq i^*} p_{t,i}\Delta_i\right)^{1-\alpha} + \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i\neq i^*} p_{t,i}\Delta_i\right)^{1-\alpha} \\
&\leq \frac{2S^{\alpha}}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i\in[K]} p_{t,i}\Delta_i\right)^{1-\alpha}.
\end{aligned}$$

Overall, we have

$$\mathbb{E}[h_t z_t] \leq \frac{(6d)^{2-\alpha}}{\alpha(1-\alpha)\Delta_{\min}} (K-1)^{1-\alpha} S^{\alpha} \mathbb{E}\left[\sum_{i=1}^{K} p_{t,i}\Delta_i\right]. \tag{6.25}$$

Furthermore, by Jensen's inequality,

$$\mathbb{E}\left[\sqrt{2^{-J} h_{\max} \sum_{t=1}^{T} z_t}\right] \leq \sqrt{\mathbb{E}\left[2^{-J} h_{\max} \sum_{t=1}^{T} z_t\right]} \leq \sqrt{2^{-J} T \frac{(K-1)^{1-\alpha}}{\alpha} \frac{(6d)^{2-\alpha} S^{\alpha}}{2(1-\alpha)}}. \tag{6.26}$$

By plugging $J = \lceil \log_2(T) \rceil$, (6.25) and (6.26) into (6.23), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right] &\leq O\left(\sqrt{\frac{\ln(T)(K^{1-\alpha}-1)S^{\alpha}\mathbb{E}[\sum_{t=1}^{T}\sum_{i=1}^{K} p_{t,i}\Delta_i]}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K^{1-\alpha}-1)S^{\alpha}}{\alpha(1-\alpha)}}\right) \\
&\leq O\left(\sqrt{\frac{\ln(T)(K^{1-\alpha}-1)S^{\alpha}(R_T + C)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K^{1-\alpha}-1)S^{\alpha}}{\alpha(1-\alpha)}}\right).
\end{aligned}$$

In summary, keeping only the dominant $\sqrt{T}$ terms, we have the following BOBW bounds that hold simultaneously:

- In adversarial regime,

$$R_T \leq O\left(\sqrt{\frac{(K^{1-\alpha}-1)S^\alpha T}{\alpha(1-\alpha)}}\right)$$

- In adversarial regime with a self-bounding constraint:

$$R_T \leq O\left(\frac{(K-1)^{1-\alpha}S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}S^\alpha}{\alpha(1-\alpha)}}\right)$$

Note that we can set α sufficiently close 1 so that $\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)} = O(\ln(K))$, $\frac{(K-1)^{1-\alpha}}{\alpha(1-\alpha)} = O(\ln(K))$ and while $\gamma = O(\sqrt{\frac{\alpha}{1-\alpha}})$ grows with K instead of T. For example, in Appendix 6.G, we show that $\alpha = 1 - \frac{1}{2\ln(K)}$ satisfies $\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)} \leq 4\ln(K)$, $\frac{(K-1)^{1-\alpha}}{\alpha(1-\alpha)} \leq 4\ln(K)$ while $\gamma \lesssim \sqrt{\ln(K)}$. This ensures that

$$\begin{aligned}\gamma K \ln(T) + \frac{\beta_1(K^{1-\alpha}-1)}{\alpha} &= \gamma K \ln(T) + \frac{4K(K^{1-\alpha}-1)}{\alpha(1-\alpha)} \\ &= O\left(K\ln(K)\ln(T)\right),\end{aligned}$$

and

$$\frac{1}{1-\alpha} = O(\ln(K))$$

everywhere, so we can safely ignore the terms that do not contain $\sqrt{T}$ (in the adversarial setting) and $\frac{\ln(T)}{\Delta_{\min}}$ (in the stochastic setting).

□

6.B.2 Stability Proofs

In this section, we prove Lemma 6.B.2 and Lemma 6.B.3. First, we state and prove a number of supporting lemmas. In the following, we let

$$g_{\beta,\gamma}(t) = \beta t^{\alpha-1} + \frac{\gamma}{t}. \tag{6.27}$$

be a function defined on $(0,1) \to \mathbb{R}_+$. Note that because $\beta > 0, \alpha \in (0,1)$ and $\gamma > 0$, this function $g_{\beta,\gamma}(t)$ is monotonically decreasing in t. We will drop the subscripts β and γ whenever they are clear from the context.

The first lemma shows that $\beta_{t+1} - \beta_t$ is sufficiently small for stabilizing the FTRL update in Algorithm 6.1.

Lemma 6.B.4. For any $t \geq 1$, Algorithm 6.1 guarantees

$$\beta_{t+1} - \beta_t \leq (1 - \frac{1}{d})\gamma q_{t*}^{-\alpha}, \tag{6.28}$$

where $q_{t*} = \min(\max_{i\in[K]} q_{t,i}, 1 - \max_{i\in[K]} q_{t,i})$.

Proof. Lemma 6.B.13 shows that $h_t \geq \frac{1-\alpha}{4\alpha} p_{t*}^{\alpha}$. By Lemma 6.B.14, we have $p_{t*}^{\alpha} \geq 2^{-\alpha} q_{t*}^{\alpha}$. This implies that $\frac{1}{h_t} \leq \frac{4\alpha}{1-\alpha} 2^{\alpha} q_{t*}^{-\alpha}$. By the definitions of β_{t+1}, z_t and h_t, we have

$$\begin{aligned}
\beta_{t+1} - \beta_t &= \frac{z_t}{\beta_t h_t} \\
&\leq \frac{4\alpha z_t}{(1-\alpha)\beta_t} 2^{\alpha} q_{t*}^{-\alpha} \\
&\leq \frac{4\alpha}{(1-\alpha)} \frac{18d^2}{\gamma} \ell_{t,I_t}^2 2^{\alpha} q_{t*}^{-\alpha} \\
&\leq (1 - \frac{1}{d})\gamma q_{t*}^{-\alpha}
\end{aligned}$$

where the last inequality uses

$$\frac{72\alpha d^2}{(1-\alpha)\gamma} \ell_{t,I_t}^2 2^{\alpha} \leq \frac{72\alpha d^2}{(1-\alpha)\gamma} 2^{\alpha} \leq (1 - \frac{1}{d})\gamma \tag{6.29}$$

for $d = 2$ and $\gamma \geq 48\sqrt{\frac{\alpha}{1-\alpha}}$. □

Lemma 6.B.5. For any $L \in \mathbb{R}^K, \beta > 0, \gamma > 0$ and $h \in [-1, 1]$, let

$$\begin{aligned}
x &= \arg\min_{p\in\Delta_K} \langle L, p\rangle + \beta \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} p_i^{\alpha})\right) - \gamma \sum_{i=1}^{K} \ln(p_i), \\
y &= \arg\min_{p\in\Delta_K} \langle L + \frac{h}{x_1'} e_1, p\rangle + \beta \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} p_i^{\alpha})\right) - \gamma \sum_{i=1}^{K} \ln(p_i).
\end{aligned}$$

Here, e_1 is the first vector in the standard basis of $\mathbb{R}^K$. If $4x_1' \geq x_1$ and $\gamma \geq 6$, then $y_1 \leq 3x_1$.

Proof. Using the Lagrange multiplier method, we have the following equalities that hold for

some $Z \in \mathbb{R}$,

$$\beta\left(y_1^{\alpha-1} - x_1^{\alpha-1}\right) + \gamma\left(\frac{1}{y_1} - \frac{1}{x_1}\right) = Z + \frac{h}{x_1'} \tag{6.30}$$

and for all $i \neq 1$,

$$\beta\left(y_i^{\alpha-1} - x_i^{\alpha-1}\right) + \gamma\left(\frac{1}{y_i} - \frac{1}{x_i}\right) = Z. \tag{6.31}$$

First, we show that Z and $y_1 - x_1$ has the opposite sign to h. We consider two cases:

- If $Z \geq 0$ then from (6.31), we have $g(y_i) - g(x_i) = Z \geq 0$. This implies $y_i \leq x_i$ and leads to $y_1 \geq x_1$. From (6.30), we have $Z + \frac{h}{x_1'} = g(y_1) - g(x_1) \leq 0$. Since $Z \geq 0$, this implies $h \leq 0$.
- If $Z \leq 0$ then by the same argument, we have $y_i \geq x_i$ and $y_1 \leq x_1$. Therefore, $Z + \frac{h}{x_1'} \geq 0$. Due to $Z \leq 0$, we must have $h \geq 0$.

In both cases, we have $Zh \leq 0$ and $Z(y_1 - x_1) \geq 0$. It follows that if $h \geq 0$ then we have $y_1 \leq x_1 \leq 2x_1$. If $h < 0$ then $y_1 \geq x_1$, and by rearranging (6.30), we obtain

$$\begin{aligned}
\frac{4}{x_1} \geq -\frac{h}{x_1'} &= \underbrace{Z}_{\geq 0} + \underbrace{\gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right)}_{\geq 0} + \underbrace{\beta(x_1^{\alpha-1} - y_1^{\alpha-1})}_{\geq 0} \\
&\geq \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\
&\geq 6\left(\frac{1}{x_1} - \frac{1}{y_1}\right),
\end{aligned}$$

where the last inequality is due to $\gamma \geq 6$. This implies that $\frac{3}{y_1} \geq \frac{1}{x_1}$, thus $y_1 \leq 3x_1$. □

Lemma 6.B.6. For any $L \in \mathbb{R}^K, \beta > 0, \beta' > 0, \gamma \geq 0$, define

$$\begin{aligned}
x &= \arg\min_{p \in \Delta_K} \langle L, p\rangle + \beta\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(p_i), \\
y &= \arg\min_{p \in \Delta_K} \langle L, p\rangle + \beta'\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(p_i).
\end{aligned}$$

Let $x_* = \min(\max_{i\in[K]} x_i, 1 - \max_{i\in[K]} x_i)$. For any constant $d \geq 2$, if

$$0 \leq \beta' - \beta \leq \left(1 - \frac{1}{d}\right)\gamma x_*^{-\alpha}, \tag{6.32}$$

then $y_i \leq dx_i$ for all $i \in [K]$.

Proof. Using the Lagrange multiplier method, we have for all $i \in [K]$,

$$L_i - \beta x_i^{\alpha-1} - \frac{\gamma}{x_i} = \lambda, \tag{6.33}$$

$$L_i - \beta' y_i^{\alpha-1} - \frac{\gamma}{y_i} = \lambda'. \tag{6.34}$$

Subtracting both sides of the two equations, we obtain

$$\begin{aligned}\lambda - \lambda' + g_\beta(x_i) &= g_{\beta'}(y_i) \\ &= g_\beta(y_i) + (\beta' - \beta)y_i^{\alpha-1}.\end{aligned}$$

If $\lambda - \lambda' < 0$, then because $\beta \leq \beta'$, we have $g_\beta(x_i) > g_\beta(y_i) + (\beta' - \beta)y_i^{\alpha-1} \geq g_\beta(y_i)$. This implies $x_i < y_i$ for all $i \in [K]$, a contradiction to $\sum_{i=1}^K x_i = \sum_{i=1}^K y_i = 1$. Hence, we have $\lambda - \lambda' \geq 0$, and thus $g_\beta(x_i) \leq g_{\beta'}(y_i)$.

For any $i \in [K]$, if $x_i > x_*$ then $x_i \geq \frac{1}{2}$ and hence $y_i \leq 1 \leq 2x_i \leq dx_i$. From the condition $\beta' - \beta \leq (1 - \frac{1}{d})\gamma x_*^{-\alpha}$, for $x_i \leq x_*$, we have

$$\begin{aligned}g_{\beta'}(y_i) &\geq g_\beta(x_i) \\ &= \beta x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &\geq (\beta' - (1 - \frac{1}{d})\gamma x_*^{-\alpha})x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &= \beta' x_i^{\alpha-1} - (1 - \frac{1}{d})\gamma x_*^{-\alpha} x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &\geq \beta' x_i^{\alpha-1} - (1 - \frac{1}{d})\gamma x_i^{-\alpha} x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &= \beta' x_i^{\alpha-1} + \frac{\gamma}{dx_i} \\ &\geq \beta'(dx_i)^{\alpha-1} + \frac{\gamma}{dx_i} \\ &= g_{\beta'}(dx_i),\end{aligned}$$

where the last inequality is due to $(d)^{\alpha-1} \leq 1$. This implies $y_i \leq dx_i$ for all $x_i \leq x_*$. □

Let $\|x\|_A = \sqrt{x^T A x}$ be the norm of a vector $x \in \mathbb{R}^K$ induced by a positive definite matrix A. The following lemma proves Lemma 6.B.2 when the chosen arm I_t satisfies $q_{t,I_t} \leq 1 - q_{t,I_t}$.

Lemma 6.B.7. For any $t \in [T]$, Algorithm 6.1 guarantees

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} p_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right) \tag{6.35}$$

Proof. Using standard local-norm analysis techniques for FTRL (for example, see Section 7.4 in [Orabona, 2023]), we have

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \frac{1}{2} \left\| \hat{\ell}_t \right\|^2_{(\nabla^2 \phi_t(z_t))^{-1}}, \tag{6.36}$$

where z_t is a point between q_t and q_{t+1}. The Hessian matrix of ϕ_t is a diagonal matrix with entries

$$\nabla^2 \phi_t(z_t) = \operatorname{diag}\left(\left(\beta_t(1-\alpha) z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2} \right)_{i=1,2,\ldots,K} \right). \tag{6.37}$$

Hence, its inverse is the following diagonal matrix

$$(\nabla^2 \phi(z_t))^{-1} = \operatorname{diag}\left(\left(\frac{1}{\beta_t(1-\alpha) z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}} \right)_{i=1,2,\ldots,K} \right). \tag{6.38}$$

It follows that

$$\begin{aligned}
\left\| \hat{\ell}_t \right\|^2_{(\nabla^2 \phi_t(z_t))^{-1}} &= \sum_{i=1}^K \hat{\ell}_{t,i}^2 \frac{1}{\beta_t(1-\alpha) z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}} \\
&\leq \min\left(\frac{1}{\beta_t(1-\alpha)} \sum_{i=1}^K z_{t,i}^{2-\alpha} \hat{\ell}_{t,i}^2, \frac{1}{\gamma} \sum_{i=1}^K z_{t,i}^2 \hat{\ell}_{t,i}^2 \right) \\
&= \min\left(\frac{1}{\beta_t(1-\alpha)} z_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{z_{t,I_t}^2 \hat{\ell}_{t,I_t}^2}{\gamma} \right),
\end{aligned} \tag{6.39}$$

where the last equality is due to $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$. Combining (6.36) and (6.39), we obtain

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min \left(\frac{1}{2\beta_t(1-\alpha)} z_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{z_{t,I_t}^2 \hat{\ell}_{t,I_t}^2}{2\gamma} \right). \tag{6.40}$$

Since z_t is between q_t and q_{t+1}, we have $z_{t,I_t} \leq \max(q_{t,I_t}, q_{t+1,I_t})$. The loss estimate in Algorithm 6.1 uses p_{t,I_t} where $2p_{t,I_t} \geq q_{t,I_t}$ by Lemma 6.B.14, therefore we can combine the results of Lemma 6.B.4, Lemma 6.B.6 and Lemma 6.B.5 and obtain $q_{t+1,I_t} \leq 3dq_{t,I_t}$. It follows that $z_{t,I_t} \leq 3dq_{t,I_t} \leq 6dp_{t,I_t}$, and as a result,

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min \left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} p_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{36d^2}{2\gamma} p_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 \right) \tag{6.41}$$

$$\leq \min \left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} p_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right), \tag{6.42}$$

where the last equality is due to $p_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 = \ell_{t,I_t}^2$. □

The next lemma proves Lemma 6.B.2 whenever the chosen arm I_t has the maximum sampling probability. The proof is largely based on Lemma 9 in [Ito et al., 2024] and Equation 22 in [Tsuchiya et al., 2023].

Lemma 6.B.8. For any $t \in [T]$, if $I_t \in \arg\max_{i \in [K]} p_{t,i}$, Algorithm 6.1 guarantees

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min \left(\frac{4}{\beta_t(1-\alpha)} (1 - p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{4\ell_{t,I_t}^2}{\gamma} \right). \tag{6.43}$$

Proof. When $I_t \in \arg\max_{i \in [K]} p_{t,i}$, we have $I_t \in \arg\max_{i \in [K]} q_{t,i}$ and thus $q_{t,I_t} \geq p_{t,I_t} \geq \frac{1}{K}$. Therefore,

$$\frac{\hat{\ell}_{t,I_t}}{\beta_t} = \frac{\ell_{t,I_t}}{p_{t,I_t}\beta_t} \leq \frac{1}{p_{t,I_t}\beta_t} \leq \frac{K}{\beta_t} \leq \frac{1-\alpha}{4} \leq \frac{1-\alpha}{4} (1 - q_{t,I_t})^{\alpha-1}, \tag{6.44}$$

where the third inequality is due to $\beta_t \geq \beta_1 \geq \frac{4K}{1-\alpha}$ by initialization, and the last inequality is from $(1 - q_{t,I_t})^{\alpha-1} \geq 1$ for $\alpha \in (0,1)$. Furthermore, for any $i \in [K] \setminus \{I_t\}$, we have $\frac{\hat{\ell}_{t,i}}{\beta_t} = 0 \geq -\frac{1-\alpha}{4} q_{t,i}^{\alpha-1}$. Therefore, by using Lemma 9 in [Ito et al., 2024] and noting that $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$, we obtain

$$\langle \frac{1}{\beta_t} \hat{\ell}_t, q_t - q_{t+1} \rangle - D_{TE}(q_{t+1}, q_t) \leq \frac{4}{\beta_t^2(1-\alpha)} (1 - q_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2. \tag{6.45}$$

Furthermore, Equation 22 in [Tsuchiya et al., 2023] states that if $\frac{q_{t,I_t}\hat{\ell}_{t,I_t}}{\gamma} \geq -1$ and $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$, then

$$\langle q_t - q_{t+1}, \hat{\ell}_t \rangle - \gamma D_{LB}(q_{t+1}, q_t) \leq \frac{q_{t,I_t}\hat{\ell}_{t,I_t}^2}{\gamma}. \tag{6.46}$$

Indeed, we have $\left|\frac{q_{t,I_t}\hat{\ell}_{t,I_t}}{\gamma}\right| = \left|\frac{q_{t,I_t}\ell_{t,I_t}}{p_{t,I_t}\gamma}\right| \leq \frac{1}{2\gamma} \leq \frac{1}{8}$ since $q_{t,I_t} \leq 2p_{t,I_t}$ by Lemma 6.B.14 and $\gamma \geq 4$ by definition. Therefore, (6.45) and (6.46) together implies that

$$\begin{aligned}
&\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \\
&= \langle \hat{\ell}_t, q_t - q_{t+1} \rangle - \beta_t D_{TE}(q_{t+1}, q_t) - \gamma D_{LB}(q_{t+1}, q_t) \\
&\leq \min\left(\beta_t\left(\langle \frac{1}{\beta_t}\hat{\ell}_t, q_t - q_{t+1}\rangle - D_{TE}(q_{t+1}, q_t)\right), \langle \hat{\ell}_t, q_t - q_{t+1}\rangle - \gamma D_{LB}(q_{t+1}, q_t)\right) \\
&\leq \min\left(\frac{4}{\beta_t(1-\alpha)}(1-q_{t,I_t})^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{q_{t,I_t}^2\hat{\ell}_{t,I_t}^2}{\gamma}\right) \\
&\leq \min\left(\frac{4}{\beta_t(1-\alpha)}(1-p_{t,I_t})^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{4\ell_{t,I_t}^2}{\gamma}\right),
\end{aligned} \tag{6.47}$$

where the first inequality is because Bregman divergences are non-negative. □

Next, we prove Lemma 6.B.2.

Proof. (Of Lemma 6.B.2) We consider two cases:

- If $I_t \notin \arg\max_{i\in[K]} p_{t,i}$ or $p_{t,I_t} \leq 1 - p_{t,I_t}$: in this case, we have $p_{t,I_t} = \min(p_{t,I_t}, 1-p_{t,I_t})$. By Lemma 6.B.7, we have

$$\begin{aligned}
\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) &\leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)}p_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\ell_{t,I_t}^2\right) \\
&= \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\ell_{t,I_t}^2\right).
\end{aligned}$$

- If $p_{t,I_t} > 1 - p_{t,I_t}$: in this case, we have $I_t \in \arg\max_{i\in[K]} q_{t,i}$. By Lemma 6.B.8,

$$\begin{aligned}
\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) &\leq \min\left(\frac{4}{\beta_t(1-\alpha)}(1-p_{t,I_t})^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{4\ell_{t,I_t}^2}{\gamma}\right) \\
&= \min\left(\frac{4}{\beta_t(1-\alpha)}\min(p_{t,I_t}, 1-p_{t,I_t})^{2-\alpha}\hat{\ell}_{t,i}^2, \frac{4\ell_{t,I_t}^2}{\gamma}\right).
\end{aligned}$$

Lemma 6.B.2 follows by noting that $\max\left(\frac{(6d)^{2-\alpha}}{2}, 4\right) = \frac{(6d)^{2-\alpha}}{2}$. □

Lemma 6.B.9. For any $L \in \mathbb{R}^K, \beta > 0, \gamma > 0$ and $h \in [-1, 1]$, let

$$x = \arg\min_{p \in \Delta_K} \langle L, p\rangle + \beta\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(p_i),$$
$$y = \arg\min_{p \in \Delta_K} \langle L + \frac{h}{x_1'} e_1, p\rangle + \beta\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(p_i),$$

where $4x_1' \geq x_1$. Fix an arbitrary $\omega \in (1, 2]$. If $\beta \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ and $\gamma \geq 6$, then $y_i \leq 3x_i$ for all $i \in [K]$.

Proof. If $h \leq 0$, then we have $x_1 \leq y_1 \leq 3x_1$ and $y_i \leq x_i \leq 3x_i$ for all $i \neq 1$ from the proof of Lemma 6.B.5. Thus, we focus on the case $h > 0$. In this case, we have $y_1 \leq x_1 \leq 3x_1$ and $x_i \leq y_i$ for $i \neq 1$. From (6.30) and (6.31), we have

$$g(x_i) - g(y_i) = -Z \geq 0$$

for all $i \neq 1$, and

$$g(y_1) - g(x_1) = Z + \frac{h}{x_1'} \geq 0.$$

The latter implies that $-Z \leq \frac{1}{x_1'}$. Let $\epsilon = \frac{1}{\beta(1-\omega^{\alpha-1})} \leq \frac{\omega-1}{4K} \leq \frac{1}{4K}$. Similar to the proof of Lemma 13 in [Ito et al., 2024], we consider two cases:

- If $x_1' \geq \epsilon$, then $-Z \leq \frac{1}{\epsilon}$. For all $i \neq 1$,

$$\begin{aligned} g(y_i) &= g(x_i) + Z \\ &\geq g(x_i) - \frac{1}{\epsilon} \\ &= \beta x_i^{\alpha-1} - \beta(1 - \omega^{\alpha-1}) + \frac{\gamma}{x_i} \\ &\geq \beta x_i^{\alpha-1} - \beta x_i^{\alpha-1}(1 - \omega^{\alpha-1}) + \frac{\gamma}{\omega x_i} \\ &= \beta(\omega x_i)^{\alpha-1} + \frac{\gamma}{\omega x_i} = g(\omega x_i), \end{aligned}$$

  where the last inequality is due to $x_i^{\alpha-1} \geq 1$ and $\omega > 1$. This implies that for all $i \neq 1$, $y_i \leq \omega x_i \leq 3x_i$ (since $\omega \leq 2$).

- If $x_1' < \epsilon$, then we have $x_1 \leq 4x_1' < \frac{1}{K}$. For any $i^* \in \arg\max_{i\in[K]} x_{t,i}$, we have $i* \neq 1$. Similar to the proof of Lemma 13 in [Ito et al., 2024], we have $i^* \neq 1$ and $1 \leq \frac{y_{i*}}{x_{i*}} \leq \omega$. It follows that

$$\begin{aligned} -Z &= g(x_{i*}) - g(y_{i*}) \\ &= \beta x_{i*}^{\alpha-1} + \frac{\gamma}{x_{i*}} - (\beta y_{i*}^{\alpha-1} + \frac{\gamma}{y_{i*}}) \\ &\leq \beta x_{i*}^{\alpha-1} + \frac{\gamma}{x_{i*}} - (\beta\omega^{\alpha-1} x_{i*}^{\alpha-1} + \frac{\gamma}{\omega x_{i*}}) \\ &= g(x_{i*}) - g(\omega x_{i*}). \end{aligned}$$

  As the function $g(t) - g(\omega t)$ is decreasing for $\omega > 1$, we conclude that $-Z \leq g(x_i) - g(\omega x_i)$ for all $i \in [K]$. Therefore, $g(y_i) = g(x_i) + Z \geq g(\omega x_i)$ for all $i \neq 1$, which implies $y_i \leq \omega x_i \leq 3x_i$.

In both cases, we have $y_i \leq 3x_i$ for all $i \neq 1$. Combining this with $y_1 \leq x_1$, we conclude that $y_i \leq 3x_i$ for all $i \in [K]$. □

The following corollary is obtained by combining Lemma 6.B.6 and Lemma 6.B.9.

Corollary 6.B.10. For any $L \in \mathbb{R}^K, \beta > 0, \gamma > 0$ and $h \in [-1, 1]$, let

$$\begin{aligned} x &= \arg\min_{p\in\Delta_K} \langle L, p\rangle + \beta\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(p_i), \\ y &= \arg\min_{p\in\Delta_K} \langle L + \frac{h}{x_1'}e_1, p\rangle + \beta'\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(p_i). \end{aligned}$$

Let $x_* = \min(\max_{i\in[K]} x_i, 1 - \max_{i\in[K]} x_i)$. For any $\omega \in (1, 2]$ and $d = 2$, if $2x_1' \geq x_1$, $\gamma \geq 6, \beta \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ and

$$0 \leq \beta' - \beta \leq (1 - \frac{1}{d})\gamma x_*^{-\alpha}, \tag{6.48}$$

then $y_i \leq 3dx_i$ for all $i \in [K]$.

Proof. Let

$$\bar{x} = \arg\min_{p\in\Delta_K} \langle L, p\rangle + \beta'\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(p_i).$$

Here, $\bar{x}$ differs from x only by the learning rates $\beta' \geq \beta$. Applying Lemma 6.B.6 with $d = 2$, we obtain $\bar{x}_i \leq dx_i$ for all $i \in [K]$. In particular, $\bar{x}_1 \leq 2x_1$. Since $x_1 \leq 2x'_1$, we have $\bar{x}_1 \leq 4x'_1$. Next, since $\bar{x}$ differs from y only by $\frac{h}{x'_1}e_1$ in the dot product, we apply Lemma 6.B.9 and obtain $y_i \leq 3\bar{x}_i$ for all $i \in [K]$. Overall, we obtain $y_i \leq 3\bar{x}_i \leq 3dx_i$ for all $i \in [K]$. □

Finally, we are now ready to prove Lemma 6.B.3.

Proof. (Of Lemma 6.B.3) Let $\omega = 2$, we have $\beta_1 = \frac{8K}{1-\alpha} \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ due to $2^\alpha \leq 1 + \alpha$ for $\alpha \in [0, 1]$. Since $z_t, h_t \geq 0$, the sequence of learning rates $(\beta_t)_t$ is increasing and hence, $\beta_t \geq \beta_1 \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ for all $t \geq 1$. Together with Lemma 6.B.4, we have $\beta_{t+1} - \beta_t \leq \left(1 - \frac{1}{d}\right)\gamma q_{t*}^{-\alpha}$ and $\beta_t \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ for all $t \geq 1$. In addition, we have $p_{t,I_t} \geq 2q_{t,I_t}$ by Lemma 6.B.14. Applying Corollary 6.B.10 for

$$q_t = \arg\min_{x \in \Delta_K} \langle L_{t-1}, x\rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(x_i),$$
$$q_{t+1} = \arg\min_{x \in \Delta_K} \langle L_{t-1} + \frac{\ell_{t,I_t}}{p_{t,I_t}} e_{I_t}, x\rangle + \beta_{t+1} \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(x_i),$$

we obtain $q_{t+1,i} \leq 3dq_{t,i}$ for all $i \in [K]$.

For the second statement, we apply Lemma 11 in [Ito et al., 2024]. □

6.B.3 Technical Lemmas

Lemma 6.B.11. For any $a > 0, b > 0$ and $x \geq 0$, we have

$$a^x + b^x \geq (a + b)^x \qquad \text{if } x \in [0, 1] \tag{6.49}$$
$$a^x + b^x \leq (a + b)^x \qquad \text{if } x \geq 1. \tag{6.50}$$

.

Proof. Consider the following function defined on $\mathbb{R}_+$:

$$f(x) = \ln(a^x + b^x) - x \ln(a + b). \tag{6.51}$$

Its derivative is

$$f'(x) = \frac{a^x \ln(a) + b^x \ln(b)}{a^x + b^x} - \ln(a+b) \tag{6.52}$$

$$= \frac{a^x \ln\left(\frac{a}{a+b}\right) + b^x \ln\left(\frac{b}{a+b}\right)}{a^x + b^x}. \tag{6.53}$$

Since $\frac{a}{a+b} \leq 1$ and $\frac{b}{a+b} \leq 1$, we have $f'(x) \leq 0$. Therefore,

- If $x \in [0,1]$: we have $f(x) \geq f(1) = 0$. This implies $\ln(a^x + b^x) \geq x \ln(a+b)$ for all $x \in [0,1]$. Equivalently, $a^x + b^x \geq e^{x \ln(a+b)} = (a+b)^x$.
- If $x \geq 1$: we have $f(x) \leq f(1) = 0$. This implies $\ln(a^x + b^x) \leq x \ln(a+b)$, which leads to $a^x + b^x \leq e^{x \ln(a+b)} = (a+b)^x$ for $x \geq 1$.

□

Lemma 6.B.12. Let $0 \leq a, b \leq 1$ and $a + b = 1$. Then, for any $x \in [0,1]$, we have

$$(\max(a,b))^x + (\min(a,b))^x 2^{x-1} \geq 1. \tag{6.54}$$

Proof. Without loss of generality, assume $a \geq b$. It follows that $2b \leq 1$. Consider the following function defined on $[0,1]$:

$$f(x) = \frac{b^x 2^x}{2} + a^x. \tag{6.55}$$

Its derivative is

$$f'(x) = \frac{1}{2}\left[b^x \ln(b) 2^x + b^x 2^x \ln(2)\right] + a^x \ln(a) \tag{6.56}$$

$$= b^x 2^{x-1} \ln(2b) + a^x \ln(a). \tag{6.57}$$

Since $2b \leq 1$ and $a \leq 1$, we have $f'(x) \leq 0$. Therefore, for all $x \in [0,1]$, we have

$$f(x) \geq f(1) = a + b = 1. \tag{6.58}$$

□

Lemma 6.B.13. For any $q \in \Delta_K$, we have

$$(-\psi_{TE}(q_t)) = \frac{1}{\alpha}\left(\sum_{i=1}^{K} q_i^\alpha - 1\right) \geq \frac{q_*^\alpha}{\alpha}(1 - 2^{\alpha-1}), \tag{6.59}$$

where $q_* = \min(\max_{i\in[K]} q_i, 1 - \max_{i\in[K]} q_i)$. This implies $(-\psi_{TE}(q_t)) \geq \frac{q_*^\alpha}{4\alpha}(1-\alpha)$.

Proof. Let $q_{\max} = \max_{i\in[K]} q_i$. We consider two cases: $q_{\max} \leq 0.5$ and $q_{\max} > 0.5$.

- When $q_{\max} \leq 0.5$: we have $q_* = q_{\max}$. For any $i_{\max} \in \arg\max_{i\in[K]} q_i$, the inequality (6.59) is equivalent to

  $$q_*^\alpha 2^{\alpha-1} + \sum_{i\neq i_{\max}} q_i^\alpha \geq 1, \tag{6.60}$$

  Using

  $$\sum_{i\neq i_{\max}} q_i^\alpha \geq (\sum_{i\neq i_{\max}} q_i)^\alpha \tag{6.61}$$

  from Lemma 6.B.11 and combining with Lemma 6.B.12 leads to the desired claim.

- When $q_{\max} > 0.5$: in this case, we have $q_* = 1 - q_{\max} = \sum_{i\neq i_{\max}} q_i$. The desired inequality is equivalent to

  $$(q_{i_{\max}}^\alpha + q_*^\alpha 2^{\alpha-1}) + \sum_{i\neq i_{\max}} q_i^\alpha \geq 1 + (\sum_{i\neq i_{\max}} q_i)^\alpha. \tag{6.62}$$

  Again, this follows directly from

  $$\sum_{i\neq i_{\max}} q_i^\alpha \geq (\sum_{i\neq i_{\max}} q_i)^\alpha \tag{6.63}$$

  and Lemma 6.B.12.

The implication statement follows by $2^{\alpha-1} \leq (\alpha+3)/4$ for $\alpha \in [0,1]$. □

Lemma 6.B.14. Let $q \in \Delta_K$ and $p = (1 - \frac{K}{T})q + \frac{1}{T}\mathbf{1}$ where $T \geq 4K$. The following properties hold:

- $q_i \leq 2p_i$ for any $i \in [K]$.

- $q_* \leq 2p_*$ where $q_* = \min(\max_{i\in[K]} q_i, 1 - \max_{i\in[K]} q_i)$ and $p_* = \min(\max_{i\in[K]} p_i, 1 - \max_{i\in[K]} p_i)$.
- $p_* \geq \frac{1}{T}$.

Proof. By the definition of p, we have

$$2p_i = 2(1 - \frac{K}{T})q_i + \frac{2}{T} \geq q_i + q_i(1 - \frac{2K}{T}) \geq q_i.$$

Thus, the first statement holds.

Next, let $k \in \arg\max_{i\in[K]} q_i$. Obviously, we have $k \in \arg\max_{i\in[K]} p_i$ due to $p_i \geq p_j$ if $q_i \geq q_j$. If $q_k \leq 0.5$, then we have $q_* = q_k$. We also have $q_k \geq \frac{1}{K}$ and therefore $p_k = q_k + \frac{1-Kq_k}{T} \leq q_k \leq 0.5$. Moreover,

$$2p_* = 2p_k \geq q_k = q_*,$$

where the inequality is from the first statement. On the other hand, if $q_k > 0.5$ then we have $q_* = 1 - q_k$ and

$$\begin{aligned} p_* &= \min(p_k, 1 - p_k) \\ &= \min\left((1 - \frac{K}{T})q_k + \frac{1}{T}, 1 - \left((1 - \frac{K}{T})q_k + \frac{1}{T}\right)\right) \\ &= \min\left(q_k + \frac{1 - Kq_k}{T}, 1 - q_k + \frac{Kq_k - 1}{T}\right). \end{aligned}$$

Since $q_k \geq \frac{1}{K}$, we have $1 - q_k + \frac{Kq_k-1}{T} \geq 1 - q_k \geq \frac{1-q_k}{2}$. In addition,

$$q_k + \frac{1 - Kq_k}{T} \geq q_k + \frac{-K}{T} > 0.5 - \frac{1}{4} = \frac{1}{4} \geq \frac{1 - q_k}{2}.$$

Hence, we conclude that $p_* \geq \frac{1-q_k}{2} = \frac{q_*}{2}$. Thus, the second statement holds. The last statement follows from $p_* \geq \min_{i\in[K]} p_i \geq \frac{1}{T}$. □

Proof. (Of Lemma 6.B.1) Lemma 6.B.13 implies that for all t, we have

$$h_t \geq \frac{(p_t)_*^\alpha}{4\alpha} \geq \frac{T^{-\alpha}}{4\alpha},$$

where the last inequality is from $(p_t)_* \geq \frac{1}{T}$ by Lemma 6.B.14. Additionally,

$$\begin{aligned}
\mathbb{E}_{I_t}[z_t] &\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}_{I_t}\left[(\tilde{p}_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2\right] \\
&= \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \sum_{i=1}^{K} \frac{(\tilde{p}_{t,i})^{2-\alpha} \ell_{t,i}^2}{p_{t,i}} \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \sum_{i=1}^{K} (\tilde{p}_{t,i})^{1-\alpha} (\ell_{t,i}^{\frac{2}{\alpha}})^{\alpha} \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} (\sum_{i=1}^{K} \tilde{p}_{t,i})^{1-\alpha} (\sum_{i=1}^{K} \ell_{t,i}^{2/\alpha})^{\alpha} \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} S^{\alpha},
\end{aligned}$$

where the last inequality uses $\sum_{i=1}^{K} \tilde{p}_{t,i} \leq \sum_{i=1}^{K} p_{t,i} = 1$ and $S \geq \|\ell_t\|_0$. □

6.C Proof of the Lower Bounds in Theorem 6.3.4

6.C.1 Stochastic Lower Bound

Let $i^* \in [K]$ be fixed. Recall that $0 < \Delta_{\min} \leq \frac{1}{4}$ and $1 \leq U \leq \frac{K^\alpha}{4}$. We pick $b = \frac{U - \Delta_{\min}}{K^\alpha}$ so that $bK^\alpha + \Delta_{\min} = U$. We then have $\frac{U}{2K^\alpha} \leq b < \frac{1}{4}$. Our construction is as follows:

$$\ell_t = \begin{cases} -\mathbf{1} & \text{with probability } b, \\ -e_{i^*} & \text{with probability } \Delta_{\min}, \\ \mathbf{0} & \text{with probability } 1 - 2\Delta_{\min}, \end{cases}$$

where e_{i^*} is the i^*-th vector in the standard basis of $\mathbb{R}^K$. The expected loss vector is

$$\mathbb{E}[\ell_t] = -b\mathbf{1} - \Delta_{\min} e_{i^*}.$$

It follows that $\Delta_i = \Delta_{\min}$ for all $i \in [K] \setminus \{i^*\}$. In addition, the losses of arm i^* follow a Bernoulli distribution $\mathrm{Ber}(b + \Delta_{\min})$ while the losses of sub-optimal arms follow a Bernoulli

distribution Ber(b). We verify that the constraint in (6.12) holds:

$$\mathbb{E}\left[\left(\sum_{i=1}^{K}|\ell_{t,i}|^{2/\alpha}\right)^{\alpha}\right] = bK^{\alpha} + \Delta_{\min} = U.$$

For any consistent algorithm such that $R_T = o(T^x)$ for any $x > 0$, by [Lai and Robbins, 1985], we have

$$\begin{aligned}\liminf_{T\to\infty}\frac{R_T}{\ln(T)} &\geq \sum_{i\neq i^*}\frac{\Delta_i}{KL(b \parallel b+\Delta_{\min})}\\ &\geq \frac{K\Delta_{\min}}{2KL(b \parallel b+\Delta_{\min})}\\ &\gtrsim \frac{Kb}{\Delta_{\min}}\\ &\gtrsim \frac{K^{1-\alpha}U}{\Delta_{\min}},\end{aligned}$$

where the second inequality is from $K - 1 \geq \frac{K}{2}$ for all $K \geq 4$, the third inequality is by Lemma 6.C.1, and the last inequality is $b \geq \frac{U}{2K^{\alpha}}$.

Lemma 6.C.1. For any $b, \Delta \in (0, \frac{1}{4}]$, we have

$$KL(b \parallel b+\Delta) \leq \frac{4\Delta^2}{3b}.$$

Proof.

$$\begin{aligned}KL(b \parallel b+\Delta) &= b\ln\left(\frac{b}{b+\Delta}\right) + (1-p)\ln\left(\frac{1-b}{1-\Delta-p}\right)\\ &= -b\ln\left(1+\frac{\Delta}{b}\right) - (1-b)\ln\left(1-\frac{\Delta}{1-b}\right)\\ &\leq b\left(\frac{\Delta^2}{b^2}-\frac{\Delta}{b}\right) + (1-b)\left(\frac{\Delta^2}{(1-b)^2}+\frac{\Delta}{1-b}\right)\\ &= \Delta^2\left(\frac{1}{b}+\frac{1}{1-b}\right) = \frac{\Delta^2}{b(1-b)} \leq \frac{4\Delta^2}{3b},\end{aligned}$$

where the first inequality is due to $\ln(1+x) \geq x - x^2$ for all $x > 0$ and $\ln(1-x) \geq -x - x^2$ for all $0 \leq x \leq \frac{1}{3}$, and the second inequality is $1 - b \geq \frac{3}{4}$ for all $b \leq \frac{1}{4}$. □

6.C.2 Adversarial Lower Bound

For the adversarial lower bound, we construct a neutral environment V_0 and K competing environments $V_1, V_2, \ldots, V_K$, where:

- On V_0, the loss function is chosen by

$$\ell_t = \begin{cases} -\mathbf{1} & \text{with probability } \eta, \\ \mathbf{0} & \text{with probability } 1-\eta. \end{cases}$$

  It follows that the losses of all arms follow a Bernoulli distribution $\mathrm{Ber}(\eta)$ on V_0.

- On V_i for $i \in [K]$, the loss function is chosen by

$$\ell_t = \begin{cases} -\mathbf{1} & \text{with probability } \eta, \\ -e_i & \text{with probability } \epsilon, \\ \mathbf{0} & \text{with probability } 1-\eta-\epsilon. \end{cases}$$

  It follows that except for arm i, the losses of all other arms follow a Bernoulli distribution $\mathrm{Ber}(\eta)$ on V_i. The loss of arm i follows $\mathrm{Ber}(\eta+\epsilon)$.

Here, η and ϵ are constants chosen to be the solution of the following system of (in)equalities:

- $\eta + \epsilon \leq \frac{1}{4}$.
- $\eta K^\alpha + \epsilon = U$.
- $\frac{T}{K}\frac{8\epsilon^2}{\eta} = 1$.
- $\eta K^\alpha \geq \frac{U}{2}$.

Note that for all $K \geq 4, T \geq 4K$ and $U \leq \frac{K^\alpha}{4}$, the solution

$$\begin{aligned} \sqrt{\eta} &= \frac{-\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^\alpha U}}{2K^\alpha}, \\ \epsilon &= \sqrt{\frac{\eta K}{8T}} \end{aligned} \tag{6.64}$$

satisfies the system of inequalities since

$$\sqrt{\eta} = \frac{-\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^{\alpha}U}}{2K^{\alpha}} \leq \frac{\sqrt{1+\frac{1}{32}}}{2K^{\alpha}} \leq \frac{1}{8},$$
$$\epsilon = \sqrt{\frac{\eta K}{8T}} \leq \sqrt{\frac{K}{64T}} \leq \frac{1}{8}.$$

Moreover, we have

$$\begin{aligned}
\eta &= \frac{1}{4K^{2\alpha}} \left(-\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^{\alpha}U} \right)^2 \\
&= \frac{1}{4K^{2\alpha}} \frac{16K^{2\alpha}U^2}{\left(\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^{\alpha}U} \right)^2} \\
&= \frac{4U^2}{\left(\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^{\alpha}U} \right)^2} \\
&\geq \frac{U^2}{4K^{\alpha}U} = \frac{1}{4}\frac{U}{K^{\alpha}} \\
&\geq \frac{K^{1-2\alpha}}{8T},
\end{aligned}$$

where the second equality is $\sqrt{a} - \sqrt{b} = \frac{a-b}{\sqrt{a}+\sqrt{b}}$, the first inequality is $\frac{K}{T} \leq \frac{K^{\alpha}U}{4}$ and the last inequality is $\frac{U}{K^{\alpha}} \geq \frac{K^{1-2\alpha}}{2T}$, both hold for all $T \geq 4K$ and $U \geq 1$. This implies that

$$\eta^2 K^{2\alpha} = \eta(\eta K^{2\alpha}) \geq \eta \frac{K^{1-2\alpha}}{8T} K^{2\alpha} = \frac{\eta K}{8T} = \epsilon^2.$$

As a result, $\eta K^{\alpha} \geq \epsilon = U - \eta K^{\alpha}$. Hence, $\eta K^{\alpha} \geq \frac{U}{2}$.

With the choice of η and ϵ in (6.64), we verify that (6.12) holds:

$$\mathbb{E}\left[\left(\sum_{i=1}^{K} |\ell_{t,i}|^{2/\alpha} \right)^{\alpha} \right] = \eta K^{\alpha} + \epsilon = U.$$

Let $\mathcal{A}$ denote the algorithm of a learner. Let $N_i = \sum_{t=1}^{T} \mathbb{1}\{I_t = i\}$ denote the number of times arm i is pulled by $\mathcal{A}$. Let P_0 and P_i denote the distribution of the observed losses on V_0 and V_i, respectively. Similarly, let $\mathbb{E}_0$ and $\mathbb{E}_i$ denote the expectation taken on V_0 and V_i, respectively.

We first run $\mathcal{A}$ on V_0. Let $a = \arg\min_{i \in [K]} \mathbb{E}_0[N_i]$ be the arm that is pulled the least in

expectation. Since $\sum_{i=1}^{K} N_i = T$, we have $\mathbb{E}_0[N_a] \leq \frac{T}{K}$.

By the standard arguments in establishing lower bounds for adversarial bandits [e.g. Auer et al., 2002a, Equation 28-30], we have

$$\begin{aligned}
\mathbb{E}_a[N_a] &\leq \mathbb{E}_0[N_a] + \frac{T}{2}\|P_a - P_0\|_1 \\
&\leq \mathbb{E}_0[N_a] + \frac{T}{2}\sqrt{2\ln(2) KL(P_0 \parallel P_a)} \\
&\leq \mathbb{E}_0[N_a] + \frac{T}{2}\sqrt{2\ln(2)\mathbb{E}_0[N_a] KL(\eta \parallel \eta+\epsilon)} \\
&\leq \frac{T}{K} + \frac{T}{2}\sqrt{2\ln(2)\frac{T}{K} KL(\eta \parallel \eta+\epsilon)} \\
&\leq \frac{T}{K} + \frac{T}{2}\sqrt{2\ln(2)\frac{T}{K}\frac{4\log_2(e)\epsilon^2}{\eta}} \\
&= \frac{T}{K} + \frac{T}{2}\sqrt{\frac{T}{K}\frac{8\epsilon^2}{\eta}},
\end{aligned}$$

where

- the second inequality is Pinsker's inequality,
- the third inequality is due to the chain rule for KL-divergence [Cover and Thomas, 2006] and the fact that V_0 and V_a differ only by the loss distribution of arm a,
- the fourth inequality is because $\mathbb{E}_0[N_a] \leq \frac{T}{K}$,
- the last inequality is the reverse Pinsker's inequality [Sason, 2015].

This further implies that $\mathbb{E}_a[N_a] \leq \frac{T}{K} + \frac{T}{2} \leq \frac{3T}{4}$ for $K \geq 4$. Hence,

$$\begin{aligned}
\mathbb{E}_a[R_{T,a}] &= \epsilon(T - \mathbb{E}_a[N_a]) \\
&\geq \epsilon(T - \frac{3T}{4}) = \frac{T\epsilon}{4} \\
&= \sqrt{\frac{TK\eta}{32}} \\
&= \Omega(\sqrt{K^{1-\alpha}UT}),
\end{aligned}$$

where the last equality is due to $\eta K^{\alpha} \geq \frac{U}{2}$.

Algorithm 6.3 Optimistic FTRL using Tsallis entropy plus log-barrier regularization for losses in $[0,1]$

Input: $K \geq 1, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{4K}{1-\alpha}, \gamma = \max(3, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $L_{0,i} = 0$ for $i \in [K]$
for each round $t = 1, \ldots, T$ **do**
  Compute $m_t \in [0,1]^K$
  Compute $q_t = \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(x_i)$
  Compute $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$
  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
  Compute loss estimate $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{I_t = i\}}{p_{t,i}}$ $\quad L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
  Compute $z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2 \right)$
  Compute $h_t = \left(\frac{1}{\alpha}(\sum_{i=1}^K p_{t,i}^\alpha - 1) \right)$
  Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$
end

6.D Proofs for Section 6.4

Before proving Theorem 6.4.1, in Appendix 6.D.1, we first establish the BOBW regret bound for an algorithm that combines real-time SPM and Optimistic FTRL without any reservoir samplings. The full procedure is given in Algorithm 6.3. Later on, in Appendix 6.D.3, we will use this regret bound in the analysis of Algorithm 6.4.

6.D.1 A General SPM-based Regret Bound for Optimistic FTRL

We consider the adversarial multi-armed bandits with losses in $[0,1]$. Note that the analysis can be trivially extended to the $[-1,1]$ case by increasing β_t and γ by a multiplicative factor of 2.

In round t, the learner computes $m_t \in [0,1]^K$ before drawing arm I_t and uses Optimistic FTRL with the hybrid regularizer

$$\begin{aligned} q_t &= \arg\min_{x \in \Delta_K} \langle m_t + \sum_{s=1}^{t-1} \hat{\ell}_s, x \rangle + \phi_t(x) \\ &= \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \psi_{TE}(x) + \gamma \psi_{LB}(x) \\ &= \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(1 - \sum_{i=1}^K x_i^\alpha \right) - \gamma \sum_{i=1}^K \ln(p_i). \end{aligned}$$

Let $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$. The learner draws $I_t \sim p_t$ and use the unbiased loss estimator

$$\hat{\ell}_{t,i} = m_{t,i} + \frac{\ell_{t,i} - m_{t,i}}{p_{t,i}} \mathbb{1}\{I_t = i\}.$$

SPM learning rates: the learning rates are set according to SPM rule [Ito et al., 2024], where

$$h_t = (-\psi_{TE})(p_t) = \frac{1}{\alpha}\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1\right),$$
$$z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2\right).$$

Details are given in Algorithm 6.3.

6.D.2 Analysis for Algorithm 6.3

Similar to [Ito et al., 2022], the analysis uses

$$\begin{aligned} r_{t+1} &= \arg\min_{x \in \Delta_K} \langle L_{t-1} + \hat{\ell}_t, x\rangle + \frac{\beta_{t+1}}{\alpha}(1 - \sum_{i=1}^{K} x_i^{\alpha}) - \gamma \sum_{i=1}^{K} \ln(x_i) \\ &= \arg\min_{x \in \Delta_K} \langle L_{t-1} + m_t + \frac{\ell_{tI_t} - m_{t,I_t}}{p_{t,I_t}} e_{I_t}, x\rangle + \frac{\beta_{t+1}}{\alpha}(1 - \sum_{i=1}^{K} x_i^{\alpha}) - \gamma \sum_{i=1}^{K} \ln(x_i), \end{aligned}$$

where $2p_{t,I_t} \geq q_{t,I_t}$. Observe that $(\ell_{t,I_t} - m_{t,I_t}) \in [-1, 1]$ and

$$\begin{aligned} \beta_{t+1} - \beta_t &= \frac{z_t}{\beta_t h_t} \\ &\leq \frac{18 d^2}{h_t \gamma} \\ &\leq \left(1 - \frac{1}{d}\right) \gamma q_{t*}^{-\alpha}, \end{aligned}$$

similar to the proof of Lemma 6.B.4. Therefore, we can invoke Corollary 6.B.10 with $\omega = 2$ and obtain $r_{t+1,i} \leq 3 q_{t,i} \leq 6 p_{t,i}$ for all $i \in [K]$. Combining this with Lemma 1 in [Ito et al.,

2022] and our Lemma 6.B.2, we obtain

$$\begin{aligned}
\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u\rangle &\le \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}(\phi_t(r_{t+1}) - \phi_{t+1}(r_{t+1})) \\
&\quad + \sum_{t=1}^{T}\langle \hat{\ell}_t - m_t, q_t - r_{t+1}\rangle - D_t(r_{t+1}, q_t) \\
&\le \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}(\beta_{t+1} - \beta_t)(-\psi_{TE}(r_{t+1})) + \sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\le \phi_{T+1}(u) - \phi_1(r_1) + 6\left(\sum_{t=1}^{T}(\beta_{t+1} - \beta_t)(-\psi_{TE}(p_{t+1})) + \sum_{t=1}^{T}\frac{z_t}{\beta_t}\right) \\
&= \phi_{T+1}(u) - \phi_1(r_1) + 6\left(\sum_{t=1}^{T}(\beta_{t+1} - \beta_t)h_t + \sum_{t=1}^{T}\frac{z_t}{\beta_t}\right) \\
&= \phi_{T+1}(u) - \phi_1(r_1) + 12\sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\le \gamma K \ln(T) + \frac{\beta_1(K^{1-\alpha} - 1)}{\alpha} + 12\sum_{t=1}^{T}\frac{z_t}{\beta_t}.
\end{aligned}$$

It follows that

$$\begin{aligned}
R_T &\le \mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u\rangle\right] + 3K \\
&\lesssim \gamma K \ln(T) + \frac{\beta_1(K^{1-\alpha} - 1)}{\alpha} + \mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right].
\end{aligned}$$

Applying (6.23) with $S = K$, we obtain the following bounds on $\sum_{t=1}^{T}\frac{z_t}{\beta_t}$ in each environment.

In adversarial regime with a self-bounding constraint:

With $J = \log_2(T)$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right] &\lesssim \sqrt{\ln(T)\sum_{t=1}^{T}\mathbb{E}[h_t z_t]} \\
&\leq \sqrt{\ln(T)\sum_{t=1}^{T}\mathbb{E}[h_t \mathbb{E}_{I_t}[z_t]]} \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)}\mathbb{E}\left[h_t\left(\sum_{i=1}^{K}(\tilde{p}_{t,i}^{1-\alpha})(\ell_{t,i}-m_{t,i})^2\right)\right] \\
&= \frac{(6d)^{2-\alpha}}{2(1-\alpha)}\mathbb{E}\left[\left(\frac{1}{\alpha}\sum_{i=1}^{K}p_{t,i}^{\alpha}-1\right)\left(\sum_{i=1}^{K}(\tilde{p}_{t,i}^{1-\alpha})(\ell_{t,i}-m_{t,i})^2\right)\right] \\
&\leq \frac{(6d)^{2-\alpha}}{2\alpha(1-\alpha)}\mathbb{E}\left[\left(\sum_{i=1}^{K}p_{t,i}^{\alpha}-1\right)\left(\sum_{\ell_{t,i}\neq 0}\tilde{p}_{t,i}^{1-\alpha}\right)\right],
\end{aligned}$$

where the last inequality is from $(\ell_{t,i}-m_{t,i})^2 \leq 1$. Observe that the last bound is exactly the bound in (6.24), hence

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right] \lesssim O\left(\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}K^{\alpha}}{\alpha(1-\alpha)}}\right) \tag{6.65}$$

holds for Optimistic FTRL as well.

In adversarial bandits:

$$\begin{aligned}
\sqrt{\mathbb{E}\left[h_{\max}\sum_{t=1}^{T}z_t\right]} &\lesssim \sqrt{\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)}\mathbb{E}\left[\sum_{t=1}^{T}p_{t,I_t}^{-\alpha}(\ell_{t,I_t}-m_{t,I_t})^2\right]} \\
&= \sqrt{\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}p_{t,i}^{1-\alpha}(\ell_{t,i}-m_{t,i})^2\right]} \\
&\leq \sqrt{\frac{(K^{1-\alpha}-1)}{\alpha(1-\alpha)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\ell_{t,i}-m_{t,i})^2\right]},
\end{aligned} \tag{6.66}$$

where the inequality is due to $p_{t,i}^{1-\alpha} \leq 1$.

6.D.3 Proof for Theorem 6.4.1

Let $\mu_t = \frac{1}{s}\sum_{s=1}^{t} \ell_s$ and

$$Q = \sum_{t=1}^{T} \|\ell_t - \mu_T\|_2^2.$$

Using the reservoir sampling technique in [Hazan and Kale, 2011], we can use a prediction vector m_t satisfying $\mathbb{E}[m_t] = \mu_t$ and $\mathrm{Var}[m_t] \le \frac{Q}{t\ln(T)}$.

Regret Analysis

Without loss of generality, assume $\ln(T) \in \mathbb{N}$ (otherwise, this increases at most a constant factor in the regret bound). For any fixed $u \in \Delta_K$, we have,

$$\sum_{t=1}^{T}\langle \ell_t, p_t - u\rangle = \sum_{t=K\ln(T)+1} \langle \ell_t, p_t - u\rangle + \sum_{t=1}^{K\ln(T)} \langle \ell_t, p_t - u\rangle \tag{6.67}$$

$$\le \sum_{t=K\ln(T)+1} \langle \ell_t, p_t - u\rangle + K\ln(T) \tag{6.68}$$

$$= \underbrace{\sum_{t=K\ln(T)+1} \mathbb{1}\{b_t = 1\}\langle \ell_t, p_t - u\rangle}_{(A)} + \underbrace{\sum_{t=1}^{T} \mathbb{1}\{b_t = 0\}\langle \ell_t, p_t - u\rangle}_{(B)} + K\ln(T). \tag{6.69}$$

Recall that $b_t = 1$ indicates a reservoir sampling round where, for $t > K\ln(T)$, the sampling probability is the uniform distribution $p_t = \frac{1}{K}\mathbf{1}$, and $b_t = 0$ indicates an FTRL round. Next, we bound the expectation of A and B in the equation above. First, we have

$$\mathbb{E}[A] \le \mathbb{E}\left[\sum_{t=K\ln(T)+1}^{T} \mathbb{1}\{b_t = 1\}\right] = \sum_{t=K\ln(T)+1}^{T} \Pr[b_t = 1] \le \sum_{t=1}^{T} \frac{K\ln(T)}{t} \le O(K(\ln(T))^2). \tag{6.70}$$

Next, the set of rounds with $b_t = 0$ are the Optimistic FTRL rounds; hence, we can apply (6.65) and (6.66). In the adversarial regime with a self-bounding constraint, we have

$$\begin{aligned}
\mathbb{E}[B] &\lesssim \sqrt{\ln(T)\mathbb{E}[\mathbb{1}\{b_t = 0\}h_t z_t]} \\
&\lesssim \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}\mathbb{E}\left[\mathbb{1}\{b_t = 0\}\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1\right)\left(\sum_{i=1}^{K} \tilde{p}_{t,i}^{1-\alpha}\right)\right]} \\
&\leq \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}\mathbb{E}\left[\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1\right)\left(\sum_{i=1}^{K} \tilde{p}_{t,i}^{1-\alpha}\right)\right]},
\end{aligned}$$

where the last inequality is from $\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1\right)\left(\sum_{i=1}^{K} \tilde{p}_{t,i}^{1-\alpha}\right) \geq 0$ in the rounds where $b_t = 1$ (in such rounds, p_t is either a one-hot vector if $t \leq K \ln T$ or $\frac{1}{K}\mathbf{1}$ if $t > K \ln T$). By (6.65), we obtain

$$\mathbb{E}[B] \lesssim O\left(\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}K^{\alpha}}{\alpha(1-\alpha)}}\right).$$

In the adversarial regime, we have

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}z_t] &\lesssim \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\sum_{i=1}^{K}(\ell_{t,i} - m_{t,i})^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\sum_{i=1}^{K}(\ell_{t,i} - \tilde{\mu}_{t,i})^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\|\ell_t - \tilde{\mu}_t\|_2^2\right] \\
&\leq \left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\|\ell_t - \mu_t\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\|\tilde{\mu}_t - \mu_t\|_2^2\right]\right) \\
&\leq \left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\|\ell_t - \mu_T\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{b_t = 0\}\|\tilde{\mu}_t - \mu_t\|_2^2\right]\right) \\
&\leq \left(Q + \sum_{t=1}^{T}\frac{Q}{t\ln(T)}\right) \leq 3Q,
\end{aligned}$$

where the first inequality is triangle inequality, the second inequality is $\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t - \mu_t\|_2^2\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t - \mu_T\|_2^2\right]$ by Lemma 10 in [Hazan and Kale, 2011], the third inequality is by

Lemma 11 in [Hazan and Kale, 2011], and the last inequality is due to $\sum_{t=1}^{T} \frac{1}{t} \leq \ln(T) + 1$. Overall, the regret for adversarial bandits is

$$R_T \lesssim \sqrt{\frac{(K^{1-\alpha} - 1)Q}{\alpha(1-\alpha)}}.$$

6.E Proofs for Section 6.5

We have

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, p_t \rangle - u] &= \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, q_t - u + \frac{\mathbf{1} - K q_t}{T} \rangle] \\
&\leq \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, q_t - u \rangle] + K \\
&= \mathbb{E}[\sum_{t=1}^{T} \langle \hat{\ell}_t, q_t - u \rangle] + K.
\end{aligned}$$

Furthermore, let

$$r_{t+1} = \arg\min_{x \in \Delta_K} \langle L_{t-1} + \hat{\ell}_t, x \rangle + \sum_{i=1}^{K} \beta_{t,i} \left(\frac{1}{\alpha}(-x_i^{\alpha}) + (1 - x_i)\ln(1 - x_i) + x_i \right) - \gamma \sum_{i=1}^{K} \ln(x_i).$$

Then,

$$
\begin{aligned}
&\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u\rangle \\
&\le \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}\sum_{i=1}^{K}(\beta_{t+1,i} - \beta_{t,i})\left(\frac{p_{t+1,i}^{\alpha}}{\alpha} + (p_{t+1,i}-1)\ln(1-p_{t+1,i}) - p_{t+1,i}\right) + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&\le \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}\frac{2}{\alpha}(\beta_{t+1,I_t} - \beta_{t,I_t})p_{t+1,I_t}^{\alpha} + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&\le \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}\frac{12}{\alpha}(\beta_{t+1,I_t} - \beta_{t,I_t})p_{t,I_t}^{\alpha} + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&= \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}12(\beta_{t+1,I_t} - \beta_{t,I_t})h_{t,I_t} + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&= \phi_{T+1}(u) - \phi_1(r_1) + 13\sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&= \phi_{T+1}(u) - \phi_1(r_1) + 13\sum_{i=1}^{K}\sum_{t=1}^{T}\frac{z_{t,i}}{\beta_{t,i}},
\end{aligned}
$$

where the first inequality is from $p_{t+1,i} \ge 0$ and Lemma 6.E.5, the second inequality is from $p_{t+1,I_t}^{\alpha} \le (6p_{t,I_t})^{\alpha} \le 6p_{t,I_t}^{\alpha}$ and the last equality is $z_{t,i} = 0$ for all $i \neq I_t$.

From the previous section, we have for all $i \in [K]$,

$$
\sum_{t=1}^{T}\frac{z_{t,i}}{\beta_{t,i}} \lesssim \min\left\{\sqrt{\mathbb{E}\left[\ln(T)\sum_{t=1}^{T}h_{t,i}z_{t,i}\right]} + \sqrt{\frac{1}{T}\mathbb{E}\left[h_{i,\max}\sum_{t=1}^{T}z_{t,i}\right]}, \sqrt{\mathbb{E}\left[h_{i,\max}\sum_{t=1}^{T}z_{t,i}\right]}\right\}. \tag{6.71}
$$

First, we have $h_{i,\max} = \max_t h_{t,i} = \frac{1}{\alpha}\max_t p_{t,i}^{\alpha} \le \frac{1}{\alpha}$. In addition, $\mathbb{E}_{I_t}[z_{t,i}] \le \mathbb{E}_{I_t}[\mathbb{1}\{I_t = i\}p_{t,i}^{-\alpha}(\ell_{t,i} - m_{t,i})^2] = p_{t,i}^{1-\alpha}(\ell_{t,i} - m_{t,i})^2 \le 1$. Therefore, the sum $\frac{1}{T}\mathbb{E}\left[h_{i,\max}\sum_{t=1}^{T}z_{t,i}\right]$ is bounded by $\frac{1}{\alpha}$. We can simplify (6.71) by

$$
\sum_{t=1}^{T}\frac{z_{t,i}}{\beta_{t,i}} \lesssim \min\left\{\sqrt{\ln(T)\mathbb{E}\left[\sum_{t=1}^{T}h_{t,i}z_{t,i}\right]} + \sqrt{\frac{1}{\alpha}}, \sqrt{\mathbb{E}\left[\sum_{t=1}^{T}z_{t,i}\right]}\right\}. \tag{6.72}
$$

A Bound for Stochastic Bandits from $\sqrt{\ln(T)\mathbb{E}\left[\sum_{t=1}^{T} h_{t,i}z_{t,i}\right]}$

We have

$$h_{t,i}z_{t,i} \lesssim \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_{t,i})^2 p_{t,i}^{\alpha} \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{p_{t,I_t}^{-\alpha}, \frac{1-p_{t,I_t}}{p_{t,I_t}^2}\right\}\right).$$

Therefore, by Lemma 6.E.9,

$$\mathbb{E}_{I_t}[h_{t,i}z_{t,i}] \lesssim \frac{1}{1-\alpha}\tilde{p}_{t,i}(\ell_{t,i} - m_{t,i})^2.$$

Denote $P_i = \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{I_t = i\}]$. Bounding $(\ell_{t,i} - m_{t,i})^2 \le 1$ for any $m_t \in [0,1]^K$, we obtain

$$\begin{aligned}
\mathrm{Reg}_T &\lesssim \sum_{i=1}^{K} \sqrt{\mathbb{E}\left[\ln(T)\sum_{t=1}^{T} h_{t,i}z_{t,i}\right]} \\
&\lesssim \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}}\left(\sum_{i\neq i^*}\sqrt{P_i} + \sqrt{(\sum_{i\neq i^*} P_i)}\right) + K\ln(T) \\
&\le \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}}\left(\sum_{i\neq i^*}\sqrt{P_i} + \frac{1}{\sqrt{K-1}}\sum_{i\neq i^*}\sqrt{P_i}\right) + K\ln(T) \\
&\lesssim \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}}\left(\sum_{i\neq i^*}\sqrt{P_i}\right) + K\ln(T).
\end{aligned}$$

Similar to [Ito et al., 2022], by using $\mathrm{Reg}_T = 2\mathrm{Reg}_T - \mathrm{Reg}_T$ and $2\sqrt{ax} - bx \le \frac{a}{b}$ for $a = \frac{\ln(T)}{\alpha(1-\alpha)}, b = \Delta_i$ and $x = P_i$, we obtain (note that we set $\alpha = \frac{1}{2}$)

$$\mathrm{Reg}_T \lesssim \frac{1}{\alpha(1-\alpha)}\sum_{i\neq i^*}\frac{\ln(T)}{\Delta_i}.$$

Bounds for Adversarial Bandits

Fix $i \in [K]$. Since $\tilde{p}_{t,i} \le p_{t,i}$, we have $h_{t,i}z_{t,i} \lesssim \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_{t,i})^2$. Therefore, by setting $m_{t,i}$ to be the output of an online learning algorithm with fully-observable squared loss as

in [Ito et al., 2022], i.e.,

$$m_{t,i} = \frac{1}{1 + \sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\}} \left(\frac{1}{2} + \sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\} \ell_{t,i} \right)$$

and then applying their Lemma 3, we obtain for any fixed $m^* \in [0,1]^K$,

$$\sum_{t=1}^{T} \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_{t,i})^2 \lesssim \sum_{t=1}^{T} \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_i^*)^2 + \ln\left(1 + \sum_{t=1}^{T} \mathbb{1}\{I_t = i\}\right).$$

As already shown in [Ito et al., 2022], for each appropriately chosen m^*, we would recover the data-dependent bounds of order $\sqrt{KQ_\infty \ln(T)}$ (with $m^* \in \arg\min_{\bar{\ell} \in \mathbb{R}^K} \sum_{t=1}^{T} \left\|\ell_t - \bar{\ell}\right\|_2^2$), $\sqrt{KL^* \ln(T)}$ (with $m^* = \mathbf{0}$) and $\sqrt{K(T - L^*)\ln(T)}$ (with $m^* = \mathbf{1}$).

On the other hand, from the quantity $\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} z_{t,i}\right]}$ and Jensen's inequality, we obtain

$$\begin{aligned}
\mathrm{Reg}_T &\lesssim \sqrt{K\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} z_{t,i}\right]} \\
&\lesssim \sqrt{K\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_{t,i}^{1-\alpha}(\ell_{t,i} - m_{t,i})^2\right]} \\
&\lesssim \sqrt{K\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_{t,i}^{1-\alpha}\right]} \\
&\leq \sqrt{K\mathbb{E}\left[\sum_{t=1}^{T} K^{\alpha}\right]} \\
&= \sqrt{K^{1+\alpha}T} = K^{\frac{1}{4}}\sqrt{KT} \qquad (\alpha = 1/2),
\end{aligned}$$

which grows with $\sqrt{T}$ in the worst-case.

6.E.1 Stability Proofs

In this section, we define the following function

$$g(x) = x_i^{\alpha-1} + \ln(1 - x). \tag{6.73}$$

Note that g is decreasing. In addition, let $d_f(y,x) = f(y) - f(x) - f'(x)(y-x)$ denote the Bregman divergence associated with a one-dimensional strictly convex function $f : \mathbb{R} \to \mathbb{R}$. Note that $d_f(y,x) \geq 0$ for all $x, y \in \mathbb{R}$.

Lemma 6.E.1. For any $L \in \mathbb{R}^K, \beta \in R_+^K, \gamma > 0$ and $d \geq 2$, let

$$x = \arg\min_{p \in \Delta_K} \langle L, p \rangle + \sum_{i=1}^K \beta_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i \right) - \gamma \sum_{i=1}^K \ln(p_i)$$
$$y = \arg\min_{p \in \Delta_K} \langle L, p \rangle + \sum_{i=1}^K \beta_i' \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i \right) - \gamma \sum_{i=1}^K \ln(p_i).$$

If $0 \leq \beta_1' - \beta_1 \leq \left(1 - \frac{1}{d}\right) \gamma x_1^{-\alpha}$ and $\beta_i' = \beta_i$ for $i > 1$, then $y_1 \leq d x_1$.

Proof. If $dx_1 \geq 1$ then $y_1 \leq 1 \leq dx_1$ trivially. Hence, we assume $dx_1 \leq 1$. By the Lagrange multiplier method, we have for $i = 2, \ldots, K$ and some $\lambda, \lambda' \in \mathbb{R}$,

$$L_i - \beta_i(x_i^{\alpha-1} + \ln(1-x_i)) - \frac{\gamma}{x_i} = \lambda,$$
$$L_i - \beta_i(x_i^{\alpha-1} + \ln(1-y_i)) - \frac{\gamma}{y_i} = \lambda'.$$

Similary, for $i = 1$, we have

$$L_1 - \beta_1(x_1^{\alpha-1} + \ln(1-x_1)) - \frac{\gamma}{x_1} = \lambda,$$
$$L_1 - \beta_1'(x_1^{\alpha-1} + \ln(1-y_1)) - \frac{\gamma}{y_1} = \lambda'.$$

Taking $Z = \lambda' - \lambda$ over all K pairs of equations, we obtain

$$\beta_i(g(x_i) - g(y_i)) + \gamma \left(\frac{1}{x_i} - \frac{1}{y_i} \right) = Z \tag{6.74}$$
$$\beta_1 g(x_1) - \beta_1' g(y_1) + \gamma \left(\frac{1}{x_1} - \frac{1}{y_1} \right) = Z. \tag{6.75}$$

If $Z \geq 0$, then since $\beta_i > 0$ and both $g(x)$ and $\frac{\gamma}{x}$ are decreasing, we have $y_i \geq x_i$ for all $i \neq 1$. This straightforwardly implies that $y_1 \leq x_1$. Thus, we focus on the case $Z < 0$. In this case, we have $y_i < x_i$ for all $i \neq 1$ and $y_1 > x_1$. We consider two cases:

- If $g(x_1) \leq 0$: from $y_1 \geq x_1$, we have $g(y_1) \leq g(x_1) \leq 0$. Hence, from $0 < \beta_1 \leq \beta_1'$, we

obtain

$$\beta_1' g(y_1) \leq \beta_1 g(y_1) \leq \beta_1 g(x_1).$$

This implies that $0 > Z = \beta_1 g(x_1) - \beta_1' g(y_1) + \gamma \left(\frac{1}{x_1} - \frac{1}{y_1}\right) \geq 0$, a contradiction.

- If $g(x_1) > 0$: in this case, (6.75) and $Z < 0$ implies $\beta_1' g(y_1) = \beta_1 g(x_1) + \gamma \left(\frac{1}{x_1} - \frac{1}{y_1}\right) - Z > 0$. Furthermore, by re-arranging, we obtain

$$\begin{aligned}
\beta_1' g(y_1) + \frac{\gamma}{y_1} &\geq \beta_1 g(x_1) + \frac{\gamma}{x_1} \\
&\geq \left(\beta_1' - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}\right)(x_1^{\alpha-1} + \ln(1-x_1)) + \frac{\gamma}{x_1} \\
&= \beta_1'(x_1^{\alpha-1} + \ln(1-x_1)) - \left(1 - \frac{1}{d}\right)\gamma x^{-1} \\
&\quad - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}\ln(1-x_1) + \frac{\gamma}{x_1} \\
&= \beta_1'(x_1^{\alpha-1} + \ln(1-x_1)) + \frac{\gamma}{dx_1} - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}\ln(1-x_1) \\
&\geq \beta_1'(x_1^{\alpha-1} + \ln(1-x_1)) + \frac{\gamma}{dx_1} \\
&\geq \beta_1'((dx_1)^{\alpha-1} + \ln(1-dx_1)) + \frac{\gamma}{dx_1} \\
&= \beta_1' g(dx_1) + \frac{\gamma}{dx_1},
\end{aligned}$$

where the second inequality is from $\beta_1 \geq \beta_1' - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}$, the third inequality is due to $\ln(1-x_1) < 0$ and the last inequality is from $d \geq 2 > 1$. Since $\beta g(x) + \frac{\gamma}{x}$ is decreasing for all $\beta > 0, \gamma > 0$, we conclude that $y_1 \leq dx_1$.

□

Lemma 6.E.2. For any $L \in \mathbb{R}^K, \beta \in R_+^K, \gamma > 0$ and $h \in [-1, 1]$, let

$$\begin{aligned}
x &= \arg\min_{p \in \Delta_K} \langle L, p\rangle + \sum_{i=1}^K \beta_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i\right) - \gamma \sum_{i=1}^K \ln(p_i) \\
y &= \arg\min_{p \in \Delta_K} \langle L + \frac{h}{x_1'}, p\rangle + \sum_{i=1}^K \beta_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i\right) - \gamma \sum_{i=1}^K \ln(p_i),
\end{aligned}$$

where $4x_1' \geq x_1$. If $\gamma \geq 6$ then $y_1 \leq 3x_1$.

Proof. Using the Lagrange multiplier method, we have the following equalities that hold for some $Z \in \mathbb{R}$,

$$\beta_1 \left(g(y_1) - g(x_1)\right) + \gamma \left(\frac{1}{y_1} - \frac{1}{x_1}\right) = Z + \frac{h}{x_1'} \tag{6.76}$$

and for all $i \neq 1$,

$$\beta_i \left(g(y_i) - g(x_i)\right) + \gamma \left(\frac{1}{y_i} - \frac{1}{x_i}\right) = Z. \tag{6.77}$$

First, we show that Z and $y_1 - x_1$ has the opposite sign to h. We consider two cases:

- If $Z \geq 0$ then from (6.77) and the monotonic decreasing property of $\beta g(x) + \frac{\gamma}{x}$, we have $y_i \leq x_i$ and this leads to $y_1 \geq x_1$. Combining $y_1 \geq x_1$ and (6.76), we have $Z + \frac{h}{x_1'} \leq 0$. Since $Z \geq 0$, this implies $h \leq 0$.
- If $Z \leq 0$ then by the same argument, we have $y_i \geq x_i$ and $y_1 \leq x_1$. Therefore, $Z + \frac{h}{x_1'} \geq 0$. Due to $Z \leq 0$, we must have $h \geq 0$.

In both cases, we have $Zh \leq 0$ and $Z(y_1 - x_1) \geq 0$. It follows that if $h \geq 0$ then we have $y_1 \leq x_1 \leq 3x_1$. If $h < 0$ then $y_1 \geq x_1$, and by rearranging (6.76), we obtain

$$\begin{aligned}
\frac{4}{x_1} \geq -\frac{h}{x_1'} &= \underbrace{Z}_{\geq 0} + \underbrace{\gamma \left(\frac{1}{x_1} - \frac{1}{y_1}\right)}_{\geq 0} + \underbrace{\beta(g(x_1) - g(y_1))}_{\geq 0} \\
&\geq \gamma \left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\
&\geq 6 \left(\frac{1}{x_1} - \frac{1}{y_1}\right),
\end{aligned}$$

where the last inequality is due to $\gamma \geq 6$. This implies that $\frac{3}{y_1} \geq \frac{1}{x_1}$, thus $y_1 \leq 3x_1$. □

By combining Lemma 6.E.1 and Lemma 6.E.2, we obtain the following corollary. The proof of this corollary is nearly identical to that of Corollary 6.B.10.

Corollary 6.E.3. For any $t \in [T]$, Algorithm 6.2 guarantees that

$$r_{t+1,I_t} \leq 3dq_{t,I_t}.$$

Lemma 6.E.4. For all $t \in [T]$, Algorithm 6.2 guarantees that

$$\langle \hat{\ell}_t - m_t, q_t - r_{t+1} \rangle - D_t(r_{t+1}, q_t) \leq \frac{z_{t,I_t}}{\beta_{t,I_t}},$$

where

$$z_{t,I_t} = (\ell_{t,I_t} - m_{t,I_t})^2 \min\left\{ \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{ p_{t,I_t}^{-\alpha}, \frac{1-p_{t,I_t}}{p_{t,I_t}^2} \right\}, \frac{\beta_{t,I_t} 18 d^2}{\gamma} \right\}.$$

Proof. Let

$$\begin{aligned} f_1(x) &= \frac{-x^\alpha}{\alpha}, \\ f_2(x) &= (1-x)\ln(1-x) + x, \\ f_3(x) &= -\ln(x). \end{aligned}$$

Since $\phi_t(x) = \sum_{i=1}^K \left(\beta_{t,i}(f_1(x_i) + f_2(x_i)) + \gamma f_3(x_i)\right)$, we have

$$\begin{aligned} D_t(r_{t+1}, q_t) &= \sum_{i=1}^K \left(\beta_{t,i}(d_{f_1}(r_{t+1,i}, q_{t,i}) + d_{f_2}(r_{t+1,i}, q_{t,i})) + \gamma d_{f_3}(r_{t+1,i}, q_{t,i})\right) \\ &\geq \beta_{t,I_t}(d_{f_1}(r_{t+1,I_t}, q_{t,I_t}) + d_{f_2}(r_{t+1,I_t}, q_{t,I_t})) + \gamma d_{f_3}(r_{t+1,I_t}, q_{t,I_t}). \end{aligned}$$

Furthermore, as $\hat{\ell}_{t,i} - m_{t,i} = 0$ for all $i \neq I_t$, we have

$$\begin{aligned} \langle \hat{\ell}_t - m_t, q_t - r_{t+1} \rangle - D_t(r_{t+1}, q_t) &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - D_t(r_{t+1}, q_t) \\ &\leq \min(A, B, C), \end{aligned}$$

where

$$\begin{aligned} A &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - \beta_{t,I_t} d_{f_1}(r_{t+1,I_t}, q_{t,I_t}), \\ B &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - \beta_{t,I_t} d_{f_2}(r_{t+1,I_t}, q_{t,I_t}), \\ C &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - \gamma d_{f_3}(r_{t+1,I_t}, q_{t,I_t}). \end{aligned}$$

Here, we used $x - (a+b+c) \leq \min(x-a, x-b, x-c)$ for $a, b, c \geq 0$.

Note that $\ell_{t,I_t} - m_{t,I_t} \in [-1, 1]$ for $0 \leq \ell_{t,I_t}, m_{t,I_t} \leq 1$. By Corollary 6.E.3 and the fact

that $r_{t+1} \in \Delta_K$, we have $0 \le r_{t+1,I_t} \le 3dq_{t,I_t}$. Combining this with Lemma 6.E.7, we have

$$\begin{aligned}
\min(A, B) &\le \frac{(3d)^{2-\alpha}(\ell_{t,I_t} - m_{t,I_t})^2}{\beta_{t,I_t} p_{t,I_t}^2} \min\left\{\frac{q_{t,I_t}^{2-\alpha}}{2(1-\alpha)}, 1 - q_{t,I_t}\right\} \\
&\le \frac{(6d)^{2-\alpha}(\ell_{t,I_t} - m_{t,I_t})^2}{\beta_{t,I_t} p_{t,I_t}^2} \min\left\{\frac{p_{t,I_t}^{2-\alpha}}{(1-\alpha)}, 2(1 - p_{t,I_t})\right\} \\
&\le \frac{(6d)^{2-\alpha}(\ell_{t,I_t} - m_{t,I_t})^2}{2(1-\alpha)\beta_{t,I_t}} \min\left\{p_{t,I_t}^{-\alpha}, \frac{(1 - p_{t,I_t})}{p_{t,I_t}^2}\right\},
\end{aligned} \tag{6.78}$$

where the second inequality is due to $q_{t,I_t} \le 2p_{t,I_t}$ and $1 - q_{t,I_t} \le 2(1 - p_{t,I_t})$ by Lemma 6.E.8 and the last inequality is $1 - \alpha \le 1$.

The second-order derivative of $\gamma f_3(x)$ is $\frac{\gamma}{x^2}$. Therefore, by Lemma 6.E.6, we have

$$C \le \frac{(\ell_{t,I_t} - m_{t,I_t})^2 v^2}{2\gamma p_{t,I_t}^2} \le \frac{(\ell_{t,I_t} - m_{t,I_t})^2 18d^2}{\gamma}, \tag{6.79}$$

where $v \le 3dq_{t,I_t} \le 6dp_{t,I_t}$ is a point between r_{t+1,I_t} and q_{t,I_t}.

Combining (6.78) and (6.79) implies the desired bound. □

6.E.2 Technical Lemmas

Lemma 6.E.5. For any $\alpha \in [0, 1]$ and $x \in [0, 1]$,

$$x^\alpha \ge (x - 1)\ln(1 - x). \tag{6.80}$$

Proof. For $\alpha \in [0, 1]$, we have $x^\alpha \ge x$. Let

$$h(x) = x + (1 - x)\ln(1 - x). \tag{6.81}$$

Its derivative is $h'(x) = 1 - \ln(1 - x) - 1 = -\ln(1 - x) \ge 0$ for all $x \in [0, 1]$. Therefore, $h(x) \ge h(0) = 0$ for all $x \in [0, 1]$. We conclude that

$$x^\alpha - (x - 1)\ln(1 - x) \ge x - (x - 1)\ln(1 - x) = h(x) \ge 0. \tag{6.82}$$

□

Recall that $d_f(y,x) = f(y) - f(x) - f'(x)(y-x)$ denote the Bregman divergence associated with a one-dimensional strictly convex function $f : \mathbb{R} \to \mathbb{R}$. The following lemma is essentially a one-dimensional local-norm analysis of FTRL, whose proof can be found in standard literature. We provide a proof here for completeness.

Lemma 6.E.6. For any $a \in \mathbb{R}, x, y \in (0,1)$ and strictly convex function $f : (0,1) \to \mathbb{R}$, we have

$$a(x-y) - d_f(y,x) \le \frac{1}{2}\frac{a^2}{f''(z)}$$

for some z between x and y.

Proof. The inequality trivally holds when $x = y$, hence we assume $x \neq y$. By Taylor's theorem, we have $d_f(y,x) = \frac{f''(z)}{2}(x-y)^2$ for some z between x and y. Note that the strict convexity of f implies $f''(z) > 0$. We have

$$\begin{aligned} a(x-y) - d_f(x,y) &= a(x-y) - \frac{f''(z)}{2}(x-y)^2 \\ &= \frac{1}{2}\left(- \left((x-y)\sqrt{f''(z)} - \frac{a}{\sqrt{f''(z)}} \right)^2 + \frac{a^2}{f''(z)} \right) \le \frac{a^2}{2f''(z)}. \end{aligned}$$

□

Next, the following two lemma establish the foundation for choosing $z_{t,i}$ in Algorithm 6.2.

Lemma 6.E.7. Let $\alpha \in (0,1), \beta > \frac{4}{1-\alpha}$ be fixed and

$$\begin{aligned} f_1(x) &= \left(\frac{-x^\alpha}{\alpha} \right), \\ f_2(x) &= ((1-x)\ln(1-x) + x). \end{aligned}$$

For any $d \ge 1, h \in [-1,1], q \in (0,1]$ and $p \ge \frac{q}{2}$, we have

$$\begin{aligned} &\min\left\{ \max_{0 \le u \le dq} \left(\frac{h}{p}(q-u) - \beta d_{f_1}(u,q) \right), \max_{u \in \mathbb{R}} \left(\frac{h}{p}(q-u) - \beta d_{f_2}(u,q) \right) \right\} \\ &\le \frac{d^{2-\alpha}h^2}{\beta p^2} \min\left\{ \frac{q^{2-\alpha}}{2(1-\alpha)}, 1-q \right\}. \end{aligned}$$

Proof. First, we bound $\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right)$. For any $u\ge 0$, by Lemma 6.E.6, we have for some v between q and u:

$$\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right) \le \frac{1}{2}\frac{h^2}{p^2}\frac{v^{2-\alpha}}{\beta(1-\alpha)} \tag{6.83}$$

$$\le \frac{h^2}{\beta p^2}\frac{d^{2-\alpha}q^{2-\alpha}}{2(1-\alpha)}, \tag{6.84}$$

where we used the fact that the second-order derivative of $\beta f_1(v)$ is $\beta(1-\alpha)v^{\alpha-2}$ and $v\le \max(q,u)\le dq$. It follows that

$$\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right) \le \frac{h^2}{\beta p^2}\frac{d^{2-\alpha}q^{2-\alpha}}{2(1-\alpha)}. \tag{6.85}$$

Next, using Lemma 5 in [Ito et al., 2022], we have

$$\begin{aligned}\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right) &= \beta\max_{u\in\mathbb{R}}\left(\frac{h}{\beta p}(q-u)-d_{f_2}(u,q)\right)\\ &= \beta(1-q)\left(\exp\left(\frac{h^2}{\beta^2p^2}\right)-\frac{h}{\beta p}-1\right).\end{aligned}$$

We consider two cases:

- If $\beta p\ge 1$: in this case, we have $\frac{h}{\beta p}\le 1$ for any $h\in[-1,1]$. From the inequality $\exp(a)-a-1\le a^2$ for $a\le 1$, we have $\exp\left(\frac{h^2}{\beta^2p^2}\right)-\frac{h}{\beta p}-1\le\frac{h^2}{\beta^2p^2}$. Therefore,

$$\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right)\le(1-q)\frac{h^2}{\beta p^2}.$$

  This implies that

$$\begin{aligned}&\min\left\{\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right),\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right)\right\}\\ &\le\frac{(d)^{2-\alpha}h^2}{\beta p^2}\min\left\{\frac{q^{2-\alpha}}{2(1-\alpha)},1-q\right\}.\end{aligned}$$

- If $\beta p<1$: in this case, we have $q\le 2p\le\frac{2}{\beta}\le\frac{1-\alpha}{2}$. This implies that $\frac{q}{1-\alpha}\le\frac{1}{2}$ and also

$q \le \frac{1}{2}$. Combining this with $q^{1-\alpha} \le 1$, we obtain

$$\frac{q^{2-\alpha}}{2(1-\alpha)} \le \frac{q}{2(1-\alpha)} \le \frac{1}{4} \le 1 - q.$$

It follows that by (6.85),

$$\begin{aligned}\max_{0 \le u \le dq} \left(\frac{h}{p}(q-u) - \beta d_{f_1}(u,q) \right) &\le \frac{h^2}{\beta p^2} \frac{(d)^{2-\alpha} q^{2-\alpha}}{2(1-\alpha)} \\ &= \frac{(d)^{2-\alpha} h^2}{\beta p^2} \min\left\{ \frac{q^{2-\alpha}}{2(1-\alpha)}, 1-q \right\}.\end{aligned}$$

In both cases, we have

$$\begin{aligned}&\min\left\{ \max_{0 \le u \le dq} \left(\frac{h}{p}(q-u) - \beta d_{f_1}(u,q) \right), \max_{u \in \mathbb{R}} \left(\frac{h}{p}(q-u) - \beta d_{f_2}(u,q) \right) \right\} \\ &\le \frac{(d)^{2-\alpha} h^2}{\beta p^2} \min\left\{ \frac{q^{2-\alpha}}{2(1-\alpha)}, 1-q \right\}.\end{aligned}$$

□

Lemma 6.E.8. For any $K \ge 3, T \ge 4K$ and $q \in [0,1]$, let

$$p = \left(1 - \frac{K}{T}\right) q + \frac{1}{T}.$$

Then, we have $1 - q \le 2(1-p)$.

Proof. The desired inequality is equivalent to $2p - q \le 1$. By the definition of p, we have

$$\begin{aligned}2p - q &= 2\left(1 - \frac{K}{T}\right) q + \frac{2}{T} - q \\ &= \left(1 - \frac{2K}{T}\right) q + \frac{2}{T} \\ &\le \left(1 - \frac{2K}{T}\right) + \frac{2}{T} \\ &\le 1.\end{aligned}$$

□

Lemma 6.E.9. Fix an index $i \in [K]$ and let $p \in \Delta_K$ be an arbitrary vector in Δ_K. Let $\alpha \in (0,1)$ be a constant and $I \sim p$ be a random variable distributed according to p. We

have

$$\mathbb{E}_{I\sim p}\left[\mathbb{1}\{I=i\}p_i^{\alpha}\min\left(p_i^{-\alpha},\frac{1-p_i}{p_i^2}\right)\right]\le 2\min(p_i,1-p_i).$$

Proof. By the definition of I, the left-hand side is equal to

$$\begin{aligned}\mathbb{E}_{I\sim p}\left[\mathbb{1}\{I=i\}p_i^{\alpha}\min\left(p_i^{-\alpha},\frac{1-p_i}{p_i^2}\right)\right] &= p_i^{1+\alpha}\min\left(p_i^{-\alpha},\frac{1-p_i}{p_i^2}\right)\\ &= \min\left(p_i,\frac{(1-p_i)}{p_i^{1-\alpha}}\right).\end{aligned}$$

We consider two cases: $p_i\le\frac{1}{2}$ and $p_i>\frac{1}{2}$.

- If $p_i\le\frac{1}{2}$: since $p_i^{1-\alpha}\le 1$, we have $\frac{1-p_i}{p_i^{1-\alpha}}\ge 1-p_i\ge p_i$. Hence, $\min\left(p_i,\frac{(1-p_i)}{p_i^{1-\alpha}}\right)=p_i\le 2p_i=2\min(p_i,1-p_i)$.
- If $p_i>\frac{1}{2}$: we then have $\frac{(1-p_i)}{p_i^{1-\alpha}}\le\frac{1-p_i}{p_i}\le 2(1-p_i)$. Therefore, $\min\left(p_i,\frac{(1-p_i)}{p_i^{1-\alpha}}\right)\le\min(p_i,2(1-p_i))\le 2\min(p_i,1-p_i)$.

□

6.F SPM for Adversarial Sleeping Bandits

Intuitively, the sparsity constraint $\|\ell_t\|_0\le S$ indicates that there are at most S arms containing non-trivial information in each round, however the learner does not know the arms with non-trivial information. In this sense, sparse bandits is conceptually more difficult than adversarial sleeping bandits [Kleinberg et al., 2010], where in each round t the learner is given, by an adversary, a set $\mathbb{A}_t\subseteq[K]$ of active arms to choose from. Note that the learner is not allowed to choose an arm in $[K]\setminus\mathbb{A}_t$. The performance of the learner is measured by its per-action regret

$$R_{T,a}=\sum_{t=1}^{T}\mathbb{1}\{a\in\mathbb{A}_t\}(\ell_{t,I_t}-\ell_{t,a}).$$

A natural question is whether Algorithm 6.1 can be extended to this adversarial sleeping bandits setting. The following theorem answers this question in the positive.

Theorem 6.F.1. For any $K \geq 4, T \geq 4k$, Algorithm 6.5 (in Appendix 6.F) guarantees that for all $a \in [K]$,

$$\mathbb{E}[R_{T,a}] \leq O\left(\sqrt{\frac{(K^{1-\alpha}-1)(\max_{t\in[T]}|\mathbb{A}_t|)^\alpha}{\alpha(1-\alpha)}T}\right),$$

Our Algorithm 6.5 is a combination of Algorithm 6.1 and the SB-EXP3 algorithm in [Nguyen and Mehta, 2024]. More specifically, Algorithm 6.5 uses the estimated cumulative *regret* (instead of losses) to compute q_t in the FTRL update. Then, the sampling probability vector p_t is obtained by a filtering step $p_{t,i} = \frac{q_{t,i}\mathbb{1}\{i\in\mathbb{A}_t\}}{\sum_{j=1}^K q_{t,j}\mathbb{1}\{j\in\mathbb{A}_t\}}$. While the bound in Theorem 6.F.1 is of the same order as in Nguyen and Mehta [2024, Theorem 2], it has the advantage of not requiring the knowledge of $\max_t |\mathbb{A}_t|$ in advance nor any complicated two-level doubling trick.

Algorithm: We use the same regularization function in Algorithm 6.1,

$$\begin{aligned}\Phi_t(p) &= \beta_t\psi_{TE}(p) - \gamma\psi_{LB}(p)\\ &= \frac{\beta_t}{\alpha}\left(1-\sum_{i=1}^K p_i^\alpha\right) - \gamma\sum_{i=1}^K \ln(p_i).\end{aligned}$$

Instead of running FTRL on the sequence of losses, we run FTRL on the sequence of *estimated regret* $R_{t,i} = \sum_{s=1}^t \mathbb{1}\{i\in\mathbb{A}_s\}(\ell_{s,I_s} - \hat{\ell}_{s,i})$, i.e.,

$$\begin{aligned}q_t = \arg\min_{x\in\Delta_K} F_t(x) &:= \arg\min_{x\in\Delta_K}\langle -R_{t-1}, x\rangle + \Phi_t(x)\\ &= \arg\min_{x\in\Delta_K}\langle -R_{t-1}, x\rangle + \beta_t\left(\frac{1}{\alpha}(1-\sum_{i=1}^K x_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(x_i).\end{aligned}$$

Given the set of active arms $\mathbb{A}_t$, the sampling probability p_t is $p_{t,i} = \frac{\mathbb{1}\{i\in\mathbb{A}_t\}}{\sum_{j=1}^K \mathbb{1}\{j\in\mathbb{A}_t\}q_{t,j}}$. An arm $I_t \sim p_t$ is drawn. The learning rates are set by SPM rules [Ito et al., 2024]: $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$, where

$$\begin{aligned}z_t &= \min\left(\frac{(4d)^{2-\alpha}}{(1-\alpha)}\sum_{i\in\mathbb{A}_t}\min(p_{t,i}, 1-p_{t,i})^{1-\alpha}, \beta_t\frac{18d^2}{\gamma}\sum_{i\in\mathbb{A}_t}\tilde{p}_{t,i}\right),\\ h_t &= (-\psi_{TE}(q_t)) = \frac{1}{\alpha}\left(\sum_{i=1}^K q_{t,i}^\alpha - 1\right).\end{aligned}$$

6.F.1 Regret Analysis

In this section, we prove Theorem 6.F.1.

Proof. Let $I_{a,t} = \mathbb{1}\{a \in \mathbb{A}_t\}$. By the definition of $\hat{\ell}_t$ and the fact that $p_{t,i} = 0$ for $i \notin \mathbb{A}_t$, for any $a \in \mathbb{A}_t$, we have

$$\mathbb{E}_{I_t}[\hat{\ell}_{t,a}] = \sum_{i=1}^{K} p_{t,i} \frac{\ell_{t,a} \mathbb{1}\{a = i\}}{p_{t,a}} = \sum_{i \in \mathbb{A}_t} p_{t,i} \frac{\ell_{t,a} \mathbb{1}\{a = i\}}{p_{t,a}} = \ell_{t,a}. \tag{6.86}$$

Therefore, the per-action regret with respect to $a \in [K]$ is

$$\begin{aligned}
R_{T,a} &= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\ell_{t,I_t} - \ell_{t,a})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\langle \hat{\ell}_t, q_t \rangle - \langle \ell_t, e_a \rangle)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\langle \hat{\ell}_t, q_t - e_a \rangle)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\langle \hat{\ell}_t - \ell_{t,I_t}\mathbf{1}, q_t - e_a \rangle)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle -r_t, q_t - e_a \rangle\right],
\end{aligned}$$

where

- the second equality uses

$$\begin{aligned}
\langle \hat{\ell}_t, q_t \rangle &= \sum_{i=1}^{K} \hat{\ell}_{t,i} q_{t,i} \\
&= \sum_{i \in \mathbb{A}_t} \hat{\ell}_{t,i} q_{t,i} + \sum_{i \notin \mathbb{A}_t} \hat{\ell}_{t,i} q_{t,i} \\
&= \hat{\ell}_{t,I_t} q_{t,I_t} + \ell_{t,I_t} \sum_{i \notin \mathbb{A}_t} q_{t,i} \\
&= \frac{\ell_{t,I_t}}{p_{t,I_t}} q_{t,I_t} + \ell_{t,I_t} \sum_{i \notin \mathbb{A}_t} q_{t,i} \\
&= \ell_{t,I_t} \sum_{j \in \mathbb{A}_t} q_{t,j} + \ell_{t,I_t} \sum_{i \notin \mathbb{A}_t} q_{t,i} \\
&= \ell_{t,I_t}.
\end{aligned}$$

- the fourth equality uses $\langle \ell_{t,I_t} \mathbf{1}, q_t - e_a \rangle = \ell_{t,I_t} (\sum_{i=1}^{K} q_{t,i} - e_{a,i}) = 0$.
- the last equality uses

$$r_{t,i} = \begin{cases} \ell_{t,I_t} - \hat{\ell}_{t,i} & i \in \mathbb{A}_t \\ 0 & i \notin \mathbb{A}_t. \end{cases}$$

Let $u_a = (1 - \frac{K}{T}) e_a + \frac{1}{T} \mathbf{1}$. We have

$$\sum_{t=1}^{T} \langle -r_t, q_t - e_a \rangle = \sum_{t=1}^{T} \langle -r_t, q_t - u_a \rangle + \sum_{t=1}^{T} \langle -r_t, u_a - e_a \rangle.$$

The expectation of the second term is bounded by

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T}\langle -r_t, u_a - e_a\rangle] &= \sum_{t=1}^{T}\langle \mathbb{E}[-r_t], u_a - e_a\rangle \\
&= \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}[-r_{t,j}](u_{a,i} - e_{a,i}) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}[\hat{\ell}_{t,j} - \ell_{t,I_t}](u_{a,i} - e_{a,i}) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{K}(\ell_{t,j} - \mathbb{E}_{i\sim p_t}[\ell_{t,i}])(u_{a,i} - e_{a,i}) \\
&\leq 2K.
\end{aligned}$$

Therefore, we only need to focus on the first term $\sum_{t=1}^{T}\langle -r_t, q_t - u_a\rangle$. Next, by the definition of $F_t(x) = \langle \sum_{s=1}^{t-1} -r_s, x\rangle + \phi_t(x)$ and $q_t = \arg\min_{x\in\Delta_K} F_t(x)$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\langle -r_t, q_t - u_a\rangle \\
&= \left(\sum_{t=1}^{T}\langle -r_t, q_t\rangle\right) - (F_{T+1}(u_a) - \phi_{T+1}(u_a)) \\
&= \left(\sum_{t=1}^{T}\langle -r_t, q_t\rangle\right) - F_1(q_1) + \left(\sum_{t=1}^{T}F_t(q_t) - F_{t+1}(q_{t+1})\right) \\
&\quad + \phi_{T+1}(u_a) + F_{T+1}(q_{T+1}) - F_{T+1}(u_a) \\
&\leq \left(\sum_{t=1}^{T}\langle -r_t, q_t\rangle\right) - F_1(q_1) + \left(\sum_{t=1}^{T}F_t(q_t) - F_{t+1}(q_{t+1})\right) + \phi_{T+1}(u_a) \\
&= \phi_{T+1}(u_a) - F_1(q_1) + \left(\sum_{t=1}^{T}F_t(q_t) - F_{t+1}(q_{t+1}) + \langle -r_t, q_t\rangle\right) \\
&\leq \gamma K\ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + \underbrace{\left(\sum_{t=1}^{T}F_t(q_t) - F_{t+1}(q_{t+1}) + \langle -r_t, q_t\rangle\right)}_{\heartsuit}.
\end{aligned}$$

We bound each term in $\heartsuit$ as follows:

$$\begin{aligned}
&F_t(q_t) - F_{t+1}(q_{t+1}) + \langle -r_t, q_t \rangle \\
&= \langle -R_{t-1}, q_t \rangle + \langle R_t, q_{t+1} \rangle + \phi_t(q_t) - \phi_{t+1}(q_{t+1}) + \langle -r_t, q_t \rangle \\
&= \langle -R_{t-1}, q_t - q_{t+1} \rangle + \phi_t(q_t) - \phi_t(q_{t+1}) + \phi_t(q_{t+1}) - \phi_{t+1}(q_{t+1}) + \langle -r_t, q_t - q_{t+1} \rangle \\
&= (\beta_{t+1} - \beta_t) h_{t+1} + (\langle -R_{t-1}, q_t - q_{t+1} \rangle + \phi_t(q_t) - \phi_t(q_{t+1})) \\
&\quad + \langle -r_t, q_t - q_{t+1} \rangle \\
&\leq (\beta_{t+1} - \beta_t) h_{t+1} + \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t),
\end{aligned}$$

where the inequality is from $\langle -R_{t-1} + \nabla \phi_t(q_t), q_{t+1} - q_t \rangle \geq 0$ by the optimality of q_t and hence,

$$\begin{aligned}
-D_t(q_{t+1}, q_t) &= \phi_t(q_t) - \phi_t(q_{t+1}) + \langle \nabla \phi_t(q_t), q_{t+1} - q_t \rangle \\
&\geq \phi_t(q_t) - \phi_t(q_{t+1}) + \langle -R_{t-1}, q_t - q_{t+1} \rangle.
\end{aligned}$$

We have $q_{t+1,i} \leq 4d q_{t,i}$ for all $i \in [K]$ from the combination of the results of Lemma 6.F.2, Lemma 6.B.6 and Lemma 6.F.3. It follows that

$$\begin{aligned}
&\sum_{t=1}^{T} \langle -r_t, q_t - u_a \rangle \\
&\leq \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) h_{t+1} + \sum_{t=1}^{T} \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \\
&\leq \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + 4d \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) h_t + \sum_{t=1}^{T} \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \\
&= \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + 4d \sum_{t=1}^{T} \frac{z_t}{\beta_t} + \sum_{t=1}^{T} \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t),
\end{aligned}$$

where the second inequality is from Lemma 6.B.3.

Using Lemma 6.F.5 and noting that β_t is fixed before round t, we have

$$\begin{aligned}
\mathbb{E}_{I_t}\left[\langle -r_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)\right] &\le \mathbb{E}_{I_t}\left[\min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\tilde{p}_{t,I_t}^2\hat{\ell}_{t,I_t}^2\right)\right] \\
&\le \min\left(\mathbb{E}_{I_t}\left[\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2\right], \mathbb{E}_{I_t}\left[\frac{18d^2}{\gamma}\tilde{p}_{t,I_t}^2\hat{\ell}_{t,I_t}^2\right]\right) \\
&\le \min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)}\sum_{i\in\mathbb{A}_t}\tilde{p}_{t,i}^{1-\alpha}, \frac{18d^2}{\gamma}\sum_{i\in\mathbb{A}_t}\tilde{p}_{t,i}\right) \\
&= z_t.
\end{aligned}$$

It follows that

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle -r_t, q_t - u_a\rangle\right] \le \gamma K\ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha}-1) + 6\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right].$$

Let

$$\begin{aligned}
z_{\max} &= \max_{t\in[T]} z_t, \\
h_{\max} &= \max_{t\in[T]} h_t \\
A &= \max_{t\in[T]} |\mathbb{A}_t|
\end{aligned}$$

The quantity $z_{\max}$ is bounded by

$$\begin{aligned}
z_{\max} &\le \frac{(4d)^{2-\alpha}}{1-\alpha}\max_{t\in[T]}\sum_{i=1}^{K}\tilde{p}_{t,i}^{1-\alpha} \\
&\le \frac{(4d)^{2-\alpha}}{1-\alpha}\max_{t\in[T]}\sum_{i\in\mathbb{A}_t}\tilde{p}_{t,i}^{1-\alpha} \\
&\le \frac{(4d)^{2-\alpha}A^{\alpha}}{1-\alpha}.
\end{aligned}$$

Hence, we can bound $\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right]$ using the same analysis for SPM learning rates in [Ito

et al., 2024] and obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t'}{\beta_t}\right] \\
&\leq O\left(\min\left\{\inf_{J\in\mathbb{N}}\mathbb{E}\left[\left\{\sqrt{8J\sum_{t=1}^{T}h_t z_t}+2\sqrt{2^{-J}Th_{\max}z_{\max}}\right\}\right],\mathbb{E}\left[\sqrt{Th_{\max}z_{\max}}\right]\right\}+\mathbb{E}\left[\frac{z_{\max}}{\beta_1}\right]\right) \\
&\leq O\left(\min\left\{\inf_{J\in\mathbb{N}}\mathbb{E}\left[\left\{\sqrt{8J\sum_{t=1}^{T}h_t z_t}+\sqrt{2^{-J}Th_{\max}z_{\max}}\right\}\right],\mathbb{E}\left[\sqrt{Th_{\max}z_{\max}}\right]\right\}\right),
\end{aligned} \tag{6.87}$$

where the second inequality is due to $\frac{z_{\max}}{\beta_1}\leq O\left(\frac{A^\alpha}{(1-\alpha)\frac{4K}{1-\alpha}}\right)=O(1)$. Here, we used

$$\begin{aligned}
h_{\max}&=\max_{t\in[T]}h_t \\
&=\frac{1}{\alpha}(\sum_{i=1}^{K}q_{t,i}^{\alpha}-1) \\
&\leq\frac{K^{1-\alpha}-1}{\alpha}.
\end{aligned}$$

In the adversarial regime, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t'}{\beta_t}\right]&\leq O(\left(\mathbb{E}[\sqrt{Th_{\max}z_{\max}}]\right)) \\
&\leq O\left(\sqrt{T\frac{(K^{1-\alpha}-1)A^\alpha}{\alpha(1-\alpha)}}\right).
\end{aligned}$$

Hence, the regret bound is of order

$$\mathbb{E}[R_{T,a}]\leq O\left(\sqrt{T\frac{(K^{1-\alpha}-1)A^\alpha}{\alpha(1-\alpha)}}\right).$$

□

6.F.2 Stability Proofs

Recall from Section 6.B.2 that the function $g : [0,1] \to \mathbb{R}_+$ defined by

$$g(x) = \beta x^{\alpha-1} + \frac{\gamma}{x}$$

is decreasing in $x \in [0,1]$ for $\beta, \gamma > 0$.

Lemma 6.F.2. For any $t \geq 1$, Algorithm 6.5 guarantees

$$\beta_{t+1} - \beta_t \leq (1 - \frac{1}{d})\gamma q_{t*}^{-\alpha}, \tag{6.88}$$

where $q_{t*} = \min(\max_{i \in \mathbb{A}_t} q_{t,i}, 1 - \max_{i \in \mathbb{A}_t} q_{t,i})$.

Proof. Equation (6.59) shows that $h_t \geq \frac{1-\alpha}{4\alpha} q_{t*}^{\alpha}$. This implies that $\frac{1}{h_t} \leq \frac{4\alpha}{1-\alpha} q_{t*}^{-\alpha}$. By the definitions of β_{t+1}, z_t and h_t, we have

$$\begin{aligned}
\beta_{t+1} - \beta_t &= \frac{z_t}{\beta_t h_t} \\
&\leq \frac{4\alpha z_t}{(1-\alpha)\beta_t} q_{t*}^{-\alpha} \\
&\leq \frac{4\alpha}{(1-\alpha)} \frac{18 d^2}{\gamma} q_{t*}^{-\alpha} \\
&\leq (1 - \frac{1}{d})\gamma q_{t*}^{-\alpha}
\end{aligned}$$

where the last inequality uses

$$\frac{72\alpha d^2}{(1-\alpha)\gamma} \leq (1 - \frac{1}{d})\gamma \tag{6.89}$$

for $d = 2$ and $\gamma \geq 48\sqrt{\frac{\alpha}{1-\alpha}}$. □

Lemma 6.F.3. For any $K \geq 3, \alpha \in (0,1), \beta > 0, \gamma \geq 0, R \in \mathbb{R}^K$ and $h \in [-1,1]$, let $S \subseteq [K]$ be a subset of $[K]$ where $1 \in S$. Let $e_S \in \{0,1\}^K$ be a vector such that $e_{S,i} = \mathbb{1}\{i \in S\}$. Define

$$\begin{aligned}
x &= \arg\min_{p \in \Delta_K} \langle -R, p \rangle + \frac{\beta}{\alpha}(1 - \sum_{i=1}^K p_i^{\alpha}) - \gamma \sum_{i=1}^K \ln(p_i) \\
y &= \arg\min_{p \in \Delta_K} \langle -R + \frac{h}{x_1'} e_1 - h e_S, p \rangle + \frac{\beta}{\alpha}(1 - \sum_{i=1}^K p_i^{\alpha}) - \gamma \sum_{i=1}^K \ln(p_i),
\end{aligned}$$

where $1 \geq x_1' \geq x_1$. Fix an $\omega \in (1, 2]$. If $\gamma \geq 6$ and $\beta \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$, then $y_i \leq 4x_i$ for all $i \in [K]$.

Proof. Using the Lagrange multiplier methods, we have the following three equalities that hold for some $Z \in \mathbb{R}$:

$$g(x_1) - g(y_1) = Z + h - \frac{h}{x_1'}, \tag{6.90}$$

$$g(x_i) - g(y_i) = Z + h \qquad \text{for } i \in S \setminus \{1\}, \tag{6.91}$$

$$g(x_i) - g(y_i) = Z \qquad \text{for } i \notin S. \tag{6.92}$$

When $h \leq 0$:

First, we prove that $Z + h \leq 0$. Assume the contrary that $Z + h \geq 0$. Since $h \in [-1, 0]$, this implies that $Z > 0$ and $Z + h + \frac{(-h)}{x_1'} > 0$. Hence, $g(x_i) > g(y_i)$ for all $i \in [K]$, which is a contradiction since both x and y are in Δ_K. Thus, we must have $Z + h \leq 0$.

For any $i \in S \setminus \{1\}$, we have $g(x_i) - g(y_i) = Z + h \leq 0$ and therefore $y_i \leq x_i$. Next, we consider two cases of Z: $Z \geq 0$ and $Z < 0$.

- If $Z \geq 0$: we have $0 \leq Z \leq -h \leq 1$. For all $i \notin S$, we have $g(y_i) = g(x_i) - Z \geq g(x_i) - 1 \geq g(2x_i)$ by Lemma 6.F.6, which implies $y_i \leq 2x_i$. Thus, we only need to show that $y_1 \leq 2x_1$. If $y_1 \leq x_1$ then this is trivially true. If $y_1 \geq x_1$,

$$\begin{aligned} \frac{2}{x_1} \geq \frac{-h}{x_1'} &= (-(Z + h)) + \beta(x^{\alpha-1} - y^{\alpha-1}) + \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\ &\geq \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\ &\geq 4\left(\frac{1}{x_1} - \frac{1}{y_1}\right), \end{aligned}$$

  which leads to $y_1 \leq 2x_1$.

- If $Z < 0$: in this case, for all $i \notin S$, we have $g(x_i) \leq g(y_i)$ which implies $x_i \geq y_i$. As $x_i \geq y_i$ for all $i \neq 1$, we must have $x_1 \leq y_1$. Therefore, we have $y_1 \leq 2x_1$ by the same argument in the previous case.

When $h \geq 0$:

First, we prove that $Z + h \geq 0$. Assume the contrary that $Z + h < 0$. Since $h \in [0, 1]$, this implies $Z < 0$ and $Z + h - \frac{h}{x_1'} < 0$. Hence, $g(x_i) < g(y_i)$ for all $i \in [K]$, which is a

contradiction since both x and y are in Δ_K. Thus, we must have $Z + h \geq 0$.

For any $i \in S \setminus \{1\}$, we have $g(x_i) - g(y_i) = Z + h \geq 0$ and therefore $x_i \leq y_i$. Next, we consider two cases of Z: $Z \geq 0$ and $Z < 0$.

- If $Z \geq 0$: in this case, due to the monotonicity of the function g, we have $x_i \leq y_i$ for $i \neq 1$ and therefore, $x_1 \geq y_1$. This implies $Z + h - \frac{h}{x_1'} \leq 0$. Let

$$\epsilon = \frac{1}{\beta(1 - \omega^{\alpha-1})} \leq \frac{\omega - 1}{4K} \leq \frac{1}{K}$$

  as in the proof of Lemma 6.B.9. We further consider two cases of x_1'.

  - If $x_1' \geq \epsilon$: we have $Z \leq Z + h \leq \frac{h}{x_1'} \leq \frac{1}{\epsilon}$. Therefore, for all $i \neq 1$,

$$\begin{aligned} g(y_i) &= g(x_i) - Z - h\mathbb{1}\{i \in S\} \\ &\geq g(x_i) - Z - h \\ &\geq g(x_i) - \frac{1}{\epsilon} \\ &= \beta x_i^{\alpha-1} - \beta(1 - \omega^{\alpha-1}) + \frac{\gamma}{x_i} \\ &\geq \beta x_i^{\alpha-1} - \beta x_i^{\alpha-1}(1 - \omega^{\alpha-1}) + \frac{\gamma}{\omega x_i} \\ &= \beta(\omega x_i)^{\alpha-1} + \frac{\gamma}{\omega x_i} = g(\omega x_i), \end{aligned}$$

    where the last inequality is due to $x_i^{\alpha-1} \geq 1$ and $\omega > 1$. This implies that for all $i \neq 1$, $y_i \leq \omega x_i \leq 3x_i$ since $\omega \leq 2$.

  - If $x_1' < \epsilon$: we have $x_1 \leq \epsilon\frac{1}{2K}$ and $\sum_{i=1}^K (y_i - x_i) = x_1 - y_1 \leq \epsilon\frac{\omega-1}{K}$. Let $i* = \arg\max_{i\in[K]} x_i$. We have $x_{i*} \geq \frac{1}{K} > \frac{1}{2K}$, hence $i* \neq 1$. Furthermore,

$$\begin{aligned} \frac{1}{K}\left(\frac{y_{i*}}{x_{i*}} - 1\right) &\leq x_{i*}\left(\frac{y_{i*}}{x_{i*}} - 1\right) \\ &= y_{i*} - x_{i*} \\ &\leq \sum_{i\neq 1}(y_i - x_i) \\ &\leq \frac{\omega - 1}{K}, \end{aligned}$$

    which implies that $y_{i*} \leq \omega x_{i*}$. Therefore, using the fact that $g(x) - g(\omega x)$ is also

decreasing in x, for all $i \neq 1$, we have

$$\begin{aligned} g(y_i) &= g(x_i) - (Z + h\mathbb{1}\{i \in S\}) \\ &\geq g(x_i) - Z - 1 \qquad \text{since } h \in [0,1] \\ &\geq g(x_i) - (g(x_{i*}) - g(y_{i*})) - 1 \\ &\geq g(x_i) - (g(x_{i*}) - g(\omega x_{i*})) - 1 \\ &\geq g(x_i) - (g(x_i) - g(\omega x_i)) - 1 \\ &= g(\omega x_i) - 1 \\ &\geq g(2\omega x_i), \end{aligned}$$

where the second inequality is from $Z \leq Z + h\mathbb{1}\{i^* \in S\} = g(x_{i*}) - g(y_{i*})$, the third inequality is from $g(y_{i*}) \geq g(\omega x_{i*})$, and the last inequality is $g(x) - 1 \geq g(2x)$ by Lemma 6.F.6. From $g(y_i) \geq g(2\omega x_i)$, we conclude that $y_i \leq 2\omega x_i \leq 4x_i$ for all $i \neq 1$.

- If $Z < 0$: since $x_1' \leq 1$ and $h \in [0,1]$, we have $Z + h - \frac{h}{x_1'} < 0$. It follows that $g(x_1) - g(y_1) < 0$, hence $x_1 \geq y_1$. Moreover, for $i \notin S$, we also have $x_i \geq y_i$ due to $0 > Z = g(x_i) - g(y_i)$. Thus, we only need to show $y_i \leq 3x_i$ for $i \in S \setminus \{1\}$. For such i, we have

$$g(y_i) = g(x_i) - (h + Z) \geq g(x_i) - h \geq g(x_i) - 1 \geq g(2x_i),$$

where the last inequality is from Lemma 6.F.6. This implies $y_i \leq 2x_i$.

□

Lemma 6.F.4. For any $t \in [T]$ and constant $c \in [0,1]$, Algorithm 6.5 guarantees

$$\sum_{i=1}^{K} (q_{t,i})^{2-c} r_{t,i}^2 \leq 2(\tilde{p}_{t,I_t})^{2-c} \hat{\ell}_{t,I_t}^2.$$

Proof. Since $r_{t,i} = 0$ for $i \notin \mathbb{A}_t$, $r_{t,i} = \ell_{t,I_t}$ for $i \in \mathbb{A}_t \setminus \{I_t\}$ and $\hat{\ell}_{t,I_t} = \frac{\ell_{t,I_t}}{p_{t,I_t}}$, we have

$$\begin{aligned}
\sum_{i=1}^{K} q_{t,i}^{2-c} r_{t,i}^2 &= \sum_{i \in \mathbb{A}_t} q_{t,i}^{2-c} r_{t,i}^2 \\
&= \ell_{t,I_t}^2 \sum_{i \in \mathbb{A}_t, i \neq I_t} q_{t,i}^{2-c} + q_{t,I_t}^{2-c} \ell_{t,I_t}^2 \left(1 - \frac{1}{p_{t,I_t}}\right)^2 \\
&= \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2} \left(p_{t,I_t}^2 \sum_{i \in \mathbb{A}_t, i \neq I_t} q_{t,i}^{2-c} + q_{t,I_t}^{2-c} \left(p_{t,I_t} - 1\right)^2 \right) \\
&\leq \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2} \left(p_{t,I_t}^2 \sum_{i \in \mathbb{A}_t, i \neq I_t} p_{t,i}^{2-c} + p_{t,I_t}^{2-c} \left(p_{t,I_t} - 1\right)^2 \right) \\
&\leq \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2} \left(p_{t,I_t}^2 \left(\sum_{i \in \mathbb{A}_t, i \neq I_t} p_{t,i} \right)^{2-c} + p_{t,I_t}^{2-c} \left(p_{t,I_t} - 1\right)^2 \right) \\
&= \hat{\ell}_{t,I_t}^2 (p_{t,I_t}(1 - p_{t,I_t}))^{2-c} \left(p_{t,I_t}^c + (1 - p_{t,I_t})^c \right) \\
&\leq 2 \hat{\ell}_{t,I_t}^2 (p_{t,I_t}(1 - p_{t,I_t}))^{2-c} \\
&\leq 2 \tilde{p}_{t,I_t}^{2-c} \hat{\ell}_{t,I_t}^2,
\end{aligned}$$

where the first inequality is due to $q_{t,i} \leq p_{t,i}$ for $i \in \mathbb{A}_t$, the second inequality is from repeatedly applying $a^x + b^x \leq (a+b)^x$ for $x = 2 - c \geq 1$ by Lemma 6.B.11, the third inequality is $p_{t,I_t}^c \leq 1$ and $(1 - p_{t,I_t})^c \leq 1$, and the last inequality is $x(1-x) \leq \min(x, 1-x)$ for $x \in [0, 1]$. □

Lemma 6.F.5. For any $t \in [T]$, Algorithm 6.5 guarantees

$$\langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min \left(\frac{(4d)^{2-\alpha}}{\beta_t (1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18 d^2}{\gamma} \tilde{p}_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 \right) \tag{6.93}$$

Proof. Using standard local-norm analysis techniques for FTRL (for example, see Section 7.4 in [Orabona, 2023]), we have

$$\langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \frac{1}{2} \|r_t\|^2_{(\nabla^2 \phi_t(z_t))^{-1}}, \tag{6.94}$$

where z_t is a point between q_t and q_{t+1}. The Hessian matrix of ϕ_t is a diagonal matrix with

entries

$$\nabla^2\phi_t(z_t) = \operatorname{diag}\left(\left(\beta_t(1-\alpha)z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}\right)_{i=1,2,\ldots,K}\right). \tag{6.95}$$

Hence, its inverse is the following diagonal matrix

$$(\nabla^2\phi(z_t))^{-1} = \operatorname{diag}\left(\left(\frac{1}{\beta_t(1-\alpha)z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}}\right)_{i=1,2,\ldots,K}\right). \tag{6.96}$$

It follows that

$$\begin{aligned}\|r_t\|^2_{(\nabla^2\phi_t(z_t))^{-1}} &= \sum_{i=1}^K r_{t,i}^2 \frac{1}{\beta_t(1-\alpha)z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}} \\ &\le \min\left(\frac{1}{\beta_t(1-\alpha)}\sum_{i=1}^K z_{t,i}^{2-\alpha}r_{t,i}^2, \frac{1}{\gamma}\sum_{i=1}^K z_{t,i}^2 r_{t,i}^2\right)\end{aligned} \tag{6.97}$$

where the last equality is due to $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$. Combining (6.94) and (6.97), we obtain

$$\langle -r_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \le \min\left(\frac{1}{\beta_t(1-\alpha)}\sum_{i=1}^K z_{t,i}^{2-\alpha}r_{t,i}^2, \frac{1}{\gamma}\sum_{i=1}^K z_{t,i}^2 r_{t,i}^2\right). \tag{6.98}$$

Since z_t is between q_t and q_{t+1}, we have $z_{t,I_t} \le \max(q_{t,I_t}, q_{t+1,I_t})$. The loss estimate in Algorithm 6.5 uses p_{t,I_t} where $p_{t,I_t} \ge q_{t,I_t}$, therefore we can combine the results of Lemma 6.F.2, Lemma 6.B.6 and Lemma 6.F.3 and obtain $q_{t+1,i} \le 4dq_{t,i}$ for all $i \in [K]$. It follows that $z_{t,i} \le 4dq_{t,i}$, and as a result,

$$\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \le \min\left(\frac{(4d)^{2-\alpha}}{2\beta_t(1-\alpha)}\sum_{i=1}^K q_{t,i}^{2-\alpha}r_{t,i}^2, \frac{9d^2}{2\gamma}\sum_{i=1}^K q_{t,i}^2 r_{t,i}^2\right) \tag{6.99}$$

$$\le \min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\tilde{p}_{t,I_t}^2\hat{\ell}_{t,I_t}^2\right), \tag{6.100}$$

where the last inequality is from applying Lemma 6.F.4 twice: once with $c = \alpha$ and once with $c = 0$. □

6.F.3 Technical Lemmas

Lemma 6.F.6. For any $x \in [0,1]$, if $\gamma \geq 2$ then $g(x) - 1 \geq g(2x)$.

Proof. We have

$$\begin{aligned} g(x) - 1 &= \beta x^{\alpha-1} + \frac{\gamma}{x} - 1 \\ &\geq \beta 2^{\alpha-1} x^{\alpha-1} + \frac{\gamma}{2x} + \frac{\gamma}{2x} - 1 \\ &= g(2x) + \frac{\gamma - 2x}{2x} \\ &\geq g(2x). \end{aligned}$$

□

6.G Setting α appropriately close to 1

Recall that we assume $K \geq 3$ in our algorithms. Let $b = 1 - \alpha$. We will set $\alpha = 1 - \frac{0.5}{\ln(K)}$, which is equivalent to setting $b = \frac{0.5}{\ln(K)}$. Note that $\alpha \geq 1 - \frac{0.5}{\ln(3)} > 0.5$. Taking exponent on both sides of

$$\ln(1 + 2b\ln(K)) = \ln(2) \geq 0.5 = b\ln(K), \tag{6.101}$$

we obtain $1 + 2b\ln(K) \geq K^b$. This implies

$$\frac{K^{1-\alpha} - 1}{\alpha(1-\alpha)} \leq \frac{2(K^b - 1)}{b} \leq 4\ln(K). \tag{6.102}$$

Furthermore,

$$\frac{(K-1)^{1-\alpha}}{\alpha(1-\alpha)} \leq \frac{2(K-1)^{1-\alpha}}{1-\alpha} = \ln(K)(K-1)^{\frac{0.5}{\ln K}} \leq \ln(K) K^{\frac{0.5}{\ln K}} \leq 2\ln K, \tag{6.103}$$

where the last inequality is due to $K^{\frac{0.5}{\ln(K)}} = (e^{\ln K})^{\frac{0.5}{\ln(K)}} = e^{0.5} < 2$. In addition,

$$\frac{\alpha}{1-\alpha} \leq \frac{1}{1-\alpha} = \frac{1}{b} = 2\ln(K). \tag{6.104}$$

This implies that $\gamma = \max\left(6, 48\sqrt{\frac{\alpha}{1-\alpha}}\right) \lesssim \sqrt{\ln(K)}$.

Algorithm 6.4 SPM with Optimistic FTRL and Reservoir Sampling for losses in $[0,1]$

Input: $K \geq 1, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}, \gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $\$_i = \emptyset, \tilde{\mu}_{0,i} = 0, L_{0,i} = 0$ for $i \in [K]$
for each round $t = 1, \ldots, T$ **do**
  Sample $b_t \sim \text{Ber}(\min(\frac{K \ln(T)}{t}, 1))$
  if $b_t = 1$ **then**
    if $t \leq K \ln(T)$ **then**
      Draw $I_t = t \mod K + 1$ and observe ℓ_{t,I_t}
      Add ℓ_{t,I_t} to the reservoir $\$_{I_t}$ of arm I_t
    end
    else
      Draw $I_t \sim \text{Unif}([K])$ and observe ℓ_{t,I_t}
      Draw a random element by $\text{Unif}(\$_{I_t})$ and replace it by ℓ_{t,I_t}
    end
    Update the mean estimate $\tilde{\mu}_{t,I_t}$ in the reservoir $\$_{I_t}$ [Hazan and Kale, 2011]
    Compute $m_t = \tilde{\mu}_t$
  end
  else
    Compute $m_t = m_{t-1}$
    Compute $q_t = \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(x_i)$
    Compute $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$
    Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
    Compute loss estimate $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{I_t = i\}}{p_{t,i}}$
    Update $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
    Compute $z_t = \min \left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2 \right)$
    Compute $h_t = \left(\frac{1}{\alpha} (\sum_{i=1}^K p_{t,i}^\alpha - 1) \right)$
    Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$
  end
end

Algorithm 6.5 SPM Approach for Fully-Adversarial Sleeping Bandits

Input: $K \geq 1, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}, \gamma = \max\left(6, 48\sqrt{\frac{\alpha}{1-\alpha}}\right), d = 2.$

Initialize $R_{0,i} = 0$ for $i \in [K]$

for each round $t = 1, \ldots, T$ **do**

  The adversary reveals the set of active arms $\mathbb{A}_t$

  Compute $q_t = \arg\min_{x \in \Delta_K} \langle -R_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} x_i^{\alpha}) \right) - \gamma \sum_{i=1}^{K} \ln(x_i)$

  for arm $i \in [K]$ **do**

    Compute $p_{t,i} = \frac{q_{t,i} \mathbb{1}\{i \in \mathbb{A}_t\}}{\sum_{j=1}^{K} q_{t,j} \mathbb{1}\{j \in \mathbb{A}_t\}}$

  end

  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}

  for active arm $i \in [K]$ **do**

    if $i \in \mathbb{A}_t$ **then**

      Compute loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\{I_t = i\}}{p_{t,i}}$

    end

    else

      Set $\hat{\ell}_{t,i} = \ell_{t,I_t}$

    end

    Compute $r_{t,i} = \ell_{t,I_t} - \hat{\ell}_{t,i}$

    Update $R_{t,i} = R_{t-1,i} + r_t$

  end

  Compute $z_t = \min\left(\frac{(4d)^{2-\alpha}}{(1-\alpha)} \sum_{i \in \mathbb{A}_t} \min(p_{t,i}, 1 - p_{t,i})^{1-\alpha}, \beta_t \frac{18 d^2}{\gamma} \sum_{i \in \mathbb{A}_t} \tilde{p}_{t,i} \right)$

  Compute $h_t = \frac{1}{\alpha}(\sum_{i=1}^{K} q_{t,i}^{\alpha} - 1)$

  Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$

end

Bibliography

Abel, D., Jinnai, Y., Guo, S. Y., Konidaris, G., and Littman, M. (2018). Policy and value transfer in lifelong reinforcement learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 20–29. PMLR.

Abernethy, J. D., Awasthi, P., Kleindessner, M., Morgenstern, J. H., Russell, C., and Zhang, J. (2022). Active sampling for min-max fairness. In *International Conference on Machine Learning*.

Abernethy, J. D., Lee, C., and Tewari, A. (2015). Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, volume 28.

Adamskiy, D., Koolen, W. M., Chernov, A., and Vovk, V. (2016). A closer look at adaptive regret. *Journal of Machine Learning Research*, 17(23):1–21.

Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P. and Ortner, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Awasthi, P., Haghtalab, N., and Zhao, E. (2023). Open problem: The sample complexity of multi-distribution learning for VC classes. In *Proceedings of Thirty Sixth Conference*

on Learning Theory, volume 195 of *Proceedings of Machine Learning Research*, pages 5943–5949. PMLR.

Azar, M. G., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

Blum, A. (1997). Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain. *Machine Learning*, 26(1):5–23.

Blum, A. and Mansour, Y. (2007). From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324.

Bouneffouf, D., Rish, I., and Aggarwal, C. (2020). Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8.

Brunskill, E. and Li, L. (2013). Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 122–131, Arlington, Virginia, USA. AUAI Press.

Brunskill, E. and Li, L. (2015). The online discovery problem and its application to lifelong reinforcement learning. *CoRR*, abs/1506.03379.

Bubeck, S., Cohen, M., and Li, Y. (2018). Sparsity, variance and curvature in multi-armed bandits. In Janoos, F., Mohri, M., and Sridharan, K., editors, *Proceedings of Algorithmic*

Learning Theory, volume 83 of *Proceedings of Machine Learning Research*, pages 111–127. PMLR.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In Mannor, S., Srebro, N., and Williamson, R. C., editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 42.1–42.23, Edinburgh, Scotland. PMLR.

Carl, B. and Stephani, I. (1990). *Entropy, Compactness and the Approximation of Operators*. Cambridge Tracts in Mathematics. Cambridge University Press.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Chang, S.-H., Cosman, P. C., and Milstein, L. B. (2011). Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944.

Chernov, A. and Vovk, V. (2009). Prediction with expert evaluators' advice. In *Algorithmic Learning Theory*, pages 8–22, Berlin, Heidelberg.

Chernov, A. V. and Vovk, V. (2010). Prediction with advice of unknown number of experts. *CoRR*, abs/1006.0475.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. (2015). Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1405–1411, Lille, France.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2818–2826, Cambridge, MA, USA. MIT Press.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5717–5727, Red Hook, NY, USA. Curran Associates Inc.

Dann, C., Wei, C.-Y., and Zimmert, J. (2023). A blackbox approach to best of both worlds in bandits and beyond. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5503–5570. PMLR.

Domingues, O. D., Flet-Berliac, Y., Leurent, E., Ménard, P., Shang, X., and Valko, M. (2021a). rlberry - A Reinforcement Learning Library for Research and Education.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021b). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In Feldman, V., Ligett, K., and Sabato, S., editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:125–165.

Freund, Y. and Schapire, R. (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Freund, Y., Schapire, R. E., Singer, Y., and Warmuth, M. K. (1997). Using and combining predictors that specialize. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '97, page 334–343, New York, NY, USA.

Gaillard, P., Saha, A., and Dan, S. (2023). One arrow, two kills: A unified framework for achieving optimal regret guarantees in sleeping bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 7755–7773.

Gaillard, P., Stoltz, G., and Erven, T. (2014). A second-order bound with excess losses. *Journal of Machine Learning Research*, 35.

Guo, Z. and Brunskill, E. (2015). Concurrent pac rl. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2624–2630. AAAI Press.

Haghtalab, N., Jordan, M., and Zhao, E. (2022). On-demand sampling: Learning optimally from multiple distributions. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 406–419. Curran Associates, Inc.

Hallak, A., Castro, D. D., and Mannor, S. (2015). Contextual markov decision processes.

Hazan, E. and Kale, S. (2011). Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(35):1287–1311.

Hazan, E. and Seshadhri, C. (2009). Efficient learning algorithms for changing environments. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 393–400, New York, NY, USA.

Herbster, M. and Warmuth, M. K. (1998). Tracking the best expert. *Machine Learning*, 32(2):151–178.

Ito, S. (2021). Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*.

Ito, S., Tsuchiya, T., and Honda, J. (2022). Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1421–1422. PMLR.

Ito, S., Tsuchiya, T., and Honda, J. (2024). Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In Agrawal, S. and Roth, A., editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2522–2563. PMLR.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Jang, T., Gao, H., Shi, P., and Wang, X. (2024). Achieving fairness through separability: a unified framework for fair representation learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Gar-

nett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kanade, V. and Steinke, T. (2014). Learning hurdles for sleeping experts. *ACM Trans. Comput. Theory*, 6(3).

Kleinberg, R., Niculescu-Mizil, A., and Sharma, Y. (2010). Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2–3):245–272.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. (2021). RL for latent MDPs: Regret guarantees and a lower bound. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Kwon, J. and Perchet, V. (2016). Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lecarpentier, E., Abel, D., Asadi, K., Jinnai, Y., Rachelson, E., and Littman, M. L. (2021). Lipschitz lifelong reinforcement learning. In *AAAI*.

Levin, D., Peres, Y., and Wilmer, E. (2008). *Markov Chains and Mixing Times*. American Mathematical Soc.

Liao, R. (2020). Notes on Rademacher complexity.

Luo, H. (2017). Lecture 13, Introduction to Online Learning. https://haipeng-luo.net/courses/CSCI699/lecture13.pdf.

Luo, H. and Schapire, R. E. (2015). Achieving all with no parameters: AdaNormalHedge. In *Annual Conference Computational Learning Theory*.

Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. (2018). Efficient contextual bandits in non-stationary worlds. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1739–1776. PMLR.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Basar, T. (2021). Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7447–7458. PMLR.

Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Neu, G. (2015). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28.

Neu, G. and Valko, M. (2014). Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, volume 27.

Nguyen, Q., Ito, S., Komiyama, J., and Mehta, N. A. (2025a). Data-dependent bounds with t-optimal best-of-both-worlds guarantees in multi-armed bandits using stability-penalty matching. In *Proceedings of the 38th Conference on Learning Theory*, COLT'25.

Nguyen, Q. and Mehta, N. A. (2023). Adversarial online multi-task reinforcement learning. In *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*. PMLR.

Nguyen, Q. and Mehta, N. A. (2024). Near-Optimal Per-Action Regret Bounds for Sleeping Bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Nguyen, Q., Mehta, N. A., and Guzmann, C. (2025b). Beyond minimax rates in group distributionally robust optimization via a novel notion of sparsity. In *International Conference on Machine Learning*.

Nielsen, F. and Nock, R. (2011). A closed-form expression for the Sharma–Mittal entropy of exponential families. *Journal of Physics A: Mathematical and Theoretical*, 45(3):032003.

Orabona, F. (2023). A modern introduction to online learning. *CoRR*, abs/1912.13213.

Peng, B. (2023). The sample complexity of multi-distribution learning. *arXiv preprint arXiv:2312.04027*.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition.

Rakhlin, A. and Sridharan, K. (2013). Online learning with predictable sequences. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 993–1019, Princeton, NJ, USA. PMLR.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally robust neural networks. In *International Conference on Learning Representations*.

Saha, A., Gaillard, P., and Valko, M. (2020). Improved sleeping bandits with stochastic actions sets and adversarial rewards. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20.

Sason, I. (2015). On reverse Pinsker inequalities. *arXiv preprint arXiv:1503.07118*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA.

Simchowitz, M. and Jamieson, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Slivkins, A. (2013). Dynamic ad allocation: Bandits with budgets. *CoRR*, abs/1306.0155.

Slivkins, A. (2014). Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568.

Soma, T., Gatmiry, K., and Jegelka, S. (2022). Optimal algorithms for group distributionally robust optimization and beyond.

Steimle, L. N., Kaufman, D. L., and Denton, B. T. (2021). Multi-model markov decision processes. *IISE Transactions*, 53(10):1124–1139.

Sun, Y. and Huang, F. (2020). Can agents learn by analogy? an inferable model for pac reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 1332–1340, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. (2021). A provably efficient sample collection strategy for reinforcement learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., and Lazaric, A. (2020). Active model estimation in markov decision processes. In *Uncertainty in Artificial Intelligence*.

Tsuchiya, T., Ito, S., and Honda, J. (2023). Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tulsiani, M. (2014). Lecture 6, lecture notes in information and coding theory.

Wei, C.-Y. and Luo, H. (2018a). More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR.

Wei, C.-Y. and Luo, H. (2018b). More adaptive algorithms for adversarial bandits. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep.*

Williamson, R. and Menon, A. (2019). Fairness risk measures. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR.

Zhang, L., Zhao, P., Yang, T., and Zhou, Z.-H. (2023a). Stochastic approximation approaches to group distributionally robust optimization.

Zhang, Z., Zhan, W., Chen, Y., Du, S. S., and Lee, J. D. (2023b). Optimal multi-distribution learning. *arXiv preprint arXiv:2312.05134*.

Zimmert, J. and Seldin, Y. (2021a). Tsallis-inf: an optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22(1).

Zimmert, J. and Seldin, Y. (2021b). Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49.