UNIVERSITY OF CALIFORNIA SAN DIEGO

Interactive Machine Learning with Heterogeneous Data

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Zhi Wang

Committee in charge:

Professor Kamalika Chaudhuri, Chair
Professor Sanjoy Dasgupta
Professor Tara Javidi
Professor Yian Ma

2024

The Dissertation of Zhi Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my parents

EPIGRAPH

我想：希望本是无所谓有，无所谓无的。这正如地上的路；
其实地上本没有路，走的人多了，也便成了路。

— 鲁迅

I think: one cannot say that hope exists, or does not exist.

It is like the roads that mark the earth;
For though there were none at first,
where footsteps have fallen, roads emerge.

— Lu Xun

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to first express my sincerest gratitude to my advisor, Kamalika Chaudhuri. I did not have any prior experience in machine learning theory when I first joined her group, but she guided me into the field, helped me develop my research skills, and provided me freedom to explore my interests. Beyond technical knowledge, her mentorship and support also helped me cultivate a wide range of soft skills, which I am certain will benefit me for the rest of my life.

I would also like to thank the rest of my committee, Sanjoy Dasgupta, Tara Javidi and Yian Ma, for their invaluable feedback and insightful comments. I had the opportunity to work with Sanjoy in classrooms throughout my last year at UCSD—he set a remarkable example as both an instructor and a researcher, from which I have drawn inspiration.

I am deeply indebted to all my wonderful collaborators. In particular, I would like to thank Chicheng Zhang, with whom I have worked on many interesting problems. I learned from him the art of transforming intuitive research ideas into rigorous mathematical analysis, among others. I would also like to thank Geelon So, with whom I have engaged in countless conversations, research and otherwise, often fruitful, and sometimes repetitive. His friendship has been invaluable throughout this journey. I would also like to thank Ramya Korlakai Vinayak, for insightful discussions and her infectious enthusiasm.

I am grateful to my friends and colleagues, Chen Chen, Gaurav Mahajan, Aditi Mavalankar, Sophia Sun, Yutong Shao, Rex Lei, Casey Meehan, Zhifeng Kong, Yao-Yuan Yang, Jacob Imola, Robi Bhattacharjee, Nicholas Rittler, Chhavi Yadav, Tatsuki Koga, Amrita Roy Chowdhury, Ruihan Wu, Pengrun Huang, and Konstantin Garov, among others, for the great time we spent at UCSD. I want to thank Yi Xu for her lovely presence, her warmth and her optimism. Her support has been a beacon of light even during the hardest time.

I had one memorable summer internship at Amazon Web Services. Many thanks to my mentor Vidyashankar Sivakumar. I also thank Sicun Gao for being my advisor during

my first year at UCSD. Looking back, I am deeply grateful to my mentors T. K. Satish Kumar and Sven Koenig at the University of Southern California for introducing me to algorithmic AI research.

Last but not least, I would like to thank my parents for their endless love, support and inspiration throughout my life.

Chapter 2 is based on the material as it appears in "Multitask Bandit Learning through Heterogeneous Feedback Aggregation" by Zhi Wang, Chicheng Zhang, Manish Kumar Singh, Laurel D. Riek, and Kamalika Chaudhuri [161]. The material was published in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. The dissertation author was a co-primary investigator and co-first author of the paper.

Chapter 3 is based on the material as it appears in "Thompson Sampling for Robust Transfer in Multi-Task Bandits" by Zhi Wang, Chicheng Zhang, and Kamalika Chaudhuri [162]. The material was published in *Proceedings of the 39th International Conference on Machine Learning*. The dissertation author was the primary investigator and first author of the paper.

Chapter 4 is based on the material as it appears in "Provably Efficient Multi-Task Reinforcement Learning with Model Transfer" by Chicheng Zhang and Zhi Wang [173]. The material was published in *Advances in Neural Information Processing Systems 34*. The dissertation author was a co-author of the paper.

Chapter 5 is based on the material as it appears in "Metric Learning from Limited Pairwise Preference Comparisons" by Zhi Wang, Geelon So, and Ramya Korlakai Vinayak [163]. The material is currently in submission. The dissertation author was the primary investigator and first author of the paper.

VITA

2017	B.S. in Computer Science, University of Southern California
2022	M.S. in Computer Science, University of California San Diego
2022	C. Phil. in Computer Science, University of California San Diego
2024	Ph.D. in Computer Science, University of California San Diego

PUBLICATIONS

Zhi Wang, Geelon So, and Ramya Korlakai Vinayak. "Metric Learning from Limited Pairwise Preference Comparisons." Manuscript in submission. 2024.

Zhi Wang, Chicheng Zhang, and Kamalika Chaudhuri. "Thompson Sampling for Robust Transfer in Multi-Task Bandits." *Proceedings of the 39th International Conference on Machine Learning*. 2022.

Chicheng Zhang and Zhi Wang. "Provably Efficient Multi-Task Reinforcement Learning with Model Transfer." *Advances in Neural Information Processing Systems 34*. 2021.

Zhi Wang*, Chicheng Zhang*, Manish Kumar Singh, Laurel D. Riek, and Kamalika Chaudhuri. "Multitask Bandit Learning through Heterogeneous Feedback Aggregation." *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. 2021. *Equal contribution.

Zhi Wang, Liron Cohen, Sven Koenig and T. K. Satish Kumar. "The Factored Shortest Path Problem and Its Applications in Robotics." *Proceedings of the 28th International Conference on Automated Planning and Scheduling*. 2018.

T. K. Satish Kumar, Zhi Wang, Anoop Kumar, Craig Milo Rogers and Craig A. Knoblock. "Load Scheduling of Simple Temporal Networks Under Dynamic Resource Pricing." *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2018.

T. K. Satish Kumar, Zhi Wang, Anoop Kumar, Craig Milo Rogers and Craig A. Knoblock. "On the Linear Programming Duals of Temporal Reasoning Problems." The 15th International Symposium on Artificial Intelligence and Mathematics. 2018.

ABSTRACT OF THE DISSERTATION

Interactive Machine Learning with Heterogeneous Data

by

Zhi Wang

Doctor of Philosophy in Computer Science

University of California San Diego, 2024

Professor Kamalika Chaudhuri, Chair

In interactive machine learning, learners utilize data collected from interacting with the environment or with humans to better achieve their goals. Real-world applications often involve heterogeneous data sources, such as a large pool of human users with diverse interests or preferences, or non-stationary environments with distribution shifts. In this dissertation, we investigate interactive machine learning in the presence of heterogeneous data. In particular, we study when and how provably efficient learning can be achieved when the heterogeneous data exhibit structure.

In the first part, we study transfer learning in sequential decision-making. We consider a setting where learners are deployed to perform tasks in similar yet nonidentical

multi-armed bandit environments. We study when and how knowledge acquired from one environment can be robustly transferred to others so as to improve the collective performance of the learners. We present two provably efficient algorithms that properly manage data collected across heterogeneous environments: one uses upper confidence bounds and the other is based on Thompson sampling. We then generalize the setting and certain results to multi-task reinforcement learning in tabular Markov decision processes.

In the second part, we study metric learning from crowdsourced preference comparisons. In particular, we consider the ideal point model in preference learning, where a user prefers an item over another if it is closer to their latent ideal point. While users may have individual preferences and distinct ideal points, our goal is to learn a common Mahalanobis distance, which provides a more accurate measure of "closeness" that aligns with human values, perception and preferences. We study when and how such a metric can be learned if we can query each user a few times, asking questions in the form of "Do you prefer item A or B?"

Chapter 1

Introduction

In many real-world artificial intelligence (AI) applications, machine learners use data collected from interacting with the environment or humans to better achieve their goals. We refer to these problems as *interactive machine learning* problems. In this dissertation, we consider two types of interactive machine learning problems. One is sequential decision-making (e.g., [90]), in which a learner adaptively interacts with the environment over time. The other is AI alignment (e.g., [114]), wherein a learner refines its model parameters using interaction data to better align with human values and preferences.

In practical scenarios, interaction data often comes from *heterogeneous* sources. For example, learners may interact with a group of humans with diverse values or preferences; and, interaction data could also be collected from a non-stationary environment where the underlying model shifts over time. An important challenge involving heterogeneous data is that it is often unclear whether a learner can *robustly* aggregate and make use of such data to achieve their goals, if at all.

Fortunately, real-world heterogeneous data commonly exhibit structure. For example, the group of humans may share similar preferences or hold common perceptions and values towards certain subjects; in a non-stationary environment, the underlying model at different times may be different but still related. In this dissertation, we study when and how provably efficient interactive machine learning can be achieved with heterogeneous

data, and what structural assumptions are needed.

In the first part of this dissertation, we study robust transfer learning in multi-task bandits and reinforcement learning. We consider settings in which a group of learners are deployed to perform tasks in *similar but not necessarily identical* environments. For example, these environments may have similar reward distributions. We characterize when and how auxiliary data collected from other environments can be leveraged to improve the performance of each learner.

In Chapter 2, we formulate the ϵ-multi-player multi-armed bandit (ϵ-MPMAB) problem, in which a set of M players concurrently interact with multi-armed bandit instances with bounded pairwise dissimilarities between their reward distributions (Section 2.2). When the dissimilarities are known, we provide nearly-matching upper and lower bounds on the collective regret of the players, which show that information sharing is only amenable on subpar arms, a notion we introduce in Section 2.3.2 that captures the intrinsic complexity of the ϵ-MPMAB problem. We present an upper confidence bound (UCB)-based algorithm (Section 2.3.1) that achieves the regret upper bounds. In comparison with a baseline that does not utilize any knowledge transfer, the collective regret bound on subpar arms can be improved by nearly a factor of M (Theorem 2.5). In Section 2.4, we present results for the more challenging setting where the dissimilarities are unknown.

In Chapter 3, we first generalize the learning protocol of the ϵ-MPMAB problem so that the players do not necessarily interact with their respective environments concurrently; this setting can also capture, for example, sequential transfer and lifelong learning (Section 3.2). We then present a Thompson sampling-style randomized exploration algorithm (Section 3.3), which is proved (also) near-optimal and shows stronger empirical performance on synthetic data in comparison with the UCB-based algorithm in Chapter 2.

In Chapter 4, we study multi-task reinforcement learning in similar tabular, episodic Markov decision processes, a generalization of the ϵ-MPMAB problem in Chapter 2. In this

setting, we show that the notion of subpar state-action pairs, which generalizes the notion of subpar arms, now captures the intrinsic complexity of the problem (Section 4.4). We present a model-based algorithm, and provide nearly-matching upper and lower bounds.

In the second part of this dissertation, we study AI alignment from crowdsourced data, where it is imperative to overcome latent variation in feedback across individuals.

In Chapter 5, we consider a setting where we are given representations of a set of items in $\mathbb{R}^d$, and we aim to learn a metric that aligns with human values, perception, and preferences, in that it more accurately captures how humans perceive the semantic relations among the items. In particular, we seek to learn this metric using human preference comparisons. We consider the ideal point model in preference learning [38], where a user prefers an item over another if it is closer to their latent ideal item in $\mathbb{R}^d$. Given users with diverse preferences, we study when and how we can recover an unknown Mahalanobis distances when each user provides $o(d)$ preference comparisons. We show that additional structural assumptions may be needed, and we provide algorithms with recovery guarantees.

Chapter 2

Multi-Task Bandits through Heterogeneous Feedback Aggregation

2.1 Introduction

Online multi-armed bandit learning has many important real-world applications [e.g., 154, 131, 94]. In practice, a group of online bandit learning agents are often deployed for similar tasks, and they learn to perform these tasks in similar yet nonidentical environments. For example, a group of assistive healthcare robots may be deployed to provide personalized cognitive training to people with dementia (PwD), e.g., by playing cognitive training games with people [82]. Each robot seeks to learn the preferences of its paired PwD so as to recommend tailored health intervention based on how the PwD reacts to and is engaged with the activities (as captured by sensors on the robots) [82]. As PwD may have similar preferences and may therefore exhibit similar reactions, one natural question arises—can the robots as a multi-agent system learn to perform their respective tasks faster through collaboration? In Chapter 2 and Chapter 3, we develop multi-agent bandit learning algorithms where each agent can robustly aggregate data from other agents to better perform its respective task.

We generalize the multi-armed bandit problem [8] and formulate the ϵ-Multi-Player Multi-Armed Bandit (ϵ-MPMAB) problem, which models *heterogeneous multi-task learning* in a multi-agent bandit learning setting. In an ϵ-MPMAB problem instance, a set of M

players are deployed to perform similar tasks—simultaneously they interact with a set of actions/arms, and for each arm, different players receive feedback from similar but not necessarily identical reward distributions. In the above assistive robotics example, each player corresponds to a robot; each arm corresponds to one of the cognitive activities to choose from; for each player and each arm, there is a separate reward distribution which reflects a PwD's personal preferences. Informally, $\epsilon \geq 0$ is a *dissimilarity parameter* that upper bounds the pairwise distances between different reward distributions for different players on the same arm (see Definition 2.1 in the next section). The players can communicate and share information among each other, with a goal of maximizing their collective reward.

Multi-player bandit learning has been studied extensively in the literature [e.g., 87, 31, 57], warm-starting bandit learning using different feedback sources has been investigated [174], and sequential transfer between similar tasks in a bandit learning setting has also been studied [11, 138]. In contrast, we model multi-task learning in a multi-player bandit learning perspective with a focus on adaptive and robust aggregation of player-dependent heterogeneous feedback. In Section 2.5, we further discuss and compare our problem formulation with related papers.

It is worth noting that naively utilizing data collected by other players may substantially hurt a player's regret [174], if there are large disparities between the sources of feedback. This is also well-known as *negative transfer* in transfer learning [123, 24].

Therefore, the main challenge of the ϵ-MPMAB problem is for the players to properly manage *when and how* to utilize auxiliary data shared by others—while auxiliary data can be useful to maintain more accurate estimates of the rewards for each player and each arm, they can also easily be inefficacious or even misleading. While transfer learning in the offline setting has been well studied, in this chapter we seek to characterize the difficulty of the more challenging problem of learning through heterogeneous feedback aggregation in a multi-player online setting.

We will first study the ϵ-MPMAB problem when the dissimilarity parameter ϵ is known, and then move on to the harder setting in which ϵ is unknown. Here is a summary of our main contributions:

- We model online multi-task bandit learning from heterogeneous data sources as the ϵ-MPMAB problem, with a goal of studying how to adaptively and robustly aggregate data to improve the collective performance of the players.

- In the setting where ϵ is known, we propose an upper confidence bound (UCB)-based algorithm, RobustAgg(ϵ), that adaptively aggregates rewards collected by different players.

  We provide (suboptimality)-gap-dependent and gap-independent upper bounds on the collective regret of RobustAgg(ϵ). Our regret bounds depend on the set of arms that admit information sharing among the players. When this set is large, RobustAgg(ϵ) can potentially improve the gap-dependent regret bound by nearly a factor of M compared to the baseline of players acting individually using UCB-1 [8].

  We complement these upper bounds with nearly matching gap-dependent and gap-independent lower bounds.

- In the setting where ϵ is unknown, we first establish a lower bound, showing that if an algorithm guarantees sublinear minimax regret with respect to all MPMAB instances, then it must be unable to significantly utilize inter-player similarity in a large collection of instances. To complement the above result, we use the framework of Corral [2, 115, 6] and present an algorithm that trades off minimax regret guarantee for adaptivity to "easy" MPMAB problem instances.

2.2 The ϵ-MPMAB Problem

We formulate the ϵ-MPMAB problem, building on the standard model of stochastic multi-armed bandits [86, 8].

Throughout, we denote by $[n] = \{1, \ldots, n\}$. An *MPMAB problem instance* consists of a set of M players, labeled as elements in $[M]$, and a set of K arms, labeled as elements in $[K]$. In addition, each player $p \in [M]$ and each arm $i \in [K]$ is associated with an unknown reward distribution $\mathcal{D}_i^p$ with support $[0, 1]$ and mean μ_i^p. If all $\mathcal{D}_i^p$'s are Bernoulli distributions, we call this instance a *Bernoulli MPMAB problem instance*; under the Bernoulli reward assumption, $\mu = (\mu_i^p)_{i \in [K], p \in [M]}$ completely specifies the instance.

The reward distributions of the same arm are not necessarily identical for different players—we consider the following notion of dissimilarity between the reward distributions of the players. Related conditions have been considered in works on multi-task bandit learning [e.g., 11, 138].

Definition 2.1. *An MPMAB problem instance is said to be an ϵ-MPMAB problem instance, if for every pair of players $p, q \in [M]$, $\max_{i \in [K]} |\mu_i^p - \mu_i^q| \leq \epsilon$, where $\epsilon \in [0, 1]$. We call ϵ the dissimilarity parameter.*

Interaction protocol.

Let $T > \max(M, K)$ be the horizon of an MPMAB (ϵ-MPMAB) problem instance. In each round $t \in [T]$, every player $p \in [M]$ pulls an arm i_t^p, and observes an independently-drawn reward $r_t^p \sim \mathcal{D}_{i_t^p}^p$. Once all the M players finish pulling arms in round t, each decision, i_t^p, together with the corresponding reward received, r_t^p, is immediately shared with all players.

Arm pulls, gaps, and performance measure.

Let $\mu_*^p = \max_{i \in [K]} \mu_i^p$ be the optimal mean reward for every player $p \in [M]$. Denote by $n_i^p(t)$ the number of pulls of arm i by player p after t rounds, and $\Delta_i^p = \mu_*^p - \mu_i^p \geq 0$ the *suboptimality gap* (abbrev. gap) between the means of the reward distributions associated with some optimal arm i_*^p and arm i for player p. For any arm $i \in [K]$, define $\Delta_i^{\min} = \min_{p \in [M]} \Delta_i^p$. To measure the performance of MPMAB algorithms, we use the following notion of regret. The expected regret of player p is defined as

$\mathbb{E}[\mathcal{R}^p(T)] = \sum_{i\in[K]} \Delta_i^p \cdot \mathbb{E}[n_i^p(T)]$, and the players' *expected collective regret* is defined as $\mathbb{E}[\mathcal{R}(T)] = \sum_{p\in[M]} \mathbb{E}[\mathcal{R}^p(T)]$.

Bandit learning algorithms.

A multi-player bandit learning algorithm $\mathcal{A}$ with horizon T is defined as a sequence of conditional probability distributions $\{\pi_t\}_{t=1}^T$, where for every t in $[T]$, π_t is the policy used in round t; specifically, $\pi_t(\cdot \mid (i_s^p, r_s^p)_{s\in[t-1],p\in[M]})$ is a conditional probability distribution of actions taken by all M players in round t, given historical data. A bandit learning algorithm is said to have *sublinear regret* for the ϵ-MPMAB (resp. MPMAB) problem, if there exists some $C > 0$ and $\alpha > 0$ such that $\mathbb{E}[\mathcal{R}(T)] \leq CT^{1-\alpha}$ for all ϵ-MPMAB (resp. MPMAB) problem instances.

Miscellaneous notations.

Throughout, we use $\tilde{\mathcal{O}}$ notation to hide logarithmic factors. Given a universe set $\mathcal{H}$ and any $\mathcal{J} \subseteq \mathcal{H}$, we use $\mathcal{J}^C$ to denote the set $\mathcal{H} \setminus \mathcal{J}$.

Baseline: Individual UCB.

We now consider a baseline algorithm that runs the UCB-1 algorithm individually for each player without communication—hereafter, we refer to it as IND-UCB. By [8, Theorem 1], and summing over the individual regret guarantees of all players, the expected collective regret of IND-UCB satisfies

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathcal{O}\bigg(\sum_{i\in[K]} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}\bigg).$$

In addition, IND-UCB has a gap-independent regret bound of $\tilde{\mathcal{O}}\left(M\sqrt{KT}\right)$ [e.g., 90, Theorem 7.2].

2.2.1 Can auxiliary data always help?

Since the interaction protocol allows information sharing among players, in any round $t > 1$, each player has access to more data than they would have without communication. Can the players always expect benefits from such auxiliary data and collectively perform better than IND-UCB?

Below we provide an example that illustrates that the role of auxiliary data depends on the dissimilarities between the player-dependent reward distributions, as indicated by ϵ, as well as the intrinsic difficulty of each multi-armed bandit problem each player faces individually, as indicated by the gaps Δ_i^p's. Specifically, we show in the example that when ϵ is much larger than the gaps Δ_i^p's, any sublinear-regret bandit learning algorithm for the ϵ-MPMAB problem cannot significantly take advantage of auxiliary data.

Example 2.2. *For a fixed $\epsilon \in (0, \frac{1}{8})$ and $\delta \leq \epsilon/4$, consider the following Bernoulli MPMAB problem instance: for each $p \in [M]$, $\mu_1^p = \frac{1}{2} + \delta$, $\mu_2^p = \frac{1}{2}$. This is a 0-MPMAB instance, hence an ϵ-MPMAB problem instance. Also, note that ϵ is at least four times larger than the gaps $\Delta_2^p = \delta$.*

Claim 2.3. *For the above example, any sublinear regret algorithm for the ϵ-MPMAB problem must have $\Omega(\frac{M \ln T}{\delta})$ regret on this instance, matching the IND-UCB regret upper bound.*

The claim follows from Theorem 2.9 in Section 2.3.3; see Appendix A.2 for details. The intuition is that any sublinear regret ϵ-MPMAB algorithm must have $\Omega\left(\frac{\ln T}{\delta^2}\right)$ pulls of arm 2 from every player; otherwise, as δ is small compared to ϵ, we can create a new ϵ-MPMAB instance such that arm 2 is optimal for some player and is sufficiently indistinguishable from the original MPMAB problem, causing the algorithm to fail its sublinear regret guarantee.

Complementary to the above negative result, in the next section, we establish

algorithms and sufficient conditions for the players to take advantage of the auxiliary data to achieve better regret guarantees.

2.3 ϵ-MPMAB with Known ϵ

In this section, we study the ϵ-MPMAB problem with the dissimilarity parameter ϵ known to the players. We first present our main algorithm RobustAgg(ϵ) in Section 2.3.1; Section 2.3.2 shows its regret guarantees; Finally, Section 2.3.3 provides nearly matching regret lower bounds. Our proofs are deferred to Appendices A.3, A.4 and A.5.

2.3.1 Algorithm: RobustAgg(ϵ)

We present RobustAgg(ϵ), an algorithm that adaptively and robustly aggregates rewards collected by different players in ϵ-MPMAB problem instances, given dissimilarity ϵ as an input parameter.

Intuitively, in any round, a player may decide to take advantage of data from other players who have similar reward distributions. Deciding how to use auxiliary data is tricky—on the one hand, they can help reduce variance and get a better mean reward estimate, but on the other hand, if the dissimilarity between players' reward distributions is large, auxiliary data can substantially bias the estimate. Our algorithm is built upon this insight of balancing bias and variance. A similar tradeoff in offline transfer learning for classification is studied in the work of Ben-David et al. [19]; we discuss the connection and differences between our work and theirs in Section 2.5.3.

Algorithm 1 provides a pseudocode of RobustAgg(ϵ). Specifically, it builds on the classic UCB-1 algorithm [8]: for each player p and arm i, it maintains an upper confidence bound $\text{UCB}_i^p(t)$ for mean reward μ_i^p over time (lines 5 to 10), such that with high probability, $\mu_i^p \leq \text{UCB}_i^p(t)$, for all t.

To achieve the best regret guarantees, we would like our confidence bounds on μ_i^p to be as tight as possible. To this end, we consider a family of confidence intervals for μ_i^p,

Algorithm 1: RobustAgg(ϵ): Robust learning in ϵ-MPMAB

Input : Distribution dissimilarity parameter $\epsilon \in [0,1]$;

1 **Initialization:** Set $n_i^p = 0$ for all $p \in [M]$ and all $i \in [K]$.

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

3 **for** $p \in [M]$ **do**

4 **for** $i \in [K]$ **do**

5 Let $m_i^p = \sum_{q \in [M]: q \neq p} n_i^q$;

6 Let $\overline{n_i^p} = \max(1, n_i^p)$ and $\overline{m_i^p} = \max(1, m_i^p)$;

7 Let

$$\zeta_i^p(t) = \frac{1}{\overline{n_i^p}} \sum_{\substack{s<t \\ i_s^p = i}} r_s^p, \qquad \eta_i^p(t) = \frac{1}{\overline{m_i^p}} \sum_{\substack{q \in [M] \\ q \neq p}} \sum_{\substack{s<t \\ i_s^q = i}} r_s^q,$$

$$\text{and } \kappa_i^p(t, \lambda) = \lambda \zeta_i^p(t) + (1-\lambda)\eta_i^p(t);$$

8 Let $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon) = 8\sqrt{13 \ln T \left[\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}\right]} + (1-\lambda)\epsilon$;

9 Compute $\lambda^* = \text{argmin}_{\lambda \in [0,1]} F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$;

10 Compute an upper confidence bound of the reward of arm i for player p:

$$\text{UCB}_i^p(t) = \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon).$$

11 Let $i_t^p = \text{argmax}_{i \in [K]} \text{UCB}_i^p(t)$;

12 Player p pulls arm i_t^p and observes reward r_t^p;

13 **for** $p \in [M]$ **do**

14 Let $i = i_t^p$ and set $n_i^p = n_i^p + 1$.

parameterized by a weighting factor $\lambda \in [0,1]$: $[\kappa_i^p(t,\lambda) \pm F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)]$.

In the above confidence interval formula, $\kappa_i^p(t,\lambda)$ estimates μ_i^p by taking a convex combination of $\xi_i^p(t)$ and $\eta_i^p(t)$, the empirical mean reward of arm i based on the player's own samples and the auxiliary samples, respectively (line 7). The width $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$ is a high-probability upper bound on $\left|\kappa_i^p(t,\lambda) - \mu_i^p\right|$ (line 8). Varying λ reveals the aforementioned bias-variance tradeoff: the first term, $8\sqrt{13 \ln T [\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}]}$, is a high probability upper bound on the deviation of $\kappa_i^p(t,\lambda)$ from its expectation $\mathbb{E}[\kappa_i^p(t,\lambda)]$; the second term, $(1-\lambda)\epsilon$, is an upper bound on the difference between $\mathbb{E}[\kappa_i^p(t,\lambda)]$ and μ_i^p. We

choose $\lambda^* \in [0,1]$ to minimize the width of our confidence interval for μ_i^p (line 9), similar to the calculation in [19, Section 6].[1]

2.3.2 Regret analysis

Subpar arms.

We first define the notion of *subpar arms*; we will show that this notion captures the complexity of the ϵ-MPMAB problem. Let

$$\mathcal{I}_\alpha = \left\{ i : \exists p \in [M], \mu_*^p - \mu_i^p > \alpha \right\}$$

be the set of α-subpar arms. In particular, we consider $\mathcal{O}(\epsilon)$-subpar arms, and specifically, $\mathcal{I}_{5\epsilon}$. Intuitively, $\mathcal{I}_{5\epsilon}$ contains the set of "easier" arms for which data aggregation between players can be *effective*. For each arm $i \in \mathcal{I}_{5\epsilon}$, the following fact shows that the gap $\Delta_i^p = \mu_*^p - \mu_i^p$ is sufficiently larger than the dissimilarity parameter ϵ for all players $p \in [M]$. This allows RobustAgg(ϵ) to exploit the "easiness" of these arms through data aggregation across players, thereby reducing avoidable individual explorations.

Fact 2.4. *$|\mathcal{I}_{5\epsilon}| \leq K - 1$. In addition, for each arm $i \in \mathcal{I}_{5\epsilon}$, $\Delta_i^{\min} > 3\epsilon$; in other words, for all players p in $[M]$, $\Delta_i^p = \mu_*^p - \mu_i^p > 3\epsilon$; consequently, arm i is suboptimal for all players p in $[M]$.*

We now present regret guarantees of RobustAgg(ϵ).

Theorem 2.5. *Let RobustAgg(ϵ) run on an ϵ-MPMAB problem instance for T rounds. Then, its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathcal{O}\left(\sum_{i \in \mathcal{I}_{5\epsilon}} \left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min} \right) + \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M]: \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} \right).$$

[1] See Appendix A.8 for an analytical solution to the optimal weighting factor λ^*.

The first term in the above bound shows that the collective regret incurred by the players for the subpar arms $\mathcal{I}_{5\epsilon}$ and the second term for arms in $\mathcal{I}_{5\epsilon}^C = [K] \setminus \mathcal{I}_{5\epsilon}$. Observe that for each subpar arm, the regret of the players *as a group* can be upper-bounded by $\mathcal{O}\left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min}\right)$, whereas for each arm in $\mathcal{I}_{5\epsilon}^C$, the regret on *each* player is $\mathcal{O}(\frac{\ln T}{\Delta_i^p})$ unless $\Delta_i^p = 0$.

Fact 2.6. *For any* $i \in \mathcal{I}_{5\epsilon}$, $\frac{1}{\Delta_i^{\min}} \leq \frac{2}{M}\sum_{p\in[M]}\frac{1}{\Delta_i^p}$.

Fallback guarantee.

The regret guarantee of RobustAgg(ϵ) by Theorem 2.5 is always no worse than that of Ind-UCB by a constant factor, as from Fact 2.6, for all i in $\mathcal{I}_{5\epsilon}$, $\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min} = \mathcal{O}\left(\sum_{p\in[M]}\frac{\ln T}{\Delta_i^p}\right)$.

Two extreme cases of $|\mathcal{I}_{5\epsilon}|$.

If $\mathcal{I}_{5\epsilon} = \emptyset$, in which case we do not expect data aggregation across players to be beneficial, the above bound can be simplified to:

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathcal{O}\left(\sum_{i\in[K]}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right).$$

In contrast, when $\mathcal{I}_{5\epsilon}$ has a larger size, namely, more arms admit data aggregation across players, RobustAgg(ϵ) has an improved regret bound. The following corollary gives regret bounds in the most favorable case when $\mathcal{I}_{5\epsilon}$ has size $K-1$. It is not hard to see that, in this case, $\mathcal{I}_{5\epsilon}^C$ is equal to a singleton set $\{i_*\}$, where arm i_* is optimal for all players p.

Corollary 2.7. *Let* RobustAgg(ϵ) *run on an ϵ-MPMAB problem instance with $|\mathcal{I}_{5\epsilon}| = K-1$ for T rounds. Then, its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathcal{O}\left(\sum_{i\neq i_*}\frac{\ln T}{\Delta_i^{\min}} + M\sum_{i\neq i_*}\Delta_i^{\min}\right).$$

It can be observed that, compared to the IND-UCB baseline, under the assumption that $|\mathcal{I}_{5\epsilon}| = K - 1$, ROBUSTAGG($\epsilon$) improves the regret bound by nearly a factor of M: if we set aside the $\mathcal{O}\left(M \sum_{i \neq i_*} \Delta_i^{\min}\right)$ term, which is of lower order than the rest under the mild assumption that $M = \mathcal{O}\left(\min_{i \neq i^*} \frac{\ln T}{(\Delta_i^{\min})^2}\right)$, then the expected collective regret in Corollary 2.7 is a factor of $\mathcal{O}(\frac{1}{M})$ times that of IND-UCB, in light of Fact 2.6.

Gap-independent upper bound.

We now provide an upper bound on the expected collective regret that is independent of the gaps Δ_i^p's.

Theorem 2.8. *Let ROBUSTAGG(ϵ) run on an ϵ-MPMAB problem instance for T rounds. Then its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{5\epsilon}|\, MT} + M\sqrt{(\left|\mathcal{I}_{5\epsilon}^C\right| - 1)T} + M\,|\mathcal{I}_{5\epsilon}|\right).$$

Recall that IND-UCB has a gap-independent bound of $\tilde{\mathcal{O}}\left(M\sqrt{KT}\right)$. By algebraic calculations, we can see that when $T = \Omega(KM)$, the regret bound of ROBUSTAGG(ϵ) is a factor of $\mathcal{O}\left(\max\left(\sqrt{\frac{\left|\mathcal{I}_{5\epsilon}^C\right|-1}{K}}, \sqrt{\frac{1}{M}}\right)\right)$ times IND-UCB's regret bound. Therefore, when $M = \omega(1)$ and $\left|\mathcal{I}_{5\epsilon}^C\right| = o(K)$, i.e., when there is a large number of players, and an overwhelming portion of subpar arms, ROBUSTAGG has a gap-independent regret bound of strictly lower order than IND-UCB.

Observe that the above bound has a term $M\sqrt{(\left|\mathcal{I}_{5\epsilon}^C\right| - 1)T}$ with a peculiar dependence on $\left|\mathcal{I}_{5\epsilon}^C\right| - 1$; this is due to the fact that in the special case of $|\mathcal{I}_{5\epsilon}| = K - 1$, i.e., $\left|\mathcal{I}_{5\epsilon}^C\right| = 1$, the contribution to the regret from arms in $\mathcal{I}_{5\epsilon}^C$ is zero. Indeed, in this case, $\mathcal{I}_{5\epsilon}^C$ is a singleton set $\{i_*\}$, where arm i_* is optimal for all players.

2.3.3 Lower bounds

Gap-dependent lower bound.

To complement our gap-dependent upper bound in Theorem 2.5, we now present a gap-dependent lower bound. We show that, for any fixed ϵ, any sublinear regret algorithm for the ϵ-MPMAB problem must have regret guarantees not much better than that of RobustAgg(ϵ) for a large family of $\frac{\epsilon}{2}$-MPMAB problem instances.

Theorem 2.9. *Fix $\epsilon \geq 0$. Let $\mathcal{A}$ be an algorithm and $C > 0, \alpha > 0$ be constants, such that $\mathcal{A}$ has $CT^{1-\alpha}$ regret in all ϵ-MPMAB environments. Then, for any Bernoulli $\frac{\epsilon}{2}$-MPMAB instance $\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ such that $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$ for all i and p, we have:*

$$\mathbb{E}_\mu[\mathcal{R}(T)] \geq \Omega\left(\sum_{i\in\mathcal{I}^C_{\epsilon/4}}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln(\Delta_i^p T^\alpha/C)}{\Delta_i^p} + \sum_{i\in\mathcal{I}_{\epsilon/4}:\Delta_i^{\min}>0}\frac{\ln(\Delta_i^{\min}T^\alpha/C)}{\Delta_i^{\min}}\right).$$

Theorem 2.9 is nearly tight compared with the upper bound presented in Theorem 2.5 with two differences. First, the upper bound is in terms of $\mathcal{I}_{5\epsilon}$, while the lower bound is in terms of $\mathcal{I}_{\epsilon/4}$; we leave the possibility of exploiting data aggregation for arms in $\mathcal{I}_{5\epsilon} \setminus \mathcal{I}_{\epsilon/4}$ as an open question. Second, the upper bound has an extra $\mathcal{O}(\sum_{i\in\mathcal{I}_{5\epsilon}} M\Delta_i^{\min})$ term, caused by the players issuing arm pulls in parallel in each round; we conjecture that it may be possible to remove this term by developing more efficient multi-player exploration strategies.

Gap-independent lower bound.

The following theorem shows that, there exists a value of ϵ (that depends on T and $|\mathcal{I}_{5\epsilon}|$), such that any algorithm must have a minimax collective regret not much lower than the upper bound shown in Theorem 2.8 in the family of all ϵ-MPMAB problems.

Theorem 2.10. *For any $K \geq 2, M, T \in \mathbb{N}$, and l, l^C in $\mathbb{N}$ such that $l \leq K-1, l+l^C = K$, there exists some $\epsilon > 0$, such that for any algorithm $\mathcal{A}$, there exists an ϵ-MPMAB problem instance, in which $|\mathcal{I}_{5\epsilon}| \geq l$, and $\mathcal{A}$ has a collective regret at least $\Omega(M\sqrt{(l^C-1)T} + \sqrt{MlT})$.*

The above lower bound is nearly tight in light of the upper bound in Theorem 2.8: as long as $T = \Omega(KM)$, the upper and lower bounds match within a constant.

2.4 ϵ-MPMAB with Unknown ϵ

We now turn to the setting when ϵ is unknown to the learner. Unlike the RobustAgg(ϵ) algorithm developed in the last section, which only has nontrivial regret guarantees for all ϵ-MPMAB instances, in this section, we aim to design algorithms that have nontrivial regret guarantees for all MPMAB instances.

2.4.1 Gap-dependent lower bound

Recall that, for all MPMAB problems, Ind-UCB achieves a gap-dependent regret bound of $\mathcal{O}\left(\sum_{i\in[K]}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right)$ without knowing ϵ. Interestingly, we show in the following theorem that any sublinear regret algorithm for the MPMAB problem must have gap-dependent lower bound not much better than Ind-UCB for a large family of MPMAB problem instances, regardless of the value of ϵ and the size of $\mathcal{I}_{5\epsilon}$ of that instance.

Theorem 2.11. *Let $\mathcal{A}$ be an algorithm and $C > 0, \alpha > 0$ be constants such that $\mathcal{A}$ has $CT^{1-\alpha}$ regret in all MPMAB problem instances. Then, for any Bernoulli MPMAB instance $\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ such that $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$ for all $i \in [K], p \in [M]$,*

$$\mathbb{E}_\mu[\mathcal{R}(T)] \geq \Omega\left(\sum_{i\in[K]}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln(T^\alpha\Delta_i^p/C)}{\Delta_i^p}\right).$$

2.4.2 Gap-independent upper bound

While we have shown gap-dependent lower bounds that nearly matches the upper bounds for Ind-UCB for sublinear regret MPMAB algorithms in Theorem 2.11, this does not rule out the possibility of achieving regret that improves upon Ind-UCB in small-gap instances. To see this, note that if Δ_i^p is of order $\mathcal{O}(T^{-\alpha})$ for all i in $[K]$ and p in $[M]$, the

above lower bound becomes vacuous. Therefore, it is still possible to get gap-independent upper bounds that improve over the $\tilde{\mathcal{O}}(M\sqrt{KT})$ upper bound by IND-UCB.

We present ROBUSTAGG-AGNOSTIC in Appendix A.6, an algorithm that achieves such guarantee: specifically, it achieves a gap-independent regret upper bound adaptive to $|\mathcal{I}_{10\epsilon}|$. In a nutshell, the algorithm aggregates over a set of ROBUSTAGG(ϵ) base learners with different values of ϵ, using the strategy of Corral [2]. We have the following theorem:

Theorem 2.12. *Let ROBUSTAGG-AGNOSTIC run on an ϵ-MPMAB problem instance with any $\epsilon \in [0, 1]$. Its expected collective regret in a horizon of T rounds satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \tilde{\mathcal{O}}\left(\left(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^{C}\right|\right)\sqrt{T} + M|\mathcal{I}_{5\epsilon}|\right).$$

Under the mild assumption that $T = \Omega(\min(K^2, M^2))$, the above regret bound becomes $\tilde{\mathcal{O}}\left(\left(|\mathcal{I}_{10\epsilon}|+M\left|\mathcal{I}_{10\epsilon}^{C}\right|\right)\sqrt{T}\right)$. If furthermore $|\mathcal{I}_{10\epsilon}| = K-o(\sqrt{K})$ and $M = \omega(\sqrt{K})$, the regret bound of ROBUSTAGG-AGNOSTIC is of lower order than IND-UCB's $\tilde{\mathcal{O}}(M\sqrt{KT})$ regret guarantee. In the most favorable case when $|\mathcal{I}_{10\epsilon}| = K-1$, ROBUSTAGG-AGNOSTIC has expected collective regret $\tilde{\mathcal{O}}\left((M+K)\sqrt{T}\right)$.

Such adaptivity of ROBUSTAGG-AGNOSTIC to unknown similarity structure comes at a price of higher minimax regret guarantee: when $\mathcal{I}_{5\epsilon} = \emptyset$, ROBUSTAGG-AGNOSTIC has a regret of $\tilde{\mathcal{O}}\left(MK\sqrt{T}\right)$, a factor of $\sqrt{K}$ higher than $\tilde{\mathcal{O}}(M\sqrt{KT})$, the worst-case regret of IND-UCB. We conjecture that this may be unavoidable due to lack of knowledge of ϵ, similar to results in adaptive Lipschitz bandits [101, 81, 61].

2.5 Related Work

2.5.1 Multi-agent bandits

We first compare existing multi-agent bandit learning problems with the ϵ-MPMAB problem. We provide a more detailed review of the literature in Appendix A.1.

A large portion of prior studies [75, 141, 87, 32, 78, 127, 160, 46, 34, 157] focuses on the setting where a set of players collaboratively work on one bandit learning problem instance, i.e., the reward distributions of an arm are identical across all players. In contrast, we study multi-agent bandit learning where the reward distributions across players can be different.

Multi-agent bandit learning with heterogeneous feedback has also been covered by previous studies. In [129], a group of players seek to find the arm with the largest average reward over all players; however, in each round, the players have to reach a consensus and choose the same arm. Cesa-Bianchi et al. [31] study a network of linear contextual bandit players with heterogeneous rewards, where the players can take advantage of reward similarities hinted by a graph. They use a Laplacian-based regularization, whereas we study when and how to use information from other players based on a dissimilarity parameter. Gentile et al. [57], Li et al. [95] assume that the players' reward distributions have a cluster structure; in addition, players that belong to one cluster share a *common* reward distribution; our setting does not assume such cluster structure. Dubey and Pentland [47] assume access to some side information for every player, and learns a reward predictor that takes both player's side information models and action as input. In comparison, our work do not assume access to such side information.

Similarities in reward distributions are explored in [133, 174] to warm start bandit learning agents. Azar et al. [11], Soare et al. [138] investigate multitask learning in bandits through *sequential transfer* between tasks that have similar reward distributions. In contrast, we study the multi-player setting, where all players learn continually and concurrently.

There are other practical formulations of multi-player bandits with player-dependent reward distributions [20, 110], where the existence of collision is assumed; i.e., two players pulling the same arm in the same round receive zero reward. In comparison, collision is not modeled in this chapter.

2.5.2 Bandits in metric spaces

Our setting and results are also related to the work of Slivkins [137] on contextual bandits in metrics spaces. Specifically, if one considers player indices as contexts, then ϵ-dissimilarity may be modeled using a metric $\rho : \big([M] \times [K]\big)^2 \to [0, 1]$ such that for any $p, q \in [M]$ and $i, j \in [K]$,

$$\left|\mu_i^p - \mu_j^q\right| \leq \rho\left((p, i), (q, j)\right),$$

where

$$\rho\left((p, i), (q, j)\right) = \begin{cases} 0, & \text{if } i = j, p = q; \\ \epsilon, & \text{else if } i = j, p \neq q; \\ 1, & \text{otherwise.} \end{cases}$$

While we obtain an $\mathcal{O}(\log T)$ upper bound (Theorem 2.5) by making more direct use of the ϵ-dissimilarity structure, it is unclear whether such bounds can be achieved by applying the ideas and analyses in [137] for general metrics ρ's. It is also worth mentioning that Slivkins [137] considers a setting where in each round, one context/player is revealed, whereas our focus lies in a multi-task setting, where the players concurrently interact with their respective environments. We leave further exploring the connections of these settings as future work.

2.5.3 Learning using weighted data aggregation

Our design of confidence interval in Section 2.3.1 has resemblance to the weighted empirical risk minimization algorithm proposed for domain adaptation by Ben-David et al. [19], but our purposes are different from theirs. Specifically, our choice of λ minimizes the length of the confidence intervals, whereas [19] find λ that minimizes classification error in the target domain. Furthermore, our setting in Section 2.4 is more challenging: in offline domain adaptation, one may use a validation set drawn from the target domain to

fine-tune the optimal weight λ^*, to adapt to unknown dissimilarity between the source and the target; however, in our setting (and online bandit learning in general), such tuning does not result in sample efficiency improvement.

The idea of assigning weights to different sources of samples has also been studied by Zhang et al. [174] for warm starting contextual bandit learning from misaligned distributions and by Russac et al. [125] for online learning in non-stationary environments. Zhu et al. [180] use a weighted compound of player-based estimator and cluster-based estimator for collaborative Thompson sampling, where the weights are given by a hyper-parameter; in contrast, we adaptively compute our weighting factor based on the numbers of samples collected by the players as well as the dissimilarity parameter ϵ.

2.6 Empirical Validation

We now validate our theoretical results with some empirical simulations using synthetic data. Specifically, we seek to answer the following questions:

1. In practice, how does our proposed algorithm compare with algorithms that either do not take advantage of adaptive data aggregation or do not execute aggregation in a robust fashion?

2. How does the performance of our algorithm change with different numbers of subpar arms?

We note that these questions are considered in the setting where the dissimilarity parameter ϵ is known to the algorithms.

2.6.1 Experimental setup

We first describe the algorithms compared in the simulations. We then discuss the procedure we used for generating synthetic data.

RobustAgg-Adapted(ϵ).

Since standard concentration bounds are loose in practice, we performed simulations on a more practical and aggressive variant of RobustAgg(ϵ), which we call RobustAgg-Adapted(ϵ). Specifically, we changed the constant coefficient $8\sqrt{13}$ to $\sqrt{2}$ in the UCBs; this constant was taken from the original UCB-1 algorithm [8], which is an ingredient of the baseline Ind-UCB, and we simply kept the default value.

Baselines.

We evaluate the following two algorithms as baselines: (a) Ind-UCB, described in Section 2.2; and (b) Naive-Agg, in which the players *naively* aggregate data assuming that their reward distributions are identical—in other words, Naive-Agg is equivalent to RobustAgg-Adapted(0).

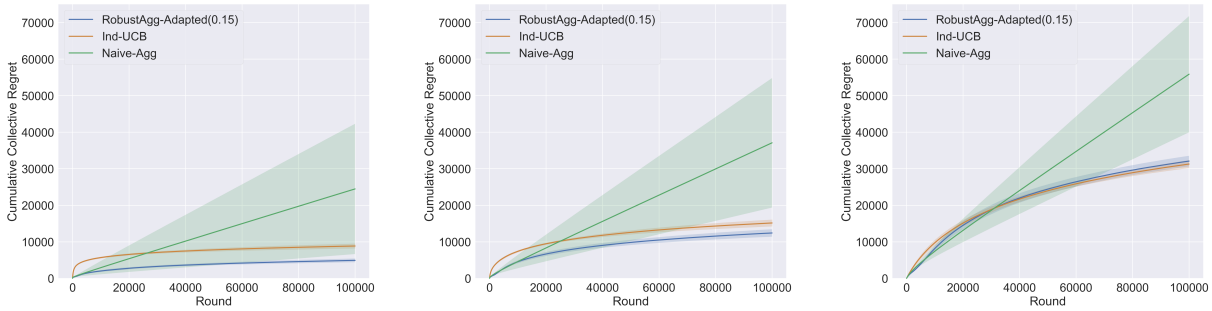
Instance generation.

We generated problem instances using the following *randomized* procedure. We first set $\epsilon = 0.15$. Then, given the number of players M, the number of arms K, and the number of subpar arms $|\mathcal{I}_{5\epsilon}| \in \{0, 1, \ldots, K-1\}$, we first sampled the means of the reward distributions for player 1:

Let $c = K - |\mathcal{I}_{5\epsilon}|$. For $i \in \{1, 2, \ldots, c\}$, we sampled $\mu_i^1 \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0.8, 0.8 + \epsilon)$, where $\mathcal{U}[a, b)$ is the uniform distribution with support $[a, b]$. Let $d = \max_{i \in [c]} \mu_i^1$. Then, for $i \in \{c+1, \ldots, K\}$, we sampled $\mu_i^1 \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0, d - 5\epsilon)$.

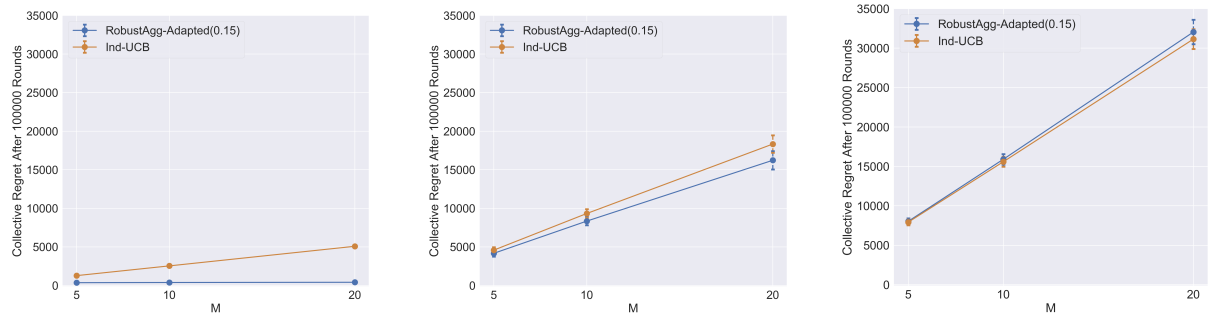
We then sampled the means of the reward distributions for players $p \in \{2, \ldots, M\}$: For each $i \in [K]$, we sampled $\mu_i^p \overset{\text{i.i.d.}}{\sim} \mathcal{U}\big[\max(0, \mu_i^1 - \frac{\epsilon}{2}), \min(\mu_i^1 + \frac{\epsilon}{2}, 1)\big)$.

Fact 2.13. *The above construction gives a Bernoulli* 0.15*-MPMAB problem instance that has exactly* $(K - c)$ *subpar arms, namely,* $\mathcal{I}_{5\epsilon} = \{i : c + 1 \le i \le K\}$.



(a) $|\mathcal{I}_{5\epsilon}| = 8$ **(b)** $|\mathcal{I}_{5\epsilon}| = 6$ **(c)** $|\mathcal{I}_{5\epsilon}| = 0$

Figure 2.1. Compares the average performance of RobustAgg-Adapted(0.15), Ind-UCB, and Naive-Agg on randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$ and $M = 20$. The x-axis shows a horizon of $T = 100,000$ rounds, and the y-axis shows the cumulative collective regret of the players.



(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 5$ **(c)** $|\mathcal{I}_{5\epsilon}| = 0$

Figure 2.2. Compares the average performance of RobustAgg-Adapted(0.15) and Ind-UCB on randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$. The x-axis shows different values of M, and the y-axis shows the cumulative collective regret of the players after $100,000$ rounds.

2.6.2 Simulations and results

We ran two sets of simulations, and the results are shown in Figure 2.1 and Figure 2.2. More detailed results are deferred to Appendix A.7.

Experiment 1.

We compare the cumulative collective regrets of the three algorithms in problem instances with *different numbers of subpar arms*. We set $M = 20$, $K = 10$ and $\epsilon = 0.15$. For each $v \in \{0, 1, 2, \ldots, 9\}$, we generated 30 Bernoulli 0.15-MPMAB problem instances,

each of which has exactly v subpar arms, i.e., we generated instances with $|\mathcal{I}_{5\epsilon}| = v$. Figures 2.1a, 2.1b and 2.1c show the average regrets in a horizon of $100,000$ rounds over these generated instances, in which $|\mathcal{I}_{5\epsilon}| = 8, 6$ and 0, respectively. In the interest of space, figures in which $|\mathcal{I}_{5\epsilon}|$ takes other values are deferred to Appendix A.7.2.

Notice that RobustAgg-Adapted(0.15) outperforms both baseline algorithms in Figures 2.1a and 2.1b when $|\mathcal{I}_{5\epsilon}| = 8$ and 6. Figure 2.1c demonstrates that when $|\mathcal{I}_{5\epsilon}| = 0$, i.e., when there is no arm that is amenable to data aggregation, the performance of RobustAgg-Adapted(0.15) is still on par with that of Ind-UCB. Also, as shown in Figure 2.1a, even when $|\mathcal{I}_{5\epsilon}^C| = 2$, i.e., when there are only two "competitive" (not subpar) arms, the collective regret of Naive-Agg can still easily be nearly linear in the number of rounds.

Experiment 2.

We study how the collective regrets of RobustAgg-Adapted(0.15) and Ind-UCB scale with the *number of players* in problem instances with different numbers of subpar arms. We set $K = 10$ and $\epsilon = 0.15$. For each combination of $M \in \{5, 10, 20\}$ and $v \in \{0, 1, 2, \ldots, 9\}$, we generated 30 Bernoulli 0.15-MPMAB problem instances with M players and exactly v subpar arms, that is, for each instance, $|\mathcal{I}_{5\epsilon}| = v$. Figures 2.2a, 2.2b and 2.2c compare the average regrets after $100,000$ rounds in instances with different numbers of players M, in which $|\mathcal{I}_{5\epsilon}|$ are set to be $9, 5$ and 0, respectively. Again, figures in which $|\mathcal{I}_{5\epsilon}|$ takes other values are deferred to Appendix A.7.2.

Note that when $|\mathcal{I}_{5\epsilon}|$ is large, the collective regret of RobustAgg-Adapted(0.15) is less sensitive to the number of players. In the extreme case when $|\mathcal{I}_{5\epsilon}| = 9$, all suboptimal arms are subpar arms, and Figure 2.2a shows that the collective regret of RobustAgg-Adapted(0.15) has negligible dependence on the number of players M.

2.6.3 Discussion

Back to the earlier questions, our simulations show that RobustAgg-Adapted(ϵ), in general, outperforms the baseline algorithms Ind-UCB and Naive-Agg. When the set of subpar arms $\mathcal{I}_{5\epsilon}$ is large, we showed that properly managing data aggregation can substantially improve the players' collective performance in an ϵ-MPMAB problem instance. When there is no subpar arm, we demonstrated the robustness of RobustAgg-Adapted(ϵ), that is, its performance is comparable with Ind-UCB, in which the players do not share information. These empirical results validate our theoretical analyses in Section 2.3.

2.7 Conclusion and Future Work

In this chapter, we studied multitask bandit learning from heterogeneous feedback. We formulated the ϵ-MPMAB problem and showed that whether inter-player information sharing can boost the players' performance depends on the dissimilarity parameter ϵ as well as the intrinsic difficulty of each individual bandit problem that the players face. In particular, in the setting where ϵ is known, we presented a UCB-based data aggregation algorithm which has near-optimal instance-dependent regret guarantees. We also provided upper and lower bounds in the setting where ϵ is unknown.

There are many avenues for future work. For example, we are interested in extending our results to contextual bandits and Markov decision processes. Another direction is to study multitask bandit learning under other interaction protocols (e.g., only a subset of players take actions in each round). In the future, we would also like to evaluate our algorithms in real-world applications such as healthcare robotics [122].

Acknowledgement.

Chapter 2 is based on the material as it appears in "Multitask Bandit Learning through Heterogeneous Feedback Aggregation" by Zhi Wang, Chicheng Zhang, Manish Kumar Singh, Laurel D. Riek, and Kamalika Chaudhuri [161]. The material was published

in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. The dissertation author was a co-primary investigator and co-first author of the paper.

Chapter 3

Thompson Sampling for Robust Transfer in Multi-Task Bandits

3.1 Introduction

In this chapter, we study an alternative approach to the ϵ-multi-player multi-armed bandit (ϵ-MPMAB) problem formulated in Chapter 2, which can be used to model multi-task bandits. We now also consider a generalized interaction protocol, where a set of players sequentially and potentially concurrently interact with a common set of arms that have player-dependent reward distributions. Each player and its associated reward distributions (data sources) are thereby regarded as a task. Again, we consider the reward distributions that the players face for each arm to be *similar* but not necessarily identical, and the level of (dis)similarity is specified by a parameter $\epsilon \in [0, 1]$.

As discussed in Chapter 2, the ϵ-MPMAB problem can be used to model important real-world applications. For example, in healthcare robotics, a set of robots, which correspond to players, can be paired with people with dementia to provide personalized cognitive training and wellness activities [83]. Each training/wellness activity corresponds to an arm in the ϵ-MPMAB problem, and people with similar preferences or symptoms may exhibit similar interests or needs—this is modeled via similarity in reward distributions of each arm. Another example can be seen in recommendation systems where learning agents are assigned to people within a social network, who may have similar interests due

to inter-network influence [120].

Despite the similarity in its reward distributions, the ϵ-MPMAB problem is still challenging for two reasons: on the one hand, misusing auxiliary data can lead to negative transfer and substantially impair a player's performance [123]; on the other hand, while auxiliary data are often immediately accessible in their entirety in offline transfer learning settings, in the ϵ-MPMAB problem, the available auxiliary data grow in time and depend on the interactions between the players and the environments.

In Chapter 2, we proposed an upper confidence bound (UCB)-based algorithm, RobustAgg(ϵ), for the ϵ-MPMAB problem. It achieves strong, near-optimal theoretical guarantees through robust data aggregation. Nevertheless, RobustAgg(ϵ)'s empirical performance can, unfortunately, be underwhelming.

Meanwhile, Thompson sampling (TS) algorithms [148], another family of bandit algorithms, have been shown superior empirically in comparison with UCB-based algorithms in standard single-task settings [e.g., 33]. In fact, we show in Section 3.7 that, for the ϵ-MPMAB problem, a baseline algorithm which employs TS for each task individually without transfer learning can outperform RobustAgg(ϵ) in many cases.

In spite of the encouraging signs from the empirical evaluations, the theoretical study of TS have lagged behind, especially in terms of *frequentist* analyses [4, 76] for data aggregation and transfer learning in the multi-task setting[1]. It is therefore imperative to design multi-task TS-type algorithms that have superior empirical performance *and* strong theoretical guarantees. Our contributions in this chapter are:

1. Inspired by prior works [31, 57, 63], we generalize the ϵ-MPMAB problem to model a wider class of multi-task bandit learning scenarios so that it covers sequential and concurrent multi-task learning as special cases.

2. We design a TS-type algorithm, RobustAgg-TS(ϵ), for the ϵ-MPMAB problem and

[1]See Section 3.6 for a discussion on related work.

provide a frequentist analysis with near-optimal performance guarantees.

3. We empirically evaluate RobustAgg-TS(ϵ) on synthetic data and show that it outperforms the UCB-based RobustAgg(ϵ) and a baseline algorithm that runs TS for each individual task without data sharing.

4. Technical highlight: frequentist analyses of Thompson sampling can be much harder to conduct than those of UCB-based algorithms (see Remark 3.4); a concentration inequality loose in logarithmic factors can result in a polynomial increase in regret guarantee (see Remark 3.9). To cope with this challenge, we prove a novel concentration inequality for multi-task data aggregation at random stopping times (Lemma 3.8), which leads to tight performance guarantees for RobustAgg-TS(ϵ). Our technique may be of independent interest for analyzing other multi-task sequential learning problems.

3.2 Preliminaries

In this section, we first revisit and generalize the problem formulation. We then briefly review the results in Chapter 2, and then introduce a new baseline algorithm based on TS.

Notations.

Throughout, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian distribution with mean μ and variance σ^2. Let $a \vee b = \max(a, b)$. For a set $A \subseteq U$, denote by $A^C = U \setminus A$ the complement of A in the universe U. We use $\tilde{\mathcal{O}}$ to hide logarithmic factors.

3.2.1 ϵ-MPMAB with generalized interaction protocol

We consider and generalize the ϵ-MPMAB problem introduced in Chapter 2. An ϵ-MPMAB problem instance comprises M players, K arms, and a dissimilarity parameter $\epsilon \in [0, 1]$. Let $[M]$ denote the set of players and $[K]$ the set of arms. For each player

$p \in [M]$ and each arm $i \in [K]$, there is an initially-unknown reward distribution $\mathcal{D}_i^p$, which has support $[0, 1]$ and has mean μ_i^p.

Reward dissimilarity.

The reward distributions for each arm are assumed to be *similar but not necessarily identical* for different players; specifically,

$$\forall i \in [K],\ p, q \in [M], \quad \left|\mu_i^p - \mu_i^q\right| \leq \epsilon. \tag{3.1}$$

Protocol.

In Chapter 2, the players interact with the arms in rounds, and within each round, all players take an action concurrently. In this chapter, inspired by the problem setup of Hong et al. [63], we generalize the interaction protocol such that it allows any subset of the players to take an action. In each round $t \in [T]$, where $T > \max(K, M)$ is the time horizon of learning, a subset of players $\mathcal{P}_t \subseteq [M]$ is chosen (called the *active player set* at round t) by an oblivious adversary; each active player $p \in \mathcal{P}_t$ then pulls an arm $i_t^p \in [K]$ and observes an independently-drawn reward $r_t^p \sim \mathcal{D}_{i_t^p}^p$. At the end of round t, the active players communicate their decisions, $\left\{i_t^p : p \in \mathcal{P}_t\right\}$, as well as their observed rewards, $\left\{r_t^p : p \in \mathcal{P}_t\right\}$, to all players. Note that, when $|\mathcal{P}_t| = 1$ for all t, the problem setting resembles the one in [31] and captures a sequential transfer bandit learning setting [e.g., 10]; when $\mathcal{P}_t = [M]$ for all t, we recover the setting in Chapter 2.

Performance metric.

The goal of the players is to minimize their expected collective regret, which we define shortly. For each player $p \in [M]$, let $\mu_*^p = \max_{j \in [K]} \mu_j^p$ denote the mean reward of an optimal arm for p; then, for each arm $i \in [K]$, let $\Delta_i^p = \mu_*^p - \mu_i^p \geq 0$ denote the (suboptimality) gap of arm i for player p. In addition, let $n_i^p(t) = \sum_{s \leq t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\}$ denote the number of pulls of arm i by player p after t rounds. Then, the individual

expected regret of any player p is defined as

$$\mathrm{Reg}^p(T) = \mathbb{E}\left[\sum_{\substack{t\in[T]:\\ p\in\mathcal{P}_t}} \mu_*^p - \mu_{i_t^p}^p\right] = \sum_{i\in[K]} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p.$$

Finally, the *expected collective regret* is defined as the sum of individual expected regret over all the players, i.e.,

$$\mathrm{Reg}(T) = \sum_{p\in[M]} \mathrm{Reg}^p(T) = \sum_{i\in[K]}\sum_{p\in[M]} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p. \tag{3.2}$$

Does one need to know ϵ?

In this chapter, we focus on the case where ϵ is *known* to the players in the ϵ-MPMAB problem. This is because in Chapter 2, we have shown that, unfortunately, not much can be done when ϵ is unknown to the players—a lower bound (Theorem 11 therein) shows that no sublinear-regret algorithms can effectively take advantage of inter-task data aggregation for every $\epsilon \in [0, 1]$ to achieve improved regret upper bounds.

3.2.2 Existing results

In the concurrent setting ($\mathcal{P}_t = [M]$ for all t), we showed in Chapter 2 that, whether data aggregation can be provably beneficial for an arm i depends on how its associated suboptimality gaps, Δ_i^p's, compare with the dissimilarity parameter, ϵ. Specifically, the problem complexity is captured by the notion *subpar arms*, $\mathcal{I}_\alpha = \left\{i : \exists p \in [M], \mu_*^p - \mu_i^p > \alpha\right\}$; see Section 2.3.2.

Upper and lower bounds are provided in Chapter 2. They characterize that, informally, the collective performance of the players can be improved by a factor of M (resp. $\sqrt{M}$) for each $\mathcal{O}(\epsilon)$-subpar arm in the (suboptimality) gap-dependent (resp. gap-independent) bounds, where we recall that M is the number of players. This improvement is in comparison with baseline algorithms in which each player runs their own instance of

a bandit algorithm individually, IND-UCB.

In Appendix B.4, we briefly recap the algorithm in Chapter 2, ROBUSTAGG(ϵ). We show that with a few small modifications, it can be extended to work in the generalized ϵ-MPMAB setting, and achieve generalized regret guarantees (see Theorem B.76).

3.2.3 Baseline: IND-TS

In this chapter, we consider another baseline algorithm, IND-TS, in which each player runs the standard TS algorithm with Gaussian priors. We now describe the TS algorithm. At a high level, every learner (player) p begins with some prior belief on the mean reward of each arm, and through interactions with the environment, the learner updates its posterior belief. Specifically, we consider TS with Gaussian product priors—a learner maintains one Gaussian posterior distribution for each arm, beginning with $\mathcal{N}(0,1)$. In each round t, the learner draws an independent sample $\theta_i^p(t)$ for each arm i from its corresponding posterior distribution, which is of form $\mathcal{N}\left(\bar{\mu}_i^p, \frac{1}{n_i^p(t-1)\vee 1}\right)$, where $\bar{\mu}_i^p = \frac{1}{n_i^p(t-1)\vee 1}\sum_{s<t:p\in\mathcal{P}_s, i_s^p=i} r_s^p$ is the empirical mean reward of player p pulling arm i. The learner then pulls the arm $i_t^p = \text{argmax}_i\, \theta_i^p(t)$, receives a reward $r_t^p \sim \mathcal{D}_{i_t^p}^p$, and updates the posterior distribution for arm i.

Based on the results of Agrawal and Goyal [4], we obtain the regret guarantees of IND-TS by summing over individual bounds: $\mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in[K]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right)$ and $\tilde{\mathcal{O}}\left(M\sqrt{KT}\right)$.

In Appendix B.4, we briefly recap the guarantees of IND-UCB and IND-TS in the generalized ϵ-MPMAB setting, where $\mathcal{P}_t$'s are not necessarily $[M]$ in every round.

3.3 Algorithm: ROBUSTAGG-TS(ϵ)

We now present a TS-type randomized exploration algorithm, ROBUSTAGG-TS(ϵ) (Algorithm 2), which can robustly leverage data collected by all the players.

In each round t, for each active player $p \in \mathcal{P}_t$ and arm i, ROBUSTAGG-TS(ϵ)

Algorithm 2: RobustAgg-TS (ϵ)

Input : Dissimilarity parameter $\epsilon \in [0,1]$, universal constants $c_1, c_2 > 0$;

1 **Initialization:** For every $i \in [K]$ and $p \in [M]$, set $n_i^p = 0$, ind-$\hat{\mu}_i^p = 0$, ind-var$_i^p = c_2$, agg-$\hat{\mu}_i^p = 0$, and agg-var$_i^p = c_2$; for every $i \in [K]$, set $n_i = 0$.

2 **for** *round* $t \in [T]$ **do**

3 Receive active set of players $\mathcal{P}_t$.

4 **for** *active player* $p \in \mathcal{P}_t$ **do**

5 **for** *arm* $i \in [K]$ **do**

6 **if** $n_i^p \geq \frac{c_1 \ln T}{\epsilon^2} + 2M$ **then**

7 $\hat{\mu}_i^p \leftarrow$ ind-$\hat{\mu}_i^p$, var$_i^p \leftarrow$ ind-var$_i^p$; // Use the individual posterior

8 **else**

9 $\hat{\mu}_i^p \leftarrow$ agg-$\hat{\mu}_i^p$, var$_i^p \leftarrow$ agg-var$_i^p$; // Use the aggregate posterior

10 $\theta_i^p(t) \sim \mathcal{N}(\hat{\mu}_i^p, \text{var}_i^p)$

11 Player p pulls arm $i_t^p = \text{argmax}_{i \in [K]} \theta_i^p(t)$ and observes reward r_t^p.

12 **for** *active player* $p \in \mathcal{P}_t$ **do**

13 Let $i = i_t^p$. Update $n_i^p \leftarrow n_i^p + 1$ and $n_i \leftarrow n_i + 1$.

14 **for** *active player* $p \in \mathcal{P}_t$ **do**

15 Let $i = i_t^p$.

// **Only** update posteriors associated with p and i_t^p

16 Update

$$\text{ind-}\hat{\mu}_i^p \leftarrow \frac{1}{n_i^p \vee 1} \sum_{s \leq t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\} r_s^p, \quad \text{ind-var}_i^p \leftarrow \frac{c_2}{n_i^p \vee 1};$$

17

$$\text{agg-}\hat{\mu}_i^p \leftarrow \frac{1}{n_i \vee 1} \sum_{s \leq t} \sum_{q \in \mathcal{P}_s} \mathbb{1}\left\{i_s^q = i\right\} r_s^q + \epsilon, \quad \text{agg-var}_i^p \leftarrow \frac{c_2}{(n_i - M) \vee 1}.$$

maintains two Gaussian "posterior" distributions. As a standard single-task TS algorithm with Gaussian priors would normally maintain [e.g. 4], $\mathcal{N}\left(\text{ind-}\hat{\mu}_i^p, \text{ind-var}_i^p\right)$, the *individual posterior* is solely based on player p's own interactions with arm i, with ind-$\hat{\mu}_i^p$ and ind-var$_i^p$ defined in line 16. In contrast, the *aggregate posterior*, $\mathcal{N}\left(\text{agg-}\hat{\mu}_i^p, \text{agg-var}_i^p\right)$, is unique to the multi-task setting—its mean, agg-$\hat{\mu}_i^p$, is the sum of the empirical mean of all players' observed rewards for arm i and a bonus term ϵ, and its variance, agg-var$_i^p$, is based on the total number of pulls of arm i by all players (line 17).

The algorithm chooses one of the posterior distributions (lines 6 to 9), i.e., decides whether to utilize data shared by other players, by balancing a bias-variance trade-off

[19, 138, see also Chapter 2]: while an inclusion of n_i reward samples collected by all players leads to a variance, agg-var$_i^p$, which can be much smaller than ind-var$_i^p$, it may also cause agg-$\hat{\mu}_i^p$ to be biased as the reward distributions for different players may be different. The algorithm then independently draws a sample, $\theta_i^p(t)$, from the chosen posterior distribution (line 10) and pulls the arm with the largest $\theta_i^p(t)$ for player p (line 11).

Specifically, in round t, for player $p \in \mathcal{P}_t$ and arm $i \in [K]$, the algorithm chooses a posterior distribution by comparing n_i^p, the number of pulls of i by p at the beginning of round t, to a threshold in terms of the dissimilarity parameter, i.e., $\frac{c_1 \ln T}{\epsilon^2} + 2M$ (line 6), where $c_1 > 0$ is some numerical constant. Intuitively, when ϵ is smaller, each player stays longer on using the aggregate posterior to perform randomized exploration, which indicates a higher degree of trust on data from other tasks.

After all players in $\mathcal{P}_t$ obtain rewards for their arm pulls, they compute and *update* their posteriors with new data. In principle, data from one player can affect the aggregate posteriors of all players. We make the design choice that this effect gets delayed: the algorithm only updates the posteriors for player p and arm i in round t, if $p \in \mathcal{P}_t$ and $i = i_t^p$ (line 15). Although our current analysis (see Sections 3.4 and 3.5 below) relies on this property to establish sharp regret guarantees, we conjecture that similar regret guarantees can be shown even if the algorithm updates the posteriors of all players and all arms in every round[2].

3.4 Main Results

We now present gap-dependent and gap-independent regret upper bounds of RobustAgg-TS(ϵ).

Recall that $\mathcal{I}_\alpha = \{i \in [K] : \exists p,\ \Delta_i^p > \alpha\}$ is the set of α-subpar arms.

[2]In Section B.5.1 of the appendix, we show that this variation induces little effect on the empirical performance of the algorithm.

Theorem 3.1 (Gap-dependent bound)**.** *There exists a setting of $c_1, c_2 > 0$, such that, the expected collective regret of* RobustAgg-TS(ϵ) *after $T > \max(K, M)$ rounds satisfies:*

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_{10\epsilon}}\sum_{\substack{p\in[M]\\ \Delta_i^p>0}}\frac{\ln T}{\Delta_i^p} + \sum_{i\in\mathcal{I}_{10\epsilon}^C}\sum_{\substack{p\in[M]\\ \Delta_i^p>0}}\frac{\ln T}{\Delta_i^p} + M^2K\right).$$

Theorem 3.2 (Gap-independent bound)**.** *There exists a setting of $c_1, c_2 > 0$, such that, the expected collective regret of* RobustAgg-TS(ϵ) *after $T > \max(K, M)$ rounds satisfies:*

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{10\epsilon}|P} + \sqrt{M\left(|\mathcal{I}_{10\epsilon}^C| - 1\right)P} + M^2K\right),$$

where $P = \sum_{t=1}^{T}|\mathcal{P}_t|$.

The proofs of Theorems 3.1 and 3.2 can be found in Appendix B.3; in Section 3.5, we also highlight several technical challenges and proof ingredients in our analysis.

Guarantees in the generalized ϵ-MPMAB setting.

Our guarantees for RobustAgg-TS(ϵ) hold under the generalized ϵ-MPMAB setting, in that $\mathcal{P}_t$'s at each round can change over time. Observe that the regret bound given by Theorem 3.1 does not depend on $\mathcal{P}_t$'s, and the regret bound given by Theorem 3.2 has the highest value when $P = MT$. In addition, recall that near-matching gap-dependent and gap-independent lower bounds have been shown in Chapter 2 for the $\mathcal{P}_t \equiv [M]$ setting (Section 3.2.2). These lower bounds indicate the near-optimality of RobustAgg-TS(ϵ)'s guarantees, modulo an additive lower-order term $\mathcal{O}(M^2K)$ which does not depend on T.

Furthermore, the gap-independent guarantee in Theorem 3.2 adapts to the value of P. This shows the flexibility of RobustAgg-TS(ϵ). Specifically, if $|\mathcal{P}_t| = 1$ (similar to the settings of [31, 57]), we have $P = T$, and

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{10\epsilon}|T} + \sqrt{M\left(|\mathcal{I}_{10\epsilon}^C| - 1\right)T} + M^2K\right).$$

Similarly, if $\mathcal{P}_t = [M]$ for all t (Chapter 2), then $P = MT$, and

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{M|\mathcal{I}_{10\epsilon}|T} + M\sqrt{\left(|\mathcal{I}_{10\epsilon}^C| - 1\right)T} + M^2K\right).$$

Comparison with baselines.

In comparison with the guarantees of the UCB-based algorithm RobustAgg(ϵ) in Appendix B.4.2, we see that RobustAgg-TS(ϵ) has competitive guarantees, except that the set of arms which benefits from data aggregation changes from $\mathcal{I}_{5\epsilon}$ to $\mathcal{I}_{10\epsilon}$.

In comparison with the guarantees of Ind-UCB and Ind-TS, the regret guarantees of RobustAgg-TS(ϵ) are never worse (modulo lower-order terms), and save factors of $\frac{1}{M}$ and $\frac{1}{\sqrt{M}}$ in $\mathcal{I}_{10\epsilon}$'s contribution in the gap-dependent and gap-independent regret guarantees, respectively.

3.5 Proof Ingredients

In this section, we highlight some of the novel proof ingredients used in our analysis of Algorithm 2, which are unique to the *multi-task* setting[3].

We begin by decomposing the regret in terms of subpar arms and non-subpar arms. It follows from Eq. (3.2) that

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sum_{i\in\mathcal{I}_{10\epsilon}} \mathbb{E}\left[n_i(T)\right]\Delta_i^{\min} + \sum_{i\in\mathcal{I}_{10\epsilon}^C}\sum_{p\in[M]} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p\right),$$

where we let $n_i(T) = \sum_{p=1}^M n_i^p(T)$ be the number of pulls of arm i by all players after T rounds; we recall that $\Delta_i^{\min} = \min_{p\in[M]}\Delta_i^p$; and we use the fact that for any subpar arm $i \in \mathcal{I}_{10\epsilon}$ and any player $p \in [M]$, $\Delta_i^p \leq 2\Delta_i^{\min}$ (Fact B.24).

In the interest of space, we focus on the analysis for subpar arms and defer the

[3]Our analysis involves various proofs by cases. Figure B.1 in the appendix provides an overview illustrating the case division rules used in our proofs.

discussion on non-subpar arms to the appendix. The following lemma provides an upper bound on $\mathbb{E}\left[n_i(T)\right]$ for $i \in \mathcal{I}_{10\epsilon}$, which can be subsequently used to derive the upper bounds on the expected collective regret incurred by the 10ϵ-subpar arms in Section 3.4.

Lemma 3.3. *For any arm $i \in \mathcal{I}_{10\epsilon}$,*

$$\mathbb{E}\left[n_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

While a similar lemma can be found for the UCB-based algorithm (see Lemma A.7), RobustAgg(ϵ), proving Lemma 3.3 requires new ingredients that we present in the rest of this section.

Let us fix an arm $i \in \mathcal{I}_{10\epsilon}$. To control $\mathbb{E}\left[n_i(T)\right] = \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}\right]$, we begin by generalizing a technique introduced by Agrawal and Goyal [4] for standard TS to the multi-task setting. In each round t and for each active player p, we consider two cases: (1) player p pulls arm i (namely, $i_t^p = i$), and $\theta_i^p(t)$ (line 10 in Algorithm 2) is greater than some threshold $y_i^p \in (\mu_i^p, \mu_*^p)$ to be defined shortly, and (2) $i_t^p = i$ and $\theta_i^p(t) \leq y_i^p$. We have

$$\begin{aligned}\mathbb{E}\left[n_i(T)\right] = &\underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \theta_i^p(t) > y_i^p, \mathcal{E}_t\right\}\right]}_{(A)} \\ &+ \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \theta_i^p(t) \leq y_i^p, \mathcal{E}_t\right\}\right]}_{(B)} + \mathcal{O}\left(1\right),\end{aligned}$$

where $\mathcal{E}_t$, informally, is a high-probability "clean" event in which $\hat{\mu}_i^p$'s maintained by Algorithm 2 in round t for each i and p concentrate towards their respective expected values.

Term (A) can be controlled because, as more pulls of arm i are made, $\left\{\theta_i^p(t) > y_i^p\right\}$ is unlikely to happen, as $\hat{\mu}_i^p$ concentrates towards a value smaller than y_i^p, and var_i^p decreases.

See Lemma B.40 in the appendix for a detailed proof.

In what follows, we focus on bounding term (B). Observe that $\left\{i_t^p = i, \theta_i^p(t) \leq y_i^p\right\}$ in (B) happens only if $\forall j \in [K]$, $\theta_j^p(t) \leq y_i^p$, including the optimal arm(s) for player p. Since in an ϵ-MPMAB problem instance, different players may have different optimal arms, we consider a common near-optimal arm $\dagger \in \mathcal{I}_{2\epsilon}^C$—see Fact B.24 in the appendix for the existence of such an arm. It can be easily verified that, for any arm $i \in \mathcal{I}_{10\epsilon}$ and player $p \in [M]$, $\delta_i^p := \mu_\dagger^p - \mu_i^p > 0$ (see Fact B.38). In other words, while $\dagger$ may not necessarily be an optimal arm for every player, it has a larger mean reward than any $i \in \mathcal{I}_{10\epsilon}$. We can now define $y_i^p := \mu_i^p + \frac{1}{2}\delta_i^p \in (\mu_i^p, \mu_\dagger^p) \subset (\mu_i^p, \mu_*^p)$.

Using a technique first introduced in [4], we will show that $\theta_\dagger^p(t)$ converges to a value greater than y_i^p *fast* enough so that $\left\{\forall j \in [K], \theta_j^p(t) \leq y_i^p\right\}$ will unlikely happen soon enough and thus (B) can be controlled.

Remark 3.4 (Comparison with UCB-based analyses)**.** *We note that controlling term* (B) *is often not required in the analyses of UCB-based algorithms. Colloquially, this term concerns the event in which arm* i *is pulled even when its sample/index value is smaller than* y_i^p*; such an event would unlikely happen for UCB-based algorithms as the* optimism in the face of uncertainty *principle ensures that, with high probability, the UCB index of an optimal arm for player* p *is greater than or equal to* $\mu_*^p \geq \mu_\dagger^p > y_i^p$.

Before we formalize the above-mentioned intuition for bounding term (B) in Lemma 3.5, we first lay out a few helpful definitions. We define $\{\mathcal{F}_t\}_{t=0}^T$ to be a filtration such that $\mathcal{F}_t = \sigma\left(\{i_s^q, r_s^q : s \leq t, q \in \mathcal{P}_s\}\right)$ is the σ-algebra generated by interactions of all players up until round t. Then, let $\phi_{i,t}^p = \Pr\left(\theta_\dagger^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right)$. Observe that if $\phi_{i,t}^p$ is large, the event $\left\{i_t^p = i, \theta_i^p(t) \leq y_i^p\right\}$ will unlikely happen.

Lemma 3.5.

$$(B) \leq \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1}{\phi^p_{i,t}}-1\right)\mathbb{1}\left\{i^p_t=\dagger,\mathcal{E}_t\right\}\right]}_{(B*)}.$$

See Lemma B.45 and its proof in the appendix for details. We now consider the following two cases: in any round t and for any active player p that pulls arm $\dagger$, i.e., $i^p_t=\dagger$, p uses *either* the individual *or* the aggregate posterior distribution associated with arm $\dagger$ (lines 6 to 9 in Algorithm 2). Let $H^p_\dagger(t)$ be the event that p uses the individual posterior distribution and $\overline{H^p_\dagger(t)}$ be the event that p uses the aggregate posterior (see Definition B.13 in the appendix for the formal definitions). We can then decompose $(B*)$ as follows:

$$(B*) = \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1}{\phi^p_{i,t}}-1\right)\mathbb{1}\left\{i^p_t=\dagger,\mathcal{E}_t,H^p_\dagger(t)\right\}\right]}_{(b1)}$$
$$+\underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1}{\phi^p_{i,t}}-1\right)\mathbb{1}\left\{i^p_t=\dagger,\mathcal{E}_t,\overline{H^p_\dagger(t)}\right\}\right]}_{(b2)}.$$

Let $m^p_\dagger(t)$ denote the aggregate number of pulls of arm $\dagger$ maintained by player p after t rounds (see Definition B.9 in the appendix). Note that, by the design choice of Algorithm 2 (line 15), $m^p_\dagger(t)$ is not necessarily the same as $n_\dagger(t)$. With foresight, let

$L = \Theta\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right)$, and let $G_t^p = \left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}\right\}$. We have

$$(b2) = \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{G_t^p, m_\dagger^p(t-1) < L\right\}\right]}_{(b2.1)}$$
$$+ \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{G_t^p, m_\dagger^p(t-1) \geq L\right\}\right]}_{(b2.2)}.$$

Both $(b1)$ and $(b2.2)$ can be bounded by $\mathcal{O}(M)$, because, informally speaking, either player p has pulled arm $\dagger$ many times when the individual posterior is used (term $(b1)$) or the players collectively have pulled $\dagger$ many times when the aggregate posterior is used (term $(b2.2)$), and $\frac{1}{\phi_{i,t}^p} - 1$ can therefore be upper bounded by $\frac{1}{T}$. See Lemma B.47 and Lemma B.52 and their proofs for details.

The main challenge in bounding $\mathbb{E}\left[n_i(T)\right]$ lies in term $(b2.1)$, for which we show the following lemma.

Lemma 3.6 (Bounding term $(b2.1)$)**.**

$$(b2.1) \leq \mathcal{O}(L) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

Proving Lemma 3.6 is *central* to our analysis and as we will see, requires special care. We begin by introducing the following notion. For any arm $j \in [K]$ and $k \in [TM]$, let

$$\tau_k(j) = \min\left\{T+1, \min\left\{t : n_j(t) \geq k\right\}\right\}$$

be the round in which arm j is pulled the k-th time by any player. Furthermore, let $\tau_0(j) = 0$ by convention. For any $j \in [K]$ and $k \in [TM]$, it is easy to verify that $\tau_k(j)$ is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$. In what follows, when circumstances permit, we

abuse the notation and denote $\tau_k(\dagger)$ by τ_k.

Invariant property.

By the construction of Algorithm 2, in any round t, a player only updates the posteriors associated with an arm if the player pulls the arm in the round t (line 15). This design choice induces an invariant property: for any arm and player, certain random variables associated with them stay invariant between consecutive pulls of the arm by the player (see Definition B.20 and a few examples in the appendix).

The invariant property allows us to bound $(b2.1)$ as follows in terms of the stopping times τ_k's (See Lemma B.48 and Lemma B.72 in the appendix):

$$(b2.1) \leq \sum_{p=1}^{M} \mathbb{E}\left[\left(\frac{1}{\phi_{i,1}^{p}} - 1\right) \mathbb{1}\left\{\overline{H_{\dagger}^{p}(1)}\right\}\right] + \sum_{k=1}^{L-1} \mathbb{E}\left[\left(\frac{1}{\phi_{i,\tau_k+1}^{p_k}} - 1\right) \mathbb{1}\left\{\tau_k \leq T, \overline{H_{\dagger}^{p}(\tau_k+1)}\right\}\right],$$

where $p_k := p_k(\dagger)$ is the player that makes the k-th pull of arm $\dagger$ (Definition B.17).

Using basic Gaussian tail bounds, we can show that $\mathbb{E}\left[\left(\frac{1}{\phi_{i,1}^{p}} - 1\right) \mathbb{1}\left\{\overline{H_{\dagger}^{p}(1)}\right\}\right] \leq \mathcal{O}(1)$ for any player p. Then, the following lemma suffices to prove Lemma 3.6.

Lemma 3.7. *For any $k \in [TM]$,*

$$\mathbb{E}\left[\left(\frac{1}{\phi_{i,\tau_k+1}^{p_k}} - 1\right) \mathbb{1}\left\{\tau_k \leq T, \overline{H_{\dagger}^{p}(\tau_k+1)}\right\}\right] \leq \mathcal{O}(1).$$

Technical highlight.

Lemma 3.7 generalizes Agrawal and Goyal [4, Lemma 2.13] for standard TS to the *multi-task* setting. A complete proof can be found in the appendix, which uses anti-concentration bounds of Gaussian random variables [59] as well as a *novel* concentration inequality for multi-task data aggregation at random stopping times $\tau_k(\dagger)$'s, which we

highlight here[4]. For any arm j, let

$$\text{agg-}\hat{\mu}_j(t) = \frac{1}{n_j(t) \vee 1} \sum_{s \leq t} \sum_{q \in \mathcal{P}_s} \mathbb{1}\left\{i_s^q = j\right\} r_s^q + \epsilon$$

be the aggregate mean reward estimate of j constructed using data by all players after t rounds, offset by ϵ.

Lemma 3.8. *For any arm $j \in [K]$ and $k \in [TM] \cup \{0\}$, denote by $\tau_k = \tau_k(j)$. Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, one of the following events happens:*

1. $\tau_k = T + 1$;
2. $\forall p \in [M],\ \mu_j^p - \text{agg-}\hat{\mu}_j(\tau_k) \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_j(\tau_k) - M\right) \vee 1}}$.

Remark 3.9. *We note that Lemma 3.8 is critical to the tight performance guarantee in Lemma 3.7 and subsequently the near-optimal regret guarantees. This result is non-trivial, as it is a concentration bound for a sequence of random variables whose length, $n_j(\tau_k(j))$, is also a random variable. Furthermore, since $\tau_k(j)$ is the round in which arm j is pulled the k-th time by any player, $n_j(\tau_k(j))$ can potentially take any integer value in $[k, k + M - 1]$ because there can be up to M pulls of arm j in round $\tau_k(j)$. We note that using the Azuma-Hoeffding inequality together with a union bound or Freedman's inequality (similar to Lemma A.4) can lead to extra $\mathcal{O}(M)$ or $\mathcal{O}(\ln T)$ terms for Lemma 3.7, respectively (see Remark B.51 in the appendix for details).*

To our best knowledge, we are not aware of any similar tight concentration bounds for data aggregation in multi-task bandits, and our technique may be of independent interest for analyzing other multi-task sequential learning problems.

[4]In the single-task case ($M = 1$), our proof technique (Lemma B.70) also simplifies the proof of the first case of Agrawal and Goyal [4, Lemma 2.13].

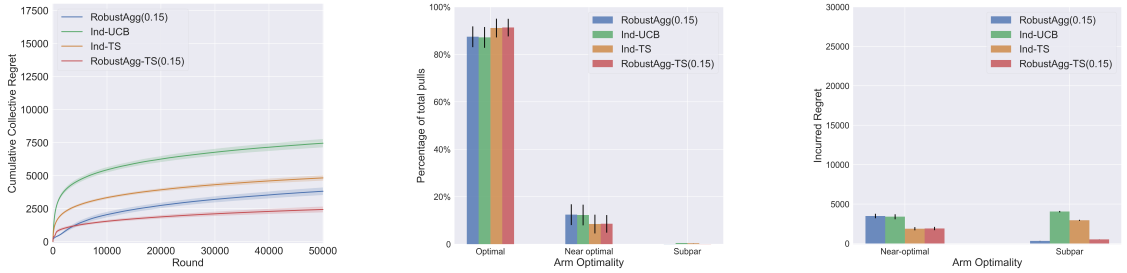
3.6 Related Work

There exist many prior works that study multi-player or multi-task bandits with heterogeneous reward distributions. For example, Cesa-Bianchi et al. [31] use Laplacian-based regularization to learn a network of bandit problem instances such that connected problems have similar parameters; Gentile et al. [57], among others, study clustering of bandit problem instances. The ϵ-MPMAB problem is introduced in Chapter 2; see Appendix A.1 for a detailed comparison with related work. In Chapter 4, we generalize the ϵ-MPMAB problem to episodic, tabular Markov decision processes. We note that while the methods in the above-mentioned works are UCB-based, we study TS-type algorithms in this chapter.

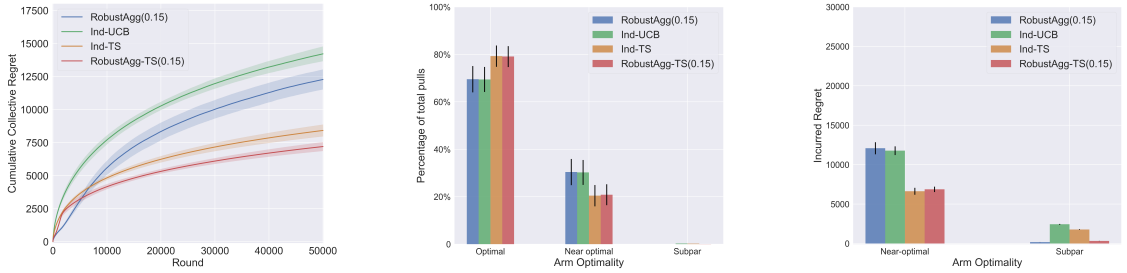
TS is initially proposed by Thompson [148] decades ago, but its frequentist analysis has not emerged until recent years [e.g., 3, 76]. Jin et al. [73] present the first minimax optimal TS-type algorithm. Our proof techniques in this chapter are mostly inspired by the work of Agrawal and Goyal [4].

TS algorithms have been studied in multi-task Bayesian bandits. For example, several recent works study the setting of interacting with a sequence of M bandit problem instances (tasks) sampled from a common, unknown prior distribution, with a goal of minimizing the M-instance Bayesian regret [16, 85, 119, 17]. The recent work of Hong et al. [63] proposes a hierarchical Bayesian bandit problem that generalizes many multi-task bandit settings, and analyzes the Bayes regret. In contrast, we use frequentist regret as our performance metric, and we do not assume a shared prior distribution over the players' problem instances/tasks. Wan et al. [155] study multi-task TS in a hierarchical Bayesian model and assume knowledge of metadata of each task; while they provide a frequentist regret bound, we study the ϵ-MPMAB problem which models task relations differently.

Similar models on *sequential* transfer between problem instances have also been studied by Azar et al. [10] and Soare et al. [138]. Zhang and Bareinboim [175], Zhang



(a) $|\mathcal{I}_{5\epsilon}| = 8$



(b) $|\mathcal{I}_{5\epsilon}| = 5$

Figure 3.1. Compares the average performance of the algorithms on 30 randomly generated problem instances with $|\mathcal{I}_{5\epsilon}| = 8$ and $|\mathcal{I}_{5\epsilon}| = 5$ in a horizon of $T = 50000$ rounds. Figures in the left column plot the cumulative collective regret over time; figures in the middle column demonstrate the percentages of pulls of optimal arms, non-subpar yet non-optimal arms (referred to as near-optimal arms), and subpar arms; figures in the right column then show the incurred cumulative regret by arm optimality.

et al. [174], Sharma et al. [130] investigate warm-starting bandits from misaligned data. In this chapter, we focus on a more general interaction protocol, under which the players may interact with the environment concurrently.

3.7 Empirical Evaluation

In this section, we present an empirical evaluation of RobustAgg-TS(ϵ) on synthetic data. We focus on the concurrent setting ($\mathcal{P}_t = [M]$ for all t), which is the setting studied in Chapter 2. Our goal is to address the following two questions:

1. How does RobustAgg-TS(ϵ) perform in comparison with the UCB-based algorithm, RobustAgg(ϵ), and the baseline algorithms without transfer learning?

2. Does the notion of subpar arms characterize the performance of the algorithms in

practice?

Experimental Setup.

We compared the performance of 4 algorithms: (1) RobustAgg-TS(ϵ) with constants $c_1 = \frac{1}{2}$ and $c_2 = 1$; (2) RobustAgg(ϵ) (Section 2.6); (3) Ind-TS, the baseline algorithm that runs TS with Gaussian priors for each player individually; and (4) Ind-UCB, the baseline algorithm that runs UCB-1 for each player individually.

The algorithms were evaluated on randomly generated 0.15-MPMAB problem instances with different numbers of subpar arms. To stay consistent with Chapter 2, we followed the same instance generation procedure and considered $\mathcal{I}_{5\epsilon}$ to be the set of subpar arms—we set the number of players $M = 20$ and the number of arms $K = 10$; then, for each integer value $v \in [0, 9]$, we generated 30 0.15-MPMAB problem instances with Bernoulli reward distributions and $|\mathcal{I}_{5\epsilon}| = v$. We ran the algorithms on each instance for a horizon of $T = 50,000$ rounds.

Results and Discussion.

Figure 3.1 compares the average performance of the algorithms on instances with $|\mathcal{I}_{5\epsilon}| = 8$ and 5. We defer the rest of the results to Appendix B.5.

From the left column, we first observe that, while the UCB-based algorithm, RobustAgg(ϵ), outperforms its counterpart, Ind-UCB, in the cumulative collective regret ($\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mu_*^p - \mu_{i_t^p}^p$), its empirical performance is underwhelming in comparison with TS algorithms. In particular, even on instances with half of the arms *subpar* ($|\mathcal{I}_{5\epsilon}| = 5$), RobustAgg($\epsilon$) is outperformed by the Ind-TS baseline without transfer learning. Importantly, we note that RobustAgg-TS(ϵ) shows a superior performance than the other algorithms.

The figures in the middle and right columns illustrate the arm selection of each algorithm. We categorize all arms into three groups: optimal arms, subpar arms, and near-optimal arms which are neither subpar nor optimal. Comparing the TS-type algorithms

with the UCB-based algorithms, we observe that the former algorithms perform better mainly because they pull near-optimal arms a smaller number of times and incur less regret on these arms.

Furthermore, we observe that RobustAgg(ϵ) and RobustAgg-TS(ϵ), when compared with their counterparts (Ind-UCB and Ind-TS, respectively), incur a similar amount of regret from near-optimal arms. Meanwhile, they make fewer pulls on subpar arms. This may be less obvious from the plots on the percentage of total pulls because none of the algorithms pull subpar arms extensively over the horizon. However, since the suboptimality gaps of subpar arms are large, we see from the figures in the right column that RobustAgg(ϵ) and RobustAgg-TS(ϵ) incur far less regret on subpar arms. These results thereby demonstrate that the notion of subpar arms can capture the amenability of transfer learning in subpar arms but not near-optimal arms.

In addition, the results show that, empirically, RobustAgg-TS(ϵ) can robustly leverage transfer for arms in $\mathcal{I}_{5\epsilon} \supseteq \mathcal{I}_{10\epsilon}$—this suggests that our upper bounds may be improved; we leave this as future work.

3.8 Conclusion

In this chapter, we studied transfer learning in multi-task bandits under the framework of a generalized version of the ϵ-MPMAB problem. We proposed a TS-type algorithm, RobustAgg-TS(ϵ), which can robustly leverage auxiliary data collected for other tasks. We showed that RobustAgg-TS(ϵ) is empirically superior when evaluated on synthetic data, and also near-optimal in gap-dependent and gap-independent frequentist guarantees. In our analysis, we also proved a novel concentration inequality for multi-task data aggregation, which can be of independent interest in the analysis of other multi-task online learning problems. For future work, we are interested in improving the lower-order terms in our regret bounds and evaluating our algorithm in real-world applications.

Acknowledgement.

Chapter 3 is based on the material as it appears in "Thompson Sampling for Robust Transfer in Multi-Task Bandits" by Zhi Wang, Chicheng Zhang, and Kamalika Chaudhuri [162]. The material was published in *Proceedings of the 39th International Conference on Machine Learning*. The dissertation author was the primary investigator and first author of the paper.

Chapter 4

Multi-Task Reinforcement Learning with Model Transfer

4.1 Introduction

In many real-world applications, reinforcement learning (RL) agents can be deployed as a group to complete similar tasks at the same time. For example, in healthcare robotics, robots are paired with people with dementia to perform personalized cognitive training activities by learning their preferences [150, 83]; in autonomous driving, a set of autonomous vehicles learn how to navigate and avoid obstacles in various environments [97]. In these settings, each learning agent alone may only be able to acquire a limited amount of data, while the agents as a group have the potential to collectively learn faster through sharing knowledge among themselves. Multi-task learning [30] is a practical framework that can be used to model such settings, where a set of learning agents share/transfer knowledge to improve their collective performance.

Despite many empirical successes of multi-task RL [e.g., 182, 98, 97] and transfer learning for RL [e.g., 93, 146], a theoretical understanding of when and how information sharing or knowledge transfer can provide benefits remains limited. Exceptions include [e.g., 60, 24, 42, 65, 117, 92], which study multi-task learning from parameter or representation transfer perspectives. However, these works still do not provide a completely satisfying answer: for example, in many application scenarios, the reward structures and the

environment dynamics are only slightly different for each task—this is, however, not captured by representation transfer [42, 65] or existing works on clustering-based parameter transfer [60, 24]. In such settings, is it possible to design provably efficient multi-task RL algorithms that have guarantees never worse than agents learning individually, while outperforming the individual agents in favorable situations?

In this work, we formulate a multi-task RL problem that is applicable to the aforementioned settings. Specifically, we generalize the results on multi-task multi-armed bandits (Chapter 2) and formulate the ϵ-Multi-Player Episodic Reinforcement Learning (abbreviated as ϵ-MPERL) problem, in which all tasks share the same state and action spaces, and the tasks are assumed to be similar—i.e., the dissimilarities between the environments of different tasks (specifically, the reward distributions and transition dynamics associated with the players/tasks) are bounded in terms of a dissimilarity parameter $\epsilon \geq 0$. This problem not only models concurrent RL [134, 60] as a special case by taking $\epsilon = 0$, but also captures richer multi-task RL settings when ϵ is nonzero. We study regret minimization for the ϵ-MPERL problem, specifically:

1. We identify a problem complexity notion named *subpar* state-action pairs, which captures the amenability of information sharing among tasks in ϵ-MPERL problem instances. As shown in the multi-task bandits literature (see Chapter 2), information sharing is *not* always helpful. Subpar state-action pairs, intuitively speaking, are clearly suboptimal for all tasks, for which we can robustly take advantage of (possibly biased) data collected for other tasks.

2. In the setting where the dissimilarity parameter ϵ is known, we design a model-based algorithm Multi-task-Euler (Algorithm 3), which is built upon state-of-the-art algorithms for learning single-task Markov decision processes (MDPs) [172, 136], as well as model transfer ideas in RL [146]. Multi-task-Euler crucially utilizes the dissimilarity assumption to robustly take advantage of information sharing among

tasks, and achieves regret upper bounds in terms of subpar state-action pairs, in both (suboptimality) gap-dependent and gap-independent fashions. Specifically, compared with a baseline algorithm that does not utilize information sharing, MULTI-TASK-EULER has a regret guarantee that: (1) is never worse, i.e., it avoids negative transfer [123]; (2) can be much superior when there are a large number of subpar state-action pairs.

3. We also present gap-dependent and gap-independent regret lower bounds for the ϵ-MPERL problem in terms of subpar state-action pairs. Together, the upper and lower bounds characterize the intrinsic complexity of the ϵ-MPERL problem.

4.2 Preliminaries

Throughout this chapter, we denote by $[n] := \{1, \ldots, n\}$. For a set A, we use A^C to denote its complement. Denote by $\Delta(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$. For functions f, g, we use $f \lesssim g$ (resp. $f \gtrsim g$) to denote that there exists some constant $c > 0$, such that $f \leq cg$ (resp. $f \geq cg$), and use $f \eqsim g$ to denote $f \lesssim g$ and $f \gtrsim g$ simultaneously. Define $a \vee b := \max(a, b)$, and $a \wedge b := \min(a, b)$. We use $\mathbb{E}$ to denote the expectation operator, and use var to denote the variance operator. Throughout, we use $\tilde{\mathcal{O}}(\cdot)$ notation to hide logarithmic factors.

Multi-task RL in episodic MDPs.

We have a set of M MDPs $\left\{\mathcal{M}_p = (H, \mathcal{S}, \mathcal{A}, p_0, \mathbb{P}_p, r_p)\right\}_{p=1}^{M}$, each associated with a player $p \in [M]$. Each MDP $\mathcal{M}_p$ is regarded as a task. The MDPs share the same episode length $H \in \mathbb{N}_+$, finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and initial state distribution $p_0 \in \Delta(\mathcal{S})$. The transition probabilities $\mathbb{P}_p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and reward distributions $r_p : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ of the players are not necessarily identical. We assume that the MDPs are layered[1], in that the state space $\mathcal{S}$ can be partitioned into disjoint

[1]This is a standard assumption [see, e.g., 168]. It is worth noting that any episodic MDP (with possibly nonstationary transition and reward) can be converted to a layered MDP with stationary transition and reward, with the state space size being H times the size of the original state space.

subsets $(\mathcal{S}_h)_{h=1}^H$, where p_0 is supported on $\mathcal{S}_1$, and for every $p \in [M]$, $h \in [H]$, and every $s \in \mathcal{S}_h, a \in \mathcal{A}$, $\mathbb{P}_p(\cdot \mid s, a)$ is supported on $\mathcal{S}_{h+1}$; here, we define $\mathcal{S}_{H+1} = \{\perp\}$ so that it contains a default terminal state $\perp$ (note that $\perp \notin \mathcal{S}$). We denote by $S := |\mathcal{S}|$ the size of the state space, and $A := |\mathcal{A}|$ the size of the action space.

Interaction process.

The interaction process between the players and the environment is as follows: at the beginning, both $(r_p)_{p=1}^M$ and $(\mathbb{P}_p)_{p=1}^M$ are unknown to the players. For each episode $k \in [K]$, each player $p \in [M]$ interacts with its respective MDP $\mathcal{M}_p$; specifically, player p starts with state $s_{1,p}^k \sim p_0$, and at every step $h \in [H]$, it chooses action $a_{h,p}^k$, transitions to next state $s_{h+1,p}^k \sim \mathbb{P}_p(\cdot \mid s_{h,p}^k, a_{h,p}^k)$ and receives a stochastic immediate reward $r_{h,p}^k \sim r_p(\cdot \mid s_{h,p}^k, a_{h,p}^k)$; after all players have finished their k-th episode, they can communicate and share information. The goal of the players is to maximize their expected collective reward $\mathbb{E}\left[\sum_{k=1}^K \sum_{p=1}^M \sum_{h=1}^H r_{h,p}^k\right]$.

Policy and value functions.

A deterministic policy π is a mapping from $\mathcal{S}$ to $\mathcal{A}$, which can be used by a player to make decisions in its respective MDP. For player p and step h, we define the value function $V_{h,p}^\pi : \mathcal{S}_h \to [0, H]$ and the action value function $Q_{h,p}^\pi : \mathcal{S}_h \times \mathcal{A} \to [0, H]$ as the expected return of player p conditioned on its being at a state at step h, and its being at a state and taking an action at step h, respectively. They satisfy the following recursive formula known as the Bellman equation:

$$\forall h \in [H]: \quad V_{h,p}^\pi(s) = Q_{h,p}^\pi(s, \pi(s)), \;\; Q_{h,p}^\pi(s, a) = R_p(s, a) + (\mathbb{P}_p V_{h+1,p}^\pi)(s, a),$$

where we use the convention that $V_{H+1,p}^\pi(\perp) = 0$, and for $f : \mathcal{S}_{h+1} \to \mathbb{R}$, $(\mathbb{P}_p f)(s, a) := \sum_{s' \in \mathcal{S}_{h+1}} \mathbb{P}_p(s' \mid s, a) f(s')$, and $R_p(s, a) = \mathbb{E}_{\hat{r} \sim r_p(\cdot \mid s, a)}[\hat{r}]$ is the expected immediate reward of player p. For player p and policy π, denote by $V_{0,p}^\pi = \mathbb{E}_{s_1 \sim p_0}\left[V_{1,p}^\pi(s_1)\right]$ its expected

reward.

For player p, we also define its optimal value function $V^\star_{h,p} : \mathcal{S}_h \to [0, H]$ and the optimal action value function $Q^\star_{h,p} : \mathcal{S}_h \times \mathcal{A} \to [0, H]$ using the Bellman optimality equation:

$$\forall h \in [H] : \quad V^\star_{h,p}(s) = \max_{a \in \mathcal{A}} Q^\star_{h,p}(s, a), \;\; Q^\star_{h,p}(s, a) = R_p(s, a) + (\mathbb{P}_p V^\star_{h+1,p})(s, a), \tag{4.1}$$

where we again use the convention that $V^\star_{H+1,p}(\perp) = 0$. For player p, denote by $V^\star_{0,p} = \mathbb{E}_{s_1 \sim p_0}\left[V^\star_{1,p}(s_1)\right]$ its optimal expected reward.

Given a policy π, as $V^\pi_{h,p}$ for different h's are only defined in the respective layer $\mathcal{S}_h$, we "collate" the value functions $(V^\pi_{h,p})_{h=1}^H$ and obtain a single value function $V^\pi_p : \mathcal{S} \to \mathbb{R}$. Formally, for every $h \in [H]$ and $s \in \mathcal{S}_h$,

$$V^\pi_p(s) := V^\pi_{h,p}(s).$$

We define $Q^\pi_p, V^\star_p, Q^\star_p$ similarly. For player p, given its optimal action value functions $Q^\star_p$, its optimal policy $\pi^\star_p : \mathcal{S} \to \mathcal{A}$ is greedy with respect to $Q^\star_p$, that is, $\pi^\star_p(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q^\star_p(s, a)$.

Suboptimality gap.

We define the suboptimality gap of state-action pair (s, a) for player p as $\mathrm{gap}_p(s, a) = V^\star_p(s) - Q^\star_p(s, a)$. We define the mininum suboptimality gap of player p as $\mathrm{gap}_{p,\min} = \min_{(s,a):\mathrm{gap}_p(s,a)>0} \mathrm{gap}_p(s, a)$, and the minimum suboptimailty gap over all players as $\mathrm{gap}_{\min} = \min_{p \in [M]} \mathrm{gap}_{p,\min}$. For player $p \in [M]$, define $Z_{p,\mathrm{opt}} := \left\{(s, a) : \mathrm{gap}_p(s, a) = 0\right\}$ as the set of optimal state-action pairs with respect to p.

Performance metric.

The performance metric of the players studied in this chapter is their collective regret, i.e., over a total of K episodes, how much extra reward they would have collected

in expectation if they were executing their respective optimal policies from the beginning. Formally, suppose for each episode k, player p executes policy $\pi^k(p)$, then the collective regret of the players is defined as:

$$\mathrm{Reg}(K) = \sum_{p=1}^{M} \sum_{k=1}^{K} \left(V_{0,p}^{\star} - V_{0,p}^{\pi^k(p)} \right).$$

Baseline: individual STRONG-EULER.

A naive baseline for multi-player RL is to let each player run a separate RL algorithm without communication. For concreteness, we choose to let each player run the state of the art STRONG-EULER algorithm [136] (see also its precursor EULER [172]), which enjoys minimax gap-independent [12, 39] and gap-dependent regret guarantees, and refer to this strategy as individual STRONG-EULER. Specifically, as it is known that STRONG-EULER has a regret of $\tilde{\mathcal{O}}(\sqrt{H^2SAK})$, individual STRONG-EULER has a collective regret of $\tilde{\mathcal{O}}(M\sqrt{H^2SAK})$. In addition, by summing up the gap-dependent regret guarantee of STRONG-EULER for the M MDPs altogether, it can be easily checked that with probability $1-\delta$, individual STRONG-EULER has a collective regret of

$$\mathrm{Reg}(K) \lesssim \ln\left(\frac{MSAK}{\delta}\right) \left(\sum_{p\in[M]} \left(\sum_{(s,a)\in Z_{p,\mathrm{opt}}} \frac{H^3}{\mathrm{gap}_{p,\min}} + \sum_{(s,a)\in Z_{p,\mathrm{opt}}^C} \frac{H^3}{\mathrm{gap}_p(s,a)} \right) + MH^3S^2A \ln\frac{MH}{\mathrm{gap}_{\min}} \right).$$

Our goal is to design multi-task RL algorithms that can achieve collective regret strictly lower than this baseline in both gap-dependent and gap-independent fashions when the tasks are similar.

Notion of similarity.

Throughout this chapter, we will consider the following notion of similarity between MDPs in the multi-task episodic RL setting.

Definition 4.1. *A collection of MDPs $(\mathcal{M}_p)_{p=1}^M$ is said to be ϵ-dissimilar, if for all $p, q \in [M]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\big|R_p(s,a) - R_q(s,a)\big| \le \epsilon,\ \|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a)\|_1 \le \frac{\epsilon}{H}.$$

If this happens, we call $(\mathcal{M}_p)_{p=1}^M$ an ϵ-Multi-Player Episodic Reinforcement Learning (abbrev. ϵ-MPERL) problem instance.

If the MDPs in $(\mathcal{M}_p)_{p=1}^M$ are 0-dissimilar, then they are identical by definition, and our interaction protocol degenerates to the concurrent RL protocol [134]. Our dissimilarity notion is complementary to those of [24, 60], in that they require the MDPs to be either identical, or have well-separated parameters for at least one state-action pair; in contrast, our dissimilarity notion allows the MDPs to be nonidentical and arbitrarily close.

We have the following intuitive lemma that shows the closeness of optimal value functions of different MDPs, in terms of the dissimilarity parameter ϵ:

Lemma 4.2. *If $(\mathcal{M}_p)_{p=1}^M$ are ϵ-dissimilar, then for every $p, q \in [M]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\Big|Q_p^\star(s,a) - Q_q^\star(s,a)\Big| \le 2H\epsilon$; consequently, $\Big|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\Big| \le 4H\epsilon$.*

4.3 Algorithm: Multi-task-Euler

We now describe our main algorithm, Multi-task-Euler (Algorithm 3). Our model-based algorithm is built upon recent works on episodic RL that provide algorithms with sharp instance-dependent guarantees in the single task setting [172, 136]. In a nutshell, for each episode k and each player p, the algorithm performs optimistic value iteration to construct high-probability upper and lower bounds for the optimal value and action value functions $V_p^\star$ and $Q_p^\star$, and uses them to guide its exploration and decision making process.

Algorithm 3: MULTI-TASK-EULER

Input : Failure probability $\delta \in (0,1)$.

Initialize: Set $V_p(\perp) = 0$ for all p in $[M]$, where $\perp$ is the only state in $\mathcal{S}_{H+1}$;

for $k = 1, 2, \ldots, K$ **do**

  for $p = 1, 2, \ldots, M$ **do**

    // Construct optimal value estimates for player p

    for $h = H, H-1, \ldots, 1$ **do**

      for $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ **do**

        Compute:

        $\overline{\text{ind-}Q}_p(s,a) = \hat{R}_p(s,a) + (\hat{\mathbb{P}}_p \overline{V}_p)(s,a) + \text{ind-}b_p(s,a)$;

        $\underline{\text{ind-}Q}_p(s,a) = \hat{R}_p(s,a) + (\hat{\mathbb{P}}_p \underline{V}_p)(s,a) - \text{ind-}b_p(s,a)$;

        $\overline{\text{agg-}Q}_p(s,a) = \hat{R}(s,a) + (\hat{\mathbb{P}} \overline{V}_p)(s,a) + \text{agg-}b_p(s,a)$;

        $\underline{\text{agg-}Q}_p(s,a) = \hat{R}(s,a) + (\hat{\mathbb{P}} \underline{V}_p)(s,a) - \text{agg-}b_p(s,a)$;

        Update optimal action value function upper and lower bound estimates:

        $\overline{Q}_p(s,a) = \min\left\{H - h + 1, \overline{\text{ind-}Q}_p(s,a), \underline{\text{ind-}Q}_p(s,a)\right\}$;

        $\underline{Q}_p(s,a) = \max\left\{0, \underline{\text{ind-}Q}_p(s,a), \underline{\text{agg-}Q}_p(s,a)\right\}$;

      for $s \in \mathcal{S}_h$ **do**

        Define $\pi^k(p)(s) = \text{argmax}_{a \in \mathcal{A}} \overline{Q}_p(s,a)$;

        Update $\overline{V}_p(s) = \overline{Q}_p(s, \pi^k(p)(s))$, $\underline{V}_p(s) = \underline{Q}_p(s, \pi^k(p)(s))$.

    // All players p interact with their respective environments, and update reward and transition estimates

    for $p = 1, 2, \ldots, M$ **do**

      Player p executes policy $\pi^k(p)$ on $\mathcal{M}_p$ and obtains trajectory $(s^k_{h,p}, a^k_{h,p}, r^k_{h,p})_{h=1}^H$.

      Update individual estimates of transition probability $\hat{\mathbb{P}}_p$, reward $\hat{R}_p$ and count $n_p(\cdot,\cdot)$.

    Update aggregate estimates of transition probability $\hat{\mathbb{P}}$, reward $\hat{R}$ and count $n(\cdot,\cdot)$.

Empirical estimates of model parameters.

For each player p, the construction of its value function bound estimates relies on empirical estimates on its transition probability and expected reward function. For both estimands, we use two estimators with complementary roles, which are at two different points of the bias-variance tradeoff spectrum: one estimator uses only the player's own data (termed *individual estimate*), which is unbiased but has large variance, the other estimator uses the data collected by all players (termed *aggregate estimate*), which is biased but has lower variance. Specifically, at the end of episode k, for every $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, the algorithm maintains its empirical count of encountering (s,a) for each player p, along with its total empirical count across all players, respectively:

$$n_p(s,a) := \sum_{l=1}^{k} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right), \; n(s,a) := \sum_{l=1}^{k}\sum_{p=1}^{M} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right). \quad (4.2)$$

The individual and aggregate estimates of immediate reward $R(s,a)$ are defined as:

$$\begin{aligned} \hat{R}_p(s,a) &:= \frac{\sum_{l=1}^{k} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right) r_{h,p}^l}{n_p(s,a)}, \\ \hat{R}(s,a) &:= \frac{\sum_{l=1}^{k}\sum_{p=1}^{M} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right) r_{h,p}^l}{n(s,a)}. \end{aligned} \quad (4.3)$$

Similarly, for every $h \in [H]$ and $(s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$, we also define the individual and aggregate estimates of transition probability as:

$$\begin{aligned} \hat{\mathbb{P}}_p(s' \mid s,a) &:= \frac{\sum_{l=1}^{k} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l, s_{h+1,p}^l) = (s,a,s')\right)}{n_p(s,a)}, \\ \hat{\mathbb{P}}(s' \mid s,a) &:= \frac{\sum_{l=1}^{k}\sum_{p=1}^{M} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l, s_{h+1,p}^l) = (s,a,s')\right)}{n(s,a)}. \end{aligned} \quad (4.4)$$

If $n(s,a) = 0$, we define $\hat{R}(s,a) := 0$ and $\hat{\mathbb{P}}(s' \mid s,a) := \frac{1}{|\mathcal{S}_{h+1}|}$; and if $n_p(s,a) = 0$, we define $\hat{R}_p(s,a) := 0$ and $\hat{\mathbb{P}}_p(s' \mid s,a) := \frac{1}{|\mathcal{S}_{h+1}|}$. The counts and reward estimates can be maintained by MULTI-TASK-EULER efficiently in an incremental manner.

Constructing value function estimates via optimistic value iteration.

For each player p, based on these model parameter estimates, MULTI-TASK-EULER performs optimistic value iteration to compute the value function estimates for states at all layers (lines 3 to 15). For the terminal layer $H+1$, $V^\star_{H+1}(\perp) = 0$ trivially, so nothing needs to be done. For earlier layers $h \in [H]$, MULTI-TASK-EULER iteratively builds its value function estimates in a backward fashion. At the time of estimating values for layer h, the algorithm has already obtained optimal value estimates for layer $h+1$. Based on the Bellman optimality equation (4.1), MULTI-TASK-EULER estimates $(Q^\star_p(s,a))_{s\in\mathcal{S}_h, a\in\mathcal{A}}$ using model parameter estimates and its estimates of $(V^\star_p(s))_{s\in\mathcal{S}_{h+1}}$, i.e., $(\overline{V}_p(s))_{s\in\mathcal{S}_{h+1}}$ and $(\underline{V}_p(s))_{s\in\mathcal{S}_{h+1}}$ (lines 5 to 12).

Specifically, MULTI-TASK-EULER constructs estimates of $(Q^\star_p(s,a))$ for all $s \in \mathcal{S}_h, a \in \mathcal{A}$ in two different ways. First, it uses the individual estimates of model of player p to construct $\underline{\text{ind-}Q}_p$ and $\overline{\text{ind-}Q}_p$, upper and lower bound estimates of $Q^\star_p$ (lines 8 and 9); this construction is reminiscent of EULER and STRONG-EULER [172, 136], in that if we were only to use $\underline{\text{ind-}Q}_p$ and $\overline{\text{ind-}Q}_p$ as our optimal action value function estimate $\overline{Q}_p$ and $\underline{Q}_p$, our algorithm becomes individual STRONG-EULER. The individual value function estimates are crucial to establishing MULTI-TASK-EULER's fall-back guarantees, ensuring that it never performs worse than the individual STRONG-EULER baseline. Second, it uses the aggregate estimate of model to construct $\underline{\text{agg-}Q}_p$ and $\overline{\text{agg-}Q}_p$, also upper and lower bound estimates of $Q^\star_p$ (lines 6 and 7); this construction is unique to the multitask learning setting, and is our new algorithmic contribution.

To ensure that $\overline{\text{agg-}Q}_p$ and $\overline{\text{ind-}Q}_p$ (resp. $\underline{\text{agg-}Q}_p$ and $\underline{\text{ind-}Q}_p$) are valid upper bounds (resp. lower bounds) of $Q^\star_p$, MULTI-TASK-EULER adds bonus terms $\text{ind-}b_p(s,a)$

and agg-$b_p(s,a)$, respectively, in the optimistic value iteration process, to account for estimation error of the model estimates against the true models. Specifically, both bonus terms comprise three parts:

$$\begin{aligned}
\text{ind-}b_p(s,a) &:= b_{\text{rw}}\left(n_p(s,a),0\right) + b_{\text{prob}}\left(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0\right) \\
&\qquad + b_{\text{str}}\left(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0\right), \\
\text{agg-}b_p(s,a) &:= b_{\text{rw}}\left(n(s,a),\epsilon\right) + b_{\text{prob}}\left(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon\right) \\
&\qquad + b_{\text{str}}\left(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon\right),
\end{aligned}$$

where

$$\begin{aligned}
b_{\text{rw}}(n,\kappa) &:= 1 \ \wedge \ \kappa + \Theta\left(\sqrt{\frac{L(n)}{n}}\right), \\
b_{\text{prob}}\left(q,n,\overline{V},\underline{V},\kappa\right) &:= H \ \wedge \ 2\kappa + \\
&\qquad \Theta\left(\sqrt{\frac{\text{var}_{s'\sim q}\left[\overline{V}(s')\right] L(n)}{n}} + \sqrt{\frac{\mathbb{E}_{s'\sim q}\left[(\overline{V}(s') - \underline{V}(s'))^2\right] L(n)}{n}} + \frac{HL(n)}{n}\right), \\
b_{\text{str}}\left(q,n,\overline{V},\underline{V},\kappa\right) &:= \kappa + \Theta\left(\sqrt{\frac{S\,\mathbb{E}_{s'\sim q}\left[(\overline{V}(s') - \underline{V}(s'))^2\right] L(n)}{n}} + \frac{HSL(n)}{n}\right),
\end{aligned}$$

and $L(n) \eqsim \ln(\frac{MSAn}{\delta})$.

The three components in the bonus terms serve for different purposes:

1. The first component accounts for the uncertainty in the reward estimation: with probability $1 - \mathcal{O}(\delta)$, $\left|\hat{R}_p(s,a) - R_p(s,a)\right| \le b_{\text{rw}}(n_p(s,a),0)$, and $\left|\hat{R}(s,a) - R_p(s,a)\right| \le b_{\text{rw}}(n(s,a),\epsilon)$.

2. The second component accounts for the uncertainty in estimating $(\mathbb{P}_p V_p^\star)(s,a)$: with

probability $1-\mathcal{O}(\delta)$, $\left|(\hat{\mathbb{P}}_p V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| \le b_{\text{prob}}(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0)$ and $\left|(\hat{\mathbb{P}} V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| \le b_{\text{prob}}(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon)$.

3. The third component accounts for the lower order terms to ensure strong optimism [136]: with probability $1-\mathcal{O}(\delta)$,
$\left|(\hat{\mathbb{P}}_p - \mathbb{P}_p)(\overline{V}_p - V_p^\star)(s,a)\right| \le b_{\text{str}}(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0)$, and
$\left|(\hat{\mathbb{P}} - \mathbb{P}_p)(\overline{V}_p - V_p^\star)(s,a)\right| \le b_{\text{prob}}(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon)$.

Based on the above concentration inequalities and the definitions of bonus terms, it can be shown inductively that, with probability $1-\mathcal{O}(\delta)$, both $\overline{\text{agg-}Q}_p$ and $\overline{\text{ind-}Q}_p$ (resp. $\underline{\text{agg-}Q}_p$ and $\underline{\text{ind-}Q}_p$) are valid upper bounds (resp. lower bounds) of $Q_p^\star$.

Finally, observe that for any $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, $Q_p^\star(s,a)$ has range $[0, H-h+1]$. By taking intersections of all confidence bounds of $Q_p^\star$ it has obtained, Multi-task-Euler constructs its final upper and lower bound estimates for $Q_p^\star(s,a)$, $\overline{Q}_p(s,a)$ and $\underline{Q}_p(s,a)$ respectively, for $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ (line 11 to 12). Similar ideas on using data from multiple sources to construct confidence intervals and guide explorations was proposed by Soare et al. [138] for multi-task linear contextual bandits. Using the relationship between the optimal value $V_p^\star(s)$ and and optimal action values $\left\{Q_p^\star(s,a) : a \in \mathcal{A}\right\}$, Multi-task-Euler also constructs upper and lower bound estimates for $V_p^\star(s)$, $\overline{V}_p(s)$ and $\underline{V}_p(s)$, respectively for $s \in \mathcal{S}_h$ (line 15).

Executing optimistic policies.

At each episode k, for each player p, its optimal action-value function upper bound estimate $\overline{Q}_p$ induces a greedy policy $\pi^k(p) : s \mapsto \text{argmax}_{a \in \mathcal{A}} \overline{Q}_p(s,a)$ (line 14); the player then executes this policy at this episode to collect a new trajectory and use this to update its individual model parameter estimates. After all players finish their episode k, the algorithm also updates its aggregate model parameter estimates (lines 16 to 19) using Equations (4.2), (4.3) and (4.4), and continues to the next episode.

4.4 Performance Guarantees

Before stating the guarantees of Algorithm 3, we define an instance-dependent complexity measure that characterizes the amenability of information sharing.

Definition 4.3. *The set of subpar state-action pairs is defined as:*

$$\mathcal{I}_\epsilon := \left\{(s,a) \in \mathcal{S} \times \mathcal{A} : \exists p \in [M], \text{gap}_p(s,a) \geq 96H\epsilon\right\},$$

where we recall that $\text{gap}_p(s,a) = V_p^\star(s) - Q_p^\star(s,a)$.

Definition 4.3 generalizes the notion of subpar arms defined for multi-task multi-armed bandit learning (chapter 2) in two ways: first, it is with regards to state-action pairs as opposed to actions only; second, in RL, suboptimality gaps depend on both immediate reward and subsequent long-term return.

To ease our later presentation, we also present the following lemma.

Lemma 4.4. *For any* $(s,a) \in \mathcal{I}_\epsilon$*, we have that: (1) for all* $p \in [M]$*,* $(s,a) \notin Z_{p,\text{opt}}$*, where we recall that* $Z_{p,\text{opt}} = \left\{(s,a) : \text{gap}_p(s,a) = 0\right\}$ *is the set of optimal state-action pairs with respect to* p*; (2) for all* $p, q \in [M]$*,* $\text{gap}_p(s,a) \geq \frac{1}{2}\text{gap}_q(s,a)$.

The lemma follows directly from Lemma 4.2; its proof can be found in the Appendix along with proofs of the following theorems. Item 1 implies that any subpar state action pair is suboptimal for all players. In other words, for every player p, the state-action space $\mathcal{S} \times \mathcal{A}$ can be partitioned to three disjoint sets: $\mathcal{I}_\epsilon, Z_{p,\text{opt}}, (\mathcal{I}_\epsilon \cup Z_{p,\text{opt}})^C$. Item 2 implies that for any subpar (s,a), its suboptimal gaps with respect to all players are within a constant of each other.

4.4.1 Upper bounds

Equipped with the above definitions, we are now ready to present the performance guarantees of Algorithm 3. We first present a gap-independent collective regret bound.

Theorem 4.5 (Gap-independent upper bound)**.** *If* $\left\{\mathcal{M}_p\right\}_{p=1}^M$ *are* ϵ*-dissimilar, then running* Multi-task-Euler*, we have with probability* $1-\delta$*,*

$$\mathrm{Reg}(K) \leq \tilde{\mathcal{O}}\left(M\sqrt{H^2|\mathcal{I}_\epsilon^C|K} + \sqrt{MH^2|\mathcal{I}_\epsilon|K} + MH^3S^2A\right).$$

We again compare this regret upper bound with individual Strong-Euler's gap independent regret bound. Recall that individual Strong-Euler guarantees that with probability $1-\delta$,

$$\mathrm{Reg}(K) \leq \tilde{\mathcal{O}}\left(M\sqrt{H^2SAK} + MH^3S^2A\right).$$

We focus on the comparison on the leading terms, i.e., the $\sqrt{K}$ terms. As $M\sqrt{H^2SAK} \approx M\sqrt{H^2|\mathcal{I}_\epsilon|\,K} + M\sqrt{H^2|\mathcal{I}_\epsilon^C|\,K}$, we see that an improvement in the collective regret bound comes from the contributions from the subpar state-action pairs: the $M\sqrt{H^2|\mathcal{I}_\epsilon|K}$ term is reduced to $\sqrt{MH^2|\mathcal{I}_\epsilon|K}$, a factor of $\tilde{\mathcal{O}}(\sqrt{\frac{1}{M}})$ improvement. Moreover, if $\left|\mathcal{I}_\epsilon^C\right| \ll SA$ and $M \gg 1$, Multi-task-Euler provides a regret bound of lower order than individual Strong-Euler.

We next present a gap-dependent upper bound on its collective regret.

Theorem 4.6 (Gap-dependent upper bound)**.** *If* $\left\{\mathcal{M}_p\right\}_{p=1}^M$ *are* ϵ*-dissimilar, then running* Multi-task-Euler*, we have with probability* $1-\delta$*,*

$$\mathrm{Reg}(K) \lesssim \ln(\frac{MSAK}{\delta})\left(\sum_{p\in[M]}\left(\sum_{(s,a)\in Z_{p,\mathrm{opt}}}\frac{H^3}{\mathrm{gap}_{p,\min}} + \sum_{(s,a)\in(\mathcal{I}_\epsilon\cup Z_{p,\mathrm{opt}})^C}\frac{H^3}{\mathrm{gap}_p(s,a)}\right) + \sum_{(s,a)\in\mathcal{I}_\epsilon}\frac{H^3}{\min_p \mathrm{gap}_p(s,a)}\right) + \ln(\frac{MSAK}{\delta})\cdot MH^3S^2A\ln\frac{MHSA}{\mathrm{gap}_{\min}},$$

where we recall that $\mathrm{gap}_{p,\min} = \min_{(s,a):\mathrm{gap}_p(s,a)>0}\mathrm{gap}_p(s,a)$*, and* $\mathrm{gap}_{\min} = \min_p \mathrm{gap}_{p,\min}$*.*

Comparing this regret bound with the regret bound obtained by the individ-

ual STRONG-EULER baseline, recall that by summing over the regret guarantees of STRONG-EULER for all players $p \in [M]$, and taking a union bound over all p, individual STRONG-EULER guarantees a collective regret bound of

$$\mathrm{Reg}(K) \lesssim \ln(\frac{MSAK}{\delta}) \left(\sum_{p\in[M]} \left(\sum_{(s,a)\in Z_{p,\mathrm{opt}}} \frac{H^3}{\mathrm{gap}_{p,\min}} + \sum_{(s,a)\in(\mathcal{I}_\epsilon \cup Z_{p,\mathrm{opt}})^C} \frac{H^3}{\mathrm{gap}_p(s,a)} \right) + \sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{p\in[M]} \frac{H^3}{\mathrm{gap}_p(s,a)} \right) + \ln(\frac{MSAK}{\delta}) \cdot MH^3S^2A \ln \frac{MHSA}{\mathrm{gap}_{\min}},$$

that holds with probability $1-\delta$. We again focus on comparing the leading terms, i.e., the terms that have polynomial dependences on the suboptimality gaps in the above two bounds. It can be seen that an improvement in the regret bound by MULTI-TASK-EULER comes from the contributions from the subpar state-action pairs: for each $(s,a) \in \mathcal{I}_\epsilon$, the regret bound is reduced from $\sum_{p\in[M]} \frac{H^3}{\mathrm{gap}_p(s,a)}$ to $\frac{H^3}{\min_p \mathrm{gap}_p(s,a)}$, a factor of $\mathcal{O}(\frac{1}{M})$ improvement. Recent work of Xu et al. [168] has shown that in the single-task setting, it is possible to replace $\sum_{(s,a)\in Z_{p,\mathrm{opt}}} \frac{H^3}{\mathrm{gap}_{p,\min}}$ with a sharper problem-dependent complexity term that depends on the multiplicity of optimal state-action pairs. We leave improving the guarantee of Theorem 4.6 in a similar manner as an interesting open problem.

4.4.2 Lower bounds

To complement the above upper bounds, we now present gap-dependent and gap-independent regret lower bounds that also depends on our subpar state-action pair notion. Our lower bounds are inspired by regret bounds for episodic RL [136, 39] and multi-task bandits (Chapter 2).

Theorem 4.7 (Gap-independent lower bound)**.** *For any $A \geq 2$, $H \geq 2$, $S \geq 4H$, $K \geq SA$, $M \in \mathbb{N}$, and $l, l^C \in \mathbb{N}$ such that $l + l^C = SA$ and $l \leq SA - 4(S + HA)$, there exists some ϵ*

such that for any algorithm Alg*, there exists an ϵ-MPERL problem instance with S states, A actions, M players and an episode length of H such that $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \geq l$, and*

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(M\sqrt{H^2 l^C K} + \sqrt{MH^2 lK}\right).$$

We also present a gap-dependent lower bound. Before that, we first formally define the notion of sublinear regret algorithms: for any fixed ϵ, we say that an algorithm Alg is a sublinear regret algorithm for the ϵ-MPERL problem if there exists some $C > 0$ and $\alpha < 1$ such that $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \leq CK^{\alpha}$.

Theorem 4.8 (Gap-dependent lower bound)**.** *Fix $\epsilon \geq 0$. For any $S \in \mathbb{N}$, $A \geq 2$, $H \geq 2$, $M \in \mathbb{N}$, such that $S \geq 2(H-1)$, let $S_1 = S - 2(H-1)$; and let $\left\{\Delta_{s,a,p}\right\}_{(s,a,p)\in[S_1]\times[A]\times[M]}$ be any set of values such that (1) each $\Delta_{s,a,p} \in [0, H/48]$, (2) for every $(s,p) \in [S_1]\times[M]$, there exists at least one action $a \in [A]$ such that $\Delta_{s,a,p} = 0$, and (3) for every $(s,a) \in [S_1]\times[A]$ and $p, q \in [M]$, $\left|\Delta_{s,a,p} - \Delta_{s,a,q}\right| \leq \epsilon/4$. There exists an ϵ-MPERL problem instance with S states, A actions, M players and an episode length of H, such that $\mathcal{S}_1 = [S_1]$, $|\mathcal{S}_h| = 2$ for all $h \geq 2$, and*

$$\mathrm{gap}_p(s,a) = \Delta_{s,a,p}, \quad \forall (s,a,p) \in [S_1]\times[A]\times[M];$$

for this problem instance, any sublinear regret algorithm Alg *for the ϵ-MPERL problem must have regret at least*

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(\ln K\left(\sum_{p\in[M]}\sum_{\substack{(s,a)\in\mathcal{I}^C_{(\epsilon/192H)}:\\ \mathrm{gap}_p(s,a)>0}}\frac{H^2}{\mathrm{gap}_p(s,a)} + \sum_{(s,a)\in\mathcal{I}_{(\epsilon/192H)}}\frac{H^2}{\min_p \mathrm{gap}_p(s,a)}\right)\right).$$

Comparing the lower bounds with Multi-task-Euler's regret upper bounds in

Theorems 4.5 and 4.6, we can see that the upper and lower bounds nearly match for any constant H. When H is large, the key difference between the upper and lower bounds is that the former are in terms of $\mathcal{I}_\epsilon$, while the latter are in terms of $\mathcal{I}_{\Theta(\frac{\epsilon}{H})}$. We conjecture that our upper bounds can be improved by replacing $\mathcal{I}_\epsilon$ with $\mathcal{I}_{\Theta(\frac{\epsilon}{H})}$—our analysis uses a clipping trick similar to [136], which may be the reason for a suboptimal dependence on H. We leave closing this gap as an open question.

4.5 Related Work

Regret minimization for MDPs.

Our work belongs to the literature of regret minimization for MDPs [e.g., 15, 68, 39, 12, 40, 71, 41, 172, 136, 179, 170, 168]. In the episodic setting, [12, 41, 172, 136, 179] achieve minimax $\sqrt{H^2SAK}$ regret bounds for general stationary MDPs. Furthermore, the Euler algorithm [172] achieves adaptive problem-dependent regret guarantees when the total reward within an episode is small or when the environmental norm of the MDP is small. Simchowitz and Jamieson [136] refine Euler, proposing Strong-Euler that provides more fine-grained gap-dependent $\mathcal{O}(\log K)$ regret guarantees. Yang et al. [170], Xu et al. [168] show that the optimistic Q-learning algorithm [71] and its variants can also achieve gap-dependent logarithmic regret guarantees. Remarkably, Xu et al. [168] achieve a regret bound that improves over that of [136], in that it replaces the dependence on the number of optimal state-action pairs with the number of non-unique state-action pairs.

Transfer and lifelong learning for RL.

A considerable portion of related works concerns transfer learning for RL tasks [see 145, 91, 181, for surveys from different angles], and many studies investigate a batch setting: given some source tasks and target tasks, transfer learning agents have access to batch data collected for the source tasks (and sometimes for the target tasks as well).

In this setting, model-based approaches have been explored in [e.g., 146]; theoretical guarantees for transfer of samples across tasks have been established in [e.g., 92, 149]. Similarly, sequential transfer has been studied under the framework of lifelong RL in [e.g., 143, 1, 56, 89]—in this setting, an agent faces a sequence of RL tasks and aims to take advantage of knowledge gained from previous tasks for better performance in future tasks; in particular, analyses on the sample complexity of transfer learning algorithms are presented in [24, 100] under the assumption that an upper bound on the total number of unique (and well-separated) RL tasks is known. We note that, in contrast, we study an online setting in which no prior data are available and multiple RL tasks are learned concurrently by RL agents.

Concurrent RL.

Data sharing between multiple RL agents that learn concurrently has also been investigated. In [e.g., 80, 135, 60, 44], a group of agents interact in parallel with *identical* environments. Another setting is studied in [60], in which agents solve different RL tasks (MDPs); however, similar to [24, 100], it is assumed that there is a finite number of unique tasks, and different tasks are well-separated, i.e., there is a minimum gap. In this work, we assume that players face similar but not necessarily identical MDPs, and we do not assume a minimum gap. Hu et al. [65] study multi-task RL with linear function approximation with representation transfer, where it is assumed that the optimal value functions of all tasks are from a low dimensional linear subspace. Our setting and results are the most similar to [117] and [48]. Pazis and Parr [117] study concurrent exploration in similar MDPs with continuous states in the PAC setting; however, their PAC guarantee does not hold for target error rate arbitrarily close to zero; in contrast, our algorithm has a fall-back guarantee, in that it always has a sublinear regret. Concurrent RL from similar *linear* MDPs has also been recently studied in [48]: under the assumption of small heterogeneity between different MDPs (a setting very similar to ours), the provided regret

guarantee involves a term that is linear in the number of episodes, whereas our algorithm in this chapter always has a sublinear regret; concurrent RL under the assumption of large heterogeneity is also studied in that work, but additional contextual information is assumed to be available for the players to ensure a sublinear regret.

Other related topics and models.

In many multi-agent RL models [177, 113], a set of learning agents interact with a common environment and have shared global states; in particular, Zhang et al. [176] study the setting with heterogeneous reward distributions, and provides convergence guarantees for two policy gradient-based algorithms. In contrast, in our setting, our learning agents interact with separate environments. Multi-agent bandits with similar, heterogeneous reward distributions are investigated in Chapter 2; herein, we generalize the multi-armed bandit setting to tabular, episodic MDPs.

4.6 Conclusion and Future Work

In this chapter, we generalize the multi-task bandit learning framework in Chapter 2 and formulate a multi-task concurrent RL problem, in which tasks are similar but not necessarily identical. We provide a provably efficient model-based algorithm that takes advantage of knowledge transfer between different tasks. Our instance-dependent regret upper and lower bounds formalize the intuition that subpar state-action pairs are amenable of information sharing among tasks.

There still remain gaps between our upper and lower bounds which can be closed by either a finer analysis or a better algorithm: first, the dependence on $\mathcal{I}_\epsilon$ in the upper bound does not match the dependence of $\mathcal{I}_{\Theta(\epsilon/H)}$ in the lower bound when H is large; second, the gap-dependent upper bound has $\mathcal{O}(H^3)$ dependence, whereas the gap-dependent lower bound only has $\Omega(H^2)$ dependence; third, the additive dependence on the number of optimal state-action pairs can potentially be removed by new algorithmic ideas [168].

Another interesting future direction is to consider more general parameter transfer for online RL, for example, in the context of function approximation.

Acknowledgement.

Chapter 4 is based on the material as it appears in "Provably Efficient Multi-Task Reinforcement Learning with Model Transfer" by Chicheng Zhang and Zhi Wang [173]. The material was published in *Advances in Neural Information Processing Systems 34*. The dissertation author was a co-author of the paper.

Chapter 5

Metric Learning from Crowdsourced Preference Comparisons

5.1 Introduction

Metric learning is commonly used to discover measures of similarity for downstream applications [e.g., 84]. In this chapter, we study metric learning from pairwise preference comparisons. In particular, we consider the ideal point model [38], in which a set of items are embedded into $\mathbb{R}^d$, and a user prefers an item x over another x' if it is *closer* to the user's latent ideal point $u \in \mathbb{R}^d$, that is,

$$\rho(x, u) < \rho(x', u),$$

for some underlying metric $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. While high-quality item embeddings have become increasingly available, for example from foundation models pre-trained on internet-scale data [e.g., 121], naively equipping these representations with the Euclidean distance may not accurately capture the semantic relations between items as perceived by humans, and therefore may not align with human values or preferences [171, 28]. Meanwhile, people often agree on their perception of item (dis)similarities [37]. In this chapter, we study when and how a shared Mahalanobis distance can be learned from a large crowd, with each user answering a few queries in the form of "Do you prefer x or x'?"

The line of work on simultaneous metric and preference learning was recently introduced by [167], who studied it under the ideal point model for a single user. They proposed an alternating minimization algorithm to recover both the Mahalanobis distance and user ideal point. After, [28] introduced a convex formulation of the problem, providing the first theoretical guarantees while extending the results to crowdsourced data. They showed that the cost of learning a Mahalanobis distance can be amortized among users; it is possible to jointly learn the metric and ideal points in $\mathbb{R}^d$ so long as sufficiently many users each provides $\Theta(d)$ preference comparisons.

However, when the representations of data are very high-dimensional, obtaining $\Omega(d)$ preference comparisons from each user can be practically infeasible. It can be expensive to ask a user more than a few queries [36] both in terms of cost and cognitive overload, and users may have concerns over their privacy [70]. Fortunately, through crowdsourcing, we often have access to preference comparisons from a *large* pool of users. In this chapter, we ask the fundamental question:

> Can we learn an *unknown* Mahalanobis distance metric in $\mathbb{R}^d$ from $o(d)$ preference comparisons per user?

We provide a twofold answer to this question. First, we show a negative result: even with infinitely many users, it is generally impossible to learn anything at all about the underlying metric when each user provides fewer than d preference comparisons. In general, there is no hope for recovering the unknown metric from preference comparisons without learning individual preference points as well.

Second, we show that the negative result does not rule out the possibility of learning the metric when the set of items are *subspace-clusterable* (Definition 5.12); that is, when they lie in a union of low-dimensional subspaces [116, 103, 50]. These subspaces may capture, for instance, different categories or classes of items; such structure has also

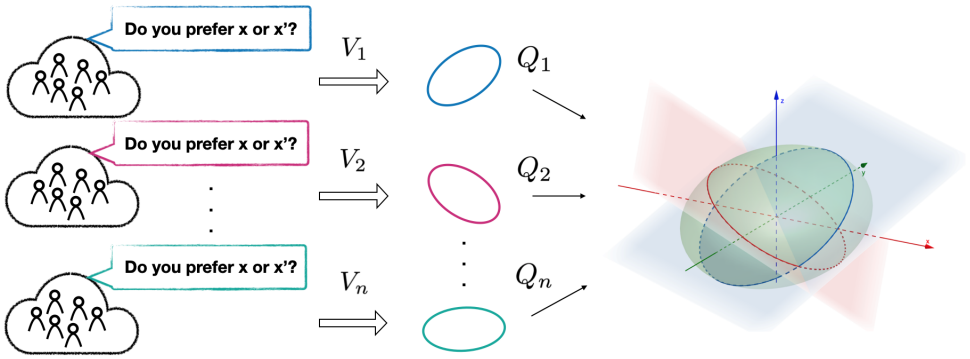


Figure 5.1. In our divide-and-conquer approach, users help us recover the metric Q_λ restricted to subspaces V_λ. We stitch these together to recover the metric M on $\mathbb{R}^d$. The ellipses visualize the low-dimensional unit spheres, which are 'slices' of the full metric.

been studied extensively in compressed sensing [102, 49] and computer vision [62], among others. Given items with subspace-clusterable structure, we show that we can learn the Mahalanobis distance using a *divide-and-conquer* approach (Figure 5.1). This involves learning the metric restricted to each subspace, which is feasible using very few comparisons per user, and then reconstructing the full metric from these subspace metrics.

Contributions.

We study the fundamental problem of learning an unknown metric with limited pairwise comparison queries, i.e, whether it is possible to learn a shared unknown metric without learning the individual preference points. Our main contributions are as follows:

1. We provide an impossibility result: nothing can be learned if the items are in general position (Section 5.3);

2. We define the notion of subspace-clusterable items and propose a divide-and-conquer approach, such that:

   - Given noiseless, unquantized comparisons that indicate how much a user prefers one item over another, we show that subspace-clusterability is necessary and sufficient for identifying the unknown metric (Section 5.4);
   - Given noisy, quantized comparisons in the form of binary responses over subspace-

clusterable items, we present recovery guarantees in terms of identification errors for our approach (Section 5.5);

3. We implement our proposed algorithm and validate our findings using synthetic data (Section 5.6).

Related work.

Metric learning from triplet comparisons or ordinal constraints has been studied extensively [84]. A line of metric learning from human feedback focuses on learning Mahalanobis distances from triplet comparisons [128, 152, 106], in which users are asked "is u closer to x or x'?" However, triplet comparisons are a specific type of feedback that is not always practical to obtain. And so, an important extension of these works is metric learning from preference comparisons, which can be seen as a variant of triplet comparisons with an unknown latent comparator u. Even though preference comparisons are a weaker form of feedback, they are also much more prevalent. For example, they can be inferred from user behavior, assuming users tend to engage more with items perceived to be more ideal. As we build directly on this line of work by [167] and [28], we now present background and existing results in greater detail. See Appendix D.1 for further discussion of related work.

5.2 Preliminaries

The ideal point model.

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$ with an unknown Mahalanobis distance ρ. Let M be its matrix representation in $\mathbb{R}^{d\times d}$. That is, M is a positive-definite (symmetric) matrix and for all $x, x' \in \mathbb{R}^d$,

$$\rho(x, x') := \sqrt{(x - x')^\top M(x - x')} = \|x - x'\|_M.$$

Suppose there is a large pool of users, and each user is associated with an unknown ideal point in $\mathbb{R}^d$. A user with ideal point u prefers an item x over another x' if and only if $\rho(x,u) < \rho(x',u)$; or whenever $\psi(x,x';u) < 0$, where:

$$\psi_M(x,x';u) := \|x-u\|_M^2 - \|x'-u\|_M^2. \tag{5.1}$$

Each user's ideal point may be distinct, but we assume that the metric ρ is a shared. We aim to recover ρ when each user provides very few preference comparisons.

We consider two types of user preference comparisons for learning the metric: *unquantized* and *quantized measurements*. From a user with ideal point u, these are of the form:

$$\underbrace{(x,x',\psi)}_{\text{unquantized}} \qquad \text{and} \qquad \underbrace{(x,x',y)}_{\text{quantized}},$$

where $\psi = \psi(x,x';u)$ is a real number that indicates the difference between the squared distances, and y is binary, taking values in $\{-1,+1\}$. When $y=-1$, x is preferred over x', and $y=+1$ indicates otherwise.

Metric learning from preference measurements.

We now review the existing algorithmic ideas for recovering the metric from preference feedback under the ideal point model. Suppose that we are given unquantized measurements from a single user with an ideal point $u \in \mathbb{R}^d$. With a little algebra [28], the measurement in Eq. (5.1) becomes:

$$\psi_M(x,x';u) = \left\langle xx^\top - x'x'^\top,\, M \right\rangle + \left\langle x-x', v \right\rangle, \tag{5.2}$$

where $v := -2Mu$. The first inner product is the trace inner product for matrices, while the second inner product is the usual inner product on $\mathbb{R}^d$. The re-parametrization v of u

is sometimes called the *pseudo-ideal point*. Thus, unquantized measurements are linear over the joint variables (M, v). Given a set of unquantized measurements from a user, one can just solve a linear system of equations to recover the matrix representation M of ρ, as described in Algorithm 7 of Appendix D.3. Since M has full rank and therefore invertible, we can then recover u from M and v [28].

As there are $\frac{d(d+1)}{2} + d$ degrees of freedom in (M, v), to recover the metric in this way requires at least that many measurements from a single user; the first term corresponds to the dimension of symmetric $d \times d$ matrices representing Mahalanobis distances, the second for the user ideal point.

When d^2 is very large, we may want to amortize learning the metric over many users. [28] show that this is possible. Let the users be indexed by elements in $[K]$. We can construct a larger linear regression problem, where each user has a separate covariate corresponding to their ideal point. Now, the joint variable is $(M, v_1, v_2, \ldots, v_K)$, which has $\frac{d(d+1)}{2} + dK$ degrees of freedom. When the population is large, it suffices to ask each user $\Theta(d + d^2/K)$ preference queries, which can be much closer to d than d^2. This procedure is given in Algorithm 8 of Appendix D.3.

However, modern representations of data may be extremely high-dimensional, and it would be too onerous for any single user to provide d measurements. In this chapter, we tackle this question: If we have access to many users but can only ask each user a much more limited number $m \ll d$ of preferences queries, can we still recover ρ? We note that with $o(d)$ pairwise queries, it is impossible to localize the ideal preference point of a user even with a known metric [69, 107]. So, our goal here is to address the open question of whether it is possible to learn an *unknown* metric with such limited queries per users given a sufficiently large pool of users.

Notation.

Let $\mathrm{Sym}(\mathbb{R}^d)$ denote the symmetric $d \times d$ matrices equipped with the trace inner product, and let $\mathrm{Sym}^+(\mathbb{R}^d)$ be the positive-definite matrices. For readability, we often make abbreviations of the form $\Delta \in \mathrm{Sym}(\mathbb{R}^d)$ and $\delta \in \mathbb{R}^d$:

$$\Delta \equiv xx^\top - x'x'^\top \qquad \text{and} \qquad \delta \equiv x - x'.$$

Then, $\Delta \oplus \delta$ is an element of $\mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d$, the direct sum of inner product spaces, and we can shorten Eq. (5.2) to:

$$\psi_M(x, x'; u) = \big\langle \Delta \oplus \delta, M \oplus v \big\rangle.$$

Following the experimental design literature, let us call a collection of such elements a *design matrix*:

Definition 5.1. *Let $\{(x_{i_0}, x_{i_1})\}_{i\in[m]}$ be a collection of item pairs. It induces the linear map $D : \mathrm{Sym}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^m$,*

$$D(A, w)_i = \big\langle \Delta_i \oplus \delta_i, A \oplus w \big\rangle,$$

where $\Delta_i = x_{i_0}x_{i_0}^\top - x_{i_1}x_{i_1}^\top$ and $\delta_i = x_{i_0} - x_{i_1}$ for $i \in [m]$. As a slight abuse of language, we call D the induced design matrix. *If item pairs are drawn from a distribution $\mathcal{P}_m$ over $(\mathbb{R}^d \times \mathbb{R}^d)^m$, we say that D is a* random design *and write $D \sim \mathcal{P}_m$. We also define $\sigma^2_{\min}(\mathcal{P}_m) = \frac{1}{m} \cdot \sigma_{\min}\big(\mathbb{E}[D^*D]\big)$.*

For additional background and notation, see Appendix D.3.

5.3 An Impossibility Result

Consider the mathematically simplified setting in which users provide *unquantized* responses. We show a negative result stating that when users provide fewer than d comparisons, we fundamentally cannot learn anything about M if the items are in general position in the following sense:

Definition 5.2. *A set $\mathcal{X} \subset \mathbb{R}^d$ has* generic pairwise relations *if for any acyclic graph $G = (\mathcal{X}, E)$ with at most d edges, the set $\{x - x' : (x, x') \in E\}$ is linearly independent.*

The geometric meaning of having generic pairwise relations is simple: if any d pairs of points are connected by lines, then those lines are linearly independent (unless they form cycles; see Figure D.1 in Appendix D.4). Proposition D.8 shows that almost all finite subsets of Euclidean space have generic pairwise relations with respect to the Lebesgue measure.

The following theorem shows that if items have generic pairwise relations, then sets of $m \leq d$ unquantized measurements from a single user provide no information about the underlying metric. In particular, suppose that M and v are the underlying matrix representation and user's pseudo-ideal point, both unknown to us. Then, for any other Mahalanobis matrix M', we can find a pseudo-ideal point v' that is also consistent with the data. In fact, the negative result holds even with infintely many users:

Theorem 5.3. *Fix $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ and $v_k \in \mathbb{R}^d$ for each $k \in \mathbb{N}$. Let $(D_k)_{k\in\mathbb{N}}$ be a collection of design matrices, each for a set of $m \leq d$ pairwise comparisons. If each set of compared items has generic pairwise relations, then for all $M' \in \mathrm{Sym}^+(\mathbb{R}^d)$, there exists $(v'_k)_{k\in\mathbb{N}} \subset \mathbb{R}^d$ such that:*

$$D_k(M, v_k) = D_k(M', v'_k), \qquad \forall k \in \mathbb{N}.$$

See Appendix D.4 for a proof of Theorem 5.3. This theorem shows that when items

have *generic pairwise relations* it is not just that we cannot recover ρ, but that we cannot glean anything at all about ρ when users each provide d or fewer comparisons, for every matrix in $\mathrm{Sym}^+(\mathbb{R}^d)$ is consistent with D. While each user provides us with more data, each also introduces new degrees of freedom—the unknown ideal points. When learning from crowds, more data does not necessarily lead to more usable information.

5.4 Exact Recovery with Low-Rank Subspace Structure

The above negative result applies to almost all finite sets of items. It seems to tell a pessimistic story for metric learning when data is embedded into high dimensions and when it is infeasible to obtain $\Omega(d)$ preference comparisons per user.

However, the story is not closed and shut yet. Real-world data often exhibit additional structure that could help us recover the metric, such as low intrinsic dimension [51]. In particular, we assume that many items of $\mathcal{X}$ lie on a *union of subspaces*. The approximate validity of this assumption is also the basis of work in manifold learning [124, 147, 18], compressed sensing [45], and sparse coding [112], among others.

In this case, we can take a divide-and-conquer approach to metric learning by identifying the metric restricted to those subspaces, before stitching them back together to recover the full metric. Let's define subspace Mahalanobis distances:

Definition 5.4. *Let V be a subspace of $\mathbb{R}^d$. A metric on V is a* subspace Mahalanobis distance *if it is a subspace metric of some Mahalanobis distance ρ on $\mathbb{R}^d$. In that case, we denote the subspace metric by $\rho\big|_V$, where for all $x, x' \in V$,*

$$\rho\big|_V(x, x') = \rho(x, x').$$

In general, we cannot hope to identify an arbitrary metric from a finite number of its subspace metrics. However, Mahalanobis distances have much more structure than

arbitrary metrics on $\mathbb{R}^d$. A Mahalanobis distance on $\mathbb{R}^d$ can be fully specified using $d(d+1)/2$ numbers. By recovering its subspace metrics, we can hope to chip away at the degrees of freedom of Mahalanobis distances. As another way of intuition, each Mahalanobis distance may be identified with its unit sphere—points that are unit distance away from the origin. These points form a $(d-1)$-dimensional ellipsoid in $\mathbb{R}^d$. To recover a subspace Mahalanobis distance on V means that we are able to determine which points of V intersect this ellipsoid (see Figure 5.1). If we do this for sufficiently many subspaces, we can determine the whole ellipsoid. To formalize this intuition, we now linear-algebraically relate a Mahalanobis distance with its subspace metrics.

5.4.1 A linear parametrization of Mahalanobis distances

To describe the linear relationship between a Mahalanobis distance and its subspace metrics, we need to parametrize the subspace metrics. To do so, we first need to fix a choice of coordinates on each $V \subset \mathbb{R}^d$. In the following, let V be an r-dimensional subspace of $\mathbb{R}^d$ and let $B \in \mathbb{R}^{d\times r}$ be an orthonormal basis of V, where $r \ll d$.

Definition 5.5. *We say V has a* canonical representation *if it is equipped with an orthonormal basis B, where the canonical representation of a vector $x \in V$ is given by $B^\top x \in \mathbb{R}^r$.*[1]

Definition 5.6. *Let $\mathrm{Sym}(V)$ and $\mathrm{Sym}^+(V)$ respectively denote the pairs $(\mathrm{Sym}(\mathbb{R}^r), B)$ and $(\mathrm{Sym}^+(\mathbb{R}^r), B)$, where V has a canonical representation given by B.*

We write $Q \in \mathrm{Sym}(V)$ to mean that $Q \in \mathrm{Sym}(\mathbb{R}^r)$, and that it carries the basis information B along with it.

Just as Mahalanobis distances on $\mathbb{R}^d$ are in one-to-one correspondence with positive-definite matrices, so too are Mahalanobis distances on V in correspondence with $\mathrm{Sym}^+(V)$.

[1] We shall always equip $\mathbb{R}^d$ with the standard basis, so that a vector is its own canonical representation in $\mathbb{R}^d$.

Furthermore, Proposition D.12 shows that the matrix representations of a Mahalanobis distance and its restriction to a subspace is given by the following linear map.

Definition 5.7. *Let V and B be as before. Define the linear map $\Pi_V : \mathrm{Sym}(\mathbb{R}^d) \to \mathrm{Sym}(V)$ by:*

$$\Pi_V(A) = B^\top AB. \tag{5.3}$$

Thus, if a Mahalanobis distance ρ on $\mathbb{R}^d$ and its restriction $\rho\big|_V$ to a subspace V have representations $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ and $Q \in \mathrm{Sym}^+(V)$, respectively, then:

$$Q = \Pi_V(M) = B^\top MB.$$

5.4.2 Learning with low-rank subspaces

To see how low-dimensional structure can help us make progress in learning the metric, consider a simple setting where all items lie in some low-dimensional subspace V. Instead of learning the full metric ρ, we could aim for a more modest goal of learning the subspace metric $\rho\big|_V$.

As before, let V be an r-dimensional subspace of $\mathbb{R}^d$ with a canonical representation. If all items and ideal points lie in V, then learning $\rho\big|_V$ immediately reduces to the usual setting of learning a Mahalanobis distance, since we can simply ignore the remaining dimensions and reparametrize the problem. But when the ideal points are not assumed to lie on V, it is not evident *a priori* that we can ignore the dimensions extending beyond the set of items. However, it turns out that for Mahalanobis distances, we may.

The next lemma shows that even if a user's ideal point $u \in \mathbb{R}^d$ falls outside of V, for items in V, there is a phantom ideal point $u_V \in V$ such that preference comparisons for items in V generated by u and u_V are equivalent.

Lemma 5.8. *Let V be an r-dimensional subspace of $\mathbb{R}^d$ with a canonical representation given by $B \in \mathbb{R}^{d\times r}$. Fix any Mahalanobis distance $M \in \mathrm{Sym}^+(\mathbb{R}^d)$, any pair of items*

$x, x' \in \mathbb{R}^d$, and ideal point $u \in \mathbb{R}^d$. Suppose that x and x' are contained in V with canonical representation $x_V = B^\top x$ and $x'_V = B^\top x'$ in $\mathbb{R}^r$. Then:

$$\psi_M\big(x, x'; u\big) = \psi_Q\big(x_V, x'_V; u_V\big),$$

where the phantom ideal point u_V of u on V satisfies $(B^\top M B)u_V = B^\top M u$, and $Q = \Pi_V(M)$ is the matrix representation in $\mathrm{Sym}^+(V)$ of the subspace metric $\rho\big|_V$.

Consequently, learning a subspace metric $\rho\big|_V$ turns into a problem of metric learning from preference comparisons in $\mathbb{R}^r$. From here, we can simply use existing algorithms to recover the matrix representation of the subspace metric. By [28], it is possible to identify the subspace metric so long as users can each provide $m \geq \Omega(r)$ preference comparisons. For this easier problem of learning $\rho\big|_V$, when $r \ll d$, we can do with $o(d)$ responses per user.

In the remainder of this section, we give a simple characterization for when a Mahalanobis distance on V can be learned from preference comparisons of items on V. The set of items needs to be sufficiently rich so that all degrees of freedom of $\mathrm{Sym}(V) \oplus V$ can be captured. We define:

Definition 5.9. *Let V be a subspace of $\mathbb{R}^d$ with canonical representation given by B. A subset $\mathcal{X}_V \subset V$* quadratically spans *V if $\mathrm{Sym}(V) \oplus V$ is linearly spanned by the set:*

$$\big\{(x_V x_V^\top - x'_V x'^\top_V) \oplus (x - x') : x, x' \in \mathcal{X}_V\big\},$$

where $x_V = B^\top x$ and $x'_V = B^\top x'$ denote the canonical representations of x and x' in V.

If we have no restriction on how many queries we can ask a user, then it is straightforward to see that quadratic spanning is a sufficient condition for recovering the underlying metric. For simplicity, let $V = \mathbb{R}^d$. If $\mathcal{X}$ quadratically spans $\mathbb{R}^d$, then we can

detect all dimensions of $M \oplus v$ corresponding to the Mahalanobis matrix and a user's pseudo-ideal point. To do so, choose any design matrix $D : \mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d \to \mathbb{R}^m$ whose rows $\{\Delta_i \oplus \delta_i : i \in [m]\}$ span $\mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d$.

When the number of queries is limited per user, the following result shows that the quadratic spanning condition is still sufficient for recovering $\rho\big|_V$, provided we can ask many users $m \geq \dim(V) + 1$ unquantized preference queries.

Proposition 5.10. *Let $\mathcal{X}$ quadratically span a subspace V of dimension r. There exists a collection $D_1, \ldots, D_K$ of design matrices, each over m pairs of items in $\mathcal{X}$, such that given a (distinct) user's response to each design, $\rho\big|_V$ can be identified when $m \geq r + 1$ and $K \geq r(r+1)/2$.*

To complement this sufficient condition, the next result shows that if $\mathcal{X}$ does not quadratically span V, then the subspace metric $\rho\big|_V$ cannot be recovered from only preference comparisons of items in $\mathcal{X} \cap V$.

Proposition 5.11. *Let $(D_k)_{k \in \mathbb{N}}$ be a set of design matrices over items in $\mathcal{X} \subset V$. If $\mathcal{X}$ does not quadratically span V, then infinitely many Mahalanobis distances on V are consistent with any set of user responses to the design matrices.*

Proofs for the above results are deferred to Appendix D.5.2.

5.4.3 Learning with subspace-clusters

We've seen how to partially learn a Mahalanobis distance given many items within a subspace. We now consider how to fully recover the metric when many items lie in a *union of subspaces* $(V_\lambda)_{\lambda \in \Lambda}$. In this case, a divide-and-conquer approach is intuitive: (i) recover each subspace metric, then (ii) reconstruct ρ from the learned subspace metrics. Recall that each subspace metric $\rho\big|_V$ is related to the full metric ρ by the linear map Π_V from Definition 5.7. Therefore, we can reconstruct ρ from its subspace metrics by solving a system of linear equations. Algorithm 4 summarizes this approach.

Algorithm 4: Metric learning from subspace clusters

Input: Unquantized measurements over items that lie in a union of subspaces $V_\lambda, \lambda \in \Lambda$

// Stage 1: learning subspace metrics

for *each subspace* $\lambda \in \Lambda$ **do**

  Recover $\hat{Q}_\lambda \in \mathrm{Sym}(\mathbb{R}^{r_\lambda})$ with respect to B_λ via reduction to Algorithm 8 [28]

// Stage 2: reconstruction

Solve the linear equations over $A \in \mathrm{Sym}(\mathbb{R}^d)$:

$$B_\lambda^\top A B_\lambda = \hat{Q}_\lambda, \quad \lambda \in \Lambda$$

Output: $\hat{A}$, the solution to the above linear equations.

In order to characterize when a Mahalanobis distance can be reconstructed from its subspace metrics, we introduce the notion of subspace-clusterability. A set of items $\mathcal{X}$ is subspace-clusterable when many of its items lie on sufficiently many item-rich subspaces. Formally:

Definition 5.12. *A set* $\mathcal{X} \subset \mathbb{R}^d$ *is* subspace-clusterable *over subspaces* $V_\lambda \subset \mathbb{R}^d$ *indexed by* $\lambda \in \Lambda$ *whenever:*

1. *each subset* $\mathcal{X} \cap V_\lambda$ *quadratically spans* V_λ.

2. $\left\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\right\}$ *linearly spans* $\mathrm{Sym}(\mathbb{R}^d)$.

By Propositions 5.10 and 5.11, the first condition is necessary and sufficient for recovering each subspace metric $\rho\big|_{V_\lambda}$. Proposition 5.13 shows that the second condition is necessary and sufficient for recovering the ρ from subspace metrics.

Proposition 5.13. *Let* ρ *be a Mahalanobis distance on* $\mathbb{R}^d$. *Let* $(V_\lambda)_{\lambda\in\Lambda}$ *be a collection of subspaces with canonical representations given by the orthonormal bases* $(B_\lambda)_{\lambda\in\Lambda}$. *The following are equivalent:*

1. $\left\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\right\}$ *spans* $\mathrm{Sym}(\mathbb{R}^d)$.

2. *Let* Π_{V_λ} *be given by Equation* (5.3). *The linear map* $\Pi : \mathrm{Sym}(\mathbb{R}^d) \to \bigoplus_{\lambda \in \Lambda} \mathrm{Sym}(V_\lambda)$ *is injective, where:*

$$\Pi(A) = \bigoplus_{\lambda \in \Lambda} \Pi_{V_\lambda}(A).$$

3. *If* $\hat{\rho}$ *is a Mahalanobis distance such that* $\hat{\rho}\big|_{V_\lambda} = \rho\big|_{V_\lambda}$ *for all* $\lambda \in \Lambda$, *then* $\hat{\rho} = \rho$.

See Appendix D.5.3 for the proof. This proposition verifies the correctness of Algorithm 4. Let $Q_\lambda \in \mathrm{Sym}^+(V)$ represent $\rho\big|_{V_\lambda}$. Then, step 3 of the algorithm specifies that $\Pi_{V_\lambda}(A) = Q_\lambda$. If Π is injective, then the only matrix $A \in \mathrm{Sym}(\mathbb{R}^d)$ consistent with the system of linear equations is the one that represents ρ.

Remark 5.14. *We can compute the number of subspaces required to identify* ρ *using Proposition 5.13. For example, when* $\dim(V_\lambda) = 1$ *for each* $\lambda \in \Lambda$, *each subspace captures one degree of freedom of* ρ, *so* $|\Lambda| \geq \frac{d(d+1)}{2}$ *is necessary. See Figure D.2 in Appendix D.5 for geometric intuition.*

5.5 Approximate Recovery from Binary Responses

Previously, we studied metric learning from unquantized preference comparisons of the form (x, x', ψ). We now consider a more realistic setting where we obtain binary responses of the form (x, x', y), where $y \in \{-1, +1\}$. Furthermore, we assume that responses are quantized and noisy, where noise can depend on the user and items, as in [106, 167, 28].

For our divide-and-conquer approach, due to the inexactness of the responses, we can no longer expect to exactly identify each subspace metric. However, we show that as long as each subspace metric can be recovered approximately, then they can be stitched together to approximately recover the full metric (Theorem 5.15). And indeed, approximate recovery in each subspace is known to be possible. In Proposition 5.18, we present a version of Theorem 4.1 of [28] adapted to subspaces; this gaurantee is provided under a probabilistic noise model that we describe shortly.

Algorithm 5: Metric learning from binary responses

Input: Quantized measurements over items that lie in a union of subspaces $V_\lambda, \lambda \in \Lambda$

// Stage 1: learning subspace metrics

for *each* $\lambda \in \Lambda$ **do**

  Recover $\hat{Q}_\lambda \in \mathrm{Sym}(\mathbb{R}^{r_\lambda})$ with respect to B_λ via reduction to Algorithm 9 [28]

// Stage 2: reconstruction

Use ordinary least squares to solve the linear regression problem over $A \in \mathrm{Sym}(\mathbb{R}^d)$:

$$\hat{M}_{\mathrm{LS}} \leftarrow \operatorname*{argmin}_{A \in \mathrm{Sym}(\mathbb{R}^d)} \sum_{\lambda \in \Lambda} \left\| \hat{Q}_\lambda - B_\lambda^\top A B_\lambda \right\|_{\mathrm{F}}^2 \tag{5.4}$$

Project $\hat{M}_{\mathrm{LS}}$ onto the set of positive semidefinite $d \times d$ matrices by solving the convex optimization problem:

$$\hat{M} \leftarrow \operatorname*{argmin}_{A \succeq 0} \left\| A - \hat{M}_{\mathrm{LS}} \right\|_{\mathrm{F}}^2 \tag{5.5}$$

Output: $\hat{M}$.

Divide-and-conquer algorithm.

Algorithm 5 generalizes our earlier algorithm for unquantized measurements. As before, say we've obtained measurements for a set of items subspace-clusterable over $(V_\lambda)_\lambda$. In the first stage, we recover the subspace metrics on each V_λ. Lemma 5.8 reduces metric learning on subspaces to metric learning on $\mathbb{R}^r$, where r is the dimension of the subspace, so we can call existing methods for metric learning from binary responses across users ([28] or Algorithm 9). Thus, we obtain an estimator $\hat{Q}_\lambda$ for each subspace metric Q_λ.

In the second stage, we approximately reconstruct the Mahalanobis matrix M from the estimators $\hat{Q}_\lambda$. When each $\hat{Q}_\lambda$ was exact, we could just solve the linear system of

equations $\Pi_{V_\lambda}(\hat{M}) = \hat{Q}_\lambda$. As this is no longer the case, we instead compute the ordinary least squares estimator $\hat{M}_{\text{LS}}$, which minimizes $\sum_\lambda \|\hat{Q}_\lambda - \Pi_{V_\lambda}(A)\|^2$ over $A \in \text{Sym}(\mathbb{R}^d)$ in Eq. (5.4) of Algorithm 5. Finally, we ensure that the reconstructed matrix corresponds to a pseudo-metric by solving a linear program to project $\hat{M}_{\text{LS}}$ onto the cone of positive semi-definite matrices [22].

5.5.1 Recovery guarantees

Reconstruction guarantee.

The following theorem gives a recovery guarantee on the full metric, given approximate recovery for each subspace metric, $\|\hat{Q}_\lambda - Q_\lambda\|_{\text{F}} \leq \varepsilon$ for some $\varepsilon > 0$. See Appendix D.6.1 for its proof.

Theorem 5.15. *Let $\mathbb{R}^d$ have a Mahalanobis distance with matrix representation $M \in \text{Sym}^+(\mathbb{R}^d)$. Let $\mathcal{X} \subset \mathbb{R}^d$ be subspace-clusterable over subspaces V_λ indexed by $\lambda \in \Lambda$, where $|\Lambda| = n$. Let $\hat{M}$ be the estimator of M and let $\hat{Q}_\lambda$ be the estimator of the subspace metric Q_λ for each λ learned from Algorithm 5. Suppose there exist $\gamma \leq \varepsilon$ such that $\big\|\mathbb{E}\big[\hat{Q}_\lambda\big] - Q_\lambda\big\|_{\text{F}} \leq \gamma$ and $\big\|\hat{Q}_\lambda - Q_\lambda\big\|_{\text{F}} \leq \varepsilon$ for each λ. Fix $p \in (0, 1]$. Then, there is a universal constant $c > 0$ such that with probability at least $1 - p$,*

$$\big\|\hat{M} - M\big\|_{\text{F}} \leq c \cdot \frac{1}{\sigma_{\min}(\Pi)} \left(\gamma\sqrt{n} + \varepsilon d \sqrt{\log \frac{2d}{p}} \right),$$

where $\sigma_{\min} > 0$ is the least singular value of Π.

Remark 5.16. *This recovery guarantee depends on three parameters: (1) $\sigma_{\min}(\Pi)$ captures how well-spread the set $\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\}$ is across $\text{Sym}(\mathbb{R}^d)$. (2) ε bounds the recovery error for each subspace metric; it decreases as the number of pairwise comparisons per user increases (Remark 5.19). (3) γ bounds the bias of the estimator $\hat{Q}_\lambda$. It can be the dominating term in the recovery bound, for example when $\sigma_{\min}(\Pi) \gg d$. While this*

bias term $\gamma \leq \varepsilon$ can be made arbitrarily small with enough comparisons per user, for data-starved regimes, bias reduction can also be applied in practice (e.g. [53]).

Recovery guarantee for subspace metrics.

For completeness, we adapt the setting and results of [28] to provide a recovery guarantee for learning each subspace metric. We assume the same probabilistic model:

Assumption 5.17 (Probabilistic model)**.** *Let $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ be the matrix representation of the Mahalanobis distance, let $v_1, \ldots, v_K \in \mathbb{R}^d$ be the pseudo-ideal points for a collection of users, and let $\mathcal{X} \subset \mathbb{R}^d$ be a set of items. We assume:*

$$\|M\|_{\mathrm{F}} \leq \zeta_M, \qquad \|v_k\| \leq \zeta_v, \qquad \sup_{x \in \mathcal{X}} \|x\| \leq 1,$$

for some $\zeta_M, \zeta_v > 0$. When asked to compare two items x and x', the kth user provides a binary response Y with:

$$\Pr[Y = y] = f\big(y \cdot \psi_M(x, x'; u_k)\big),$$

where $f : \mathbb{R} \to [0, 1]$ is a strictly increasing link function *such that $f(z) = 1 - f(-z)$, and where u_k is the corresponding ideal point. On the domain $|z| \leq 2(\zeta_M + \zeta_v)$, let f have lower bounded derivative $f'(z) \geq c_f$ and let the map $z \mapsto -\log f(z)$ have Lipschitz constant L.*

Algorithm 9 estimates $(M, v_1, \ldots, v_K)$ by using the users' measurements to construct an optimization program over the parameters; when the loss function supplied to the algorithm is $\ell(z) = -\log f(z)$, the procedure is equivalent to maximum likehlihood estimation. As noted above, it suffices to consider learning Mahalanobis distances on $\mathbb{R}^r$. The following proposition proves correctness of Algorithm 9.

Proposition 5.18 (Theorem 4.1, [28])**.** *Suppose that $\mathbb{R}^r$ has a Mahalanobis distance with representation $Q \in \mathrm{Sym}^+(\mathbb{R}^r)$ where $\|Q\|_{\mathrm{F}} \leq \zeta_M$. Let each user $k \in [K]$ have pseudo-ideal*

point $v_k \in \mathbb{R}^r$ where $v_k \leq \zeta_v$. Let $\mathcal{P}_m$ be a distribution over designs of size m over $\mathbb{R}^r$ (Definition 5.1). For each user, let $D_k \sim \mathcal{P}_m$ be an i.i.d. random design, and let $\mathcal{D}_k = \{(x_{i_0}, x_{i_1}, y_{i;k})\}_{i\in[m]}$ be the user's responses under Assumption 5.17. Fix $p \in (0, 1]$. Given loss function $\ell(z) = -\log f(z)$, Algorithm 9 returns $\hat{Q} \in \mathrm{Sym}^+(\mathbb{R}^r)$, where with probability at least $1 - p$,

$$\|\hat{Q} - Q\|_{\mathrm{F}}^2 \leq \frac{16L}{c_f^2 \cdot \sigma_{\min}^2(\mathcal{P}_m)} \sqrt{\frac{(\zeta_M^2 + K\zeta_v^2) \log \frac{4}{p}}{mK}}.$$

The proof of Proposition 5.18 is deferred to Appendix D.6.2

Remark 5.19. *We can simplify the bound if we assume that M has bounded entries, say $\|M\|_\infty \leq 1$. Let's also assume that user ideal points are contained in the unit ball, so that $\|u_k\|_2 \leq 1$ for each user. Then, we can set $\zeta_M \leq r$ and $\zeta_v \leq 2\sqrt{r}$ since $v_k = -2Mu_k$. Remark D.14 shows that given access to a subspace-clusterable set of items, we can construct a sequence of random designs $(\mathcal{P}_m)_m$ over those items such that $\sigma_{\min}^2(\mathcal{P}_m) = \Omega(1)$. Suppressing the confidence parameter p, we obtain the recovery guarantee:*

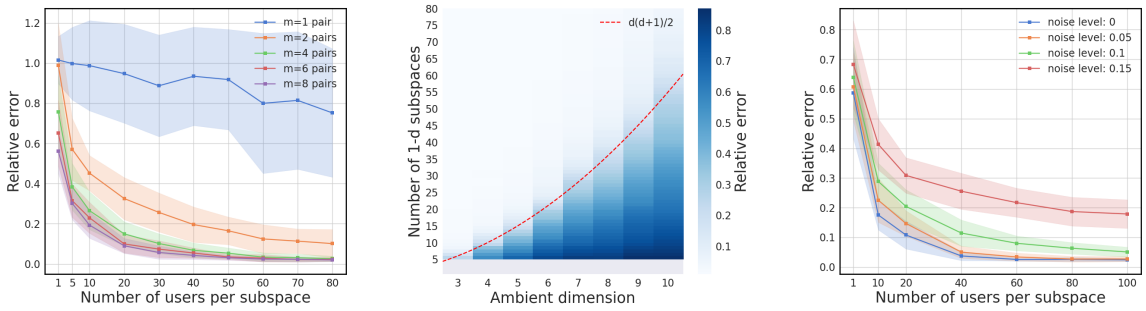
$$\|\hat{Q} - Q\|_{\mathrm{F}}^2 = \mathcal{O}\left(\sqrt{\frac{r^2 + Kr}{mK}}\right).$$

5.6 Empirical Validation

In this section, we empirically validate our findings using synthetic data. We aim to address the following questions:

1. Given limited noisy, quantized preference comparisons on subspace-clusterable items, can our proposed divide-and-conquer algorithm recover an unknown metric M?

2. Does the performance of our algorithm improve if we have access to more subspace-clusters, users or preference comparisons per user?



(a) **(b)** **(c)**

Figure 5.2. (a) Shows the average relative errors for varying numbers of users per subspace and preference comparisons per user, where items lie in a union of 80 1-dimensional subspaces of $\mathbb{R}^{10}$. (b) shows the average relative errors given increasing numbers of 1-dimensional subspaces to reconstruct $\hat{M}$; for each subspace, 60 users each provides 4 preference comparisons. The dotted red curve illustrates the dimension-counting argument in Remark 5.14. (c) shows the average relative errors for varying subspace noise levels, where items lie approximately in a union of 80 1-dimensional subspaces of $\mathbb{R}^{10}$; each user provides 8 preference comparisons. The error bars in (a) and (c) represent one standard deviation from the mean.

3. When items in $\mathcal{X}$ lie *approximately* in a union of subspaces, can we still recover M?

Experimental setup.

For each run, we generate a random ground-truth metric $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ from the standard Wishart distribution $W(I_d, d)$, a collection of uniform-at-random r-dimensional subspaces [140], and a set of user ideal points drawn i.i.d. from the Gaussian $\mathcal{N}(0, \frac{1}{d}I_d)$. Within each subspace, items are drawn i.i.d. from $\mathcal{N}(0, \frac{1}{r}BB^\top)$, where $B \in \mathbb{R}^{d\times r}$ is an orthonormal basis of that subspace. Given a user and a pair of items, a binary response is sampled according to the probabilistic model in Assumption 5.17, where the link function is chosen to be the logistic sigmoid, $f(x;\beta) = 1/\left(1+\exp(-\beta x)\right)$.

To evaluate the learned metric $\hat{M}$, we compute its relative error, $\|\hat{M} - M\|_{\mathrm{F}}/\|M\|_{\mathrm{F}}$. We observed that Huber regression [66, 118] generally leads to better performance over least squares regression within Algorithm 5 (Stage 2, line 3). In the following, we report results obtained using this robust variant of linear regression.

We ran three experiments each for 30 runs, where we set the subspace dimension

to $r = 1$, and we set $\beta = 4$ which corresponds to "medium" response noise in [28]. See Appendix D.7 for experiments for $r = 2$.

Relative error vs number of comparisons.

In the first experiment, we set the ambient dimension to $d = 10$ and generated data that lie in a union of 80 subspaces (by Remark 5.14, at least $\dim(\mathrm{Sym}(\mathbb{R}^{10})) = 55$ subspaces are needed for recovery). We ran Algorithm 5 for different combinations of K and m, where K is the number of users per subspace and m is the number of preference comparisons per user. Figure 5.2a compares the average relative errors for varying K and m. This experiment shows that with more preference comparisons, recovery within each subspace improves and we achieve better recovery of the full metric; this supports Theorem 5.15 and Proposition 5.18. This experiment also suggests that given 1-dimensional subspaces, even asking for only two measurements per user is sufficient to achieve good empirical performance for metric recovery.

Relative error vs number of subspaces.

In the second experiment, we set $K = 60$ and $m = 4$. For ambient dimensions $d = 3, 4, \ldots, 10$, we consider the relative error for reconstructing $\hat{M}$ using an increasing number of subspaces, $n = 5, 6, \ldots, 80$. Figure 5.2b shows the average relative errors. For each d, average relative error decreases as n increases. Furthermore, even in this non-idealized setting where users provide noisy, binary responses, we can obtain non-trivial relative error when the number of subspaces n exceeds the information-theoretic bound $d(d+1)/2$. This corroborates the dimension-counting argument in Remark 5.14 beyond unquantized measurements.

Recovery when items approximately lie in subspaces.

In the third experiment, we empirically study how our approach works when the subspace clusterable assumption only approximately holds. For a subspace V, we sample items near V from $\mathcal{N}(0, \frac{1}{r}BB^\top + \frac{\sigma^2}{d-r}B_\perp B_\perp^\top)$, where $\sigma > 0$ is a given noise level, $B \in \mathbb{R}^{d\times r}$

and $B_\perp \in \mathbb{R}^{d\times(d-r)}$ are orthonormal bases of V and its orthogonal complement $V^\perp$, respectively. The way user preference responses are generated remains the same as before. For each subspace V, we preprocess the items by running singular value decomposition on the nearby items to recover an r-dimensional subspace $\hat{V}$. We project these items to $\hat{V}$, before running Algorithm 5 with these approximate representations. We set $d = 10$ and $m = 8$. For each subspace noise level σ, we ran our approach on items that lie approximately in 80 subspaces for varying K; Figure 5.2c shows the average relative errors. When the noise level σ is low, we can still recover the metric well. As expected, this breaks down as σ increases; indeed, when $\sigma = 1$, there is no subspace structure at all.

5.7 Conclusion and Future Work

We studied crowd-based metric learning from very few preference comparisons per user. In general, we showed nothing can be learned. However, when the items exhibit low-rank subspace-clusterable structure, we proposed a divide-and-conquer approach and provided recovery guarantees. Interestingly, this chapter suggests that when training of foundation models, there is reason to favor learning general-purpose representations with low-rank structures, as this may reduce the cost of downstream fine-tuning and alignment.

Our experiments show that even when the items do not exactly lie on the subspaces, but instead only exhibit approximate subspace structure, our method can still recover the metric. We leave establishing theoretical gurantees for this setting for future work. Our results has implications for alignment of representations from foundation models to human preferences and we defer building an algorithmic framework that finds subspace clusters before learning metrics, and evaluating it with real-world item embeddings and human preference feedback for future work.

Acknowledgement.

Chapter 5 is based on the material as it appears in "Metric Learning from Limited Pairwise Preference Comparisons" by Zhi Wang, Geelon So, and Ramya Korlakai Vinayak. The material is currently in submission. The dissertation author was the primary investigator and first author of the paper.

Appendix A

Supplementary Material for Chapter 2

A.1 Related Work and Comparisons

We review the literature on multi-player bandit problems (see also Landgren [88, Section 1.3.2] for a survey), and we comment on how existing problem formulations/approaches compare with ours studied in this chapter.

Identical reward distributions.

A large portion of prior studies focuses on the setting where a group of players collaboratively work on one bandit learning problem instance, i.e., for each arm/action, the reward distribution is identical for every player.

For example, Kar et al. [75] study a networked bandit problem, in which only one agent observes rewards, and the other agents only have access to its sampling pattern. Peer-to-peer networks are explored by Szörényi et al. [141], in which limited communication is allowed based on an overlay network. Landgren et al. [87] apply running consensus algorithms to study a distributed cooperative multi-armed bandit problem. Kolla et al. [78] study collaborative stochastic bandits over different structures of social networks that connect a group of agents. Wang et al. [160] study communication cost minimization in multi-agent multi-armed bandits. Multi-agent bandit with a gossip-style protocol that has a communication budget is investigated in [127, 34]. Dubey and Pentland [46] investigate multi-agent bandits with heavy-tailed rewards. Wang et al. [157] present an approach with

a "parsimonious exploration principle" to minimize regret and communication cost. We note that, in contrast, we study multi-player bandit learning where the reward distributions can be different across players .

Player-dependent reward distributions.

Multi-agent bandit learning with *heterogeneous feedback* has also been covered by previous studies.

- Cesa-Bianchi et al. [31] study a network of linear contextual bandit players with heterogeneous rewards, where the players can take advantage of reward similarities hinted by a graph. In [165, 156, 159], reward distributions of each player are *generated* based on social influence, which is modeled using preferences of the player's neighbors in a graph. These papers use regularization-based methods that take advantage of graph structures; in contrast, we study *when and how* to use information from other players based on a dissimilarity parameter.

- Gentile et al. [57], Bresler et al. [23], Song et al. [139], Li et al. [95], Korda et al. [79], Li et al. [96], among others, assume that the players' reward distributions have a cluster structure and players that belong to one cluster share a *common* reward distribution; we do not assume such cluster structure.

- Nguyen and Lauw [111] investigate dynamic clustering of players with independent reward distributions and provides an empirical validation of their algorithm; Zhu et al. [180] present an algorithm that combines dynamic clustering and Thompson sampling. In contrast, in this chapter, we develop a UCB-based approach that has a fallback guarantee[1].

- In the work of Shahrampour et al. [129], a group of players seek to find the arm with

[1]In Zhu et al. [180], it is unclear how to tune the hyper-parameter β apriori to ensure a sublinear fall-back regret guarantee, even if the "similarity" parameter γ is known.

the largest average reward over all players; and, in each round, the players have to reach a consensus and choose the same arm.

- Dubey and Pentland [47] assume access to some side information for every player, and learns a reward predictor that takes both player's side information models and action as input. In comparison, our work do not assume access to such side information.

- Further, similarities in reward distributions are explored in the work of Zhang et al. [174], which studies a *warm-start* scenario, in which data are provided as history [133] for an learning agent to explore faster. Azar et al. [11], Soare et al. [138] investigate multitask learning in bandits through *sequential transfer* between tasks that have similar reward distributions. In contrast, we study the multi-player setting, where all players learn continually and concurrently.

Collisions in multi-player bandits.

Multi-player bandit problems with collisions [e.g., 99, 74, 21, 25, 132, 26, 157] are also well-studied. In such models, two players pulling the same arm in the same round *collide* and receive zero reward. These models have a wide range of practical applications (e.g., cognitive radio), and some assume *player-dependent heterogeneous* reward distributions [20, 110]; in comparison, collision is not modeled in this work.

Side information.

Models in which learning agents observe side information have also been studied in prior works—one can consider data collected by other players in multi-player bandits as side observations [88]. In some models, a player observes side information for some arms that are not chosen in the current round: stochastic models with such side information are studied in [29, 27, 166], and adversarial models in [105, 5]; Similarities/closeness among arms in one bandit problem are studied in [43, 169, 158]. We note that our problem formulation is different, because in these models, auxiliary data are from arms in the same bandit problem instance instead of from other players.

Upper and lower bounds on the means of reward distributions are used as side information in [130]. Loss predictors [164] can also be considered as side information. In contrast, we do not leverage such information.

Side information can also be encoded using a distance metric; see Section 2.5.2 for a discussion on contextual bandits in similarity spaces [137].

Other multi-player bandit learning topics.

Many other multi-player bandit learning topics have also been explored. For example, Awerbuch and Kleinberg [9], Vial et al. [153] study multi-player models in which some of the players are malicious. Christakopoulou and Banerjee [35] study collaborative bandits with applications such as top-K recommendations. Nonstochastic multi-armed bandit models with communicating agents are studied in [13, 32]. Privacy protection in decentralized exploration is investigated in [52]. We note that, in this chapter, our goal does not align closely with these topics.

A.2 Proof of Claim 2.3

We first restate Example 2.2 and Claim 2.3.

Example 2.2. *For a fixed $\epsilon \in (0, \frac{1}{8})$ and $\delta \leq \epsilon/4$, consider the following Bernoulli MPMAB problem instance: for each $p \in [M]$, $\mu_1^p = \frac{1}{2} + \delta$, $\mu_2^p = \frac{1}{2}$. This is a 0-MPMAB instance, hence an ϵ-MPMAB problem instance. Also, note that ϵ is at least four times larger than the gaps $\Delta_2^p = \delta$.*

Claim 2.3. *For the above example, any sublinear regret algorithm for the ϵ-MPMAB problem must have $\Omega(\frac{M \ln T}{\delta})$ regret on this instance, matching the IND-UCB regret upper bound.*

Proof of Claim 2.3. Suppose $\mathcal{A}$ is a sublinear-regret algorithm for the ϵ-MPMAB problem; i.e., there exist $C > 0$ and $\alpha > 0$ such that $\mathcal{A}$ has $CT^{1-\alpha}$ regret in all ϵ-MPMAB instances.

Recall that we consider the Bernoulli ϵ-MPMAB instance $\mu = (\mu_i^p)_{i\in[2],p\in[M]}$ such that $\mu_1^p = \frac{1}{2} + \delta$ and $\mu_2^p = \frac{1}{2}$ for all p. As $\epsilon \in (0, \frac{1}{8})$ and $\delta \leq \frac{\epsilon}{4}$, it can be directly verified that all μ_i^p's are in $[\frac{15}{32}, \frac{17}{32}]$. In addition, since for all p, $\Delta_2^p = \delta \leq \frac{\epsilon}{4} = 5 \cdot \frac{\epsilon}{20}$, we have $\mathcal{I}_{5\epsilon/4} = \emptyset$, i.e., $\mathcal{I}_{5\epsilon/4}^C = \{1, 2\}$.

From Theorem 2.9, we conclude that for this MPMAB instance μ, $\mathcal{A}$ has regret lower bounded as follows:

$$\mathbb{E}_\mu \left[\mathcal{R}(T)\right] \geq \Omega\left(\frac{M \ln(T^\alpha \Delta_2^p / C)}{\Delta_2^p}\right) = \Omega\left(\frac{M \ln(T^\alpha \delta / C)}{\delta}\right) = \Omega\left(\frac{M \ln T}{\delta}\right),$$

for sufficiently large T. □

A.3 Basic Properties of $\mathcal{I}_{5\epsilon}$ for ϵ-MPMAB Instances

In Section 2.3, we presented the following two facts about properties of $\mathcal{I}_{5\epsilon}$ for ϵ-MPMAB problem instances:

Fact 2.4. *$|\mathcal{I}_{5\epsilon}| \leq K - 1$. In addition, for each arm $i \in \mathcal{I}_{5\epsilon}$, $\Delta_i^{\min} > 3\epsilon$; in other words, for all players p in $[M]$, $\Delta_i^p = \mu_*^p - \mu_i^p > 3\epsilon$; consequently, arm i is suboptimal for all players p in $[M]$.*

Fact 2.6. *For any $i \in \mathcal{I}_{5\epsilon}$, $\frac{1}{\Delta_i^{\min}} \leq \frac{2}{M} \sum_{p\in[M]} \frac{1}{\Delta_i^p}$.*

Here, we will present and prove a more complete collection of facts about the properties of $\mathcal{I}_{5\epsilon}$ which covers every statement in Fact 2.4 and Fact 2.6. Before that, we first prove the following fact.

Fact A.1. *For an ϵ-MPMAB problem instance, for any $i \in [K]$, and $p, q \in [M]$, $|\Delta_i^p - \Delta_i^q| \leq 2\epsilon$.*

Proof. Fix any player $p \in [M]$, let $j \in [K]$ be an optimal arm for p such that $\mu_j^p = \mu_*^p$. We first show that, for any player $q \in [M]$, $|\mu_*^q - \mu_j^p| \leq \epsilon$.

- $\mu_*^q \geq \mu_j^p - \epsilon$ is trivially true because $\mu_j^q \geq \mu_j^p - \epsilon$ by Definition 2.1 and $\mu_*^q \geq \mu_j^q$ by the definition of μ_*^q;

- $\mu_*^q \leq \mu_j^p + \epsilon$ is true because if there exists an arm $k \in [K]$ such that $\mu_k^q > \mu_j^p + \epsilon$, then by Definition 2.1 we must have $\mu_k^p \geq \mu_k^q - \epsilon > \mu_j^p$ which contradicts with the premise that j is an optimal arm for player p.

We have shown that $|\mu_*^q - \mu_*^p| \leq \epsilon$. Since $|\mu_i^q - \mu_i^p| \leq \epsilon$ by Definition 2.1, it follows from the triangle inequality that $|\Delta_i^p - \Delta_i^q| \leq 2\epsilon$. □

We now present a set of basic properties of $\mathcal{I}_{5\epsilon}$.

Fact A.2 (Basic properties of $\mathcal{I}_{5\epsilon}$)**.** *Let $\Delta_i^{\max} = \max_{p\in[M]} \Delta_i^p$. For an ϵ-MPMAB problem instance, for each arm $i \in \mathcal{I}_{5\epsilon}$,*

(a) $\Delta_i^p > 3\epsilon$ for all players $p \in [M]$; in other words, $\Delta_i^{\min} > 3\epsilon$;

(b) arm i is suboptimal for all players $p \in [M]$, i.e., for any player $p \in [M]$, $\mu_i^p < \mu_*^p$;

(c) $\frac{\Delta_i^p}{\Delta_i^q} < 2$ for any pair of players $p, q \in [M]$; consequently, $\frac{\Delta_i^{\max}}{\Delta_i^{\min}} < 2$;

(d) $\frac{1}{\Delta_i^{\min}} \leq \frac{2}{M} \sum_{p\in[M]} \frac{1}{\Delta_i^p}$;

(e) $|\mathcal{I}_{5\epsilon}| \leq K - 1$.

Proof. We prove each item one by one.

(a) For each arm $i \in \mathcal{I}_{5\epsilon}$, by definition, there exists $p \in [M]$, $\Delta_i^p > 5\epsilon$. It follows from Fact A.1 that for any $q \in [M]$, $\Delta_i^q \geq \Delta_i^p - 2\epsilon > 3\epsilon$. $\Delta_i^{\min} > 3\epsilon$ then follows straightforwardly.

(b) For each arm $i \in \mathcal{I}_{5\epsilon}$, it follows from item (a) that for any $p \in [M]$, $\Delta_i^p > 3\epsilon \geq 0$. Therefore, i is suboptimal for all player $p \in [M]$.

(c) By Fact A.1, for any $i \in \mathcal{I}_{5\epsilon} \subseteq [K]$ and any $p, q \in [M]$, $\Delta_i^p \leq \Delta_i^q + 2\epsilon$, which implies $\frac{\Delta_i^p}{\Delta_i^q} \leq 1 + \frac{2\epsilon}{\Delta_i^q}$. Since by item (a), $\Delta_i^q > 3\epsilon$, it follows that $\frac{\Delta_i^p}{\Delta_i^q} \leq 1 + \frac{2\epsilon}{\Delta_i^q} < 2$. $\frac{\Delta_i^{\max}}{\Delta_i^{\min}} < 2$ then follows straightforwardly.

(d) For each arm $i \in \mathcal{I}_{5\epsilon}$, it follows from item (c) that for any $p \in [M]$, $\Delta_i^p \in [\Delta_i^{\min}, 2\Delta_i^{\min}]$. Therefore, we have $\frac{1}{\Delta_i^p} \in [\frac{1}{2\Delta_i^{\min}}, \frac{1}{\Delta_i^{\min}}]$, as $\Delta_i^p > 0$ for all p. It then follows that $\frac{2}{M} \sum_{p\in[M]} \frac{1}{\Delta_i^p} \geq \frac{2}{M} \sum_{p\in[M]} \frac{1}{2\Delta_i^{\min}} = \frac{1}{\Delta_i^{\min}}$.

(e) Pick an arm i that is optimal with respect to player 1; i cannot be in $\mathcal{I}_{5\epsilon}$ because of item (b). Therefore, $\mathcal{I}_{5\epsilon} \subseteq [K] \setminus \{i\}$, which implies that it has size at most $K - 1$. □

A.4 Proof of Upper Bounds in Section 2.3

A.4.1 Proof overview

In Appendix A.4.2 and A.4.3, we focus on showing that in a "clean" event $\mathcal{E}$ (defined in A.4.3), the upper confidence bound $\text{UCB}_i^p(t) = \kappa_i^p(t, \lambda) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$ (line 10 of Algorithm 1)[2] holds for every $t \in [T], i \in [K], p \in [M]$ and $\lambda \in [0, 1]$; and the "clean" event $\mathcal{E}$ occurs with $1 - 4MK/T^4$ probability.

Then, in Appendix A.4.4, we provide a proof of the gap-dependent upper bound in Theorem 2.5. In Appendix A.4.5, we provide a proof of the gap-independent upper bound in Theorem 2.8.

A.4.2 Event $\mathcal{Q}_i(t)$

Recall that $n_i^p(t-1)$ is the number of pulls of arm i by player p after the first $(t-1)$ rounds. Let $m_i^p(t-1) = \sum_{q\in[M]:q\neq p} n_i^q(t-1)$.

We now define the following event.

[2]Recall that $\overline{z} = \max\{z, 1\}$.

Definition A.3. *Let*

$$\mathcal{Q}_i(t) = \left\{ \forall p, \left|\zeta_i^p(t) - \mu_i^p\right| \le 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}, \left|\eta_i^p(t) - \sum_{q\neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \le 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}} \right\},$$

where

$$\zeta_i^p(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_t^p = i\} r_t^p}{\overline{n_i^p}(t-1)},$$

and

$$\eta_i^p(t) = \frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}\{q \neq p,\ i_t^q = i\} r_t^q}{\overline{m_i^p}(t-1)}.$$

Lemma A.4.

$$\Pr(\mathcal{Q}_i(t)) \ge 1 - 4MT^{-5}.$$

Proof. For any fixed player p, we discuss the two inequalities separately. Lemma A.4 then follows by a union bound over the two inequalities and over all $p \in [M]$.

We first discuss the concentration of $\zeta_i^p(t)$. We define a filtration $\{\mathcal{B}_t\}_{t=1}^T$, where

$$\mathcal{B}_t = \sigma(\left\{i_s^{p'}, r_s^{p'} : s \in [t], p' \in [M]\right\} \cup \left\{i_{t+1}^{p'} : p' \in [M]\right\})$$

is the σ-algebra generated by the historical interactions up to round t and the arm selection of all players at round $t+1$.

Let random variable $X_t = \mathbb{1}\{i_t^p = i\}\left(r_t^p - \mu_i^p\right)$. We have $\mathbb{E}\left[X_t \mid \mathcal{B}_{t-1}\right] = 0$; in addition, $\operatorname{var}\left[X_t \mid \mathcal{B}_{t-1}\right] = \mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid \mathcal{B}_{t-1}])^2 \mid \mathcal{B}_{t-1}\right] \le \mathbb{E}\left[(\mathbb{1}\{i_t^p = i\} r_t^p)^2 \mid \mathcal{B}_{t-1}\right] \le \mathbb{1}\{i_t^p = i\}$ and $|X_t| \le 1$.

Applying Freedman's inequality [14, Lemma 2] with $\sigma = \sqrt{\sum_{s=1}^{t-1} \operatorname{var}\left[X_s \mid \mathcal{B}_{s-1}\right]}$ and $b = 1$, and using $\sigma \le \sqrt{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\}}$, we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1} X_s\right| \le 4\sqrt{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\} \cdot \ln(T^5 \log_2 T) + 2\ln(T^5 \log_2 T)}. \tag{A.1}$$

We consider two cases:

1. If $n_i^p(t-1) = \sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\} = 0$, we have $\overline{n_i^p}(t-1) = 1$ and $\zeta_i^p(t) = 0$. In this case, we trivially have
$$\left|\zeta_i^p(t) - \mu_i^p\right| \leq 1 \leq 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}.$$

2. Otherwise, $n_i^p(t-1) \geq 1$. In this case, we have $\overline{n_i^p}(t-1) = n_i^p(t-1)$. Divide both sides of Eq. (A.1) by $n_i^p(t-1)$, and use the fact that $\log T \leq T$, we have
$$\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\} r_s^p}{n_i^p(t-1)} - \mu_i^p\right| \leq 4\sqrt{\frac{6\ln T}{n_i^p(t-1)}} + \frac{12\ln T}{n_i^p(t-1)}.$$
If $\frac{12\ln T}{n_i^p(t-1)} \geq 1$, $\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p=i\} r_s^p}{n_i^p(t-1)} - \mu_i^p\right| \leq 8\sqrt{\frac{3\ln T}{n_i^p(t-1)}}$ is trivially true. Otherwise, $\frac{12\ln T}{n_i^p(t-1)} \leq 2\sqrt{\frac{3\ln T}{n_i^p(t-1)}}$, which implies that $\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p=i\} r_s^p}{n_i^p(t-1)} - \mu_i^p\right| \leq (4\sqrt{6}+2\sqrt{3})\sqrt{\frac{\ln T}{n_i^p(t-1)}} \leq 8\sqrt{\frac{3\ln T}{n_i^p(t-1)}}$.

In summary, in both cases, with probability at least $1 - 2T^{-5}$, we have
$$\left|\zeta_i^p(t-1) - \mu_i^p\right| \leq 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}.$$

A similar application of Freedman's inequality also shows the concentration of $\eta_i^p(t)$. Similarly, we define a filtration $\{\mathcal{G}_{t,q}\}_{t\in[T],q\in[M]}$, where
$$\mathcal{G}_{t,q} = \begin{cases} \sigma\left(\left\{i_s^{p'}, r_s^{p'} : s \in [t-1], p' \in [M]\right\} \bigcup \left\{i_t^{p'}, r_t^{p'} : p' \in [M], p' \leq q\right\} \bigcup \left\{i_t^{q+1}\right\}\right), & q < M; \\ \sigma\left(\left\{i_s^{p'}, r_s^{p'} : s \in [t], p' \in [M]\right\} \bigcup \left\{i_{t+1}^1\right\}\right), & q = M. \end{cases}$$
is the σ-algebra generated by (1) the historical interactions up until round $t-1$, (2) the arm selections and observed rewards of players up until player q in round t, and (3) the

arm selection of the next player. We have

$$\mathcal{G}_{1,1} \subset \mathcal{G}_{1,2} \subset \ldots \subset \mathcal{G}_{1,M} \subset \mathcal{G}_{2,1} \subset \ldots \subset \mathcal{G}_{2,M} \subset \ldots \subset \mathcal{G}_{T,M}.$$

By convention, let $\mathcal{G}_{t,0} = \mathcal{G}_{t-1,M}$.

Now, let $Y_{t,q} = \mathbb{1}\{q \neq p,\ i_t^q = i\}\left(r_t^q - \mu_i^q\right)$. We have $\mathbb{E}\left[Y_{t,q} \mid \mathcal{G}_{t,q-1}\right] = 0$; in addition, $\operatorname{var}\left[Y_{t,q} \mid \mathcal{G}_{t,q-1}\right] = \mathbb{E}\left[Y_{t,q}^2 \mid \mathcal{G}_{t,q-1}\right] \leq \mathbb{1}\{q \neq p,\ i_t^q = i\}$, and $\left|Y_{t,q}\right| \leq 1$.

Similarly, applying Freedman's inequality [14, Lemma 2] with
$\sigma = \sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \operatorname{var}\left[Y_{s,q} \mid \mathcal{G}_{s,q-1}\right]}$ and $b = 1$, and using
$\sigma \leq \sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}\{q \neq p, i_s^q = i\}}$, we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1}\sum_{q=1}^{M} Y_{s,q}\right| \leq 4\sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}\{q \neq p,\ i_s^q = i\} \cdot \ln(T^5 \log_2(TM)) + 2\ln(T^5 \log_2(TM))}. \quad \text{(A.2)}$$

Again, we consider two cases. If $m_i^p(t-1) = 0$, then we have $\eta_i^p(t-1) = 0$ and

$$\left|\eta_i^p(t-1) - \sum_{q \neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| = 0 \leq 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}}.$$

Otherwise, we have $\overline{m_i^p}(t-1) = m_i^p(t-1)$. Divide both sides of Eq. (A.2) by $m_i^p(t-1)$, and use the fact that $\log_2(TM) \leq T^2$, we have

$$\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}\{q \neq p,\ i_s^q = i\} r_s^q}{m_i^p(t-1)} - \sum_{q \neq p} \frac{n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{7\ln T}{m_i^p(t-1)}} + \frac{14\ln T}{m_i^p(t-1)}.$$

If $\frac{14\ln T}{m_i^p(t-1)} \geq 1$, $\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}\{q\neq p,\ i_s^q=i\} r_s^q}{m_i^p(t-1)} - \sum_{q\neq p} \frac{n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\ln T}{m_i^p(t-1)}}$ is trivially true.

Otherwise, $\frac{14\ln T}{m_i^p(t-1)} \leq \sqrt{\frac{14\ln T}{m_i^p(t-1)}}$, which implies that

$$\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q\neq p,\ i_s^q=i\}r_s^q}{m_i^p(t-1)}-\sum_{q\neq p}\frac{n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq (4\sqrt{7}+\sqrt{14})\sqrt{\frac{\ln T}{m_i^p(t-1)}} \leq 4\sqrt{\frac{14\ln T}{m_i^p(t-1)}}.$$

In summary, in both cases, with probability at least $1-2T^{-5}$, we have

$$\left|\eta_i^p(t-1)-\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}}.$$

The lemma follows by taking a union bound over these two inequalities for each fixed p, and over all $p\in[M]$. □

A.4.3 Event $\mathcal{E}$

Let $\mathcal{E}=\cap_{t=1}^{T}\cap_{i=1}^{K}\mathcal{Q}_i(t)$. We present the following corollary and lemma regarding event $\mathcal{E}$.

Corollary A.5. *It follows from Lemma A.4 that* $\Pr[\mathcal{E}]\geq 1-\frac{4MK}{T^4}$.

Lemma A.6. *If $\mathcal{E}$ occurs, we have that for every $t\in[T]$, $i\in[K]$, $p\in[M]$, for all $\lambda\in[0,1]$,*

$$\left|\kappa_i^p(t,\lambda)-\mu_i^p\right| \leq 8\sqrt{13\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)}+\frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)}+(1-\lambda)\epsilon,$$

where $\kappa_i^p(t,\lambda)=\lambda\zeta_i^p(t)+(1-\lambda)\eta_i^p(t)$.

Proof. If $\mathcal{E}$ occurs, for every $t\in[T]$ and $i\in[K]$, by the definition of event $\mathcal{Q}_i(t)$, we have

$$\left|\zeta_i^p(t)-\mu_i^p\right|<8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}},\text{and }\left|\eta_i^p(t)-\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}}.$$

As $\kappa_i^p(t,\lambda) = \lambda\zeta_i^p(t) + (1-\lambda)\eta_i^p(t)$, we have:

$$\begin{aligned}\left|\kappa_i^p(t,\lambda) - \left[\lambda\mu_i^p + (1-\lambda)\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right]\right| \leq & 8\lambda\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}} + 4(1-\lambda)\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}} \\ \leq & 8\sqrt{13\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)},\end{aligned} \tag{A.3}$$

where the second inequality uses the elementary facts that $\sqrt{A}+\sqrt{B} \leq \sqrt{2(A+B)}$.

Furthermore, from Definition 2.1, we have

$$\left|\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q - \mu_i^p\right| \leq \sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\left|\mu_i^q - \mu_i^p\right| \leq \epsilon.$$

This shows that

$$\left|\mu_i^p - (\lambda\mu_i^p + (1-\lambda)\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q)\right| \leq (1-\lambda)\epsilon.$$

Combining the above inequality with Eq. (A.3), we get

$$\left|\kappa_i^p(t,\lambda) - \mu_i^p\right| \leq 8\sqrt{13\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)} + (1-\lambda)\epsilon.$$

This completes the proof. □

A.4.4 Proof of Theorem 2.5

We first restate Theorem 2.5.

Theorem 2.5. *Let* RobustAgg(ϵ) *run on an ϵ-MPMAB problem instance for T rounds.*

Then, its expected collective regret satisfies

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathcal{O}\left(\sum_{i \in \mathcal{I}_{5\epsilon}} \left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min} \right) + \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M]: \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} \right).$$

Recall that the expected collective regret is defined as $\mathbb{E}[\mathcal{R}(T)] = \sum_{i \in [K]} \sum_{p \in [M]} \Delta_i^p \cdot \mathbb{E}[n_i^p(T)]$. Before we prove Theorem 2.5, we first present the following two lemmas, which provides an upper bound for (1) the total number of arm pulls for arm i, for i in $\mathcal{I}_{5\epsilon}$ and (2) the individual number of arm pulls for arm i and player p, for i in $\mathcal{I}_{5\epsilon}^C$, conditioned on $\mathcal{E}$ happening.

Lemma A.7. *Denote $n_i(T) = \sum_{p \in [M]} n_i^p(T)$ as the total number of pulls of arm i by all the players after T rounds. Let RobustAgg(ϵ) run on an ϵ-MPMAB problem instance for T rounds. Then, for each $i \in \mathcal{I}_{5\epsilon}$, we have*

$$\mathbb{E}[n_i(T)|\mathcal{E}] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M \right).$$

Lemma A.8. *Let RobustAgg(ϵ) run on an ϵ-MPMAB problem instance for T rounds. Then, for each $i \in \mathcal{I}_{5\epsilon}^C$ and player $p \in [M]$ such that $\Delta_i^p > 0$, we have*

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} \right).$$

Proof of Lemma A.7. We first note that it follows from item (b) of Fact A.2 that every arm $i \in \mathcal{I}_{5\epsilon}$ is suboptimal for all players $p \in [M]$.

We have

$$\begin{aligned} n_i(T) &= \sum_{t=1}^{T} \sum_{p=1}^{M} \mathbb{1}\{i_t^p = i\} \\ &\leq M + \tau + \sum_{t=1}^{T} \sum_{p=1}^{M} \mathbb{1}\{i_t^p = i, n_i(t-1) > \tau\}. \end{aligned} \tag{A.4}$$

Here, $\tau \geq 1$ is an arbitrary integer. The term M is due to parallel arm pulls in the ϵ-MPMAB problem: Let s be the first round such that after round s, the total number of pulls $n_i(s) > \tau$. This implies that $n_i(s-1) \leq \tau$. Then in round s, there can be up to M pulls of arm i by all the players, which means that in round $(s+1)$ when the third term in Eq. (A.4) can first start counting, there could have been up to $\tau + M$ pulls of the arm i.

It then follows that

$$n_i(T) \leq M + \tau + \sum_{t=1}^{T} \sum_{p=1}^{M} \mathbb{1}\left\{\mathrm{UCB}_{i_*^p}^p(t) \leq \mathrm{UCB}_i^p(t),\ n_i(t-1) > \tau\right\}. \tag{A.5}$$

Recall that $\Delta_i^{\min} = \min_p \Delta_i^p$, and for each $i \in \mathcal{I}_{5\epsilon}$, we have $\Delta_i^p \geq \Delta_i^{\min} > 3\epsilon$ by item (a) of Fact A.2.

With foresight, we choose $\tau = \lceil \frac{3328 \ln T}{(\Delta_i^{\min} - 2\epsilon)^2} \rceil$. Conditional on $\mathcal{E}$, we show that, *for any arm* $i \in \mathcal{I}_{5\epsilon}$, the event $\left\{\mathrm{UCB}_{i_*^p}^p(t) \leq \mathrm{UCB}_i^p(t), n_i(t-1) > \tau\right\}$ never happens. It suffices to show that if $n_i(t-1) > \tau$,

$$\mathrm{UCB}_{i_*^p}^p(t) \geq \mu_*^p, \tag{A.6}$$

and

$$\mathrm{UCB}_i^p(t) < \mu_*^p \tag{A.7}$$

happen simultaneously.

Eq. (A.6) follows straightforwardly from the definition of $\mathcal{E}$ along with Lemma A.6.

For Eq. (A.7), we have the following upper bound on $\mathrm{UCB}_i^p(t)$:

$$\begin{aligned}
\mathrm{UCB}_i^p(t) &= \kappa_i^p(t,\lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\
&\le \mu_i^p + 2F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\
&= \mu_i^p + 2\Big[\min_{\lambda\in[0,1]} 8\sqrt{13\ln T[\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon\Big] \\
&\le \mu_i^p + 2\Big[8\sqrt{\frac{13\ln T}{\overline{n_i^p}(t-1) + \overline{m_i^p}(t-1)}} + \epsilon\Big] \\
&\le \mu_i^p + 2\Big[8\sqrt{\frac{13\ln T}{n_i(t-1)}} + \epsilon\Big] \\
&< \mu_i^p + 2\Big[8\sqrt{\frac{13\ln T(\Delta_i^p - 2\epsilon)^2}{3328\ln T}} + \epsilon\Big] = \mu_i^p + \Delta_i^p = \mu_*^p,
\end{aligned}$$

where the first inequality is from the definition of $\mathcal{E}$ and Lemma A.6; the second inequality is from choosing $\lambda = \frac{\overline{n_i^p}(t-1)}{\overline{n_i^p}(t-1)+\overline{m_i^p}(t-1)}$; the third inequality is from the simple facts that $n_i^p(t-1) \le \overline{n_i^p}(t-1)$, $m_i^p(t-1) \le \overline{m_i^p}(t-1)$, and $n_i(t-1) = n_i^p(t-1) + m_i^p(t-1)$; the last inequality is from the premise that $n_i(t-1) > \tau \ge \frac{3328\ln T}{(\Delta_i^{\min}-2\epsilon)^2} \ge \frac{3328\ln T}{(\Delta_i^p-2\epsilon)^2}$.

Continuing Eq. (A.5), it then follows that, for each $i \in \mathcal{I}_{5\epsilon}$,

$$\mathbb{E}[n_i(T)|\mathcal{E}] \le \lceil \frac{3328\ln T}{(\Delta_i^{\min}-2\epsilon)^2} \rceil + M \le \frac{3328\ln T}{(\Delta_i^{\min}-2\epsilon)^2} + (M+1). \tag{A.8}$$

Now, by item (a) of Fact A.2, for each $i \in \mathcal{I}_{5\epsilon}$, $\Delta_i^{\min} > 3\epsilon$. We then have $\frac{\Delta_i^{\min}}{\Delta_i^{\min}-2\epsilon} = \frac{\Delta_i^{\min}-2\epsilon+2\epsilon}{\Delta_i^{\min}-2\epsilon} = 1 + \frac{2\epsilon}{\Delta_i^{\min}-2\epsilon} < 3$. It follows that

$$\frac{3328\ln T}{(\Delta_i^{\min}-2\epsilon)^2} = \frac{3328\ln T}{(\Delta_i^{\min})^2} \cdot \left(\frac{\Delta_i^{\min}}{\Delta_i^{\min}-2\epsilon}\right)^2 < \frac{29952\ln T}{(\Delta_i^{\min})^2}.$$

Therefore, continuing Eq. (A.8), for each $i \in \mathcal{I}_{5\epsilon}$, we have

$$\mathbb{E}[n_i(T)|\mathcal{E}] < \frac{29952 \ln T}{(\Delta_i^{\min})^2} + (M+1) \leq \frac{29952 \ln T}{(\Delta_i^{\min})^2} + 2M.$$

where the second inequality follows from the fact that $M \geq 1$.

It then follows that

$$\mathbb{E}[n_i(T)|\mathcal{E}] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

This completes the proof of Lemma A.7. □

Proof of Lemma A.8. Let's now turn our attention to arms in $\mathcal{I}_{5\epsilon}^C = [K] \setminus \mathcal{I}_{5\epsilon}$. For each arm $i \in \mathcal{I}_{5\epsilon}^C$ and for each player $p \in [M]$ such that $\mu_i^p < \mu_*^p$, we seek to bound the expected number of pulls of arm i by p in T rounds, under the assumption that the event $\mathcal{E}$ occurs. Since the optimal arm(s) may be different for different players, we treat each player separately.

Fix a player $p \in [M]$ and a suboptimal arm $i \in \mathcal{I}_{5\epsilon}^C$ such that $\Delta_i^p > 0$. Recall that $n_i^p(t-1)$ is the number of pulls of arm i by player p after $(t-1)$ rounds. We have

$$\begin{aligned} n_i^p(T) &= \sum_{t=1}^{T} \mathbb{1}\{i_t^p = i\} \\ &\leq \tau + \sum_{t=\tau+1}^{T} \mathbb{1}\{i_t^p = i, n_i^p(t-1) > \tau\}, \end{aligned} \tag{A.9}$$

where $\tau \geq 1$ is an arbitrary integer. It then follows that

$$n_i^p(T) \leq \tau + \sum_{t=\tau+1}^{T} \mathbb{1}\{\mathrm{UCB}_{i_*^p}^p(t) \leq \mathrm{UCB}_i^p(t), n_i^p(t-1) > \tau\}.$$

With foresight, let $\tau = \lceil \frac{3328 \ln T}{(\Delta_i^p)^2} \rceil$. Conditional on $\mathcal{E}$, we show that, *for any* $i \in \mathcal{I}_{5\epsilon}^C$ *such that* $\Delta_i^p > 0$, the event $\{\mathrm{UCB}_{i_*^p}^p(t) \leq \mathrm{UCB}_i^p(t), n_i^p(t-1) > \tau\}$ never happens. It

suffices to show that if $n_i^p(t-1) > \tau$,

$$\mathrm{UCB}_{i_*^p}^p(t) \geq \mu_*^p, \tag{A.10}$$

and

$$\mathrm{UCB}_i^p(t) < \mu_*^p \tag{A.11}$$

happen simultaneously.

Eq. (A.10) follows straightforwardly from the definition of $\mathcal{E}$ along with Lemma A.6. For Eq. (A.11), we have the following upper bound on $\mathrm{UCB}_i^p(t)$:

$$\begin{aligned}
\mathrm{UCB}_i^p(t) &= \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\
&\leq \mu_i^p + 2F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\
&= \mu_i^p + 2\Big[\min_{\lambda\in[0,1]} 8\sqrt{13\ln T[\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon\Big] \\
&\leq \mu_i^p + 2\Big[8\sqrt{\frac{13\ln T}{\overline{n_i^p}(t-1)}}\Big] \\
&\leq \mu_i^p + 2\Big[8\sqrt{\frac{13\ln T}{n_i^p(t-1)}}\Big] \\
&< \mu_i^p + 2\Big[8\sqrt{\frac{13\ln T(\Delta_i^p)^2}{3328\ln T}}\Big] = \mu_i^p + \Delta_i^p = \mu_*^p,
\end{aligned}$$

where the first inequality is from the definition of event $\mathcal{E}$ and Lemma A.6; the second inequality is from choosing $\lambda = 1$; the third inequality uses the basic fact that $n_i^p(t-1) \leq \overline{n_i^p}(t-1)$; the fourth inequality is by our premise that $n_i^p(t-1) > \tau \geq \frac{3328\ln T}{(\Delta_i^p)^2}$.

It follows that conditional on $\mathcal{E}$, the second term in Eq. (A.9) is always zero, i.e., player p would not pull arm i again. Therefore, for any $i \in \mathcal{I}_{5\epsilon}^C$ such that $\Delta_i^p > 0$, we have

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq \lceil\frac{3328\ln T}{(\Delta_i^p)^2}\rceil \leq \frac{3328\ln T}{(\Delta_i^p)^2} + 1 \leq \frac{3328\ln T}{(\Delta_i^p)^2}\cdot 2 = \frac{6656\ln T}{(\Delta_i^p)^2}. \tag{A.12}$$

It then follows that

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

This completes the proof of Lemma A.8. □

Proof of Theorem 2.5.

We now prove Theorem 2.5.

Proof. We have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathbb{E}[\mathcal{R}(T)|\overline{\mathcal{E}}]\Pr[\overline{\mathcal{E}}] \\ &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + (TM)\frac{4MK}{T^4} \\ &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathcal{O}(1) \end{aligned} \tag{A.13}$$

where the second inequality uses the fact that $\mathbb{E}[\mathcal{R}(T)|\overline{\mathcal{E}}] \leq TM$, as the instantaneous regret for each player in each round is bounded by 1; and the last inequality follows under the premise that $T > \max(M, K)$.

Let $\Delta_i^{\max} = \max_p \Delta_i^p$. We have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{i\in[K]}\sum_{p\in[M]} \mathbb{E}[n_i^p(T)|\mathcal{E}] \cdot \Delta_i^p \\ &\leq \sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^{\max} + \sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}] \cdot \Delta_i^p, \end{aligned} \tag{A.14}$$

where the inequality holds because the instantaneous regret for any arm i and any player p is bounded by $\Delta_i^{\max}$.

Now, it follows from Lemma A.7 that there exists some constant $C_1 > 0$ such that for each $i \in \mathcal{I}_{5\epsilon}$,

$$\mathbb{E}[n_i(T)|\mathcal{E}] \leq C_1\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right),$$

and it follows from Lemma A.8 that there exists some constant $C_2 > 0$ such that for each $i \in \mathcal{I}_{5\epsilon}^C$ and $p \in [M]$ with $\Delta_i^p > 0$,

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq C_2 \left(\frac{\ln T}{(\Delta_i^p)^2} \right).$$

Then, continuing Eq. (A.14), we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\leq \sum_{i \in \mathcal{I}_{5\epsilon}} C_1 \left(\frac{\ln T}{(\Delta_i^{\min})^2} + M \right) \cdot \Delta_i^{\max} + \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M]: \Delta_i^p > 0} C_2 \left(\frac{\ln T}{(\Delta_i^p)^2} \right) \cdot \Delta_i^p \\
&\leq 2C_1 \sum_{i \in \mathcal{I}_{5\epsilon}} \left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min} \right) + C_2 \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M]: \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p},
\end{aligned}$$

where the second inequality follows from item (c) of Fact A.2 which states that $\forall i \in \mathcal{I}_{5\epsilon}, \Delta_i^{\max} < 2\Delta_i^{\min}$.

It then follows from Eq. (A.13) that

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T)] &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathcal{O}(1) \\
&\leq \mathcal{O}\left(\sum_{i \in \mathcal{I}_{5\epsilon}} \left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min} \right) + \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M]: \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} \right),
\end{aligned}$$

This completes the proof of Theorem 2.5. □

A.4.5 Proof of Theorem 2.8

We first restate Theorem 2.8.

Theorem 2.8. *Let* RobustAgg(ϵ) *run on an ϵ-MPMAB problem instance for T rounds. Then its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{5\epsilon}|\, MT} + M\sqrt{(|\mathcal{I}_{5\epsilon}^C| - 1)T} + M\,|\mathcal{I}_{5\epsilon}| \right).$$

Proof. From the earlier proof of Theorem 2.5, we have

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathcal{O}(1). \tag{A.15}$$

Recall that $\Delta_i^{\max} = \max_p \Delta_i^p$. We also have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{i\in[K]} \sum_{p\in[M]} \mathbb{E}[n_i^p(T)|\mathcal{E}] \cdot \Delta_i^p \\ &\leq \sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^{\max} + \sum_{i\in\mathcal{I}_{5\epsilon}^C} \sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}] \cdot \Delta_i^p \end{aligned} \tag{A.16}$$

Again, it follows from Lemma A.7 that there exists some constant $C_1 > 0$ such that for each $i \in \mathcal{I}_{5\epsilon}$,

$$\mathbb{E}[n_i(T)|\mathcal{E}] \leq C_1 \left(\frac{\ln T}{(\Delta_i^{\min})^2} + M \right), \tag{A.17}$$

and it follows from Lemma A.8 that there exists some constant $C_2 > 0$ such that for each $i \in \mathcal{I}_{5\epsilon}^C$ and $p \in [M]$ with $\Delta_i^p > 0$,

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq C_2 \left(\frac{\ln T}{(\Delta_i^p)^2} \right). \tag{A.18}$$

Now let us bound the two terms in Eq. (A.16) separately, using the technique from [90, Theorem 7.2].

For the first term, with foresight, let us set $\delta_1 = \sqrt{\frac{C_1|\mathcal{I}_{5\epsilon}|\ln T}{MT}}$. If $|\mathcal{I}_{5\epsilon}| = 0$, we have $\sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^{\max} = 0$ trivially. Otherwise, $\delta_1 > 0$ because $T > \max(M, K) \geq 1$.

Then, we have

$$
\begin{aligned}
&\sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} \\
&\le 2\sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\min} \\
&\le 2\left(\sum_{i\in\mathcal{I}_{5\epsilon}:\Delta_i^{\min}\in(0,\delta_1)} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\min} + \sum_{i\in\mathcal{I}_{5\epsilon}:\Delta_i^{\min}\in[\delta_1,1]} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\min}\right) \\
&\le 2\left(MT\delta_1 + \sum_{i\in\mathcal{I}_{5\epsilon}:\Delta_i^{\min}\in[\delta_1,1]} C_1\left(\frac{\ln T}{(\Delta_i^{\min})^2}+M\right)\Delta_i^{\min}\right) \\
&\le 2\left(MT\delta_1 + \sum_{i\in\mathcal{I}_{5\epsilon}:\Delta_i^{\min}\in[\delta_1,1]} \frac{C_1\ln T}{\Delta_i^{\min}} + C_1\sum_{i\in\mathcal{I}_{5\epsilon}:\Delta_i^{\min}\in[\delta_1,1]} M\Delta_i^{\min}\right) \\
&\le 2\left(MT\delta_1 + \frac{C_1|\mathcal{I}_{5\epsilon}|\ln T}{\delta_1} + C_1\sum_{i\in\mathcal{I}_{5\epsilon}} M\Delta_i^{\min}\right) \\
&\le 4\sqrt{C_1|\mathcal{I}_{5\epsilon}|MT\ln T} + 2C_1M|\mathcal{I}_{5\epsilon}|, \qquad \text{(A.19)}
\end{aligned}
$$

where the first inequality follows from item (c) of Fact A.2; the third inequality follows from Eq. (A.17) and the fact that $\sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}] \le MT$ as M players each pulls one arm in each of T rounds; and the last inequality follows from our premise that $\delta_1 = \sqrt{\frac{C_1|\mathcal{I}_{5\epsilon}|\ln T}{MT}}$.

For the second term, we consider two cases:

Case 1: $|\mathcal{I}_{5\epsilon}^C| = 1$**.**

In this case, as we have discussed in the paper, $\mathcal{I}_{5\epsilon}^C$ is a singleton set $\{i_*\}$ where arm i_* is optimal for all players p; that is, $\Delta_{i_*}^p = 0$ for all $p \in [M]$. We therefore have

$$
\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p = 0 = 4M\sqrt{C_2\big(\big|\mathcal{I}_{5\epsilon}^C\big|-1\big)T\ln T}. \qquad \text{(A.20)}
$$

Case 2: $|\mathcal{I}_{5\epsilon}^C| \geq 2$**.**

With foresight, let us set $\delta_2 = \sqrt{\frac{C_2|\mathcal{I}_{5\epsilon}^C| \ln T}{T}}$.

$$
\begin{aligned}
&\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \\
\leq &\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p\in(0,\delta_2)} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p + \sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p\in[\delta_2,1]} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \\
\leq &MT\delta_2 + \sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p\in[\delta_2,1]} C_2\left(\frac{\ln T}{(\Delta_i^p)^2}\right)\Delta_i^p \\
\leq &MT\delta_2 + \sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p\in[\delta_2,1]} \left(\frac{C_2\ln T}{\Delta_i^p}\right) \\
\leq &MT\delta_2 + \frac{C_2 M|\mathcal{I}_{5\epsilon}^C|\ln T}{\delta_2} \\
\leq &4M\sqrt{C_2(|\mathcal{I}_{5\epsilon}^C|-1)T\ln T},
\end{aligned} \tag{A.21}
$$

where the second inequality follows from Eq. (A.18) and the fact that $\sum_{i\in\mathcal{I}_{5\epsilon}^C}\mathbb{E}[n_i(T)|\mathcal{E}] \leq MT$ as M players each pulls one arm in each of T rounds; and the last inequality follows from our premise that $\delta_2 = \sqrt{\frac{C_2|\mathcal{I}_{5\epsilon}^C|\ln T}{T}}$ and $|\mathcal{I}_{5\epsilon}^C| \leq 2(|\mathcal{I}_{5\epsilon}^C|-1)$.

In summary, from Eqs. (A.20) and (A.21), we have in both cases,

$$
\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \leq 4M\sqrt{C_2(|\mathcal{I}_{5\epsilon}^C|-1)T\ln T}. \tag{A.22}
$$

Combining Eq. (A.19) and Eq. (A.22), we have

$$
\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq 4\sqrt{C_1|\mathcal{I}_{5\epsilon}|MT\ln T} + 2C_1M|\mathcal{I}_{5\epsilon}| + 4M\sqrt{C_2(|\mathcal{I}_{5\epsilon}^C|-1)T\ln T}
$$

It then follows from Eq. (A.15) that

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\leq 4\sqrt{C_1|\mathcal{I}_{5\epsilon}|MT\ln T} + 4M\sqrt{C_2(|\mathcal{I}_{5\epsilon}^C|-1)T\ln T} + 2C_1M|\mathcal{I}_{5\epsilon}| + \mathcal{O}(1)\\ &\leq \mathcal{O}\left(\sqrt{|\mathcal{I}_{5\epsilon}|\,MT\ln T} + M\sqrt{(|\mathcal{I}_{5\epsilon}^C|-1)T\ln T} + M\,|\mathcal{I}_{5\epsilon}|\right)\\ &\leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{5\epsilon}|\,MT} + M\sqrt{(|\mathcal{I}_{5\epsilon}^C|-1)T} + M\,|\mathcal{I}_{5\epsilon}|\right).\end{aligned}$$

This completes the proof of Theorem 2.8. □

A.5 Proof of the Lower Bounds

A.5.1 Gap-independent lower bound with known ϵ

We first restate Theorem 2.10.

Theorem 2.10. *For any $K \geq 2, M, T \in \mathbb{N}$ such that $T \geq K$, and l, l^C in $\mathbb{N}$ such that $l \leq K-1, l+l^C = K$, there exists some $\epsilon > 0$, such that for any algorithm $\mathcal{A}$, there exists an ϵ-MPMAB problem instance, in which $|\mathcal{I}_{5\epsilon}| \geq l$, and $\mathcal{A}$ has a collective regret at least $\Omega(M\sqrt{(l^C-1)T} + \sqrt{MlT})$.*

Proof. Fix algorithm $\mathcal{A}$. We consider two cases regarding the comparison between l and $M(l^C-1)$.

Case 1: $l > M(l^C-1)$.

To simplify notations, define $\Delta = \sqrt{\frac{l+1}{24MT}}$. Observe that $\Delta \leq \frac{1}{4}$ as $T \geq K \geq l+1$. We will set $\epsilon = 0$.

We will now define l different Bernoulli ϵ-MPMAB instances, and show that under at least one of them, $\mathcal{A}$ will have a collective regret at least $\frac{1}{96}\sqrt{MlT} \geq \frac{1}{192}(M\sqrt{(l^C-1)T} + \sqrt{MlT})$.

For j in $[l+1]$, define a Bernoulli MPMAB instance E_j to be such that for all players p in $[M]$, the expected reward of arm i,

$$\mu_i^p = \begin{cases} \frac{1}{2} + \Delta & i = j \\ \frac{1}{2} & i \in [l+1] \setminus \{j\} \\ 0 & i \notin [l+1] \end{cases}.$$

We first verify that for every instance E_j, it (1) is an ϵ-MPMAB instance, and (2) $[l+1] \setminus \{j\} \subseteq \mathcal{I}_{5\epsilon}$ and therefore $\mathcal{I}_{5\epsilon}$ has size $\geq l$:

1. For item (1), observe that for any fixed i, we have μ_i^p share the same value across all player p's. Therefore, the is trivially ϵ-dissimilar.

2. For item (2), for all i in $[l+1] \setminus \{j\}$, we have $\Delta_i^p = \Delta > 5\epsilon = 0$ for all p; this implies that $[l+1] \setminus \{j\}$ is a subset of $\mathcal{I}_{5\epsilon}$.

We will now argue that

$$\mathbb{E}_{j \sim \text{Unif}([l+1])} \mathbb{E}_{E_j} \left[\mathcal{R}(T) \right] \geq \frac{1}{96} \sqrt{MlT}. \tag{A.23}$$

To this end, it suffices to show

$$\mathbb{E}_{j \sim \text{Unif}([l+1])} \mathbb{E}_{E_j} \left[MT - n_j(T) \right] \geq \frac{MT}{4}. \tag{A.24}$$

To see why Eq. (A.24) implies Eq. (A.23), recall that under instance E_j, j is the optimal arm for all players. In this instance, $\mathcal{R}(T) = \sum_{i \neq j} n_i(T) \Delta_i^1$. As under E_j, for all $i \neq j$ and all p, $\Delta_i^1 \geq \frac{\Delta}{4}$, we have $\mathcal{R}(T) \geq \frac{\Delta}{4} \cdot (MT - n_j(T))$. Eq. (A.23) follows from combining this inequality with Eq. (A.24), along with some algebra.

We now come back to the proof of Eq. (A.24). First, we define a helper instance

E_0, such that for all players p in $[M]$, the expected reward of arm i is defined as:

$$\mu_i^p = \begin{cases} \frac{1}{2} & i \in [l+1] \\ 0 & i \notin [l+1] \end{cases}$$

In addition, for all i in $\{0\} \cup [l+1]$, define $\mathbb{P}_i$ as the joint distribution of the interaction logs (arm pulls and rewards) for all M players over a horizon of T; furthermore, denote by $\mathbb{E}_i$ expectation with respect to $\mathbb{P}_i$.

For every i in $[l+1]$, we have

$$\begin{aligned} d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) &= \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_i\|_1 \\ &\leq \frac{1}{2}\sqrt{2\,\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)} \\ &\leq \frac{1}{2}\sqrt{2\,\mathrm{KL}(\mathrm{Ber}(0.5, 0.5+\Delta))\mathbb{E}_0[n_i(T)]} \\ &\leq \sqrt{\frac{3}{2}\mathbb{E}_0[n_i(T)]\Delta^2} \\ &= \frac{1}{4}\sqrt{\frac{l+1}{MT}\mathbb{E}_0[n_i(T)]} \end{aligned}$$

where the first equality is from $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\|\mathbb{P} - \mathbb{Q}\|_1$ for any two distributions $\mathbb{P}$, $\mathbb{Q}$; the first inequality uses Pinsker's inequality; the second inequality is from the well-known divergence decomposition lemma (e.g. [90], Lemma 15.1); the third inequality uses Lemma A.12; and the last equality is by recalling that $\Delta \in [0, \frac{1}{4}]$ and algebra.

Now, applying Lemma A.11 with $m = l + 1 \geq 2$, $N_i = n_i(T)$ for all i in $[l+1]$, and $B = MT$, Eq. (A.24) is proved. This in turn finishes the proof of the regret lower bound.

Case 2: $M(l^C - 1) \geq l$**.**

To simplify notations, define $\Delta = \sqrt{\frac{l^C}{24T}}$. Observe that $\Delta < \frac{1}{4}$ as $T \geq K \geq l^C$. In addition, we must have $l^C \geq 2$ in this case, as if $l^C = 1$, $M(l^C - 1) = 0 < K = l$. We set $\epsilon = \frac{\Delta}{2}$.

We will now define $[l^C]^M$ different Bernoulli ϵ-MPMAB instances, and show that under at least one of them, $\mathcal{A}$ will have a collective regret at least $\frac{1}{24}M\sqrt{l^C T} \geq \frac{1}{192}(M\sqrt{(l^C-1)T} + \sqrt{MlT})$.

For $i_1, \ldots, i_M \in [l^C]^M$, define Bernoulli MPMAB instance $E_{i_1,\ldots,i_M}$ to be such that for p in $[M]$ and i in $[K]$, the expected reward of player p on pulling arm i is

$$\mu_i^p = \begin{cases} \frac{1}{2} + \Delta & i = i_p \\ \frac{1}{2} & i \in [l^C] \setminus \{i_p\} \\ 0 & i \notin [l^C] \end{cases}$$

We first verify that for every $i_1, \ldots, i_M$, instance $E_{i_1,\ldots,i_M}$ (1) is an ϵ-MPMAB instance, and (2) $[K] \setminus [l^C] \subset \mathcal{I}_{5\epsilon}$, and therefore, $\mathcal{I}_{5\epsilon}$ has size $\geq l$:

1. For item (1), observe that for all i in $[l^C]$ and all p in $[M]$, $\mu_i^p \in \left\{\frac{1}{2}, \frac{1}{2} + \Delta\right\}$; therefore, for every p, q, $\left|\mu_i^p - \mu_i^q\right| \leq \Delta = \epsilon$. Meanwhile, for all i in $[K] \setminus [l^C]$ and all p in $[M]$, $\mu_i^p = 0$, implying that for every p, q, $\left|\mu_i^p - \mu_i^q\right| = 0 \leq \epsilon$. Therefore $E_{i_1,\ldots,i_M}$ is ϵ-dissimilar.

2. For item (2), for all i in $[K] \setminus [l^C]$ and all p, $\Delta_i^p = \frac{1}{2} + \Delta > \frac{5}{2}\Delta = 5\epsilon$. This implies that all elements of $[K] \setminus [l^C]$ are in $\mathcal{I}_{5\epsilon}$.

We will now argue that

$$\mathbb{E}_{(i_1,\ldots,i_M)\sim\text{Unif}([l^C]^M)}\mathbb{E}_{E_{i_1,\ldots,i_M}}\left[\mathcal{R}(T)\right] \geq \frac{M\sqrt{l^C T}}{24}.$$

As the roles of all M players are the same, by symmetry, it suffices to show that the expected regret of player 1 satisfies

$$\mathbb{E}_{(i_1,\ldots,i_M)\sim\text{Unif}([l^C]^M)}\mathbb{E}_{E_{i_1,\ldots,i_M}}\left[\mathcal{R}^1(T)\right] \geq \frac{\sqrt{l^C T}}{24}. \tag{A.25}$$

It therefore suffices to show,

$$\mathbb{E}_{(i_1,\ldots,i_M)\sim\mathrm{Unif}([l^C]^M)}\mathbb{E}_{E_{i_1,\ldots,i_M}}\left[T - n_{i_1}^1(T)\right] \geq \frac{T}{4}. \tag{A.26}$$

This is because, recall that when i_1 is the optimal arm for player 1, $\mathcal{R}(T) = \sum_{i=1}^K n_i^1(T)\Delta_i^1 = \sum_{i\neq i_1} n_i^1(T)\Delta_i^1$; in addition, for all $i \neq i_1$, $\Delta_i^1 \geq \Delta$. This implies that $\mathcal{R}^1(T) \geq \Delta(T - n_{i_1}^1(T))$. Eq. (A.25) follows from the above inequality, Eq. (A.26), and the definition of Δ.

We now come back to the proof of Eq. (A.26). To this end, we define the following set of "helper" instances to facilitate our reasoning. Given $i_2, \ldots, i_M \in [K]^{M-1}$, define instance $E_{0,i_2,\ldots,i_M}$ such that its reward distribution is identical to $E_{i_1,i_2,\ldots,i_M}$ except for player 1 on arm i_1. Formally, it has the following expected reward profile:

$$\text{for } p = 1, \mu_i^1 = \begin{cases} \frac{1}{2} & i \in [l^C] \\ 0 & i \notin [l^C] \end{cases} \quad \text{for } p \neq 1, \mu_i^p = \begin{cases} \frac{1}{2} + \Delta & i = i_p \\ \frac{1}{2} & i \in [l^C] \setminus \left\{i_p\right\} \\ 0 & i \notin [l^C] \end{cases}$$

In addition, for all $i_1, \ldots, i_M$ in $(\{0\} \cup [l^C]) \times [l^C]^{M-1}$, define $\mathbb{P}_{i_1,\ldots,i_M}$ as the joint distribution of the interaction logs (arm pulls and rewards) for all M players over a horizon of T; furthermore, for i in $(\{0\} \cup [l^C])$, define $\mathbb{P}_i = \frac{1}{(l^C)^{M-1}} \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \mathbb{P}_{i,i_2,\ldots,i_M}$, and denote by $\mathbb{E}_i$ expectation with respect to $\mathbb{P}_i$. In this notation, Eq. (A.26) can be rewritten as

$$\frac{1}{l^C}\sum_{i=1}^{l^C} \mathbb{E}_i[T - n_i^1(T)] \geq \frac{T}{2}.$$

For every i in $[l^C]$,

$$\begin{aligned}
d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) &= \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_i\|_1 \\
&= \frac{1}{2}\left\| \frac{1}{l^{C^{M-1}}} \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} (\mathbb{P}_{0,i_2,\ldots,i_M} - \mathbb{P}_{i,i_2,\ldots,i_M}) \right\|_1 \\
&\leq \frac{1}{(l^C)^{M-1}} \cdot \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \frac{1}{2}\|\mathbb{P}_{0,i_2,\ldots,i_M} - \mathbb{P}_{i,i_2,\ldots,i_M}\|_1 \\
&\leq \frac{1}{(l^C)^{M-1}} \cdot \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathrm{Ber}(0.5, 0.5+\Delta)) \cdot \mathbb{E}_{0,i_2,\ldots,i_M}[N_i^1(T)]} \\
&\leq \frac{1}{(l^C)^{M-1}} \cdot \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \sqrt{\frac{3}{2}\Delta^2 \cdot \mathbb{E}_{0,i_2,\ldots,i_M}[N_i^1(T)]} \\
&\leq \sqrt{\frac{3}{2}\Delta^2 \cdot \mathbb{E}_0[N_i^1(T)]}
\end{aligned}$$

where the first equality is from $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\|\mathbb{P} - \mathbb{Q}\|_1$ for any two distributions $\mathbb{P}$, $\mathbb{Q}$; the second equality is from the definition of $\mathbb{P}_i$, $i \in \{0\} \cup [l^C]$; the first inequality is from triangle inequality of ℓ_1 norm; the second inequality is from Pinsker's inequality, and the divergence decomposition lemma ([90], Lemma 15.1); the third inequality is from Lemma A.12 and recalling that $\Delta \in [0, \frac{1}{4}]$; the last inequality is from Jensen's inequality, and the definition of $\mathbb{P}_0$.

Applying Lemma A.11 with $m = l^C \geq 2$, $N_i = n_i^1(T)$ for all i in $[l^C]$ and $B = T$, Eq. (A.26) is proved. This in turn finishes the proof of the regret lower bound. □

A.5.2 Gap-dependent lower bounds with known ϵ

We restate Theorem 2.9 here with specifications of exact constants in the lower bound.

Theorem A.9 (Restatement of Theorem 2.9)**.** *Fix $\epsilon \geq 0$ and $\alpha, C > 0$. Let $\mathcal{A}$ be an algorithm such that $\mathcal{A}$ has at most $CT^{1-\alpha}$ regret in all ϵ-MPMAB problem instances. Then,*

for any Bernoulli $\frac{\epsilon}{2}$-MPMAB instance $\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ such that $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$ for all i and p, we have:

$$\mathbb{E}_\mu[\mathcal{R}(T)] \geq \sum_{i\in\mathcal{I}_{5\epsilon/4}^C} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln\left(\Delta_i^p T^\alpha/8C\right)}{12\Delta_i^p} + \sum_{i\in\mathcal{I}_{5\epsilon/4}:\Delta_i^{\min}>0} \frac{\ln\left(\Delta_i^{\min} T^\alpha/8C\right)}{12\Delta_i^{\min}}.$$

Proof. We will first prove the following two claims:

1. For any i_0 in $[K]$ such that $\Delta_{i_0}^{\min} > 0$, $\mathbb{E}_\mu\left[n_{i_0}(T)\right] \geq \frac{\ln\left(\Delta_{i_0}^{\min} T^\alpha/8C\right)}{12(\Delta_{i_0}^{\min})^2}$.

2. For any i_0 in $\mathcal{I}_{5\epsilon/4}^C$ and any p_0 in $[M]$ such that $\Delta_{i_0}^{p_0} > 0$, $\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \geq \frac{\ln\left(\Delta_{i_0}^{p_0} T^\alpha/8C\right)}{12(\Delta_{i_0}^{p_0})^2}$.

The proof of these two claims are as follows:

1. Fix i_0 in $[K]$ such that $\Delta_{i_0}^{\min} > 0$, i.e., $\Delta_{i_0}^p > 0$ for all p in $[M]$. Define $p_0 = \operatorname{argmin}_{p\in[M]} \Delta_{i_0}^p$.

   We consider a new Bernoulli MPMAB instance, with mean reward defined as follows:

$$\forall p \in [M], \quad \nu_i^p = \begin{cases} \mu_i^p + 2\Delta_{i_0}^{p_0}, & i = i_0, \\ \mu_i^p & \text{otherwise} \end{cases}$$

   We have the following key observations:

   (a) ν is an ϵ-MPMAB instance; this is because $\nu_i^p - \nu_i^q = \mu_i^p - \mu_i^q$ for any p, q in $[M]$ and i in $[K]$, and μ is an $\frac{\epsilon}{2}$-MPMAB instance. By our assumption that $\mathcal{A}$ has $CT^{1-\alpha}$ regret on all ϵ-MPMAB environments, we have

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}, \quad \mathbb{E}_\nu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}. \tag{A.27}$$

(b) By the divergence decomposition lemma ([90], Lemma 15.1),

$$\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) = \sum_{p=1}^{M} \mathbb{E}_\mu \left[n_{i_0}^p(T) \right] \mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^p), \mathrm{Ber}(\mu_{i_0}^p + 2\Delta_{i_0}^{p_0}) \right), \tag{A.28}$$

As for all p, $\mu_{i_0}^p \in [\frac{15}{32}, \frac{17}{32}]$, and $\Delta_{i_0}^p \leq \frac{1}{16}$, using Lemma A.12, we have that for all p,

$$\mathrm{KL}\left(\mathrm{Ber}(\mu_i^p), \mathrm{Ber}(\mu_i^p + 2\Delta_{i_0}^{p_0}) \right) \leq 3 \cdot (2\Delta_{i_0}^{p_0})^2 = 12(\Delta_{i_0}^{p_0})^2.$$

Plugging into Eq. (A.28), we get

$$\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) \leq \sum_{p=1}^{M} (\mathbb{E}_\mu \left[n_{i_0}^p(T) \right] \cdot 12(\Delta_{i_0}^{p_0})^2) = 12\mathbb{E}_\mu \left[n_{i_0}(T) \right] (\Delta_{i_0}^{p_0})^2. \tag{A.29}$$

(c) Under MPMAB instance μ, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0} n_{i_0}^{p_0}(T) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) \geq \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\mu \left[\mathcal{R}(T) \right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}). \tag{A.30}$$

Likewise, under MPMAB instance ν, for player p_0, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0}(T - n_{i_0}^{p_0}(T)) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) < \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\nu \left[\mathcal{R}(T) \right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}). \tag{A.31}$$

Adding up Eq. (A.30) and Eq. (A.31), we have

$$\mathbb{E}_\nu \left[\mathcal{R}(T) \right] + \mathbb{E}_\mu \left[\mathcal{R}(T) \right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \left(\mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}) + \mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}) \right). \tag{A.32}$$

From Eq. (A.27), we have the left hand side is at most $2CT^{1-\alpha}$. By Bretagnolle-Huber inequality (see Lemma A.13) and Eq. (A.29), we have that $\mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}) +$

$\mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\mu)) \geq \frac{1}{2}\exp(-12\mathbb{E}_\mu\left[n_{i_0}(T)\right](\Delta_{i_0}^{p_0})^2)$. Plugging these to Eq. (A.32), we get

$$2CT^{1-\alpha} \geq \frac{\Delta_{i_0}^{p_0} T}{4}\exp(-12\mathbb{E}_\mu\left[n_{i_0}(T)\right](\Delta_{i_0}^{p_0})^2).$$

Solving for $\mathbb{E}_\mu\left[n_{i_0}(T)\right]$, we conclude that

$$\mathbb{E}_\mu\left[n_{i_0}(T)\right] \geq \frac{1}{12(\Delta_{i_0}^{p_0})^2}\cdot\ln\left(\frac{\Delta_{i_0}^{p_0}T^\alpha}{8C}\right) = \frac{1}{12(\Delta_{i_0}^{\min})^2}\cdot\ln\left(\frac{\Delta_{i_0}^{\min}T^\alpha}{8C}\right).$$

2. Fix i_0 in $\mathcal{I}_{5\epsilon/4}^C$ and $p_0 \in [M]$ such that $\Delta_{i_0}^{p_0} > 0$. By definition of $\mathcal{I}_{5\epsilon/4}^C$, we also have $\Delta_{i_0}^{p_0} = \mu_*^{p_0} - \mu_{i_0}^{p_0} \leq \epsilon/4$.

   We consider a new MPMAB environment ν, with mean reward defined as follows:

$$\nu_i^p = \begin{cases} \mu_i^p + 2\Delta_{i_0}^{p_0} & i = i_0, p = p_0 \\ \mu_i^p & \text{otherwise} \end{cases}$$

   Same as before, we have the following three key observations:

   (a) ν is an ϵ-MPMAB instance; this is because $(\nu_i^p - \nu_i^q) - (\mu_i^p - \mu_i^q) \in \left\{-\frac{\epsilon}{2}, 0, \frac{\epsilon}{2}\right\}$ for any p, q in $[M]$ and i in $[K]$, and μ is an $\frac{\epsilon}{2}$-MPMAB instance. By our assumption that $\mathcal{A}$ has $CT^{1-\alpha}$ regret on all ϵ-MPMAB problem instances, we have

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}, \quad \mathbb{E}_\nu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}. \tag{A.33}$$

   (b) By the divergence decomposition lemma ([90], Lemma 15.1),

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) &= \mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right]\mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}), \mathrm{Ber}(\mu_{i_0}^{p_0} + 2\Delta_{i_0}^{p_0})\right) \\ &\leq 12\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right](\Delta_{i_0}^{p_0})^2,\end{aligned} \tag{A.34}$$

where the second equality uses the following observation: $\mu_{i_0}^{p_0} \in [\frac{15}{32}, \frac{17}{32}]$, and $\Delta_{i_0}^{p_0} \leq \frac{1}{16}$, using Lemma A.12, $\mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}), \mathrm{Ber}(\mu_{i_0}^{p_0} + 2\Delta_{i_0}^{p_0})\right) \leq 3 \cdot (2\Delta_{i_0}^{p_0})^2 = 12(\Delta_{i_0}^{p_0})^2$.

(c) Under MPMAB instance μ, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0} n_{i_0}^{p_0}(T) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) \geq \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}). \tag{A.35}$$

Likewise, under MPMAB instance ν, for player p_0, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0}(T - n_{i_0}^{p_0}(T)) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) < \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\nu\left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}). \tag{A.36}$$

Same as the proof of item 1, combining Equations (A.33), (A.34), (A.35), (A.36), and using Bretagnolle-Huber inequality, we get

$$\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \geq \frac{1}{12(\Delta_{i_0}^{p_0})^2} \cdot \ln\left(\frac{\Delta_{i_0}^{p_0} T^\alpha}{8C}\right).$$

We now use the above two claims to conclude the proof. Recall that $\mathbb{E}_\mu\left[\mathcal{R}(T)\right] = \sum_{i\in[K]} \sum_{p\in[M]} \Delta_i^p \mathbb{E}_\mu\left[n_i^p(T)\right]$. For i in $\mathcal{I}_{5\epsilon/4}$ such that $\Delta_i^{\min} > 0$, item 1 implies:

$$\sum_{p\in[M]} \Delta_i^p \mathbb{E}_\mu\left[n_i^p(T)\right] \geq \Delta_i^{\min} \sum_{p\in[M]} \mathbb{E}_\mu\left[n_i^p(T)\right] \geq \frac{1}{12\Delta_i^{\min}} \cdot \ln\left(\frac{\Delta_i^{\min} T^\alpha}{8C}\right).$$

For i in $\mathcal{I}_{5\epsilon/4}^C$, item 2 implies:

$$\sum_{p\in[M]} \Delta_i^p \mathbb{E}_\mu\left[n_i^p(T)\right] \geq \sum_{p\in[M]:\Delta_i^p>0} \frac{1}{12\Delta_i^p} \cdot \ln\left(\frac{\Delta_i^p T^\alpha}{8C}\right).$$

Summing over all i in $[K]$ on the above two inequalities, we have

$$\begin{aligned}\mathbb{E}_\mu\left[\mathcal{R}(T)\right] &= \sum_{i\in[K]}\sum_{p\in[M]}\Delta_i^p\mathbb{E}_\mu\left[n_i^p(T)\right] \\ &\geq \sum_{i\in\mathcal{I}^C_{5\epsilon/4}}\sum_{p\in[M]:\Delta_i^p>0}\frac{1}{12\Delta_i^p}\cdot\ln\left(\frac{\Delta_i^p T^\alpha}{8C}\right)+\sum_{i\in\mathcal{I}_{5\epsilon/4}:\Delta_i^{\min}>0}\frac{1}{12\Delta_i^{\min}}\cdot\ln\left(\frac{\Delta_i^{\min}T^\alpha}{8C}\right).\end{aligned}$$

□

Remark.

The above lower bound argument aligns with our intuition that arms that are near-optimal with respect to some players (i.e., those in $\mathcal{I}^C_{5\epsilon/4}$) are harder for information sharing: in addition to a lower bound on the collective number of pulls to it across all players (item 1 of the claim), we are able to show a stronger lower bound on the number of pulls to it from *each player* (item 2 of the claim).

A.5.3 Gap-dependent lower bounds with unknown ϵ

We restate Theorem 2.11 here with specifications of exact constants in the lower bound.

Theorem A.10 (Restatement of Theorem 2.11)**.** *Fix $\alpha, C > 0$. Let $\mathcal{A}$ be an algorithm such that $\mathcal{A}$ has at most $CT^{1-\alpha}$ regret in all MPMAB problem instances. Then, for any Bernoulli MPMAB instance $\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ such that $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$ for all i and p, we have:*

$$\mathbb{E}_\mu[\mathcal{R}(T)] \geq \sum_{i\in[K]}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln\left(\Delta_i^p T^\alpha/8C\right)}{12\Delta_i^p}.$$

Proof. Recall that $\mathbb{E}_\mu\left[\mathcal{R}(T)\right] = \sum_{i=1}^K\sum_{p=1}^M\Delta_i^p\mathbb{E}_\mu\left[N_i^p(T)\right]$; it suffices to show that for any i_0 in $[K]$ and any p_0 in $[M]$ such that $\Delta_{i_0}^{p_0} > 0$, $\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \geq \frac{\ln\left(\Delta_{i_0}^{p_0}T^\alpha/8C\right)}{12(\Delta_{i_0}^{p_0})^2}$.

The proof of this claim is almost identical to the proof of the the second claim in the previous theorem, except that we have more flexibility to choose the "alternative instances" ν's, because $\mathcal{A}$ is assumed to have sublinear regret in all MPMAB instances; specifically, ν no longer needs to be an ϵ-MPMAB instance. We include the argument here for completeness. Fix i_0 in $[K]$ and $p_0 \in [M]$ such that $\Delta_{i_0}^{p_0} > 0$. We consider a new Bernoulli MPMAB instance ν, with mean reward defined as follows:

$$\nu_i^p = \begin{cases} \mu_i^p + 2\Delta_{i_0}^{p_0} & i = i_0, p = p_0 \\ \mu_i^p & \text{otherwise} \end{cases}$$

We have the following three key observations:

(a) ν is still a valid Bernoulli MPMAB instance; this is because for all p in $[M]$ and i in $[K]$, $(\nu_i^p - \mu_i^p) \in [-\frac{1}{8}, \frac{1}{8}]$, and $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$, implying that $\nu_i^p \in [0, 1]$. By our assumption that $\mathcal{A}$ has $CT^{1-\alpha}$ regret on all Bernoulli MPMAB instances, we have

$$\mathbb{E}_\mu \left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}, \quad \mathbb{E}_\nu \left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}. \tag{A.37}$$

(b) By the divergence decomposition lemma ([90], Lemma 15.1),

$$\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) = \mathbb{E}_\mu \left[n_{i_0}^{p_0}(T)\right] \mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}), \mathrm{Ber}(\mu_{i_0}^{p_0} + 2\Delta_{i_0}^{p_0})\right) \leq 12\mathbb{E}_\mu \left[n_{i_0}^{p_0}(T)\right] (\Delta_{i_0}^{p_0})^2, \tag{A.38}$$

where the second equality uses the following observation: $\mu_{i_0}^{p_0} \in [\frac{15}{32}, \frac{17}{32}]$, and $\Delta_{i_0}^{p_0} \leq \frac{1}{16}$, using Lemma A.12, $\mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}), \mathrm{Ber}(\mu_{i_0}^{p_0} + 2\Delta_{i_0}^{p_0})\right) \leq 3 \cdot (2\Delta_{i_0}^{p_0})^2 = 12(\Delta_{i_0}^{p_0})^2$.

(c) Under MPMAB instance μ, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0} n_{i_0}^{p_0}(T) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) \geq \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\mu \left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}). \tag{A.39}$$

Likewise, under MPMAB instance ν, for player p_0, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0}(T - n_{i_0}^{p_0}(T)) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) < \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\nu \left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}). \tag{A.40}$$

Combining Equations (A.37), (A.38), (A.39), (A.40), and using Bretagnolle-Huber inequality, we get

$$\mathbb{E}_\mu \left[n_{i_0}^{p_0}(T)\right] \geq \frac{1}{12(\Delta_{i_0}^{p_0})^2} \cdot \ln\left(\frac{\Delta_{i_0}^{p_0} T^\alpha}{8C}\right).$$

This concludes the proof of the claim, and in turn concludes the proof of the theorem. □

A.5.4 Auxiliary lemmas

The following lemma is well known for proving gap-independent lower bounds in single player K-armed bandits. We will be using the following convention: for probability distribution $\mathbb{P}_i$, denote by $\mathbb{E}_i$ its induced expectation operator.

Lemma A.11. *Suppose m, B are positive integers and $m \geq 2$; there are $m+1$ probability distributions $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_m$, and m random variables $N_1, \ldots, N_m$, such that: (1) Under any of the $\mathbb{P}_i$'s, $N_1, \ldots, N_m$ are non-negative and $\sum_{i=1}^m N_i \leq B$ with probability 1; (2) for all i in $[m]$, $d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) \leq \frac{1}{4}\sqrt{\frac{m}{B} \cdot \mathbb{E}_0[N_i]}$. Then,*

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_i[B - N_i] \geq \frac{B}{4}.$$

Proof. For every i in $[m]$, as N_i is a random variable that takes values in $[0, B]$, we have,

$$\left|\mathbb{E}_i[N_i] - \mathbb{E}_0[N_i]\right| \leq B \cdot d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i).$$

By item (2) and algebra, this implies that

$$\mathbb{E}_i[N_i] \le \mathbb{E}_0[N_i] + \frac{1}{4}\sqrt{mB\mathbb{E}_0[N_i]}.$$

Averaging over i in $[m]$ and using Jensen's inequality, we have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_i[N_i] &\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_0[N_i] + \frac{1}{4m}\sum_{i=1}^{m}\sqrt{mB\mathbb{E}_0[N_i]} \\ &\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_0[N_i] + \frac{1}{4}\sqrt{mB\cdot\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_0[N_i]\right)}\end{aligned}$$

Noting that item (2) implies $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_0[N_i] \le \frac{B}{m}$; plugging this into the above inequality, we have

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_i[N_i] \le \frac{B}{m} + \frac{1}{4}\sqrt{mB\cdot\frac{B}{m}} \le \frac{B}{2} + \frac{B}{4} = \frac{3B}{4},$$

where the second inequality uses the assumption that $m \ge 2$. The lemma is concluded by negating and adding B on both sides. □

Lemma A.12. *Suppose a, b are both in $[\frac{1}{4}, \frac{3}{4}]$. Then,* $\mathrm{KL}(\mathrm{Ber}(a), \mathrm{Ber}(b)) \le 3(b-a)^2$.

Proof. Define $h(x) = x\ln\frac{1}{x} + (1-x)\ln\frac{1}{1-x}$. One can easily verify that $\mathrm{KL}(\mathrm{Ber}(a), \mathrm{Ber}(b)) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$, which in turn equals $h(a) - h(b) - h'(b)(a-b)$. By Taylor's theorem, there exists some $\xi \in [a, b] \subseteq [\frac{1}{4}, \frac{3}{4}]$ such that

$$h(a) - h(b) - h'(b)(a-b) = \frac{h''(\xi)}{2}(b-a)^2 = \frac{1}{2\xi(1-\xi)}(b-a)^2.$$

The lemma is concluded by verifying that $\frac{1}{2\xi(1-\xi)} \le 3$ for ξ in $[\frac{1}{4}, \frac{3}{4}]$. □

Lemma A.13 (Bretagnolle-Huber)**.** *For any two distributions $\mathbb{P}$ and $\mathbb{Q}$ and an event A,*

$$\mathbb{P}(A) + \mathbb{Q}(A^C) \ge \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}, \mathbb{Q})).$$

A.6 Upper Bounds with Unknown ϵ

In this section, we provide a description of RobustAgg-Agnostic, an algorithm that has regret adaptive to $\mathcal{I}_{5\epsilon}$ in all MPMAB environments with unknown ϵ.

To ensure sublinear regret in all MPMAB environments, RobustAgg uses the aggregation-based framework named Corral [2, see also Lemma A.16 below], which we now briefly review. The Corral meta-algorithm allows one to combine multiple online bandit learning algorithms (called base learners) into one master algorithm that has performance competitive with all base learners'. For different environments, different base learners may stand out as the best, and therefore the master algorithm exhibits some degree of adaptivity. We refer readers to [2] for the full description of Corral.

In the context of MPMAB problems, recall that we have developed RobustAgg(ϵ) that has good regret guarantees for all ϵ-MPMAB instances. The central idea of RobustAgg-Agnostic is to apply the Corral algorithm over a series of baser learners, i.e., $\left\{\text{RobustAgg}(\epsilon_b)\right\}_{b=1}^{B}$, where $E = \{\epsilon_b\}_{b=1}^{B}$ is a covering of the $[0,1]$ interval. With an appropriate setting of E, for any ϵ-MPMAB instance, there exists some b_0 in $[B]$ such that ϵ_{b_0} is not much larger than ϵ, and running RobustAgg(ϵ_{b_0}) would achieve regret guarantee competitive to RobustAgg(ϵ). As Corral achieves online performance competitive with all RobustAgg(ϵ_b)'s [2], it must be competitive with RobustAgg(ϵ_{b_0}), and therefore can inherit the adaptive regret guarantee of RobustAgg(ϵ_{b_0}).

We now provide important technical details of RobustAgg-Agnostic:

1. $B = \lceil \log(MT) \rceil + 1$ is the number of base learners, and $E = \left\{\epsilon_b = 2^{-b+1} : b \in [B]\right\}$ is the grid of ϵ to be aggregated. Corral uses master learning rate $\eta = \frac{1}{M\sqrt{T}}$.

2. For each base learner that runs RobustAgg(ϵ) for some ϵ, we require them to take a new parameter $\rho \geq 1$ as input, to accommodate for the fact that it may not be selected by the Corral master all the time. Specifically, it performs bandit learning interaction

with an environment whose returned rewards are unbiased but *importance weighted*: at time step t, when player p pulls arm i, instead of directly receiving reward drawn from $r \sim \mathcal{D}_i^p$, it receives $\hat{r} = \frac{W_t}{w_t} r$, where $w_t \in [1, \frac{1}{\rho}]$ is a random number, and conditioned on w_t, $W_t \sim \text{Ber}(w_t)$ is an independently-drawn Bernoulli random variable. Observe that $\hat{r}$ has conditional mean μ_i^p, lies in the interval $[0, \rho]$, and has conditional variance at most ρ.

We call an environment that has the above analytical form a *ρ-importance weighted environment*; in the special case of $\rho = 1$, $w_t = 1$ and $W_t = 1$ with probability 1 for all t, and therefore a 1-importance weighted environment is the same as the original bandit learning environment.

Under an ρ-importance weighted environment, the rewards are no longer bounded in $[0, 1]$, therefore, the constructions of the UCB's of the mean rewards in the original RobustAgg(ϵ) becomes invalid. Instead, we will rely on the following lemma (analogue of Lemma A.4) for constructing valid UCB's:

Lemma A.14. *With probability at least $1 - 4MT^{-5}$, we have*

$$\left|\zeta_i^p(t-1) - \mu_i^p\right| \leq 8\sqrt{\frac{3\rho \ln T}{\overline{n_i^p}(t-1)}},$$

$$\left|\eta_i^p(t-1) - \sum_{q \neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)} \mu_i^q\right| \leq 4\sqrt{\frac{14\rho \ln T}{\overline{m_i^p}(t-1)}}$$

holding for all p in $[M]$, where $\zeta_i^p(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_t^p = i\} \hat{r}_t^p}{\overline{n_i^p}(t-1)}$, and $\eta_i^p(t) = \frac{\sum_{s=1}^{t-1} \sum_{q=1}^{M} \mathbb{1}\{q \neq p,\ i_t^q = i\} \hat{r}_t^q}{\overline{m_i^p}(t-1)}$.

According to the above concentration bounds, changing the definition of confidence interval width to $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon) = 8\sqrt{13\rho \ln T \left[\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}\right]} + (1-\lambda)\epsilon$ would maintain the validity of the UCB's in ρ-importance weighted environments; henceforth, we incorporate this modification in RobustAgg(ϵ).

We have the following important analogue of Theorem 2.8, which establishes a gap-independent regret upper bound when RobustAgg(ϵ) is run in a ρ-importance weighted ϵ-MPMAB environment. This shows RobustAgg(ϵ) enjoys stability: the regret of the algorithm degrades gracefully with increasing ρ.[3]

Lemma A.15. *Let RobustAgg(ϵ) run on a ρ-importance weighted ϵ-MPMAB problem instance for T rounds. Then its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \tilde{\mathcal{O}}\left(\sqrt{\rho\left|\mathcal{I}_{5\epsilon}\right| MT} + M\left|\mathcal{I}_{5\epsilon}\right| + \min\left(M\sqrt{\rho\left|\mathcal{I}_{5\epsilon}^{C}\right| T}, \epsilon MT\right)\right).$$

The proof of Lemmas A.14 and A.15 can be found at the end of this section.

3. Corral maintains a probability distribution on base learners $q_t = (q_{t,b} : b \in [B])$ over time. At time step t, each base learner b proposes their own arm pull decisions $(i_t^p(b) : p \in [M])$; the Corral master chooses a base learner with probability according to q_t, that is, $i_t^p = i_t^p(b_t)$ for all p, where $b_t \sim q_t$. After the arm pulls, learner b receives feedback $\hat{r}_t^p(b) = \frac{\mathbb{1}\{b_t=b\}}{q_{t,b}} r_t^p$, which is equivalent to interacting with an importance weighted environment discussed before—$q_{t,b}$ and $\mathbb{1}\{b_t = b\}$ correspond to w_t and W_t, respectively; when $b_t = b$, r_t^p is drawn from $\mathcal{D}_{i_t^p}^p$ for all p in $[M]$.

   Corral also uses the above feedback to update q_{t+1}, its weighting of the base learners: define $\ell_{t,b} = \frac{\mathbb{1}\{b_t=b\}}{q_{t,b}} \mathbb{1}\{b_t = b\}(\sum_{p=1}^{M}(1 - r_t^p))$ to be the importance weighted loss of base learner b at time step t; q_t is updated to q_{t+1} using $(\ell_{t,b} : b \in [B])$, with online mirror descent with the log-barrier regularizer and learning rate $\eta > 0$. A small complication of directly applying the existing results of Corral is that Corral originally assumes that the losses suffered by the base learner from each round have range $[0, 1]$. In the

[3] See an elegant definition of $(R(T), \alpha)$-(weak) stability for bandit algorithms in [2]. Our guarantee on RobustAgg(ϵ) in Lemma A.15 is slightly stronger than the $(\tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{5\epsilon}|\, MT} + M\left|\mathcal{I}_{5\epsilon}\right| + \min\left(M\sqrt{\left|\mathcal{I}_{5\epsilon}^{C}\right| T}, \epsilon MT\right)\right), \frac{1}{2})$-weak stability, in that the regret bound has terms that are unaffected by ρ.

multi-player setting, the losses suffered by the base learner is the sum of the losses of all players, which has range $[0, M]$. Nevertheless, we can obtain a similar guarantee. Denote by ρ_b be the final value of ρ of base learner b (see also the next item). A slight modification of Agarwal et al. [2, Lemma 13] shows that for all base learner b,

$$\sum_{t=1}^{T}\sum_{b'=1}^{B} q_{t,b'}\ell_{t,b'} - \sum_{t=1}^{T}\ell_{t,b} \leq \mathcal{O}\left(\frac{B}{\eta} + \eta M^2 T\right) - \frac{\rho_b}{40\eta \ln T}.$$

Taking expectation on both sides, and observing that

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{b'=1}^{B} q_{t,b'}\ell_{t,b'}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}(1-\mu^p_{i^p_t})\right],$$

and $\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,b}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}(1-\mu^p_{i^p_t(b)})\right]$, along with some algebra, we get the following lemma.

Lemma A.16. *Suppose* RobustAgg-Agnostic *is run for T rounds. Then, for every b in $[B]$, we have that the regret of the master algorithm with respect to base learner b is bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t(b)} - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t}\right] \leq \mathcal{O}\left(\frac{B}{\eta} + \eta M^2 T\right) - \frac{\mathbb{E}\left[\rho_b\right]}{40\eta \ln T}.$$

4. Following [2], a doubling trick is used for maintaining the value of ρ's for all base learners over time. Specifically, each base learner b maintains a separate guess of ρ, an upper bound of $\max_{s=1}^{t}\frac{1}{q_{s,b}}$; if the upper bound is violated, its ρ gets doubled and the base learner restarts. As Corral initializes ρ as $2B$ for each base learner, and maintains the invariant that $\rho \leq BT$, the number of doublings/restarts for each base learner is at most $\lceil \log T \rceil$. For a fixed b, summing over the regret guarantees between different restarts of base learner b, we have the following regret guarantee.

Lemma A.17. *Suppose $\epsilon_b \geq \epsilon$, and* RobustAgg(ϵ_b) *is run as a base learner of* RobustAgg-Agnostic, *on an ϵ-MPMAB problem instance for T rounds. Denote by ρ_b the final value of ρ. Then its expected collective regret satisfies*

$$\mathbb{E}\left[T\sum_{p=1}^{M}\mu_*^p - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu_{i_t^p(b)}^p\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\rho_b\right]\left|\mathcal{I}_{5\epsilon_b}\right| MT} + \min\left(M\sqrt{\mathbb{E}\left[\rho_b\right]\left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon MT\right) + M\left|\mathcal{I}_{5\epsilon_b}\right|\right).$$

The proof of this lemma can be found at the end of this section; we also refer the reader to [2, Appendix D] for details.

Combining all the lemmas above, we are now ready to prove Theorem 2.12, restated below for convenience.

Theorem 2.12. *Let* RobustAgg-Agnostic *run on an ϵ-MPMAB problem instance with any $\epsilon \in [0,1]$. Its expected collective regret in a horizon of T rounds satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \tilde{\mathcal{O}}\left(\left(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|\right)\sqrt{T} + M|\mathcal{I}_{5\epsilon}|\right).$$

Proof of Theorem 2.12. Suppose RobustAgg-Agnostic interacts with an ϵ-MPMAB problem instance. Let $b_0 = \max\left\{b \in [B] : \epsilon_b \geq \epsilon\right\}$. From the definition of $E = \left\{1, 2^{-1}, \ldots, 2^{-B+1}\right\}$ and $\epsilon \in [0,1]$, b_0 is well-defined.

We present the following technical claim that elucidates the guarantee provided by learner b_0 based on Lemma A.17; we defer its proof after the proof of the theorem.

Claim A.18. *Let b_0 be defined above.* RobustAgg(ϵ_{b_0}) *is run as a base learner of* RobustAgg-Agnostic, *on a ϵ-MPMAB problem instance for T rounds. Denote by ρ_{b_0}*

the final value of ρ. Then its expected collective regret satisfies

$$\mathbb{E}\left[T\sum_{p=1}^{M}\mu_*^p - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu_{i_t^p(b_0)}^p\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\rho_{b_0}\right]MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|)} + M\left|\mathcal{I}_{5\epsilon}\right|\right).$$

Combining Claim A.18 and Lemma A.16 with $b = b_0$, we have the following regret guarantee for RobustAgg-Agnostic:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T)\right] &= \mathbb{E}\left[T\sum_{p=1}^{M}\mu_*^p - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu_{i_t^p}^p\right] \\
&= \mathbb{E}\left[T\sum_{p=1}^{M}\mu_*^p - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu_{i_t^p(b_0)}^p\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}\mu_{i_t^p(b_0)}^p - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu_{i_t^p}^p\right] \\
&\leq \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\rho_{b_0}\right]MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|)} + M\left|\mathcal{I}_{5\epsilon}\right| + \frac{B}{\eta} + \eta M^2 T\right) - \frac{\mathbb{E}\left[\rho_{b_0}\right]}{40\eta\ln T} \\
&\leq \tilde{\mathcal{O}}\left(\eta MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|) + M\left|\mathcal{I}_{5\epsilon}\right| + \frac{B}{\eta} + \eta M^2 T\right),
\end{aligned}$$

where the first inequality is from Claim A.18 and Lemma A.16; the second inequality is from the AM-GM inequality that $\sqrt{\mathbb{E}\left[\rho_{b_0}\right]MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|)} \leq \mathcal{O}(\eta MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|) + \frac{\mathbb{E}\left[\rho_{b_0}\right]}{\eta\ln T})$ and algebra (canceling out the second term in the last expression with $-\frac{\mathbb{E}\left[\rho_{b_0}\right]}{40\eta\ln T}$). As RobustAgg-Agnostic chooses Corral's master learning rate $\eta = \frac{1}{M\sqrt{T}}$, and $B = \tilde{\mathcal{O}}(1)$, we have that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T)\right] &\leq \tilde{\mathcal{O}}\left((M + |\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|)\sqrt{T} + M\left|\mathcal{I}_{5\epsilon}\right|\right) \\
&\leq \tilde{\mathcal{O}}\left((|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}_{10\epsilon}^C\right|)\sqrt{T} + M\left|\mathcal{I}_{5\epsilon}\right|\right),
\end{aligned}$$

where the second inequality uses the fact that $\left|\mathcal{I}_{10\epsilon}^C\right| \geq 1$. □

Proof of Claim A.18. As $\epsilon_b \geq \epsilon$ always holds, $M\left|\mathcal{I}_{5\epsilon_{b_0}}\right| \leq M\left|\mathcal{I}_{5\epsilon}\right|$. It remains to check by

algebra that

$$\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{5\epsilon_{b_0}}\right| MT} + \min\left(M\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}^C_{5\epsilon_{b_0}}\right| T}, MT\epsilon_{b_0}\right) = \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}^C_{10\epsilon}\right|)}\right). \tag{A.41}$$

We consider two cases:

1. $\epsilon_{b_0} \le 2\epsilon$. In this case, we have $\mathcal{I}_{5\epsilon_{b_0}} \subset \mathcal{I}_{10\epsilon}$. We have the following derivation:

$$\begin{aligned}\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{5\epsilon_{b_0}}\right| MT} + M\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}^C_{5\epsilon_{b_0}}\right| T} &\le 2\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT \cdot (\left|\mathcal{I}_{5\epsilon_b}\right| + M\left|\mathcal{I}^C_{5\epsilon_b}\right|)} \\ &\le 2\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}^C_{10\epsilon}\right|)}\end{aligned}$$

   where the first inequality is from the basic fact that $\sqrt{A} + \sqrt{B} \le 2\sqrt{A+B}$ for positive A, B; the second inequality is from the fact that $\left|\mathcal{I}_{5\epsilon_b}\right| + M\left|\mathcal{I}^C_{5\epsilon_b}\right| \le |\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}^C_{10\epsilon}\right|$, as $\left|\mathcal{I}_{5\epsilon_b}\right| \ge |\mathcal{I}_{10\epsilon}|$, $M \ge 1$, and $|\mathcal{I}_\alpha| + \left|\mathcal{I}^C_\alpha\right| = K$ for any α. This verifies Eq. (A.41).

2. $\epsilon_{b_0} > 2\epsilon$. In this case, $b_0 = B = 1 + \lceil \log(MT) \rceil$ and $\epsilon_{b_0} \le \frac{1}{MT}$. Although we no longer have $\mathcal{I}_{5\epsilon_{b_0}} \subset \mathcal{I}_{10\epsilon}$, we can still upper bound the left hand side as follows.

   First, the second term, $\min\left(M\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}^C_{5\epsilon_{b_0}}\right| T}, MT\epsilon_{b_0}\right) \le MT \cdot \frac{1}{MT} = 1$.

   Moreover, the first term, $\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{5\epsilon_{b_0}}\right| MT} \le \sqrt{\mathbb{E}\left[\rho_{b_0}\right] KMT}$. As $|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}^C_{10\epsilon}\right| \ge K$, we have $\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{5\epsilon_{b_0}}\right| MT} \le \sqrt{\mathbb{E}\left[\rho_{b_0}\right](|\mathcal{I}_{10\epsilon}| + M\left|\mathcal{I}^C_{10\epsilon}\right|)MT}$. Combining the above two, Eq. (A.41) is proved. □

Proof sketch of Lemma A.14. Since the proof of Lemma A.4 can be almost directly carried over here, we only sketch the proof by pointing out the major differences. We also refer the reader to [6, Appendix C.3] for a similar reasoning.

We first consider the concentration of $\zeta_i^p(j,t)$. We define a filtration $\{\mathcal{B}_t\}_{t=1}^T$, where

$$\mathcal{B}_t = \sigma(\{w_s, i_s^{p'}, \hat{r}_s^{p'} : s \in [t], p' \in [M]\} \cup \left\{i_{t+1}^{p'} : p' \in [M]\right\})$$

is the σ-algebra generated by the history (including that of w_s's) up to round t and the arm selection of all players at time step $t+1$

Let $X_t = \mathbb{1}\{i_t^p = i\}\left(\hat{r}_t^p - \mu_i^p\right)$. We have $\mathbb{E}\left[X_t \mid \mathcal{B}_{t-1}\right] = 0$. In addition,

$$\begin{aligned}
\text{var}\left[X_t \mid \mathcal{B}_{t-1}\right] &= \mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid \mathcal{B}_{t-1}])^2 \mid \mathcal{B}_{t-1}\right] \\
&= \mathbb{E}\left[X_t^2 \mid \mathcal{B}_{t-1}\right] \\
&\leq \mathbb{1}\{i_t^p = i\}\mathbb{E}\left[w_t\left(\frac{r_t^p}{w_t} - \mu_i^p\right)^2 + (1-w_t)0 \mid \mathcal{B}_{t-1}\right] \\
&\leq \mathbb{1}\{i_t^p = i\}\mathbb{E}\left[w_t\left(\frac{r_t^p}{w_t}\right)^2 \mid \mathcal{B}_{t-1}\right] \\
&\leq \mathbb{1}\{i_t^p = i\}\rho.
\end{aligned}$$

Also, $|X_t| \leq \rho$ with probability 1. Applying Freedman's inequality [14, Lemma 2] with $\sigma = \sqrt{\sum_{s=1}^{t-1} \text{var}\left[X_s \mid \mathcal{B}_{s-1}\right]}$ and $b = \rho$, and using $\sigma \leq \sqrt{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\}\rho}$, we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1} X_s\right| \leq 4\sqrt{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\}\rho \cdot \ln(T^5 \log_2 T)} + 2\rho \ln(T^5 \log_2 T). \tag{A.42}$$

We can then show that

$$\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\}\hat{r}_s^p}{\overline{n_i^p}(t-1)} - \mu_i^p\right| \leq 8\sqrt{\frac{3\rho \ln T}{\overline{n_i^p}(t-1)}}.$$

following the same strategy in the proof for Lemma A.4.

Similarly, we show the concentration of $\eta_i^p(t)$. We define a filtration $\{\mathcal{G}_{t,q}\}_{t\in[T],q\in[M]}$,

where

$$\mathcal{G}_{t,q} = \sigma(\Big\{w_s, i_s^{p'}, \hat{r}_s^{p'} : s \in [t], p' \in [M], i \in [K]\Big\} \cup \Big\{i_{t+1}^{p'} : p' \in [M], p' \leq q\Big\})$$

is the σ-algebra generated by the history (including that of w_s's) up to round t and the arm selection of players $1, 2, \ldots, q$ at round $t+1$. By convention, let $\mathcal{G}_{t,0} = \mathcal{G}_{t-1,M}$.

Let random variable $Y_{t,q} = \mathbb{1}\{q \neq p,\ i_t^q = i\}\left(\hat{r}_s^q - \mu_i^q\right)$. We have $\mathbb{E}\left[Y_{t,q} \mid \mathcal{G}_{t,q-1}\right] = 0$; in addition, $\text{var}\left[Y_t \mid \mathcal{G}_{t,q-1}\right] = \mathbb{E}\left[Y_{t,q}^2 \mid \mathcal{G}_{t,q-1}\right] \leq \mathbb{1}\{q \neq p,\ i_t^q = i\}\rho$ and $\left|Y_{t,q}\right| \leq \rho$.

Again, applying Freedman's inequality [14, Lemma 2], we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1}\sum_{q=1}^{M} Y_{s,q}\right| \leq 4\sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}\{q \neq p,\ i_s^q = i\}\rho \cdot \ln(T^5 \log_2(TM))} + 2\rho \ln(T^5 \log_2(TM)). \tag{A.43}$$

Using the same strategy from the proof for Lemma A.4, we can show that

$$\left|\eta_i^p(t-1) - \sum_{q \neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\rho \ln T}{\overline{m_i^p}(t-1)}}.$$

The lemma then follows by applying the union bound. □

Proof sketch of Lemma A.15. Similar to the proof of Theorem 2.8, we define $\mathcal{E} = \cap_{t=1}^{T} \cap_{i=1}^{K} \mathcal{Q}_i(t)$, where

$$\mathcal{Q}_i(t) = \left\{\forall p, \left|\zeta_i^p(t) - \mu_i^p\right| \leq 8\sqrt{\frac{3\rho \ln T}{\overline{n_i^p}(t-1)}}, \left|\eta_i^p(t) - \sum_{q \neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\rho \ln T}{\overline{m_i^p}(t-1)}}\right\};$$

note that the new definition of $\mathcal{Q}_i(t)$ has a dependence on ρ.

Similar to the proof of Theorem 2.5, we have,

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathcal{O}(1),$$

and

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{i\in[K]} \sum_{p\in[M]} \mathbb{E}[n_i^p(T)|\mathcal{E}] \cdot \Delta_i^p \\ &\leq \sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^{\max} + \sum_{i\in\mathcal{I}_{5\epsilon}^C} \sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}] \cdot \Delta_i^p\end{aligned}$$

We bound these two terms respectively, applying the technique from [90, Theorem 7.2].

1. We can show the following analogue of Lemma A.7: there exists some constant $C_1 > 0$ such that for each $i \in \mathcal{I}_{5\epsilon}$,

   $$\mathbb{E}[n_i(T)|\mathcal{E}] \leq C_1 \left(\frac{\rho \ln T}{(\Delta_i^{\min})^2} + M \right).$$

   Using the above fact, and from a similar calculation of Equation (A.19) in the proof of Theorem 2.8, we get

   $$\sum_{i\in\mathcal{I}_{5\epsilon}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^{\max} \leq 4\sqrt{C_1 \rho |\mathcal{I}_{5\epsilon}| MT \ln T} + 2C_1 M |\mathcal{I}_{5\epsilon}|.$$

2. We can show the following analogue of Lemma A.8: there exists some constant $C_2 > 0$ such that for each $i \in \mathcal{I}_{5\epsilon}^C$ and $p \in [M]$ with $\Delta_i^p > 0$,

   $$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq C_2 \left(\frac{\ln T}{(\Delta_i^p)^2} \right).$$

   Using the above fact, and from a similar calculation of Equation (A.22) in the proof of

Theorem 2.8, we get

$$\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0}\mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p\le 2M\sqrt{C_2\rho|\mathcal{I}_{5\epsilon}^C|T\ln T}.$$

On the other hand, we trivially have that for all i in $\mathcal{I}_{5\epsilon}^C$, $\Delta_i^p\le 5\epsilon$; therefore,

$$\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0}\mathbb{E}[n_i^p(T)]\cdot\Delta_i^p\le 5\epsilon MT.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[\mathcal{R}(T)]\\
\le&\left(4\sqrt{C_1\rho|\mathcal{I}_{5\epsilon}|MT\ln T}+2C_1M|\mathcal{I}_{5\epsilon}|\right)+\min\left(2M\sqrt{C_2\rho|\mathcal{I}_{5\epsilon}^C|T\ln T},5\epsilon MT\right)+\mathcal{O}(1)\\
\le&\tilde{\mathcal{O}}\left(\sqrt{\rho\left|\mathcal{I}_{5\epsilon}\right|MT}+M\left|\mathcal{I}_{5\epsilon}\right|+\min\left(M\sqrt{\rho\left|\mathcal{I}_{5\epsilon}^C\right|T},\epsilon MT\right)\right).
\end{aligned}$$

□

Proof of Lemma A.17. The proof closely follows [2, Theorem 15]; we cannot directly repeat that proof here, because Lemma A.15 is not precisely a weak stability statement (see footnote 3).

For base learner b, suppose that its ρ gets doubled n_b times throughout the process, where n_b is a random number in $[\lceil\log T\rceil]$. For every $i\in[n_b]$, denote by random variable t_i the i-th time step where the value of ρ gets doubled. In addition, denote by $t_0=0$ and $t_{n_b+1}=T$. In this notation, for all $t\in\{t_i+1,t_i+2,\ldots,t_{i+1}\}$, the value of ρ is equal to $\rho^i=2B\cdot 2^i$; in addition, $\rho_b=2B\cdot 2^{n_b}$.

Therefore, we have:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T) \mid n_b = n\right] &= \sum_{i=0}^{n} \mathbb{E}\left[\sum_{t=t_i+1}^{t_{i+1}} \left(\sum_{p=1}^{M} \mu_*^p - \sum_{p=1}^{M} \mu_{i_t^p(b)}^p\right) \mid n_b = n\right] \\
&= \sum_{i=0}^{n} \tilde{\mathcal{O}}\left(\sqrt{\rho^i \left|\mathcal{I}_{5\epsilon_b}\right| MT} + M\left|\mathcal{I}_{5\epsilon_b}\right| + \min\left(M\sqrt{\rho^i \left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\rho^n \left|\mathcal{I}_{5\epsilon_b}\right| MT} + M\left|\mathcal{I}_{5\epsilon_b}\right| + \min\left(M\sqrt{\rho^n \left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right),
\end{aligned}$$

where the first equality of by the definition of $\mathcal{R}(T)$, and $[T] = \cup_{i=1}^{n} \{t_i + 1, t_i + 2, \ldots, t_{i+1}\}$; the second equality is from Lemma A.17's guarantee in each time interval $\{t_i + 1, t_i + 2, \ldots, t_{i+1}\}$ and $\epsilon_b \geq \epsilon$; and the third equality is by algebra.

As $n_b = n$ is equivalent to $\rho^n = \rho_b$, this implies that

$$\mathbb{E}\left[\mathcal{R}(T) \mid \rho_b\right] = \tilde{\mathcal{O}}\left(\sqrt{\rho_b \left|\mathcal{I}_{5\epsilon_b}\right| MT} + M\left|\mathcal{I}_{5\epsilon_b}\right| + \min\left(M\sqrt{\rho_b \left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right);$$

observe that the expression inside $\tilde{\mathcal{O}}$ in the last line is a concave function of ρ_b.

Now, by the law of total expectation,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathcal{R}(T) \mid \rho_b\right]\right] \\
&= \mathbb{E}\left[\tilde{\mathcal{O}}\left(\sqrt{\rho_b \left|\mathcal{I}_{5\epsilon_b}\right| MT}\right) + M\left|\mathcal{I}_{5\epsilon_b}\right| + \min\left(M\sqrt{\rho_b \left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right] \\
&= \tilde{\mathcal{O}}\left(\mathbb{E}\left[\sqrt{\rho_b \left|\mathcal{I}_{5\epsilon_b}\right| MT} + M\left|\mathcal{I}_{5\epsilon_b}\right| + \min\left(M\sqrt{\rho_b \left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right]\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\rho_b\right] \left|\mathcal{I}_{5\epsilon_b}\right| MT} + M\left|\mathcal{I}_{5\epsilon_b}\right| + \min\left(M\sqrt{\mathbb{E}\left[\rho_b\right] \left|\mathcal{I}_{5\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right),
\end{aligned}$$

where the third equality is by algebra, and the last equality uses Jensen's inequality. □

A.7 Experimental Details

In Appendix A.7.1, we provide a proof of Fact 2.13 which is about the instance generation procedure. Then, in Appendix A.7.2, we present comprehensive results from the simulations we performed.

A.7.1 Proof of Fact 2.13

Proof of Fact 2.13. For every i, as $\mu_i^p \in [\mu_i^1 - \frac{\epsilon}{2}, \mu_i^1 + \frac{\epsilon}{2}]$ for all p in $[M]$, we have that for all p, q in $[M]$, $\left|\mu_i^p - \mu_i^q\right| \leq \epsilon$. This proves that $\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ is indeed a Bernoulli ϵ-MPMAB instance.

Recall that $d = \max_{i\in[c]} \mu_i^1 = \max_{i\in[K]} \mu_i^1$ is the optimal mean reward for player 1. We now show that $\mathcal{I}_{5\epsilon} = \{c+1, \ldots, K\}$ by a case analysis:

1. First, we show that for all i in $\{c+1, \ldots, K\}$, i is in $\mathcal{I}_{5\epsilon}$. This is because μ_i^1 is chosen from $[0, d - 5\epsilon)$, which implies that $\Delta_i^1 > 5\epsilon$.

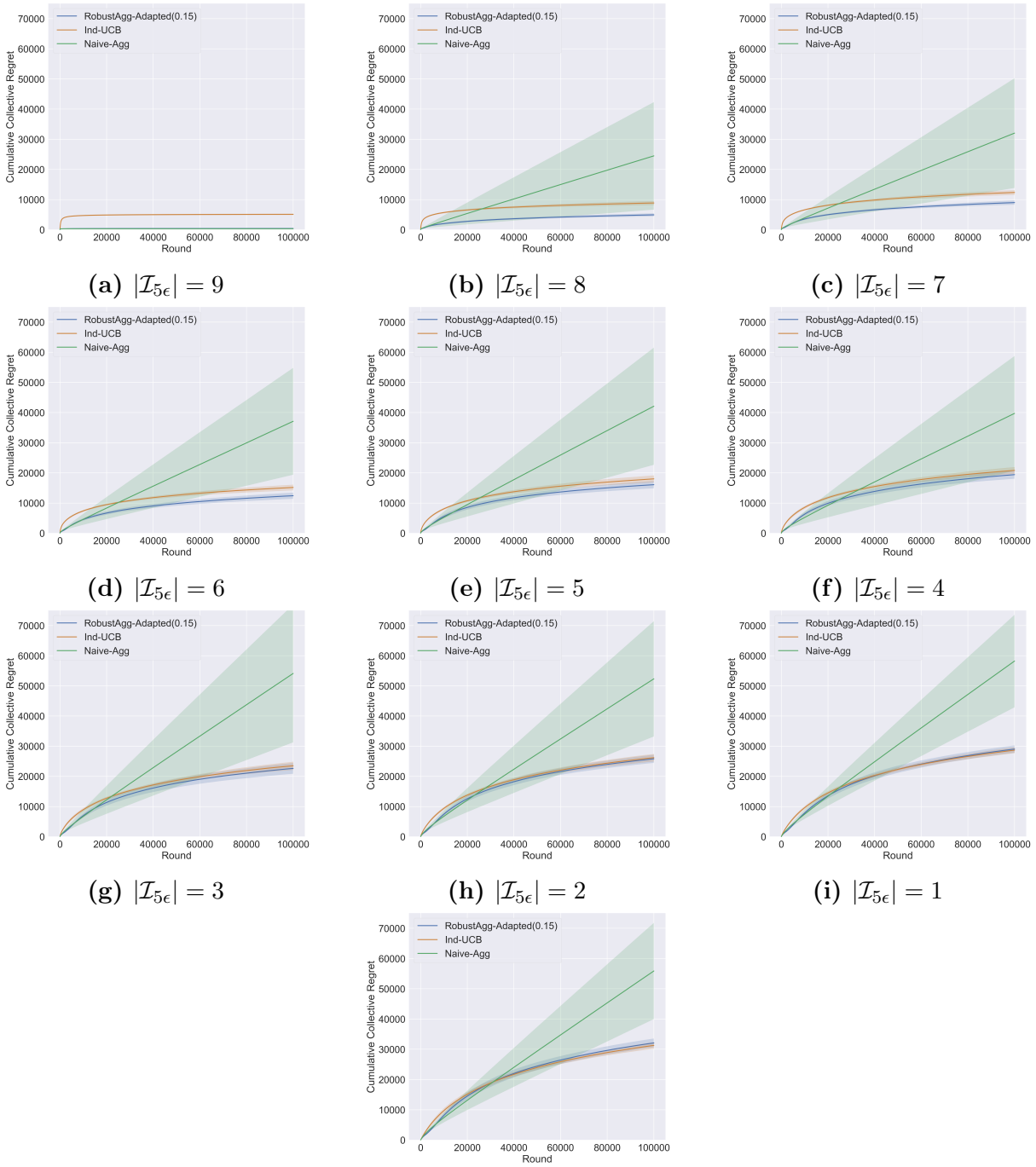
2. Second, for all i in $\{1, \ldots, c\}$, we claim that $i \notin \mathcal{I}_{5\epsilon}$. To this end, we show that for all p, $\Delta_i^p \leq 5\epsilon$.

   We start with the observation that $\mu_i^1 \in (d - \epsilon, d]$, which implies that $\Delta_i^1 = d - \mu_i^1 \leq \epsilon$. Now, it follows from Fact A.1 in Appendix A.3 that for any $i \in [K]$ and $p \in [M]$, $|\Delta_i^p - \Delta_i^1| \leq 2\epsilon$. Therefore, we have $\Delta_i^p \leq 3\epsilon$ for all p, which implies that any $i \in [c]$ cannot be in $\mathcal{I}_{5\epsilon}$.

□

A.7.2 Extended results

Here, we present comprehensive results from the simulations we performed.



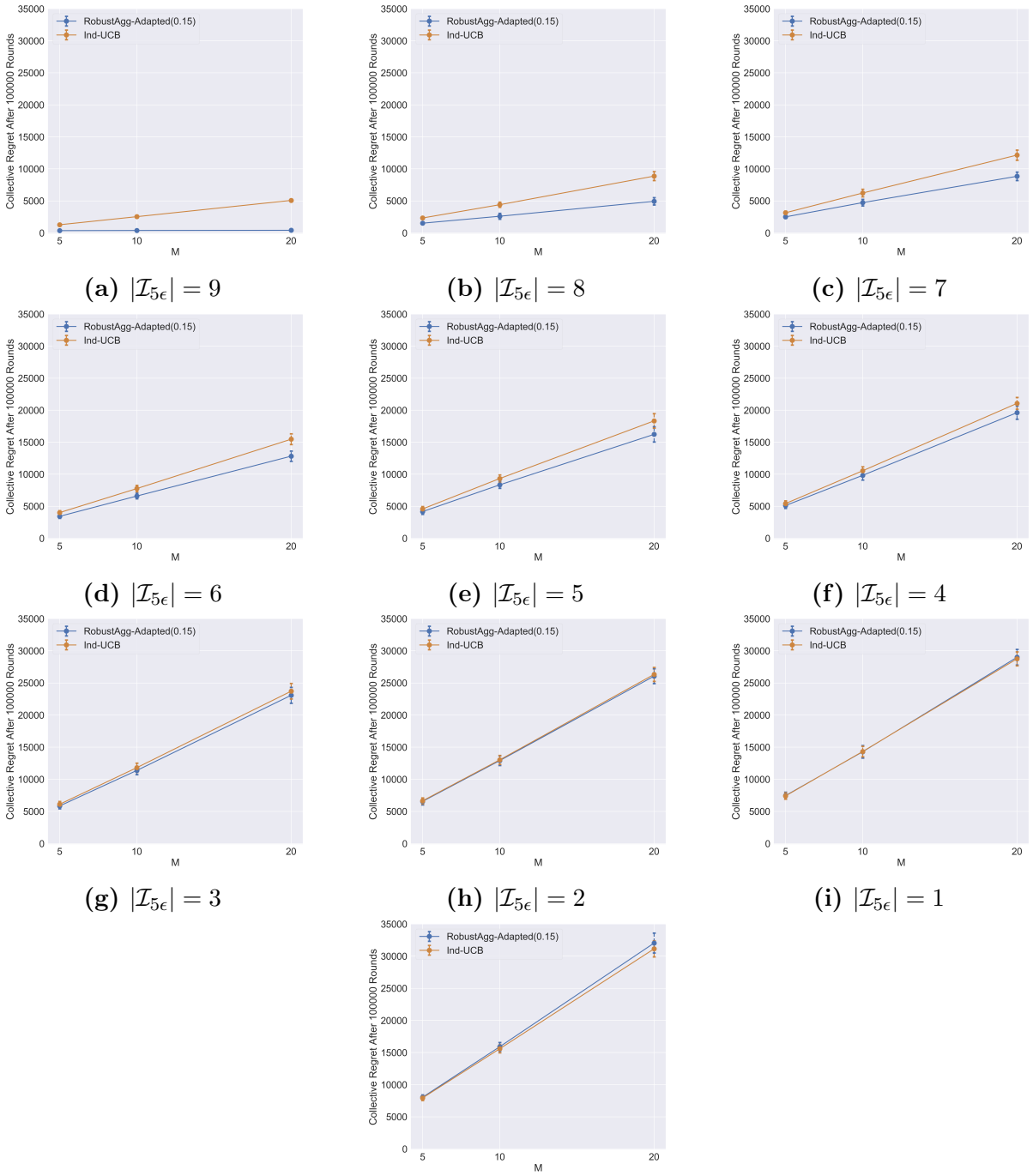
(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 8$ **(c)** $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$ **(e)** $|\mathcal{I}_{5\epsilon}| = 5$ **(f)** $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$ **(h)** $|\mathcal{I}_{5\epsilon}| = 2$ **(i)** $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure A.1. Compares the average performance of RobustAgg-Adapted(0.15), Ind-UCB, and Naive-Agg on randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$ and $M = 20$. The x-axis shows a horizon of $T = 100,000$ rounds, and the y-axis shows the cumulative collective regret of the players.



(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 8$ **(c)** $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$ **(e)** $|\mathcal{I}_{5\epsilon}| = 5$ **(f)** $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$ **(h)** $|\mathcal{I}_{5\epsilon}| = 2$ **(i)** $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure A.2. Compares the average performance of RobustAgg-Adapted(0.15) and Ind-UCB on randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$ and $M \in \{5, 10, 20\}$. The x-axis shows different values of M, and the y-axis shows the cumulative collective regret of the players after $100,000$ rounds.

Experiment 1.

Recall that for each $v \in \{0, 1, 2, \ldots, 9\}$, we generated 30 Bernoulli 0.15-MPMAB problem instances such that $|\mathcal{I}_{5\epsilon}| = v$. Figure A.1 compares the average cumulative collective regrets of the three algorithms in a horizon of $100,000$ rounds over instances with different values of $|\mathcal{I}_{5\epsilon}|$:

- Notice that RobustAgg-Adapted(0.15) outperforms both baseline algorithms when $|\mathcal{I}_{5\epsilon}| \in [2, 8]$, as shown in Figures A.1b, A.1c, ..., A.1h, especially when $|\mathcal{I}_{5\epsilon}|$ is large.
- Figure A.1a shows that when $|\mathcal{I}_{5\epsilon}| = 9$—i.e., when one arm is optimal for all players and the other arms are all subpar arms—Naive-Agg and RobustAgg-Adapted(0.15) perform much better than Ind-UCB with little difference between themselves. However, note that as long as there are more than one "competitive" arms—e.g., in Figure A.1b when $|\mathcal{I}^C_{5\epsilon}| = 2$—the collective regret of Naive-Agg can easily be nearly linear in the number of rounds.
- Figure A.1i and Figure A.1j demonstrate that when there are very few arms or even no arm that is amenable to data aggregation, RobustAgg-Adapted(0.15) has performance that is still on par with that of Ind-UCB.

Experiment 2.

Recall that for each $M \in \{5, 10, 20\}$ and $v \in \{0, 1, 2, \ldots, 9\}$, we generated 30 Bernoulli 0.15-MPMAB problem instances with M players such that $|\mathcal{I}_{5\epsilon}| = v$. Figure A.2 shows and compares the average collective regrets of RobustAgg-Adapted(0.15) and Ind-UCB after $100,000$ rounds in problem instances with $M = 5, 10$, and 20, and in each subfigure, $|\mathcal{I}_{5\epsilon}|$ takes a different value.

Observe that when $|\mathcal{I}_{5\epsilon}|$ is large (e.g., in Figures A.2a, A.2b,..., A.2e), the collective regret of RobustAgg-Adapted(0.15) is less sensitive to the number of players M, in comparison with Ind-UCB. Especially, in the extreme case when $|\mathcal{I}_{5\epsilon}| = 9$—i.e.,

all suboptimal arms are subpar arms—Figure A.2a shows that the collective regret of RobustAgg-Adapted(0.15) has negligible dependence on M.

In conclusion, our empirical evaluation validate our theoretical results in Section 2.3.

A.8 Analytical Solution to λ^*

We first present the following proposition simliar to the results in [19, Section 6 thereof]. The original solution in [19] has a $\min(1, \cdot)$ operation in the second case; we slightly simplify that result by showing that this operation is unnecessary.[4]

Proposition A.19. *Suppose $\beta \in (0, 1)$. Define function*

$$f(\alpha) = 2B\sqrt{\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + 2(1-\alpha)A,$$

Then, $\alpha^* = \operatorname{argmin}_{\alpha\in[0,1]} f(\alpha)$ has the following form:

$$\alpha^* = \begin{cases} 1 & \beta \geq \frac{B^2}{A^2}, \\ \beta\left(1 + \frac{1-\beta}{\sqrt{\frac{B^2}{A^2} - \beta(1-\beta)}}\right) & \beta < \frac{B^2}{A^2}. \end{cases}$$

Observe that when $\beta < \frac{B^2}{A^2}$, $\frac{B^2}{A^2} - \beta(1-\beta) > 0$, so the expression in the second case is well defined.

Proof. First, observe that f is a strictly convex function, and therefore has at most one stationary point in $\mathbb{R}$; and if it exists, it must be f's global minimum.

Second, we study the monotonicity property of f in $\mathbb{R}$. To this end, we calculate

[4] In [19]'s notation, this can also be seen directly by observing that when $m_T \geq D^2$, $v = \frac{m_T}{m_T+m_S} \cdot \left(1 + \frac{m_S}{\sqrt{D^2(m_S+m_T) - m_S m_T}}\right) \leq \frac{m_T}{m_T+m_S} \cdot (1 + \frac{m_S}{m_T}) = 1$.

α_0, the stationary point of f. We have

$$f'(\alpha) = 2B\frac{\frac{\alpha}{\beta} - \frac{1-\alpha}{1-\beta}}{\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}} - 2A$$

By algebraic calculations, $f'(\alpha) = 0$ is equivalent to

$$\frac{\alpha - \beta}{\beta(1-\beta)} = \frac{A}{B}\sqrt{\frac{\alpha^2 - 2\beta\alpha + 1}{\beta(1-\beta)}}.$$

This yields the following quadratic equation:

$$\left(\frac{B^2}{A^2} - \beta(1-\beta)\right)\alpha^2 - 2\beta\left(\frac{B^2}{A^2} - \beta(1-\beta)\right)\alpha + \beta^2\left(\frac{B^2}{A^2} - (1-\beta)\right) = 0,$$

with the constraint that $\alpha > \beta$. The discriminant of the above quadratic equation is $\Delta = 4\beta^2(1-\beta)^2(\frac{B^2}{A^2} - \beta(1-\beta))$. If $\Delta \geq 0$, the stationary point is

$$\alpha_0 = \frac{2\beta(\frac{B^2}{A^2} - \beta(1-\beta)) + \sqrt{\Delta}}{2(\frac{B^2}{A^2} - \beta(1-\beta))} = \beta\left(1 + \frac{1-\beta}{\sqrt{\frac{B^2}{A^2} - \beta(1-\beta)}}\right)$$

We now consider two cases:

1. If $\beta(1-\beta) > \frac{B^2}{A^2}$, it can be checked that $\Delta < 0$, and consequently $f'(\alpha) < 0$ for all $\alpha \in \mathbb{R}$, i.e., f is monotonically decreasing in $\mathbb{R}$.

2. $\beta(1-\beta) \leq \frac{B^2}{A^2}$, we have that f is monotonically decreasing in $(-\infty, \alpha_0]$, and monotonically increasing in $[\alpha_0, +\infty)$.

We are now ready to calculate $\alpha^* = \operatorname{argmin}_{\alpha\in[0,1]} f(\alpha)$.

1. If $\beta(1-\beta) > \frac{B^2}{A^2}$, as f is monotonically decreasing in $\mathbb{R}$, $\alpha^* = 1$.

2. If $\beta(1-\beta) \leq \frac{B^2}{A^2}$ and $\beta > \frac{B^2}{A^2}$, it can be checked that $\alpha_0 \geq 1$. As f is monotonically decreasing in $(-\infty, \alpha_0] \supset [0, 1]$, we also have $\alpha^* = 1$.

3. If $\beta \leq \frac{B^2}{A^2}$, $\alpha_0 \in [0,1]$. Therefore, $\alpha^* = \alpha_0 = \beta\left(1 + \frac{1-\beta}{\sqrt{\frac{B^2}{A^2} - \beta(1-\beta)}}\right)$.

In summary, we have the expression of α^* as desired. □

Algorithm 1's line 9 computes

$$\lambda^* = \underset{\lambda \in [0,1]}{\operatorname{argmin}} \; 8\sqrt{13(\ln T)[\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon;$$

we now use Proposition A.19 to give its analytical form. For notational simplicity, let $n = \overline{n_i^p}(t-1)$ and $m = \overline{m_i^p}(t-1)$. Applying Proposition A.19 with $A = \frac{\epsilon}{2}, B = 4\sqrt{\frac{13(\ln T)}{n+m}}$, and $\beta = \frac{n}{n+m}$, we have

$$\lambda^* = \begin{cases} 1 & \epsilon > 0 \text{ and } n \geq \frac{832(\ln T)}{\epsilon^2}, \\ \frac{n}{n+m}\left(1 + \epsilon m\sqrt{\frac{1}{832(\ln T)(n+m) - \epsilon^2 nm}}\right) & \text{otherwise.} \end{cases}$$

Appendix B

Supplementary Material for Chapter 3

B.1 Basic Definitions and Facts

In this section, we revisit and introduce a few basic definitions, facts and additional notations that are useful in our proofs.

Definition B.1 (Constants used in the analysis)**.** *In the analysis, we set*

$$c_1 = 40, c_2 = 4$$

to be the constants used in Algorithm 2.[1]

Definition B.2 (Number of pulls)**.** *Recall that*

$$n_i^p(t) = \sum_{s \leq t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\}$$

is the number of pulls of arm i by player p after t rounds. We define

$$n_i(t) = \sum_{p \in [M]} n_i^p(t)$$

[1]If we choose c_1 to some other positive number, we can still show guarantees similar to Theorems 3.1 and 3.2, except that $\mathcal{I}_{10\epsilon}$ needs to be changed to $\mathcal{I}_{\mathcal{O}\left(\sqrt{\frac{1}{c_1}}\epsilon\right)}$—the analysis of case $(A1)$ needs to be changed accordingly. On the other hand, it is also possible to change c_2 to any constant > 1 and establish similar regret guarantees, by tightening the exponents of the concentration inequalities (Corollaries B.28 and B.30) and Lemma B.70. We leave the details to interested readers.

to be the total of number of pulls of arm i by all the players after t rounds.

Definition B.3 (Individual mean estimate)**.** *For any $i \in [K]$, $p \in [M]$, and $t \in [T] \cup \{0\}$, let*

$$\text{ind-}\hat{\mu}_i^p(t) = \frac{1}{n_i^p(t) \vee 1} \sum_{s \le t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\} r_s^p$$

be the empirical mean computed for arm i using player p's own data from the first t rounds.

Definition B.4. *Define*

$$\text{ind-var}_i^p(t) = \frac{4}{n_i^p(t) \vee 1}.$$

Remark B.5 (mean and variance of the individual posteriors)**.** *By the construction of Algorithm 2, we have that, in any round $t \in [T]$, for any active player $p \in \mathcal{P}_t$ and arm i,* $\text{ind-}\hat{\mu}_i^p(t-1)$ *and* $\text{ind-var}_i^p(t-1)$ *are the mean and variance of the individual posterior associated with arm i and player p in round t, respectively.*

Definition B.6 (Aggregate mean estimate)**.** *For any $i \in [K]$ and $t \in [T] \cup \{0\}$, let*

$$\text{agg-}\hat{\mu}_i(t) = \frac{1}{n_i(t) \vee 1} \sum_{s \le t} \sum_{q : q \in \mathcal{P}_s} \mathbb{1}\left\{i_s^q = i\right\} r_s^q + \epsilon$$

be the empirical mean computed for arm i using all players' data from the first t rounds, offset by the dissimilarity parameter ϵ. Note that the definition of $\text{agg-}\hat{\mu}_i(t)$ *does not depend on the identity of a specific player p.*

Definition B.7 (Most recent pull)**.** *In any round $t \in [T] \cup \{0\}$, for any player $p \in [M]$ and arm $i \in [K]$, we define*

$$u_i^p(t) = \begin{cases} \max\left\{s \le t : p \in \mathcal{P}_s, i_s^p = i\right\}, & n_i^p(t) > 0 \\ 0, & n_i^p(t) = 0 \end{cases}$$

to be the round in which player p most recently pulled arm i (including round t); we let

$u_i^p(t) = 0$ *by convention if player* p *has not yet pulled arm* i*.*

Definition B.8 (Aggregate mean estimate maintained by player p)**.** *For any* $t \in [T] \cup \{0\}$*,* $p \in [M]$*, and* $i \in [K]$*, define*

$$\text{agg-}\hat{\mu}_i^p(t) = \text{agg-}\hat{\mu}_i(u_i^p(t)).$$

Note that the superscript p *differentiates this player-dependent aggregate mean estimate from* $\text{agg-}\hat{\mu}_i(t)$ *in Definition B.6, which does not depend on any individual player.*

Definition B.9 (Aggregate number of pulls maintained by player p)**.** *For any* $t \in [T] \cup \{0\}$*,* $p \in [M]$*, and* $i \in [K]$*, define*

$$m_i^p(t) = n_i(u_i^p(t))$$

to be the total number of pulls of arm i *by all the players until the round in which player* p *last pulled arm* i*.*

Definition B.10. *Define*

$$\text{agg-var}_i^p(t) = \frac{4}{(m_i^p(t) - M) \vee 1}.$$

Remark B.11 (mean and variance of the aggregate posteriors)**.** *By the construction of Algorithm 2, in any round* $t \in [T]$*, for any active player* $p \in \mathcal{P}_t$ *and arm* i*, we have that* $\text{agg-}\hat{\mu}_i^p(t-1)$ *and* $\text{agg-var}_i^p(t-1)$ *are the mean and variance of the aggregate posterior associated with arm* i *and player* p *in round* t*, respectively.*

Definition B.12 (Filtration)**.** *Let* $\{\mathcal{F}_t\}_{t=0}^T$ *be a filtration such that*

$$\mathcal{F}_t = \sigma\left(\{i_s^q, r_s^q : s \leq t, q \in \mathcal{P}_s\}\right)$$

is the σ*-algebra generated by interactions of all players up until and including round* t*.*

Definition B.13. *Let*

$$H_i^p(t) = \left\{ n_i^p(t-1) \geq \frac{40 \ln T}{\epsilon^2} + 2M \right\}$$

be the event that in round t, for arm i, player p uses the individual *posterior distribution; correspondingly, let*

$$\overline{H_i^p(t)} = \left\{ n_i^p(t-1) < \frac{40 \ln T}{\epsilon^2} + 2M \right\}$$

be the event that in round t, for arm i, player p uses the aggregate *posterior distribution. See lines 6 to 9 in Algorithm 2.*

Remark B.14. *With the above notations,*

$$\hat{\mu}_i^p(t-1) = \text{agg-}\hat{\mu}_i^p(t-1) \cdot \mathbb{1}(\overline{H_i^p(t)}) + \text{ind-}\hat{\mu}_i^p(t-1) \cdot \mathbb{1}(H_i^p(t)),$$

and

$$\text{var}_i^p(t-1) = \text{agg-var}_i^p(t-1) \cdot \mathbb{1}(\overline{H_i^p(t)}) + \text{ind-var}_i^p(t-1) \cdot \mathbb{1}(H_i^p(t)).$$

Stopping times.

In our analysis, we will frequently use the following notions of stopping times:

Definition B.15. *For any arm $i \in [K]$ and $k \in [TM]$, let*

$$\tau_k(i) = \min \left\{ T+1, \min \left\{ t : n_i(t) \geq k \right\} \right\}$$

be the round in which arm i is pulled the k-th time by any player. Furthermore, as a convention, let $\tau_0(i) = 0$.

Remark B.16. *For any $i \in [K]$ and $k \in [TM]$, $\tau_k(i)$ is a stopping time with respect to*

$\{\mathcal{F}_t\}_{t=0}^T$. *Indeed, for any* $t \leq T$,

$$\left\{\tau_k(i) \leq t\right\} = \left\{\sum_{s\in[t]} \sum_{p:p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\} \geq k\right\} \in \mathcal{F}_t.$$

Definition B.17. *For any arm* $i \in [K]$ *and* $k \in [TM]$, *such that* $\tau_k(i) \leq T$, *let* $p_k(i)$ *be the unique* $p \in [M]$ *such that* $i_{\tau_k(i)}^p = i$ *and*

$$\sum_{s=1}^{\tau_k(i)-1} \sum_{q\in\mathcal{P}_s} \mathbb{1}\left\{i_s^q = i\right\} + \sum_{q\in\mathcal{P}_{\tau_k(i)}:q\leq p} \mathbb{1}\left\{i_s^q = i\right\} = k.$$

In words, $p_k(i)$ *is the player that makes the* k*-th pull of arm* i, *where arm pulls within a round are ordered by the indices of active players in that round.*

Definition B.18. *For any arm* $i \in [K]$, *player* $p \in [M]$, *and* $k \in [T]$, *let*

$$\pi_k(i,p) = \min\left\{T+1, \min\left\{t : n_i^p(t) \geq k\right\}\right\}$$

be the round in which arm i *is pulled the* k*-th time by player* p. *In addition, let* $\pi_0(i,p) = 0$ *by convention.*

Remark B.19. *For any* $i \in [K]$ *and* $k \in [T]$, $\pi_k(i,p)$ *is a stopping time with respect to* $\{\mathcal{F}_t\}_{t=0}^T$. *Indeed, for any* $t \leq T$,

$$\left\{\pi_k(i,p) \leq t\right\} = \left\{\sum_{s\in[t]:p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\} \geq k\right\} \in \mathcal{F}_t.$$

The following property, namely, the invariant property, will also be useful for our analysis.

Definition B.20 (Invariant property)**.** *We say that:*

1. *a set of random variables* $\{g_t : t \in [T]\}$ *satisfies the* invariant property with respect to arm $i \in [K]$ and player $p \in [M]$, *if* g_t *stays constant/invariant between two consecutive pulls of arm* i *by player* p, *i.e., for any* $s \in [T]$ *such that* $\pi_s(i,p) \leq T$, g_t *is constant for all* $t \in [\pi_{s-1}(i,p)+1, \pi_s(i,p)]$. *In other words, for any* $s \in [T]$ *such that* $\pi_s(i,p) \leq T$,

$$g_{\pi_{s-1}(i,p)+1} = g_{\pi_{s-1}(i,p)+2} = \ldots = g_{\pi_s(i,p)}.$$

2. *a set of random variables* $\{f_t^p : t \in [T], p \in [M]\}$ *satisfies the* invariant property with respect to arm $i \in [K]$, *if for every player* $p \in [M]$, $\{f_t^p : t \in [T]\}$ *satisfy the invariant property with respect to* (i,p).

Example B.21. *By the construction of Algorithm 2, in any round* t, *a player only updates the posteriors associated with an arm if the player pulls the arm in round* t *(line 15). It is easy to verify that for any arm* $i \in [K]$ *and* $p \in [M]$, $\{H_i^p(t) : t \in [T]\}$ *satisfies the invariant property with respect to* (i,p). *Specifically, for any* $s \in [T]$ *such that* $\pi_s(i,p) \leq T$,

$$H_i^p(\pi_{s-1}(i,p)+1) = H_i^p(\pi_{s-1}(i,p)+2) = \ldots = H_i^p(\pi_s(i,p)).$$

Consequently, $\{H_i^p(t) : t \in [T], p \in [M]\}$ *satisfies the invariant property with respect to* i.

Example B.22. *For any arm* $i \in [K]$ *and any player* $p \in [M]$, $\{n_i^p(t-1) : t \in [T]\}$ *and* $\{m_i^p(t-1) : t \in [T]\}$ *both satisfy the invariant property with respect to* (i,p) *(see Definition B.2 and Definition B.9, respectively). Specifically, for any player* p *and any* $s \in [T]$ *such that* $\pi_s(i,p) \leq T$,

$$n_i^p(\pi_{s-1}(i,p)) = n_i^p(\pi_{s-1}(i,p)+1) = \ldots = n_i^p(\pi_s(i,p)-1) = s-1,$$

$$m_i^p(\pi_{s-1}(i,p)) = m_i^p(\pi_{s-1}(i,p)+1) = \ldots = m_i^p(\pi_s(i,p)-1) = n_i(\pi_{s-1}(i,p))$$

However, $\{n_i^p(t) : t \in [T]\}$ *and* $\{m_i^p(t) : t \in [T]\}$ *do* not *necessarily satisfy the*

invariant property with respect to i.

Similarly, the sets $\left\{\text{ind-}\hat{\mu}_i^p(t-1) : t \in [T]\right\}$, $\left\{\text{ind-var}_i^p(t-1) : t \in [T]\right\}$, $\left\{\text{agg-}\hat{\mu}_i^p(t-1) : t \in [T]\right\}$, $\left\{\text{agg-var}_i^p(t-1) : t \in [T]\right\}$ *all satisfy the invariant property with respect to* (i,p).

Example B.23. *For any arm* $i \in [K]$ *and any player* $p \in [M]$, $\left\{\hat{\mu}_i^p(t-1) : t \in [T]\right\}$ *satisfy the invariant property with respect to* (i,p). *This follows from Eq.* (B.14) *and the above two examples that*

$$\left\{\text{ind-}\hat{\mu}_i^p(t-1) : t \in [T]\right\}, \left\{\text{agg-}\hat{\mu}_i^p(t-1) : t \in [T]\right\}, \text{ and } \left\{H_i^p(t) : t \in [T]\right\}$$

all satisfy the invariant property with respect to (i,p).

Following a similar reasoning, $\left\{\text{var}_i^p(t-1) : t \in [T]\right\}$ *satisfy the invariant property with respect to* (i,p).

Facts about Subpar Arms.

We now present some facts about subpar arms.

Fact B.24 (Properties of subpar arms, see also Fact A.2)**.** *The following are true:*

1. *for any* $i \in [K]$ *and* $p, q \in [M]$, $\left|\Delta_i^p - \Delta_i^q\right| \le 2\epsilon$ *(see Fact A.1);*
2. *For any* $i \in \mathcal{I}_{10\epsilon}$ *and* $p \in [M]$, $\Delta_i^p > 8\epsilon$, *which means that* $\Delta_i^{\min} > 8\epsilon$.
3. $\left|\mathcal{I}_{2\epsilon}^C\right| \ge 1$;
4. *Let* $\Delta_i^{\max} = \max_{p\in[M]} \Delta_i^p$. *For any* $i \in \mathcal{I}_{10\epsilon} \subseteq \mathcal{I}_{5\epsilon}$, $\Delta_i^{\max} \le 2\Delta_i^{\min}$; *furthermore,* $\frac{1}{\Delta_i^{\min}} \le \frac{2}{M}\sum_{p\in[M]} \frac{1}{\Delta_i^p}$ *(see Fact A.2).*

Proof. For item 2, by the definition of $\mathcal{I}_{10\epsilon}$, there exists p such that $\Delta_i^p > 10\epsilon$. Then, for all $q \in [M]$, we have $\Delta_i^q > 8\epsilon$ by item 1.

For item 3, using a similar argument, we have, for each $i \in \mathcal{I}_{2\epsilon}$ and $p \in [M]$, $\Delta_i^p > 0$. Let j be an optimal for player 1 such that $\Delta_j^p = 0$. Then $j \notin \mathcal{I}_{2\epsilon}$. □

Additional notations.

- Denote by $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ the cumulative distribution function (CDF) of the standard Gaussian distribution.
- Let $\overline{\Phi}(x) = 1 - \Phi(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ denote the complementary CDF of the standard Gaussian distribution.
- Denote by $(z)_+ = z \vee 0$.
- For any arm $i \in [K]$, player $p \in [M]$ and $t \in [T] \cup \{0\}$, let

$$\overline{n_i^p}(t) := n_i^p(t) \vee 1,$$

and

$$\overline{m_i^p}(t) := (m_i^p(t) - M) \vee 1.$$

B.2 Concentration Bounds

B.2.1 Novel concentration inequality for multi-task data aggregation at random stopping time τ_k's

We begin by introducing the following definition.

Definition B.25 (Mixture expected reward at t)**.** *For any arm $i \in [K]$ and $t \in [T]$, define*

$$\tilde{\mu}_i(t) = \frac{1}{n_i(t) \vee 1} \sum_{s \le t} \sum_{q \in \mathcal{P}_s} \mathbb{1}\left\{i_s^q = i\right\} \mu_i^q + \epsilon$$

to be the ϵ-offset mixture expected reward of arm i up to round t.

In what follows, we will consider $\tilde{\mu}_i(\tau_k(i))$ for any $i \in [K]$ and $k \in [TM]$, where the definition of $\tau_k(i)$ can be found in Definition B.15.

Lemma B.26. *For any arm $i \in [K]$ and $k \in [TM]$, denote by $\tau_k = \tau_k(i)$. If $\tau_k \leq T$, then for every player $p \in [M]$, we have*

$$\begin{aligned}
&\text{agg-}\hat{\mu}_i(\tau_k) - \mu_i^p \leq \text{agg-}\hat{\mu}_i(\tau_k) - \tilde{\mu}_i(\tau_k) + 2\epsilon; \text{ and} \\
&\mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \leq \tilde{\mu}_i(\tau_k) - \text{agg-}\hat{\mu}_i(\tau_k).
\end{aligned}$$

Proof. For every $t \in [T]$, observe that

$$\tilde{\mu}_i(t) = \frac{1}{n_i(t) \vee 1} \sum_{s \leq t} \sum_{\substack{q \in \mathcal{P}_s: \\ i_s^q = i}} \mu_i^q + \epsilon = \sum_{q \in [M]} \frac{n_i^q(t) \cdot \mu_i^q}{n_i(t) \vee 1} + \epsilon.$$

It can be easily verified that, if $n_i(t) > 0$, for every player $p \in [M]$,

$$\tilde{\mu}_i(t) - \mu_i^p \leq 2\epsilon \text{ and } \mu_i^p - \tilde{\mu}_i(t) \leq 0,$$

where we note that the asymmetry comes from the additive term ϵ in $\tilde{\mu}_i(t)$. Therefore, for $k \in [TM]$, if $\tau_k \leq T$, then $n_i(\tau_k) \geq k > 0$ and we have

$$\tilde{\mu}_i(\tau_k) - \mu_i^p \leq 2\epsilon \text{ and } \mu_i^p - \tilde{\mu}_i(\tau_k) \leq 0.$$

It then follows that, for every player $p \in [M]$,

$$\begin{aligned}
&\text{agg-}\hat{\mu}_i(\tau_k) - \mu_i^p \leq \text{agg-}\hat{\mu}_i(\tau_k) - \tilde{\mu}_i(\tau_k) + 2\epsilon, \text{ and} \\
&\mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \leq \tilde{\mu}_i(\tau_k) - \text{agg-}\hat{\mu}_i(\tau_k).
\end{aligned}$$

□

We are now ready to present Lemma B.27, our novel concentration bound (see also Lemma 3.8).

Lemma B.27. *For any arm $i \in [K]$ and $k \in [TM] \cup \{0\}$, denote by $\tau_k = \tau_k(i)$; for $\delta \in (0, 1]$, we have*

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{ \{\tau_k \le T\} \cap \left\{ \forall p \in [M],\ \text{agg-}\hat{\mu}_i(\tau_k) - \mu_i^p \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}} + 2\epsilon \right\} \right\} \right) > 1 - \delta; \tag{B.1}$$

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{ \{\tau_k \le T\} \cap \left\{ \forall p \in [M],\ \mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}} \right\} \right\} \right) > 1 - \delta. \tag{B.2}$$

The following corollary is an equivalent form of Equation (B.2):

Corollary B.28. *For any arm $i \in [K]$ and $k \in [TM] \cup \{0\}$, denote by $\tau_k = \tau_k(i)$. Equivalently, for any $z \ge 0$, we have*

$$\Pr\left((\tau_k \le T) \wedge \left(\exists p \in [M],\ \mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \ge z\sqrt{\frac{4}{\left(n_i(\tau_k) - M\right) \vee 1}} \right) \right) \le 2e^{-2z^2}. \tag{B.3}$$

Proof of Corollary B.28. If $z \le \sqrt{\frac{1}{2}\ln 2}$, Equation (B.3) holds trivially as $2e^{-2z^2} \ge 1$. Otherwise $z > \sqrt{\frac{1}{2}\ln 2}$. In this case, let $\delta = 2e^{-2z^2} \in (0, 1]$ in Equation (B.2), and using De Morgan's law, we also obtain Equation (B.3). □

Proof of Lemma B.27. Fix any arm $i \in [K]$. For $k = 0$, we have $\tau_0 = 0$; both Eq. (B.1) and Eq. (B.2) hold trivially because for all $p \in [M]$ and $\delta \in (0, 1]$, $\left|\text{agg-}\hat{\mu}_i(\tau_0) - \mu_i^p\right| \leq 1 \leq \sqrt{2\ln 2} \leq \sqrt{2\ln(\frac{2}{\delta})}$.

We now focus on $k \in [TM]$. By Lemma B.26, it suffices to show that

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{\{\tau_k \leq T\} \cap \left\{\text{agg-}\hat{\mu}_i(\tau_k) - \tilde{\mu}_i(\tau_k) \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}}\right\}\right\}\right) > 1 - \delta; \text{ and,} \tag{B.4}$$

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{\{\tau_k \leq T\} \cap \left\{\tilde{\mu}_i(\tau_k) - \text{agg-}\hat{\mu}_i(\tau_k) \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}}\right\}\right\}\right) > 1 - \delta.$$

To avoid redundancy, we only prove Eq. (B.4); the other inequality follows by symmetry.

Now, for $t \in [T] \cup \{0\}$, consider $Z_t = \sum_{s=1}^{t} \sum_{p \in \mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\}(r_s^p - \mu_i^p)$. Furthermore, for $t \in [T] \cup \{0\}$ and $\lambda > 0$, let

$$w_t(\lambda) = \exp\left(\lambda Z_t - n_i(t)\frac{\lambda^2}{8}\right).$$

We now show that $\left\{w_t(\lambda)\right\}_{t=0}^{T}$ is a nonnegative supermartingale with respect to $\{\mathcal{F}_t\}_{t=0}^{T}$ for all $\lambda > 0$. Since $\mathbb{E}\left[\left|w_t(\lambda)\right|\right] < \infty$ and $w_t(\lambda) \geq 0$ for all $t \in [T] \cup \{0\}$, it suffices to show

that, for all $t \in [T]$,

$$
\begin{aligned}
&\mathbb{E}\left[w_t(\lambda) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\exp\left(\sum_{s\in[t]}\sum_{p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\Bigg[\exp\left(\sum_{s\in[t-1]}\sum_{p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \cdot \\
&\qquad\qquad \exp\left(\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}\left(\lambda(r_t^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\Bigg] \\
=&\exp\left(\sum_{s\in[t-1]}\sum_{p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \cdot \\
&\qquad\qquad \mathbb{E}\left[\exp\left(\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}\left(\lambda(r_t^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
=&w_{t-1}(\lambda) \cdot \mathbb{E}\left[\exp\left(\lambda\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}(r_t^p - \mu_i^p)\right)\exp\left(-\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \\
\leq&w_{t-1}(\lambda),
\end{aligned}
$$

where the last inequality uses the law of iterated expectation along with Hoeffding's lemma,

i.e.,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}\left(r_t^p-\mu_i^p\right)\right)\cdot\exp\left(-\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}\frac{\lambda^2}{8}\right)\mid\mathcal{F}_{t-1}\right]\\
\leq&\mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}\left(r_t^p-\mu_i^p\right)\right)\mid\mathcal{F}_{t-1},(i_t^p)_{p\in\mathcal{P}_t}\right]\cdot\right.\\
&\qquad\left.\exp\left(-\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}\frac{\lambda^2}{8}\right)\mid\mathcal{F}_{t-1}\right]\\
\leq&\mathbb{E}\left[\prod_{p\in\mathcal{P}_t}\exp\left(\frac{\lambda^2\cdot\left(\mathbb{1}\left\{i_t^p=i\right\}\right)^2}{8}\right)\cdot\exp\left(-\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}\frac{\lambda^2}{8}\right)\mid\mathcal{F}_{t-1}\right]\leq 1
\end{aligned}$$

Recall from Remark B.16 that τ_k is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$ and $\tau_k\leq T+1<\infty$ almost surely, it follows that, by the optional sampling theorem, for all $\lambda>0$,

$$\mathbb{E}\left[\mathbb{1}\left\{\tau_k\leq T\right\}\cdot w_{\tau_k}(\lambda)\right]\leq\mathbb{E}\left[w_0(\lambda)\right]=1. \tag{B.5}$$

Rewriting Eq. (B.5), we have

$$\mathbb{E}\left[\mathbb{1}\left\{\tau_k\leq T\right\}\cdot\exp\left(\lambda Z_{\tau_k}-n_i(\tau_k)\frac{\lambda^2}{8}\right)\right]\leq 1.$$

It then follows that, by Markov's inequality, for any $\delta > 0$,

$$\begin{aligned}
&\Pr\left(\mathbb{1}\left\{\tau_k \leq T\right\} \cdot \exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right) \geq \frac{1}{\delta}\right) \\
\leq\ & \frac{\mathbb{E}\left[\mathbb{1}\left\{\tau_k \leq T\right\} \cdot \exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right)\right]}{\frac{1}{\delta}} \\
\leq\ & \delta;
\end{aligned}$$

therefore,

$$\Pr\left(\left\{\tau_k \leq T\right\} \cap \left\{\exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right) \geq \frac{1}{\delta}\right\}\right) \leq \delta.$$

Rearranging the terms in the above inequality, we have, for any $\lambda > 0$,

$$\Pr\left(\left\{\tau_k = T+1\right\} \cup \left\{\left\{\tau_k \leq T\right\} \cap \left\{\frac{1}{n_i(\tau_k)} Z_{\tau_k} - \frac{\lambda}{8} < \frac{\ln\left(\frac{1}{\delta}\right)}{n_i(\tau_k) \cdot \lambda}\right\}\right\}\right) > 1 - \delta,$$

where we use the elementary fact that for sets A and B, $\neg(A \cap B) = \neg A \cup (A \cap \neg B)$.

Choosing $\lambda = \sqrt{\frac{\ln(\frac{1}{\delta})}{k}}$ and using the fact that $n_i(\tau_k) \geq k$, we have

$$\Pr\left(\left\{\tau_k = T+1\right\} \cup \left\{\left\{\tau_k \leq T\right\} \cap \left\{\frac{1}{n_i(\tau_k)} Z_{\tau_k} < \sqrt{\frac{2\ln(\frac{1}{\delta})}{k}}\right\}\right\}\right) > 1 - \delta;$$

it then follows that

$$\Pr\left(\left\{\tau_k = T+1\right\} \cup \left\{\left\{\tau_k \leq T\right\} \cap \left\{\frac{1}{n_i(\tau_k)} Z_{\tau_k} < \sqrt{\frac{2\ln(\frac{2}{\delta})}{k}}\right\}\right\}\right) > 1 - \delta. \tag{B.6}$$

We now consider two cases:

1. $n_i(\tau_k) \leq M$. We have $\frac{1}{n_i(\tau_k)} Z_{\tau_k} \leq 1 < \sqrt{\frac{2\ln(\frac{2}{\delta})}{(n_i(\tau_k)-M)\vee 1}} = \sqrt{2\ln(\frac{2}{\delta})}$ trivially for $\delta \in (0,1]$.

2. $n_i(\tau_k) \geq M+1$. Since $k \geq n_i(\tau_k) - M$, we have $\sqrt{\frac{2\ln(\frac{2}{\delta})}{k}} \leq \sqrt{\frac{2\ln(\frac{2}{\delta})}{n_i(\tau_k)-M}} = \sqrt{\frac{2\ln(\frac{2}{\delta})}{(n_i(\tau_k)-M)\vee 1}}$.

Eq. (B.4) then follows from Eq. (B.6) and the elementary fact that $A \subseteq B$ if $(A \cap C) \subseteq B$ and $(A \cap \neg C) \subseteq B$. This completes the proof. □

B.2.2 Other concentration bounds

Recall the definition of stopping times $\pi_k(i,p)$ for any arm i and player p (see Definition B.18).

Lemma B.29. *For any $i \in [K]$, $p \in [M]$, $k \in [T] \cup \{0\}$, and $\delta \in (0,1]$, we have*

$$\Pr\left(\begin{array}{l} \big\{\pi_k(i,p) = T+1\big\} \cup \\ \\ \left\{\big\{\pi_k(i,p) \leq T\big\} \cap \left\{\left|\text{ind-}\hat{\mu}_i^p(\pi_k(i,p)) - \mu_i^p\right| \leq \sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{n_i^p(\pi_k(i,p)) \vee 1}}\right\}\right\} \end{array}\right) > 1-\delta. \tag{B.7}$$

Corollary B.30. *For any $i \in [K]$, $p \in [M]$, $k \in [T] \cup \{0\}$, and $z \geq 0$, we have*

$$\Pr\left(\big(\pi_k(i,p) \leq T\big) \wedge \left(\left|\mu_i^p - \text{ind-}\hat{\mu}_i^p(\pi_k(i,p))\right| \geq z\sqrt{\frac{4}{n_i^p(\pi_k(i,p)) \vee 1}}\right)\right) \leq 4e^{-2z^2}. \tag{B.8}$$

Proof of Corollary B.30. If $z \leq \sqrt{\frac{1}{2}\ln 4}$, Equation (B.8) holds trivially as $4e^{-2z^2} \geq 1$. Otherwise $z > \sqrt{\frac{1}{2}\ln 4}$. In this case, let $\delta = 4e^{-2z^2} \in (0,1]$ in Equation (B.7), and using De Morgan's law, we also obtain Equation (B.8). □

Proof of Lemma B.29. This proof is largely similar to the one for Lemma B.27. Therefore, we omit some details here to avoid redundancy. See the proof of Lemma B.27 for full details.

Let us fix any arm $i \in [K]$ and player $p \in [M]$. Throughout this proof, to ease the exposition, we use π_k to denote $\pi_k(i,p)$.

We first observe that when $k = 0$, we have $\pi_k = 0$, ind-$\hat{\mu}_i^p(0) = 0$, and $n_i^p(0) = 0$. It follows that $\left|\text{ind-}\hat{\mu}_i^p(\pi_k) - \mu_i^p\right| \leq 1 \leq \sqrt{2\ln\left(\frac{4}{\delta}\right)}$ trivially.

It then suffices to only consider the case when $k \in [T]$. Note that $n_i^p(\pi_k) = k \geq 1$. We will show that

$$\Pr\left(\{\pi_k = T+1\} \cup \left\{\{\pi_k \leq T\} \cap \left\{\text{ind-}\hat{\mu}_i^p(\pi_k) - \mu_i^p \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{n_i^p(\pi_k)}}\right\}\right\}\right) > 1 - \delta. \quad \text{(B.9)}$$

For $t \in [T] \cup \{0\}$, let $X_t = \sum_{s\in[t]} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\}(r_s^p - \mu_i^p)$; and for $\lambda > 0$, further define $\xi_t(\lambda) = \exp\left(\lambda X_t - n_i^p(t)\frac{\lambda^2}{8}\right)$. It can be verified that $\left\{\xi_t(\lambda)\right\}_{t=0}^T$ is a nonnegative supermartingale with respect to $\{\mathcal{F}_t\}_{t=0}^T$ for all $\lambda > 0$:

1. $\mathbb{E}\left[\left|\xi_t(\lambda)\right|\right] < \infty$ for all $t \in [T] \cup \{0\}$;

2. $\xi_t(\lambda) \geq 0$ for all $t \in [T] \cup \{0\}$;

3. $\mathbb{E}\left[\xi_t(\lambda) \mid \mathcal{F}_{t-1}\right] \leq \xi_{t-1}(\lambda)$ for all $t \in [T]$.

Item 3 is true because

$$
\begin{aligned}
&\mathbb{E}\left[\xi_t(\lambda) \mid \mathcal{F}_{t-1}\right] \\
&= \exp\left(\sum_{s=1}^{t-1} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \cdot \\
&\qquad\qquad \mathbb{E}\left[\exp\left(\mathbb{1}\left\{p \in \mathcal{P}_t, i_t^p = i\right\}\left(\lambda(r_t^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
&= \xi_{t-1}(\lambda) \cdot \mathbb{E}\Bigg[\exp\left(\lambda \cdot \mathbb{1}\left\{p \in \mathcal{P}_t, i_t^p = i\right\}(r_t^p - \mu_i^p)\right) \cdot \\
&\qquad\qquad \exp\left(-\mathbb{1}\left\{p \in \mathcal{P}_t, i_t^p = i\right\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\Bigg] \\
&= \xi_{t-1}(\lambda) \cdot \mathbb{E}\Bigg[\mathbb{E}\left[\exp\left(\lambda \cdot \mathbb{1}\left\{p \in \mathcal{P}_t, i_t^p = i\right\}(r_t^p - \mu_i^p)\right) \mid \mathcal{F}_{t-1}, i_t^p\right] \cdot \\
&\qquad\qquad \exp\left(-\mathbb{1}\left\{p \in \mathcal{P}_t, i_t^p = i\right\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\Bigg] \\
&\leq \xi_{t-1}(\lambda),
\end{aligned}
$$

where we use the law of total expectation, the observation that $\xi_{t-1}(\lambda)$ is $\mathcal{F}_{t-1}$-measurable, and Hoeffding's Lemma.

Recall from Remark B.19 that π_k is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$ and $\pi_k \leq T + 1 < \infty$ almost surely. Then, by the optional sampling theorem, for all $\lambda > 0$,

$$
\mathbb{E}\left[\mathbb{1}\left\{\pi_k \leq T\right\} \cdot \xi_{\pi_k}(\lambda)\right] \leq \mathbb{E}\left[\xi_0(\lambda)\right] = 1. \tag{B.10}
$$

In other words,

$$\mathbb{E}\left[\mathbb{1}\left\{\pi_k \leq T\right\} \cdot \exp\left(\lambda X_{\pi_k} - n_i^p(\pi_k)\frac{\lambda^2}{8}\right)\right] \leq 1.$$

By Markov's inequality, we have

$$\Pr\left(\mathbb{1}\left\{\pi_k \leq T\right\} \cdot \exp\left(\lambda X_{\pi_k} - n_i^p(\pi_k)\frac{\lambda^2}{8}\right) \geq \frac{1}{\delta}\right) \leq \delta;$$

and thus,

$$\Pr\left(\left\{\pi_k \leq T\right\} \cap \left\{\exp\left(\lambda X_{\pi_k} - n_i^p(\pi_k)\frac{\lambda^2}{8}\right) \geq \frac{1}{\delta}\right\}\right) \leq \delta.$$

Using the elementary fact that for sets A and B, $\neg(A \cap B) = \neg A \cup (A \cap \neg B)$, we have, for any $\lambda > 0$,

$$\Pr\left(\left\{\pi_k = T+1\right\} \cup \left\{\left\{\pi_k \leq T\right\} \cap \left\{\frac{1}{n_i^p(\pi_k)}X_{\pi_k} - \frac{\lambda}{8} < \frac{\ln\left(\frac{1}{\delta}\right)}{n_i^p(\pi_k)\cdot\lambda}\right\}\right\}\right) > 1-\delta,$$

where we slightly rearrange the terms.

Choose $\lambda = \sqrt{\frac{\ln(\frac{1}{\delta})}{k}}$ and observe that $n_i^p(\pi_k) = k$. It follows that

$$\Pr\left(\left\{\pi_k = T+1\right\} \cup \left\{\left\{\pi_k \leq T\right\} \cap \left\{\frac{1}{n_i^p(\pi_k)}X_{\pi_k} < \sqrt{\frac{2\ln(\frac{1}{\delta})}{n_i^p(\pi_k)}}\right\}\right\}\right) > 1-\delta.$$

Eq. (B.9) follows trivially by the observation that $\ln(\frac{2}{\delta}) > \ln(\frac{1}{\delta})$. By symmetry, it

can also be shown that the following inequality is true:

$$\Pr\left(\{\pi_k = T+1\} \cup \left\{\{\pi_k \leq T\} \cap \left\{\mu_i^p - \text{ind-}\hat{\mu}_i^p(\pi_k) \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{n_i^p(\pi_k)}}\right\}\right\}\right) > 1-\delta.$$

The proof is then completed by applying the union bound. □

Definition B.31. *For any $\delta \in (0,1]$, let*

$$E_{agg}(\delta) = \Bigg\{\forall i \in [K], \forall k \in [TM] \cup \{0\}, \left(\tau_k(i) = T+1\right) \vee \Bigg(\left(\tau_k(i) \leq T\right) \wedge \Bigg(\forall p \in [M],\ \text{agg-}\hat{\mu}_i(\tau_k(i)) - \mu_i^p \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k(i)) - M\right) \vee 1}} + 2\epsilon,$$
$$\mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k(i)) \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k(i)) - M\right) \vee 1}}\Bigg)\Bigg)\Bigg\},$$

and

$$E_{ind}(\delta) = \Bigg\{\forall i \in [K], \forall p \in [M], \forall k \in [T] \cup \{0\}, \left(\pi_k(i,p) = T+1\right) \vee \Bigg(\left(\pi_k(i,p) \leq T\right) \wedge \left(\left|\text{ind-}\hat{\mu}_i^p(\pi_k(i,p)) - \mu_i^p\right| \leq \sqrt{\frac{2\ln(\frac{4}{\delta})}{n_i^p(\pi_k(i,p)) \vee 1}}\right)\Bigg)\Bigg\}.$$

Furthermore, let

$$E(\delta) = E_{agg}(\delta) \cap E_{ind}(\delta).$$

Corollary B.32. *For $\delta \in (0, 1]$,*

$$\Pr(E(\delta)) \geq 1 - 6T^3\delta.$$

Proof. By the union bound, Lemma B.27, Lemma B.29, and the assumption that $T \geq \max(K, M)$, we have

$$\Pr(E_{\text{agg}}(\delta)) \geq 1 - K(TM + 1)(2\delta) \geq 1 - 4T^3\delta.$$
$$\Pr(E_{\text{ind}}(\delta)) \geq 1 - KM(T + 1)\delta \geq 1 - 2T^3\delta.$$

The corollary then follows by the union bound. □

B.2.3 Clean event

We now define our notion of "clean" event for each t.

Definition B.33. *For any $t \in [T + 1]$, let*

$$\mathcal{E}_t = \left\{ \forall p \in [M], \forall i \in [K], \ \left|\text{ind-}\hat{\mu}_i^p(t-1) - \mu_i^p\right| \leq \sqrt{\frac{10 \ln T}{\overline{n_i^p}(t-1)}}, \right.$$
$$\text{agg-}\hat{\mu}_i^p(t-1) - \mu_i^p \leq \sqrt{\frac{10 \ln T}{\overline{m_i^p}(t-1)}} + 2\epsilon,$$
$$\left. \mu_i^p - \text{agg-}\hat{\mu}_i^p(t-1) \leq \sqrt{\frac{10 \ln T}{\overline{m_i^p}(t-1)}} \right\},$$

where we recall that $\overline{n_i^p}(t-1) = n_i^p(t-1) \vee 1$, $\overline{m_i^p}(t-1) = (m_i^p(t-1) - M) \vee 1$. Furthermore, let $\overline{\mathcal{E}_t}$ denote the complement of $\mathcal{E}_t$.

The following lemma shows that the clean event happens with high probability.

Lemma B.34.

$$\Pr(\mathcal{E}_t) > 1 - \frac{24}{T^2}.$$

Proof. The proof of Lemma B.34 follows from Corollary B.32. It suffices to show that, for any t, $E(\frac{4}{T^5}) \subseteq \mathcal{E}_t$. To this end, we will show that if $E(\frac{4}{T^5})$ happens, then $\mathcal{E}_t$ must happen.

For any $t \in [T+1]$, $i \in [K]$, $p \in [M]$, let $u = u_i^p(t-1)$ be the round in which player p last pulls arm i (see Definition B.7). In addition, let $s = n_i^p(u) \in ([T] \cup \{0\})$ and $k = n_i(u) \in ([TM] \cup \{0\})$. Note that $\pi_s(i,p) = u \leq T$ and $\tau_k(i) = u \leq T$.

It then follows by definition that,

$$\text{ind-}\hat{\mu}_i^p(t-1) = \text{ind-}\hat{\mu}_i^p(\pi_s(i,p)), \quad n_i^p(t-1) = n_i^p(\pi_s(i,p));$$
$$\text{agg-}\hat{\mu}_i^p(t-1) = \text{agg-}\hat{\mu}_i(\tau_k(i)), \quad m_i^p(t-1) = n_i(\tau_k(p)).$$

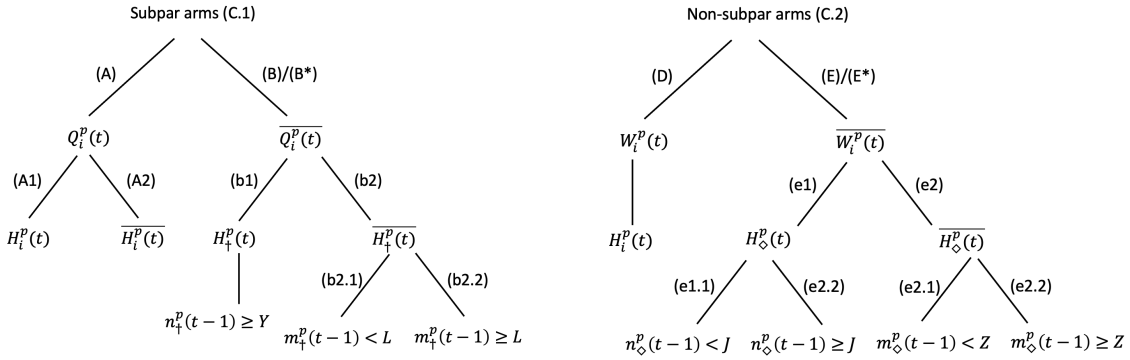
The proof is then completed straightforwardly by the definition of $E(\frac{4}{T^5})$, which indicates that for all $s \in [T] \cup \{0\}$ and $k \in [TM] \cup \{0\}$,

$$\left|\text{ind-}\hat{\mu}_i^p(\pi_s(i,p)) - \mu_i^p\right| \leq \sqrt{\frac{10 \ln T}{n_i^p(\pi_s(i,p)) \vee 1}},$$
$$\text{agg-}\hat{\mu}_i(\tau_k(i)) - \mu_i^p \leq \sqrt{\frac{10 \ln T}{\big(n_i(\tau_k(p)) - M\big) \vee 1}} + 2\epsilon, \text{ and}$$
$$\mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k(i)) \leq \sqrt{\frac{10 \ln T}{\big(n_i(\tau_k(p)) - M\big) \vee 1}}.$$

□

B.3 Proofs of Theorem 3.1 and Theorem 3.2

The following lemmas are central to our proofs of Theorem 3.1 and Theorem 3.2. In Section B.3.1, we prove Lemma B.35. In Section B.3.2, we prove Lemma B.36. We then conclude our proofs in Section B.3.3.



(a) Subpar arms (Section B.3.1) **(b)** Non-subpar arms (Section B.3.2)

Figure B.1. Illustrates the case division rules used in the proofs of Theorem 3.1 and Theorem 3.2, respectively. Formal definitions of the notions used in the figure can be found in Section B.1, Section B.3.1 and Section B.3.2.

Lemma B.35 (Subpar arms)**.** *For any arm $i \in \mathcal{I}_{10\epsilon}$,*

$$\mathbb{E}\left[n_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right),$$

where we recall that $\Delta_i^{\min} = \min_{p\in[M]} \Delta_i^p$.

Lemma B.36 (Non-subpar arms)**.** *For any arm $i \in \mathcal{I}_{10\epsilon}^C$ and player $p \in [M]$,*

$$\mathbb{E}\left[n_i^p(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Our analysis in the following Section B.3.1 and Section B.3.2 involve various proofs by cases. Figure B.1 provides an overview of the case division rules used in our analysis.

B.3.1 Subpar arms

In this section, we prove Lemma B.35.

Fix any subpar arm $i \in \mathcal{I}_{10\epsilon}$ and an arm $\dagger \in \mathcal{I}_{2\epsilon}^C$. See Fact B.24 for the existence of such an arm. We first consider the following definitions.

Definition B.37. *For any arm $i \in \mathcal{I}_{10\epsilon}$ and any player p, let*

$$\delta_i^p = \mu_\dagger^p - \mu_i^p > 0.$$

Fact B.38. *For any $i \in \mathcal{I}_{10\epsilon}$ and player $p \in [M]$,*

$$\frac{3}{4}\Delta_i^p < \delta_i^p \leq \Delta_i^p.$$

Proof. For any player $p \in [M]$, since $\dagger \in \mathcal{I}_{2\epsilon}^C$, we have $\Delta_\dagger^p = \mu_*^p - \mu_\dagger^p \leq 2\epsilon$ by the definition of $\mathcal{I}_{2\epsilon}^C$. Furthermore, for any $i \in \mathcal{I}_{10\epsilon}$, $\Delta_i^p = \mu_*^p - \mu_i^p > 8\epsilon$. Therefore, we have

1. $\delta_i^p = \mu_\dagger^p - \mu_i^p \leq \mu_*^p - \mu_i^p = \Delta_i^p$;

2. Note that $\frac{\mu_*^p - \mu_\dagger^p}{\mu_*^p - \mu_i^p} \leq \frac{2\epsilon}{8\epsilon} \leq \frac{1}{4}$. This implies that $\frac{\delta_i^p}{\Delta_i^p} = 1 - \frac{\mu_*^p - \mu_\dagger^p}{\mu_*^p - \mu_i^p} \geq \frac{3}{4}$.

□

Definition B.39. *For any player p, let $y_i^p = \mu_i^p + \frac{1}{2}\delta_i^p$ be a threshold; in any round t, further define*

$$Q_i^p(t) = \left\{\theta_i^p(t) > y_i^p\right\}$$

to be the event that the sample $\theta_i^p(t)$ from the posterior distribution associated with arm i and player p in round t is greater than the threshold y_i^p. In addition, let $\overline{Q_i^p(t)} = \left\{\theta_i^p(t) \leq y_i^p\right\}$.

Subpar Arms—Decomposition

We can then decompose $\mathbb{E}\left[n_i(T)\right]$ as follows.

$$
\begin{aligned}
&\mathbb{E}\left[n_i(T)\right] \\
&= \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p = i\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t\right\}\right] + \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right] + \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{\overline{\mathcal{E}_t}\right\}\right] \\
&\leq \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t\right\}\right]}_{(A)} + \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right]}_{(B)} + \mathcal{O}\left(1\right), \quad \text{(B.11)}
\end{aligned}
$$

where the second inequality follows from Lemma B.34. In the following two subsections, we bound term (A) and (B), respectively.

Bounding Term (A)

The following lemma provides an upper bound on term (A).

Lemma B.40.

$$
(A) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right), \tag{B.12}
$$

where we recall that $\Delta_i^{\min} = \min_{p\in[M]} \Delta_i^p$.

Proof of Lemma B.40. Recall the definition of $\mathcal{E}_t$ in Definition B.33 and the definition of $H_i^p(t)$ in Definition B.13, we have

$$
(A) = \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, H_i^p(t)\right\}\right]}_{(A1)} + \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}\right\}\right]}_{(A2)}.
$$

We first consider term $(A1)$. Recall that, for simplicity, we let $\overline{n_i^p}(t-1)$ denote $n_i^p(t-1) \vee 1$; also recall that $\overline{\Phi}(\cdot)$ is the complementary CDF of the standard Gaussian distribution, and $(z)_+ = z \vee 0$. We have

$$
\begin{aligned}
(A1) &\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{Q_i^p(t), \mathcal{E}_t, H_i^p(t)\right\}\right] \\
&= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Q_i^p(t), \mathcal{E}_t, H_i^p(t)\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \overline{\Phi}\left(\sqrt{\overline{n_i^p}(t-1)/4}\left(y_i^p - \text{ind-}\hat{\mu}_i^p(t-1)\right)\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \exp\left(-\frac{n_i^p(t-1)(y_i^p - \text{ind-}\hat{\mu}_i^p(t-1))_+^2}{8}\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \exp\left(-\frac{n_i^p(t-1)(\mu_i^p + \frac{3}{8}\Delta_i^p - \mu_i^p - \frac{1}{16}\Delta_i^p)_+^2}{8}\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \exp\left(-\frac{n_i^p(t-1)(\Delta_i^p)^2}{8(16)}\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \frac{1}{T^2} = \mathcal{O}(1).
\end{aligned}
$$

where the first inequality drops the indicator $\mathbb{1}\left\{i_t^p = i\right\}$; the first equality uses the law of total expectation; the second equality follows from the observation that $\mathcal{E}_t$ and $H_i^p(t)$ are $\mathcal{F}_{t-1}$-measurable; the third equality follows from the observation that when $H_i^p(t)$ happens, $\mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right] = \mathbb{P}\left(\theta_i^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right) = \overline{\Phi}\left(\frac{y_i^p - \text{ind-}\hat{\mu}_i^p(t-1)}{\sqrt{4/\overline{n_i^p}(t-1)}}\right)$; the second inequality is from Lemma B.69 and that $\overline{n_i^p}(t-1) \geq n_i^p(t-1)$; the third inequality follows from the facts that when $\mathcal{E}_t$ and $H_i^p(t)$ happen,

1. $\overline{n_i^p}(t-1) \geq n_i^p(t-1) \geq \frac{40\ln T}{\epsilon^2} \geq \frac{2560\ln T}{(\Delta_i^p)^2}$ (see Fact B.24),

2. ind-$\hat{\mu}_i^p(t-1) \leq \mu_i^p + \sqrt{\frac{10 \ln T}{n_i^p(t-1)}} \leq \mu_i^p + \frac{1}{16}\Delta_i^p$ (see Definition B.33), and

3. $y_i^p = \mu_i^p + \frac{1}{2}\delta_i^p > \mu_i^p + \frac{3}{8}\Delta_i^p$ (see Fact B.38);

the fourth inequality is by algebra; and the fifth inequality again uses the observation that when $H_i^p(t)$ happens, $n_i^p(t-1) \geq \frac{2560 \ln T}{(\Delta_i^p)^2}$.

We now turn our attention to term $(A2)$. With foresight, let $l = \frac{10240 \ln T}{\left(\Delta_i^{\min}\right)^2} + M$. We have

$$
\begin{aligned}
(A2) &= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}\right\}\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) < l\right\}\right] \\
&\qquad\qquad + \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\}\right] \\
&\leq (l+M) + \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\}\right]. \qquad \text{(B.13)}
\end{aligned}
$$

To see why Eq. (B.13) is true, it suffices to show that, with probability 1,

$$
\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} \leq l + M.
$$

Indeed, let us define $\iota = \min\left\{t : n_i(t) = \sum_{s\in[t]} \sum_{p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\} \geq l\right\}$. The above summa-

tion can be simplified as

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i, m_i^p(t-1)<l\right\}\\
=&\sum_{t=1}^{\iota-1}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i, m_i^p(t-1)<l\right\}+\sum_{t=\iota}^{T}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i, m_i^p(t-1)<l\right\}\\
\leq&\sum_{t=1}^{\iota-1}\sum_{p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}+\sum_{p\in[M]}\sum_{t\geq\iota:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i, m_i^p(t-1)<l\right\}\\
\leq&(l-1)+M,
\end{aligned}
$$

where the $\sum_{p\in[M]}\sum_{t\geq\iota:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i, m_i^p(t-1)<l\right\}\leq M$ follows from the observation that, once the total number of pulls of arm i by all players has reached l, any player p cannot pull arm i more than once before the aggregate number of pulls of i maintained by p is updated to a value $\geq l$ (see Definition B.9).

Remark B.41. *Eq.* (B.13) *can also be deducted from the more general Lemma B.72 in Section B.3.4, by taking* $f_t^p=1$ *for all* t,p.

Now, recall that we denote $\left(m_i^p(t-1)-M\right)\vee 1$ by $\overline{m_i^p}(t-1)$. And again, recall that $\overline{\Phi}(\cdot)$ is the complementary CDF of the standard Gaussian distribution, and $(z)_+=z\vee 0$.

It follows from Eq. (B.13) that

$$
\begin{aligned}
(A2) &\leq (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\} \mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\Bigg[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\} \cdot \\
&\qquad\qquad \overline{\Phi}\left(\sqrt{\overline{m_i^p}(t-1)/4}\left(y_i^p - \text{agg-}\hat{\mu}_i^p(t-1)\right)\right)\Bigg] \\
&\leq (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\Bigg[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\} \cdot \\
&\qquad\qquad \exp\left(-\frac{\overline{m_i^p}(t-1)\left(y_i^p - \text{agg-}\hat{\mu}_i^p(t-1)\right)_+^2}{8}\right)\Bigg] \\
&\leq (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\Bigg[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\} \cdot \\
&\qquad\qquad \exp\left(-\frac{\overline{m_i^p}(t-1)\left(\mu_i^p + \frac{3}{8}\Delta_i^p - \mu_i^p - \frac{9}{32}\Delta_i^p\right)_+^2}{8}\right)\Bigg] \\
&\leq (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\Bigg[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\} \cdot \\
&\qquad\qquad \exp\left(-\frac{\overline{m_i^p}(t-1)\left(\Delta_i^{\min}\right)^2}{(8)(256)}\right)\Bigg] \\
&\leq (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \frac{1}{T^2} \\
&= \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right),
\end{aligned}
$$

where the first inequality is from Eq. (B.13), dropping the indicator $\mathbb{1}\left\{i_t^p = i\right\}$ and using the law of total expectation; the first equality follows from the observation that $\mathcal{E}_t$,

$\overline{H_i^p(t)}$, and $\left\{m_i^p(t-1) \geq l\right\}$ are $\mathcal{F}_{t-1}$-measurable; the second equality follows from the observation that when $\overline{H_i^p(t)}$ happens, $\mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right] = \mathbb{P}\left(\theta_i^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right) = \overline{\Phi}\left(\frac{y_i^p - \text{agg-}\hat{\mu}_i^p(t-1)}{\sqrt{4/\overline{m_i^p}(t-1)}}\right)$; the second inequality follows from Lemma B.69; the third inequality uses the facts that

1. when $\left\{m_i^p(t-1) \geq l\right\}$ happens, $\overline{m_i^p}(t-1) \geq m_i^p(t-1) - M \geq l - M = \frac{10240 \ln T}{\left(\Delta_i^{\min}\right)^2}$,
2. $y_i^p = \mu_i^p + \frac{1}{2}\delta_i^p > \mu_i^p + \frac{3}{8}\Delta_i^{\min}$ (see Fact B.38), and
3. when $\mathcal{E}_t$ happens, $\text{agg-}\hat{\mu}_i^p(t-1) \leq \mu_i^p + \sqrt{\frac{10 \ln T}{\overline{m_i^p}(t-1)}} + 2\epsilon < \mu_i^p + \frac{1}{32}\Delta_i^{\min} + \frac{1}{4}\Delta_i^{\min} = \mu_i^p + \frac{9}{32}\Delta_i^{\min}$ (see Definition B.33 and Fact B.24);

the fourth inequality is by algebra; and the fifth inequality again uses the fact that when $\left\{m_i^p(t-1) \geq l\right\}$ happens, $\overline{m_i^p}(t-1) \geq m_i^p(t-1) - M \geq \frac{10240 \ln T}{\left(\Delta_i^{\min}\right)^2}$.

In summary, we have

$$(A) \leq (A1) + (A2) + \mathcal{O}(1) \leq \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right).$$

□

Bounding Term (B)

We now bound term (B) in Eq. (B.11).

Lemma B.42.

$$(B) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

Proof. Lemma B.42 follows from Lemmas B.45 and B.46, which we present shortly. □

Consider the following definition.

Definition B.43. *In any round $t \in [T]$, for any active player $p \in \mathcal{P}_t$, define*

$$\phi_{i,t}^p = \Pr\left(\theta_\dagger^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right).$$

Remark B.44. *Recall that $\overline{\Phi}(\cdot)$ denotes the complementary CDF of the standard Gaussian distribution; and recall $\overline{n_i^p}(t-1) = n_i^p(t-1) \vee 1$, and $\overline{m_i^p}(t-1) = (m_i^p(t-1) - M) \vee 1$. $\phi_{i,t}^p$ can be explicitly written as:*

$$\begin{aligned}\phi_{i,t}^p =& \overline{\Phi}\left(\frac{y_i^p - \hat{\mu}_\dagger^p(t-1)}{\sqrt{\mathrm{var}_\dagger^p(t-1)}}\right) && \text{(B.14)}\\ =& \overline{\Phi}\left((y_i^p - \text{ind-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{n_\dagger^p}(t-1)/4}\right)\cdot \mathbb{1}\left\{H_\dagger^p(t)\right\} + \\ & \qquad \overline{\Phi}\left((y_i^p - \text{agg-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{m_\dagger^p}(t-1)/4}\right)\cdot \mathbb{1}\left\{\overline{H_\dagger^p(t)}\right\}. && \text{(B.15)}\end{aligned}$$

Proof of Remark B.44. We have

$$\begin{aligned}\phi_{i,t}^p =& \Pr\left(\theta_\dagger^p(t) > y_i^p \mid \hat{\mu}_\dagger^p(t-1), \mathrm{var}_\dagger^p(t-1)\right)\\ =& 1 - \Pr\left(\theta_\dagger^p(t) \le y_i^p \mid \hat{\mu}_\dagger^p(t-1), \mathrm{var}_\dagger^p(t-1)\right)\\ =& 1 - \Phi\left(\frac{y_i^p - \hat{\mu}_\dagger^p(t-1)}{\sqrt{\mathrm{var}_\dagger^p(t-1)}}\right) = \overline{\Phi}\left(\frac{y_i^p - \hat{\mu}_\dagger^p(t-1)}{\sqrt{\mathrm{var}_\dagger^p(t-1)}}\right).\end{aligned}$$

Eq. (B.15) now follows by observing that:

1. if $H_\dagger^p(t)$ happens, then $\hat{\mu}_\dagger^p(t-1) = \text{ind-}\hat{\mu}_\dagger^p(t-1)$ and $\mathrm{var}_\dagger^p(t-1) = \frac{4}{n_\dagger^p(t-1)\vee 1}$;
2. if $\overline{H_\dagger^p(t)}$ happens, then $\hat{\mu}_\dagger^p(t-1) = \text{agg-}\hat{\mu}_\dagger^p(t-1)$ and $\mathrm{var}_\dagger^p(t-1) = \frac{4}{(m_\dagger^p(t-1)-M)\vee 1}$.

□

We now present the following lemma, which is inspired by a technique introduced in the work of [4].

Lemma B.45.

$$(B) \leq \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi^p_{i,t}}{\phi^p_{i,t}}\right)\mathbb{1}\left\{i^p_t=\dagger,\mathcal{E}_t\right\}\right]}_{(B*)}.$$

Proof. In any round t and for any active player $p\in\mathcal{P}_t$, consider

$$\begin{aligned}
&\Pr\left(i^p_t=i,\overline{Q^p_i(t)},\mathcal{E}_t\mid\mathcal{F}_{t-1}\right)\\
=&\Pr\left(i^p_t=i,\theta^p_i(t)\leq y^p_i\mid\mathcal{F}_{t-1}\right)\cdot\mathbb{1}\left\{\mathcal{E}_t\right\}\\
\leq&\Pr\left(i^p_t=\dagger\mid\mathcal{F}_{t-1}\right)\cdot\frac{\Pr\left(\theta^p_\dagger(t)\leq y^p_i\mid\mathcal{F}_{t-1}\right)}{\Pr\left(\theta^p_\dagger(t)>y^p_i\mid\mathcal{F}_{t-1}\right)}\cdot\mathbb{1}\left\{\mathcal{E}_t\right\}\\
=&\left(\frac{1-\phi^p_{i,t}}{\phi^p_{i,t}}\right)\cdot\Pr\left(i^p_t=\dagger\mid\mathcal{F}_{t-1}\right)\cdot\mathbb{1}\left\{\mathcal{E}_t\right\}\\
=&\left(\frac{1-\phi^p_{i,t}}{\phi^p_{i,t}}\right)\Pr\left(i^p_t=\dagger,\mathcal{E}_t\mid\mathcal{F}_{t-1}\right),
\end{aligned}\tag{B.16}$$

where the first equality follows from the definition of $Q^p_i(t)$ and that $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable; the first inequality uses Lemma B.74 with $l=\dagger$ and $z=y^p_i$; the second equality inequality is from the definition of $\phi^p_{i,t}$; and the last equality is again because $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable.

Finally, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right] &= \mathbb{E}\left[\Pr\left(i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&\leq \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\Pr\left(i_t^p = \dagger, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\}\right],
\end{aligned}$$

where we use the law of total expectation and Eq. (B.16). The lemma follows by summing over all t, p's. □

With foresight, let $L = \frac{2560 \ln T}{\left(\Delta_i^{\min}\right)^2} + M$. We further decompose term $(B*)$ as follows.

$$
\begin{aligned}
&(B*)\\
&=\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\mathcal{E}_t\right\}\right]\\
&=\underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\mathcal{E}_t,H_\dagger^p(t)\right\}\right]}_{(b1)}+\\
&\qquad\qquad\underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\mathcal{E}_t,\overline{H_\dagger^p(t)}\right\}\right]}_{(b2)},\\
&=(b1)+\underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\mathcal{E}_t,\overline{H_\dagger^p(t)},m_\dagger^p(t-1)<L\right\}\right]}_{(b2.1)}+\\
&\qquad\qquad\underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\mathcal{E}_t,\overline{H_\dagger^p(t)},m_\dagger^p(t-1)\geq L\right\}\right]}_{(b2.2)}.
\end{aligned}
\tag{B.17}
$$

where the inequality uses Lemma B.45.

Lemma B.46.

$$
(B*)\leq\mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2}+M\right).
$$

Proof. Lemma B.46 follows directly from Eq. (B.17) and the following Lemma B.47, Lemma B.48 and Lemma B.52, which provide upper bounds on terms $(b1)$, $(b2.1)$ and $(b2.2)$, respectively. □

Lemma B.47 (Bounding term $(b1)$)**.**

$$(b1) \leq \mathcal{O}(M).$$

Proof of Lemma B.47. For any player $p \in [M]$ and $t \in [T]$, recall that $\overline{n_\dagger^p}(t-1) = n_\dagger^p(t-1) \vee 1$ and $(z)_+ = z \vee 0$. When $\mathcal{E}_t$ and $H_\dagger^p(t)$ happen, $n_\dagger^p(t-1) \geq \frac{40 \ln T}{\epsilon^2} =: Y$; we have:

$$\begin{aligned}
&1 - \phi_{i,t}^p \\
=&\Pr\left(\theta_\dagger^p(t) \leq y_i^p \mid \mathcal{F}_{t-1}\right) \\
=&\Phi\left((y_i^p - \text{ind-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{n_\dagger^p}(t-1)/4}\right) \\
\leq&\exp\left(-\frac{\overline{n_\dagger^p}(t-1)(\text{ind-}\hat{\mu}_\dagger^p(t-1) - y_i^p)_+^2}{8}\right) \\
\leq&\exp\left(-\frac{n_\dagger^p(t-1)(\mu_\dagger^p - \frac{1}{4}\Delta_i^p - \mu_\dagger^p + \frac{3}{8}\Delta_i^p)_+^2}{8}\right) \\
\leq&\exp\left(-\frac{n_\dagger^p(t-1)(\Delta_i^p)^2}{8(64)}\right) \\
\leq&\frac{1}{T+1},
\end{aligned}$$

where the second equality uses Remark B.44; the first inequality uses Lemma B.69; the second inequality follows from the observations that, when $\mathcal{E}_t$ and $H_\dagger^p(t)$ happen:

1. $\overline{n_\dagger^p}(t-1) \geq n_\dagger^p(t-1) \geq Y = \frac{40 \ln T}{\epsilon^2} \geq \frac{2560 \ln T}{(\Delta_i^p)^2}$ (see Fact B.38),
2. $\text{ind-}\hat{\mu}_\dagger^p(t-1) \geq \mu_\dagger^p - \sqrt{\frac{10 \ln T}{\overline{n_\dagger^p}(t-1)}} \geq \mu_\dagger^p - \frac{1}{4}\Delta_i^p$ (see Definition B.33), and
3. $y_i^p = \mu_\dagger^p - \frac{1}{2}\delta_i^p < \mu_\dagger^p - \frac{3}{8}\Delta_i^p$;

the third inequality is by algebra; and the last inequality follows because, again, when $H_\dagger^p(t)$ happens, $n_\dagger^p(t-1) \geq Y = \frac{40 \ln T}{\epsilon^2} \geq \frac{2560 \ln T}{(\Delta_i^p)^2} \geq \frac{1280 \ln(T+1)}{(\Delta_i^p)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t$ and $H_\dagger^p(t)$ happen, $\phi_{i,t}^p \geq \frac{T}{T+1}$ and $\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p} \leq \frac{1}{T}$. Hence,

$$(b1) \leq \sum_{p\in[M]} \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{\mathcal{E}_t, H_\dagger^p(t)\right\}\right] \leq M.$$

□

Lemma B.48 (Bounding term $(b2.1)$)**.**

$$(b2.1) \leq \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right).$$

The remark below is useful for proving Lemma B.48.

Remark B.49 (Invariant property)**.** *Recall from Example B.21 that*

$$\left\{H_\dagger^p(t) : t \in [T], p \in [M]\right\}$$

satisfies the invariant property with respect to †*.*

Moreover, the construction of Algorithm 2 enforces that $\left\{\phi_{i,t}^p : t \in [T], p \in [M]\right\}$ *satisfies the invariant property with respect to* † *(note that it does not necessarily satisfy the invariant property with respect to* i*). Indeed, this follows from Eq.* (B.14)*, along with Example B.23 which shows that the posterior parameters,*

$$\left\{(\hat{\mu}_\dagger^p(t-1), \mathrm{var}_\dagger^p(t-1)) : t \in [T], p \in [M]\right\},$$

satisfy the invariant property with respect to †*.*

Combining the two observations above, $\left\{\left(\frac{1}{\phi_{i,t}^p} - 1\right) \mathbb{1}\left\{\overline{H_\dagger^p(t)}\right\} : t \in [T], p \in [M]\right\}$ *also satisfies the invariant property with respect to arm* †*.*

Proof of Lemma B.48. Proving Lemma B.48 requires more special care. Recall that

$$(b2.1) = \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\mathcal{E}_t,\overline{H_\dagger^p(t)},m_\dagger^p(t-1)<L\right\}\right]$$
$$\leq \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\dagger,\overline{H_\dagger^p(t)},m_\dagger^p(t-1)<L\right\}\right].$$

Also recall the definition of stopping time $\tau_k(\dagger)$ (Definition B.15), the round in which $\dagger$ is pulled the k-th time by any player. To ease exposition, we abuse the notation and denote $\tau_k(\dagger)$ by τ_k. Similarly, let $p_k := p_k(\dagger)$ denote the player that issues the k-th pull of arm $\dagger$ (recall Definition B.17).

Since $\left\{\left(\frac{1}{\phi_{i,t}^p}-1\right)\mathbb{1}\left\{\overline{H_\dagger^p(t)}\right\} : t\in[T], p\in[M]\right\}$ satisfies the invariant property with respect to arm $\dagger$, by Lemma B.72, we have

$$(b2.1)$$
$$\leq \sum_{p=1}^{M}\mathbb{E}\left[\left(\frac{1}{\phi_{i,1}^p}-1\right)\mathbb{1}\left\{\overline{H_\dagger^p(1)}\right\}\right] + \sum_{k=1}^{L-1}\mathbb{E}\left[\left(\frac{1}{\phi_{i,\tau_k+1}^{p_k}}-1\right)\mathbb{1}\left\{\tau_k\leq T,\overline{H_\dagger^p(\tau_k+1)}\right\}\right], \tag{B.18}$$

where we also use the linearity of expectations.

Since the variance of the aggregate posteriors are initialized as the constant $c_2 = 4$ in RobustAgg-TS(ϵ), we have that $\left(\frac{1}{\phi_{i,1}^p}-1\right)\mathbb{1}\left\{\overline{H_\dagger^p(1)}\right\} \leq \mathcal{O}(1)$ with probability 1. Therefore,

$$\sum_{p=1}^{M}\mathbb{E}\left[\left(\frac{1}{\phi_{i,1}^p}-1\right)\mathbb{1}\left\{\overline{H_\dagger^p(1)}\right\}\right] \leq \mathcal{O}(M). \tag{B.19}$$

It then suffices to bound the second term in Eq. (B.18)—it follows straightforwardly from Lemma B.50, which we present shortly, that the second term is bounded by $\mathcal{O}(L)$. It then

follows from Eq. (B.18), Eq. (B.19), and Lemma B.50 that $(b2.1) \leq \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right)$. $\square$

Lemma B.50. *For any $k \in [TM]$,*

$$\mathbb{E}\left[\left(\frac{1}{\phi_{i,\tau_k+1}^{p_k}} - 1\right) \mathbb{1}\left\{\tau_k \leq T, \overline{H_\dagger^p(\tau_k+1)}\right\}\right] \leq \mathcal{O}(1),$$

where we recall that $\tau_k = \tau_k(\dagger)$ and $p_k = p_k(\dagger)$ is the player that issues the k-th pull of arm $\dagger$.

Proof. Using Remark B.44, we observe that

$$\begin{aligned}\phi_{i,\tau_k+1}^{p_k} = &\left[\overline{\Phi}\left(\frac{y_i^p - \text{ind-}\hat{\mu}_\dagger^p(\tau_k)}{2}\sqrt{\left(n_\dagger^p(\tau_k)\right) \vee 1}\right)\right] \cdot \mathbb{1}\left\{H_\dagger^p(\tau_k+1)\right\} \\ &+ \left[\overline{\Phi}\left(\frac{y_i^p - \text{agg-}\hat{\mu}_\dagger^p(\tau_k)}{2}\sqrt{\left(m_\dagger^p(\tau_k) - M\right) \vee 1}\right)\right] \cdot \mathbb{1}\left\{\overline{H_\dagger^p(\tau_k+1)}\right\}. \qquad \text{(B.20)}\end{aligned}$$

We have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\phi_{i,\tau_k+1}^{p_k}}-1\right)\mathbb{1}\left\{\tau_k\le T,\overline{H_\dagger^{p_k}(\tau_k+1)}\right\}\right]\\
=&\mathbb{E}\left[\left(\frac{1}{\overline{\Phi}\left(\left(y_i^{p_k}-\text{agg-}\hat{\mu}_\dagger^{p_k}(\tau_k)\right)\sqrt{\left(\left(m_\dagger^{p_k}(\tau_k)-M\right)\vee 1\right)/4}\right)}-1\right)\cdot\right.\\
&\qquad\left.\mathbb{1}\left\{\tau_k\le T,\overline{H_\dagger^{p_k}(\tau_k+1)}\right\}\right]\\
\le&\mathbb{E}\left[\frac{1}{\overline{\Phi}\left(\left(\mu_\dagger^{p_k}-\text{agg-}\hat{\mu}_\dagger(\tau_k)\right)\sqrt{\left(\left(n_\dagger(\tau_k)-M\right)\vee 1\right)/4}\right)}\mathbb{1}\left\{\tau_k\le T\right\}\right],
\end{aligned}
\tag{B.21}
$$

where the last inequality uses the observations that $y_i^{p_k}\le\mu_\dagger^{p_k}$, agg-$\hat{\mu}_\dagger^{p_k}(\tau_k)=$ agg-$\hat{\mu}_\dagger(\tau_k)$ and $m_\dagger^{p_k}(\tau_k)=n_\dagger(\tau_k)$, as well as the monotonic increasing property of $z\mapsto\frac{1}{\overline{\Phi}(z)}$.

Observe that, from Corollary B.28, for any $z\ge 1$,

$$
\begin{aligned}
&\Pr\left((\tau_k\le T)\wedge\left(\mu_\dagger^{p_k}-\text{agg-}\hat{\mu}_\dagger(\tau_k)\ge z\sqrt{\frac{4}{(n_\dagger(\tau_k)-M)\vee 1}}\right)\right)\\
\le&\Pr\left((\tau_k\le T)\wedge\left(\exists p\in[M],\ \mu_\dagger^{p}-\text{agg-}\hat{\mu}_\dagger(\tau_k)\ge z\sqrt{\frac{4}{(n_\dagger(\tau_k)-M)\vee 1}}\right)\right)\\
\le&2e^{-2z^2},
\end{aligned}
$$

Applying Lemma B.70 with $X = \left(\text{agg-}\hat{\mu}_\dagger(\tau_k) - \mu_\dagger^{p_k}\right)\sqrt{\left(\left(n_\dagger(\tau_k) - M\right) \vee 1\right)/4}$ and $E = \{\tau_k \leq T\}$, we conclude the proof. □

Remark B.51. *Note that it follows from our novel concentration inequality (Corollary B.28) that*

$$\Pr\left(\tau_k \leq T, \mu_\dagger^p - \text{agg-}\hat{\mu}_\dagger(\tau_k) > \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_\dagger(\tau_k) - M\right) \vee 1}}\right) < \delta;$$

this tight bound enables us to bound Eq. (B.21) *by* $\mathcal{O}(1)$*, which is essential to our proof of Lemma B.50.*

Since $n_i(\tau_k) \leq [k, k + M - 1]$*, using the Azuma-Hoeffding inequality and the union bound, one can obtain*

$$\Pr\left(\tau_k \leq T, \mu_\dagger^p - \text{agg-}\hat{\mu}_\dagger(\tau_k) > \mathcal{O}\left(\sqrt{\frac{\ln\left(\frac{M}{\delta}\right)}{\left(n_\dagger(\tau_k) - M\right) \vee 1}}\right)\right) < \delta;$$

and using Freedman's inequality (see, e.g., Lemma A.4), one can obtain

$$\Pr\left(\tau_k \leq T, \mu_\dagger^p - \text{agg-}\hat{\mu}_\dagger(\tau_k) > \mathcal{O}\left(\sqrt{\frac{\ln\left(\frac{\ln T}{\delta}\right)}{\left(n_\dagger(\tau_k) - M\right) \vee 1}}\right)\right) < \delta.$$

However, naively combining the above bounds with Lemma B.70, one needs to set C_1 *in Lemma B.70 to be* $\mathcal{O}(M)$ *or* $\mathcal{O}(\ln T)$*, which incurs extra (undesirable)* $\mathcal{O}(M)$ *or* $\mathcal{O}(\ln T)$ *factors for bounding Eq.* (B.21)*.*

Lemma B.52 (Bounding term (b2.2))**.**

$$(b2.2) \leq \mathcal{O}(M).$$

Proof of Lemma B.52. For any player $p \in [M]$ and $t \in [T]$, recall that $\overline{m_\dagger^p}(t-1) = (m_\dagger^p(t-1) - M) \vee 1$ and $(z)_+ = z \vee 0$. When $\mathcal{E}_t, \left\{m_\dagger^p(t-1) \geq L\right\}$ and $\overline{H_\dagger^p(t)}$ happen,

$$
\begin{aligned}
& 1 - \phi_{i,t}^p \\
=& \Pr\left(\theta_\dagger^p(t) \leq y_i^p \mid \mathcal{F}_{t-1}\right) \\
=& \Phi\left((y_i^p - \text{agg-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{m_\dagger^p}(t-1)/4}\right) \\
\leq& \exp\left(-\frac{\overline{m_\dagger^p}(t-1)(\text{agg-}\hat{\mu}_\dagger^p(t-1) - y_i^p)_+^2}{8}\right) \\
\leq& \exp\left(-\frac{\overline{m_\dagger^p}(t-1)(\mu_\dagger^p - \frac{1}{4}\Delta_i^p - \mu_\dagger^p + \frac{3}{8}\Delta_i^p)_+^2}{8}\right) \\
\leq& \exp\left(-\frac{\overline{m_\dagger^p}(t-1)(\Delta_i^p)^2}{8(64)}\right) \\
\leq& \frac{1}{T+1},
\end{aligned}
$$

where the second equality uses Remark B.44; the first inequality uses Lemma B.69; the second inequality follows from the observations that, when $\mathcal{E}_t, \left\{m_\dagger^p(t-1) \geq L\right\}$ and $\overline{H_\dagger^p(t)}$ happen:

1. $\overline{m_\dagger^p}(t-1) \geq m_\dagger^p(t-1) - M \geq L - M \geq \frac{2560 \ln T}{(\Delta_i^p)^2}$,
2. $\text{agg-}\hat{\mu}_\dagger^p(t-1) \geq \mu_\dagger^p - \sqrt{\frac{10 \ln T}{\overline{m_\dagger^p}(t-1)}} \geq \mu_\dagger^p - \frac{1}{4}\Delta_i^p$ (see Definition B.33), and
3. $y_i^p = \mu_\dagger^p - \frac{1}{2}\delta_i^p < \mu_\dagger^p - \frac{3}{8}\Delta_i^p$;

the third inequality is by algebra; and the last inequality follows from the observation that $\overline{m_\dagger^p}(t-1) \geq m_\dagger^p(t-1) - M \geq L - M \geq \frac{2560 \ln T}{(\Delta_i^p)^2} \geq \frac{1280 \ln(T+1)}{(\Delta_i^p)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t, \left\{m_\dagger^p(t-1) \geq L\right\}$ and $\overline{H_\dagger^p(t)}$ happen, $\phi_{i,t}^p \geq \frac{T}{T+1}$ and

$\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p} \leq \frac{1}{T}$. Hence,

$$(b2.2) \leq \sum_{p\in[M]} \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{\mathcal{E}_t, \overline{H_\dagger^p(t)}, m_\dagger^p(t-1) \geq L\right\}\right] \leq M.$$

□

B.3.2 Non-subpar arms

In this section, we provide a proof for Lemma B.36.

Let us fix any player $p \in [M]$ and any suboptimal arm $i \in \mathcal{I}_{10\epsilon}^C$ for player p such that $\Delta_i^p > 0$. In the rest of this section, let us also fix an optimal arm for player p, $\diamond_p$, and we abbreviate it by $\diamond$. We have $\mu_\diamond^p = \mu_*^p = \max_{j\in[K]} \mu_j^p$.

Definition B.53. *Let $z_i^p = \mu_i^p + \frac{1}{2}\Delta_i^p$ be a threshold. In any round t, define*

$$W_i^p(t) = \left\{\theta_i^p(t) > z_i^p\right\}$$

to be the event that the sample $\theta_i^p(t)$ from the posterior distribution associated with arm i and player p in round t is greater than the threshold z_i^p. Therefore, $\overline{W_i^p(t)} = \left\{\theta_i^p(t) \leq z_i^p\right\}$.

Non-subpar Arms—Decomposition

We consider the following decomposition.

$$
\begin{aligned}
&\mathbb{E}\left[n_i^p(T)\right] \\
=&\mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i\right\}\right] \\
=&\mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i,W_i^p(t),\mathcal{E}_t\right\}\right]+\mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i,\overline{W_i^p(t)},\mathcal{E}_t\right\}\right]+ \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\sum_{t:p\in\mathcal{P}_t}\mathbb{E}\left[\mathbb{1}\left\{i_t^p=i,\overline{\mathcal{E}_t}\right\}\right] \\
\leq&\underbrace{\mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i,W_i^p(t),\mathcal{E}_t\right\}\right]}_{(D)}+\underbrace{\mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\mathbb{1}\left\{i_t^p=i,\overline{W_i^p(t)},\mathcal{E}_t\right\}\right]}_{(E)}+\mathcal{O}(1), \qquad \text{(B.22)}
\end{aligned}
$$

where the last inequality follows from the observation that $\mathbb{E}\left[\mathbb{1}\left\{i_t^p=i,\overline{\mathcal{E}_t}\right\}\right]\leq\mathbb{E}\left[\mathbb{1}\left\{\overline{\mathcal{E}_t}\right\}\right]$ and Lemma B.34.

Following this decomposition, Lemma B.36 is proved straightforwardly given Lemma B.54 and Lemma B.55 which we present in what follows.

Bounding Term (D)

We first bound term (D) in Eq. (B.22).

Lemma B.54.

$$(D)\leq\mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}+M\right).$$

Proof of Lemma B.54. With foresight, let $h=\frac{4000\ln T}{(\Delta_i^p)^2}+2M$. Recall that $H_i^p(t)$ is the event that the individual posterior is used in round t by active player p for arm i (see

Definition B.13). We have

$$
\begin{aligned}
(D) &= \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, W_i^p(t), \mathcal{E}_t\right\}\right] \\
&\leq h + \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, W_i^p(t), \mathcal{E}_t, n_i^p(t-1) \geq h\right\}\right] \\
&= h + \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, W_i^p(t), \mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\}\right]}_{(d)},
\end{aligned}
$$

where the last equality follows from the observation that $\left\{n_i^p(t-1) \geq h\right\}$ implies that $H_i^p(t)$ happening. To see why this is true, recall that $H_i^p(t) = \left\{n_i^p(t-1) \geq \frac{40\ln T}{\epsilon^2} + 2M\right\}$; and observe that for non-subpar arm $i \in \mathcal{I}_{10\epsilon}^C$ and player p, $\left\{n_i^p(t-1) \geq h = \frac{4000\ln T}{(\Delta_i^p)^2} + 2M\right\}$ implies $\left\{n_i^p(t-1) \geq \frac{40\ln T}{\epsilon^2} + 2M\right\}$ because $\Delta_i^p \leq 10\epsilon$.

It therefore suffices to bound term (d). We have

$$
\begin{aligned}
(d) &\leq \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{W_i^p(t), \mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\}\right] \\
&= \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\} \mathbb{E}\left[\mathbb{1}\left\{W_i^p(t)\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\} \overline{\Phi}\left((z_i^p - \text{ind-}\hat{\mu}_i^p(t-1))\sqrt{\overline{n_i^p}(t-1)/4}\right)\right] \\
&\leq \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\} \exp\left(-\frac{\overline{n_i^p}(t-1)(z_i^p - \text{ind-}\hat{\mu}_i^p(t-1))_+^2}{8}\right)\right] \\
&\leq \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\} \exp\left(-\frac{n_i^p(t-1)(\mu_i^p + \frac{1}{2}\Delta_i^p - \mu_i^p - \frac{1}{16}\Delta_i^p)_+^2}{8}\right)\right] \\
&\leq \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \geq h\right\} \exp\left(-\frac{n_i^p(t-1)(\Delta_i^p)^2}{8(16)}\right)\right] \\
&\leq \mathcal{O}(1).
\end{aligned}
$$

where the first inequality drops the indicator $\mathbb{1}\left\{i_t^p = i\right\}$; the first equality uses the law of total expectation and the observation that $\mathcal{E}_t$, $H_i^p(t)$ and $\left\{n_i^p(t-1) \geq h\right\}$ are $\mathcal{F}_{t-1}$-measurable; the second inequality follows from Lemma B.69; the third inequality is from the observations that when $\mathcal{E}_t$ and $H_i^p(t)$ happen:

1. $\overline{n_i^p}(t-1) \geq n_i^p(t-1) \geq h = \frac{4000 \ln T}{(\Delta_i^p)^2} + 2M$,
2. ind-$\hat{\mu}_i^p(t-1) \leq \mu_i^p + \sqrt{\frac{10 \ln T}{\overline{n_i^p}(t-1)}} \leq \mu_i^p + \frac{1}{16}\Delta_i^p$ (see Definition B.33), and
3. $z_i^p = \mu_i^p + \frac{1}{2}\Delta_i^p$;

the fourth inequality is by algebra; and the last inequality is from the observation that when $n_i^p(t-1) \geq h$, $\exp\left(-\frac{n_i^p(t-1)(\Delta_i^p)^2}{8(16)}\right) \leq \frac{1}{T}$.

In summary, $(D) \leq h + (d) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right)$. □

Bounding Term (E)

We now bound (E) in Eq. (B.22):

Lemma B.55.

$$(E) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Proof. Lemma B.55 follows from Lemma B.58, Eq. (B.25), Lemma B.59, and Lemma B.64 which we present shortly. □

We begin with the following definition, similar to the notion of $\phi_{i,t}^p$ used for subpar arms.

Definition B.56. *Recall that p is a fixed player, i is a fixed suboptimal arm for p, and $\diamond$ is a fixed optimal arm for p. In any round t, define*

$$\psi_{i,t}^p = \Pr\left(\theta_\diamond^p(t) > z_i^p \mid \mathcal{F}_{t-1}\right).$$

Remark B.57. *Recall that* $\overline{n^p_\diamond}(t-1) = n^p_\diamond(t-1) \vee 1$ *and* $\overline{m^p_\diamond}(t-1) = \left(m^p_\diamond(t-1) - M\right) \vee 1$. $\psi^p_{i,t}$ *can be explicitly written as:*

$$\begin{aligned}\psi^p_{i,t} =& \overline{\Phi}\left(\frac{z^p_i - \hat{\mu}^p_\diamond(t-1)}{\sqrt{\mathrm{var}^p_\diamond(t-1)}}\right) \\ =& \overline{\Phi}\left((z^p_i - \text{ind-}\hat{\mu}^p_\diamond(t-1))\sqrt{\overline{n^p_\diamond}(t-1)/4}\right) \cdot \mathbb{1}\left\{H^p_\diamond(t)\right\} \\ &+ \overline{\Phi}\left((z^p_i - \text{agg-}\hat{\mu}^p_\diamond(t-1))\sqrt{\overline{m^p_\diamond}(t-1)/4}\right) \cdot \mathbb{1}\left\{\overline{H^p_\diamond}(t)\right\}.\end{aligned} \tag{B.23}$$

The proof for the above remark is omitted, as it is very similar to that of Remark B.44.

We now present the following lemma.

Lemma B.58.

$$(E) = \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} \mathbb{1}\left\{i^p_t = i, \overline{W^p_i(t)}, \mathcal{E}_t\right\}\right] \leq \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right)\mathbb{1}\left\{i^p_t = \diamond, \mathcal{E}_t\right\}\right]}_{(E*)}$$

Proof. The proof largely follows the same outline as that of Lemma B.45.

In any round t and such that $p \in \mathcal{P}_t$, consider

$$\begin{aligned}&\Pr\left(i^p_t = i, \overline{Q^p_i(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right) \\ =& \Pr\left(i^p_t = i, \theta^p_i(t) \leq z^p_i \mid \mathcal{F}_{t-1}\right) \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\ \leq& \Pr\left(i^p_t = \diamond \mid \mathcal{F}_{t-1}\right) \cdot \frac{\Pr\left(\theta^p_\diamond(t) \leq z^p_i \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta^p_\diamond(t) > z^p_i \mid \mathcal{F}_{t-1}\right)} \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\ =& \left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right) \cdot \Pr\left(i^p_t = \diamond \mid \mathcal{F}_{t-1}\right) \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\ =& \left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right) \Pr\left(i^p_t = \diamond, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right),\end{aligned} \tag{B.24}$$

where the first equality follows from the definition of $Q_i^p(t)$ and that $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable; the first inequality uses Lemma B.74 with $l = \diamond$ and $z = z_i^p$; and the second equality inequality is from the definition of $\psi_{i,t}^p$; the last equality is again because $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable.

Finally, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right] &= \mathbb{E}\left[\Pr\left(i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&\leq \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\Pr\left(i_t^p = \diamond, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t\right\}\right],
\end{aligned}$$

where we use the law of total expectation and Eq. (B.24). The lemma follows by summing over all t's. □

Let us further decompose $(E*)$ as follows.

$$\begin{aligned}
(E*) = &\underbrace{\sum_{t:p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, H_\diamond^p(t)\right\}\right]}_{(e1)} + \\
&\underbrace{\sum_{t:p\in\mathcal{P}_t}\mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}\right\}\right]}_{(e2)}.
\end{aligned} \tag{B.25}$$

We first consider term $(e1)$.

Lemma B.59.

$$(e1) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

Proof of Lemma B.59. With foresight, let $J = \frac{640 \ln T}{(\Delta_i^p)^2}$. We have

$$(e1) = \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, H_\diamond^p(t), n_\diamond^p(t-1) < J\right\}\right]}_{(e1.1)} +$$
$$\underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, H_\diamond^p(t), n_\diamond^p(t-1) \geq J\right\}\right]}_{(e1.2)}.$$

Lemma B.59 follows straightforwardly from Lemma B.60 and Lemma B.63, which bound $(e1.1)$ and $(e1.2)$, respectively. □

Lemma B.60.

$$(e1.1) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

To prove Lemma B.60, we first present the following Remark B.49.

Remark B.61 (Invariant Property)**.** *Similar to Remark B.49, by the construction of Algorithm 2, we have that for any arm $i \in [K]$, and player $p \in [M]$, $\left\{\psi_{i,t}^p : t \in [T]\right\}$ and $\left\{H_\diamond^p(t) : t \in [T]\right\}$ satisfy the invariant property with respect to $(\diamond, p)$ (Definition B.20). Indeed, the former follows from Eq.* (B.23)*, along with Example B.23 that shows that the posterior parameters, $\left\{(\hat{\mu}_\diamond^p(t-1), \mathrm{var}_\diamond^p(t-1)) : t \in [T]\right\}$, satisfy the invariant property with respect to $(\diamond, p)$; and the latter is from Example B.21.*

Proof of Lemma B.60. We start by rewriting $(e1.1)$ as follows, where we drop $\mathcal{E}_t$.

$$
\begin{aligned}
(e1.1) \leq & \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p=\diamond, H_\diamond^p(t), n_\diamond^p(t-1)<J\right\}\right] \\
= & \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} g_t \mathbb{1}\left\{i_t^p=\diamond, n_\diamond^p(t-1)<J\right\}\right],
\end{aligned}
$$

where in the second line, we introduce the notation $g_t := \left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{H_\diamond^p(t)\right\}$;

We now focus on the sum inside the expectation. Recall that $\pi_s(\diamond, p)$ is the round in which player p pulls arm $\diamond$ the s-th time. Here, we abuse the notation and denote $\pi_s(\diamond, p)$ by π_s. By Remark B.61, $\left\{g_t : t\in[T]\right\}$ satisfies the invariant property with respect to $(\diamond, p)$. Applying Lemma B.71 on $\left\{g_t : t\in[T]\right\}$'s, we have that the term inside the above expectation is at most:

$$
\sum_{s=1}^{J-1}\left(\frac{1}{\psi_{i,\pi_s+1}^p}-1\right)\mathbb{1}\left\{\pi_s\leq T, H_\diamond^p(\pi_s+1)\right\},
$$

where we also use the observation that $\left(\frac{1}{\psi_{i,1}^p}-1\right)\mathbb{1}\left\{H_\diamond^p(1)\right\}=0$.

Therefore, by the linearity of expectation, we have

$$
(e1.1)\leq\sum_{s=1}^{J-1}\mathbb{E}\left[\left(\frac{1}{\psi_{i,\pi_s+1}^p}-1\right)\mathbb{1}\left\{\pi_s\leq T, H_\diamond^p(\pi_s+1)\right\}\right].
$$

Therefore, the following Lemma B.62 suffices to prove Lemma B.60, which we prove next. □

Lemma B.62. *For any $s\in[T]$,*

$$
\mathbb{E}\left[\left(\frac{1}{\psi_{i,\pi_s+1}^p}-1\right)\mathbb{1}\left\{\pi_s\leq T, H_\diamond^p(\pi_s+1)\right\}\right]\leq\mathcal{O}(1),
$$

where we recall that $\pi_s = \pi_s(\diamond, p)$ is the round in which player p pulls arm $\diamond$ the s-th time.

Proof of Lemma B.62. We note that this proof is similar to that of Lemma B.50. We have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\psi^p_{i,\pi_s+1}} - 1\right)\mathbb{1}\left\{\pi_s \le T, H^p_\diamond(\pi_s+1)\right\}\right] \\
=&\mathbb{E}\left[\left(\frac{1}{\overline{\Phi}\left(\left(z^p_i - \text{ind-}\hat{\mu}^p_\diamond(\pi_s)\right)\sqrt{\overline{n^p_\diamond}(\pi_s)/4}\right)} - 1\right)\mathbb{1}\left\{\pi_s \le T, H^p_\diamond(\pi_s+1)\right\}\right] \\
\le&\mathbb{E}\left[\frac{1}{\overline{\Phi}\left(\left(\mu^p_\diamond - \text{ind-}\hat{\mu}^p_\diamond(\pi_s)\right)\sqrt{\overline{n^p_\diamond}(\pi_s)/4}\right)}\mathbb{1}\left\{\pi_s \le T\right\}\right],
\end{aligned}$$

where the inequality drops $H^p_\diamond(\tau_s+1)$ and uses the observation that $z^p_i \le \mu^p_\diamond$, and the monotonic increasing property of $z \mapsto \frac{1}{\overline{\Phi}(z)}$. Now, using Lemma B.70 and Corollary B.30, we conclude that this is at most $\mathcal{O}(1)$. □

Lemma B.63.

$$(e1.2) \le \mathcal{O}(1).$$

Proof. Recall that

$$(e1.2) = \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right)\mathbb{1}\left\{i^p_t = \diamond, \mathcal{E}_t, H^p_\diamond(t), n^p_\diamond(t-1) \ge J\right\}\right].$$

Dropping $\mathbb{1}\left\{i^p_t = i^p_\diamond\right\}$, we have

$$(e1.2) \le \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right)\mathbb{1}\left\{\mathcal{E}_t, H^p_\diamond(t), n^p_\diamond(t-1) \ge J\right\}\right],$$

When $\mathcal{E}_t$, $H^p_\diamond(t)$, and $\left\{n^p_\diamond(t-1) \geq J\right\}$ happen, we have

$$\begin{aligned}
&1-\psi^p_{i,t}\\
=&\Pr\left(\theta^p_\diamond(t) \leq z^p_i \mid \mathcal{F}_{t-1}\right)\\
=&\Phi\left((z^p_i - \text{ind-}\hat{\mu}^p_\diamond(t-1))\sqrt{\overline{n^p_\diamond}(t-1)/4}\right)\\
\leq&\exp\left(-\frac{\overline{n^p_\diamond}(t-1)\left(\text{ind-}\hat{\mu}^p_\diamond(t-1) - z^p_i\right)^2_+}{8}\right)\\
\leq&\exp\left(-\frac{n^p_\diamond(t-1)\left(\mu^p_\diamond - \frac{1}{4}\Delta^p_i - \mu^p_\diamond + \frac{1}{2}\Delta^p_i\right)^2_+}{8}\right)\\
\leq&\exp\left(-\frac{n^p_\diamond(t-1)(\Delta^p_i)^2}{8(16)}\right)\\
\leq&\frac{1}{T+1},
\end{aligned}$$

where the first inequality uses Lemma B.69; the second inequality uses the observations that, when $\mathcal{E}_t$ and $\left\{n^p_\diamond(t-1) \geq J\right\}$ happen:

1. $\overline{n^p_\diamond}(t-1) \geq n^p_\diamond(t-1) \geq J = \frac{640\ln T}{(\Delta^p_i)^2}$,
2. $\text{ind-}\hat{\mu}^p_\diamond(t-1) \geq \mu^p_\diamond - \sqrt{\frac{10\ln T}{\overline{n^p_\diamond}(t-1)}} \geq \mu^p_\diamond - \frac{1}{4}\Delta^p_i$ (see Definition B.33), and
3. $z^p_i = \mu^p_\diamond - \frac{1}{2}\Delta^p_i$;

the third inequality is by algebra; and the last inequality follows as when $\left\{n^p_\diamond(t-1) \geq J\right\}$ happens, $n^p_\diamond(t-1) \geq \frac{640\ln T}{(\Delta^p_i)^2} \geq \frac{320\ln(T+1)}{(\Delta^p_i)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t$ and $\left\{n^p_\diamond(t-1) \geq J\right\}$ happen, $\psi^p_{i,t} \geq \frac{T}{T+1}$ and $\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}} \leq \frac{1}{T}$. Hence, $(e1.2) \leq 1$. □

We now consider term $(e2)$. Recall that

$$(e2) = \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right)\mathbb{1}\left\{i^p_t = \diamond, \mathcal{E}_t, \overline{H^p_\diamond(t)}\right\}\right]$$

Lemma B.64.

$$(e2) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Proof of Lemma B.64. With foresight, let $Z = \frac{640 \ln T}{(\Delta_i^p)^2} + M$. We have

$$(e2) = \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\}\right]}_{(e2.1)} +$$
$$\underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) \geq Z\right\}\right]}_{(e2.2)}.$$

The proof follows straightforwardly from Lemma B.65 and Lemma B.67 which we present subsequently. □

Lemma B.65.

$$(e2.1) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Proof of Lemma B.65. We have

$$\begin{aligned}(e2.1) &\leq \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} \left(\frac{1}{\psi_{i,t}^p} - 1\right) \mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\}\right] \\ &\leq \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} \frac{1}{\psi_{i,t}^p} \mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\}\right],\end{aligned}$$

where we drop $\mathcal{E}_t$ and use the observation that $\frac{1}{\psi_{i,t}^p} - 1 \leq \frac{1}{\psi_{i,t}^p}$.

We now focus on sum inside the expectation. We denote $\tau_k(\diamond)$ by τ_k and the player that makes the k's pull of $\diamond$ by $p_k := p_k(\diamond)$. Recall that we use $\overline{m_\diamond^p}(t-1)$ to denote

$\left(m_\diamond^p(t-1) - M\right) \vee 1$. We have

$$
\begin{aligned}
&\sum_{t:p\in\mathcal{P}_t} \frac{1}{\psi_{i,t}^p} \mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\} \\
&= \sum_{t:p\in\mathcal{P}_t} \frac{1}{\overline{\Phi}\left(\left(z_i^p - \text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)} \mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\} \\
&\leq \sum_{t:p\in\mathcal{P}_t} \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^p - \text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)} \mathbb{1}\left\{i_t^p = \diamond, m_\diamond^p(t-1) < Z\right\} \qquad \text{(B.26)} \\
&\leq \sum_{t\in[T]}\sum_{q\in\mathcal{P}_t} \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^q - \text{agg-}\hat{\mu}_\diamond^q(t-1)\right)\sqrt{\overline{m_\diamond^q}(t-1)/4}\right)} \mathbb{1}\left\{i_t^q = \diamond, m_\diamond^q(t-1) < Z\right\}, \qquad \text{(B.27)}
\end{aligned}
$$

where the first equality uses Remark B.57; the first inequality drops $\overline{H_\diamond^p(t)}$ and uses the observation that $z_i^p \leq \mu_\diamond^p$ (see Definition B.53), along with the monotonic increasing property of $z \mapsto \frac{1}{\overline{\Phi}(z)}$.; the second inequality adds similar terms for other players $q \neq p$.

Now, define $\left\{f_t^q : t \in [T], q \in [M]\right\}$ where $f_t^q = \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^q - \text{agg-}\hat{\mu}_\diamond^q(t-1)\right)\sqrt{\overline{m_\diamond^q}(t-1)/4}\right)}$; recall from Example B.22 that $\left\{\text{agg-}\hat{\mu}_\diamond^q(t-1) : t \in [T]\right\}$ and $\left\{m_\diamond^q(t-1) : t \in [T]\right\}$ both satisfy the invariant property with respect to $(\diamond, q)$; therefore, $\left\{f_t^q : t \in [T], q \in [M]\right\}$ satisfies the invariant property with respect to $\diamond$. Applying Lemma B.72 to it, we have that

$$
\text{(B.27)} \leq \sum_{q\in[M]} \frac{1}{\overline{\Phi}(0)} + \sum_{k=1}^{Z-1} \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^{p_k} - \text{agg-}\hat{\mu}_\diamond^{p_k}(\tau_k)\sqrt{\overline{m_\diamond^{p_k}}(\tau_k)/4}\right)\right)} \mathbb{1}\left\{\tau_k \leq T\right\}.
$$

Since $\sum_{q\in[M]} \frac{1}{\overline{\Phi}(0)} \leq \mathcal{O}(M)$, it then suffices to show that for every $k \in \mathbb{N}$,

$$
\mathbb{E}\left[\frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^{p_k} - \text{agg-}\hat{\mu}_\diamond^{p_k}(\tau_k)\right)\sqrt{\overline{m_\diamond^{p_k}}(\tau_k)/4}\right)} \mathbb{1}\left\{\tau_k \leq T\right\}\right] \leq \mathcal{O}(1). \qquad \text{(B.28)}
$$

Note that $\overline{m_\diamond^{p_k}}(\tau_k) = \big(n_\diamond(\tau_k) - M\big) \vee 1$. Directly applying Corollary B.28 and Lemma B.70 with $X = (\text{agg-}\hat{\mu}_\diamond^{p_k}(\tau_k) - \mu_\diamond^{p_k})\sqrt{\overline{m_\diamond^{p_k}}(\tau_k)/4}$ and $E = \{\tau_k \le T\}$ proves Eq. (B.28). □

Remark B.66. *In the above proof, we relaxed Eq.* (B.26) *to Eq.* (B.27) *by adding the corresponding terms for all other players $q \neq p$. Alternatively, we could use the observation that $n_\diamond^p(t-1) \le m_\diamond^p(t-1)$ to bound Eq.* (B.26) *by*

$$\sum_{t:p\in\mathcal{P}_t} \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^p - \text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)} \mathbb{1}\left\{i_t^p = \diamond, n_\diamond^p(t-1) < Z\right\},$$

and apply Lemma B.71 and subsequently Lemma B.70. However, right now, we do not have tight-enough concentration inequalities for $\text{agg-}\hat{\mu}_\diamond^p(\pi_k(\diamond, p))$*—the best known inequality here is Freedman's inequality, which incurs an undesirable extra $\mathcal{O}(\ln T)$ factor in the bound for* $(e2.1)$.

Lemma B.67.

$$(e2.2) \le \mathcal{O}(1).$$

Proof of Lemma B.67. Recall that

$$(e2.2) = \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) \ge Z\right\}\right].$$

Recall that $\overline{m_\diamond^p}(t-1) = \big(m_\diamond^p(t-1) - M\big) \vee 1$. When $\mathcal{E}_t, \overline{H_\diamond^p(t)}$ and $\big\{m_\diamond^p(t-1) \ge Z\big\}$

happen simultaneously,

$$
\begin{aligned}
&1-\psi_{i,t}^p\\
=&\Pr\left(\theta_\diamond^p(t)\le z_i^p\mid\mathcal{F}_{t-1}\right)\\
=&\Phi\left(\left(z_i^p-\text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)\\
\le&\exp\left(-\frac{\overline{m_\diamond^p}(t-1)\left(\text{agg-}\hat{\mu}_\diamond^p(t-1)-z_i^p\right)_+^2}{8}\right)\\
\le&\exp\left(-\frac{\overline{m_\diamond^p}(t-1)\left(\mu_\diamond^p-\frac{1}{4}\Delta_i^p-\mu_\diamond^p+\frac{1}{2}\Delta_i^p\right)_+^2}{8}\right)\\
\le&\exp\left(-\frac{\overline{m_\diamond^p}(t-1)(\Delta_i^p)^2}{8(16)}\right)\\
\le&\frac{1}{T+1},
\end{aligned}
$$

where the first inequality uses Lemma B.69; the second inequality uses the observations that when $\mathcal{E}_t, \overline{H_\diamond^p(t)}$ and $\left\{m_\diamond^p(t-1)\ge Z\right\}$ happen:

1. $\overline{m_\diamond^p}(t-1)\ge m_\diamond^p(t-1)-M\ge Z-M\ge\frac{640\ln T}{(\Delta_i^p)^2}$,

2. $\text{agg-}\hat{\mu}_\diamond^p(t-1)\ge\mu_\diamond^p-\sqrt{\frac{10\ln T}{m_\diamond^p(t-1)}}\ge\mu_\diamond^p-\frac{1}{4}\Delta_i^p$ (see Definition B.33), and

3. $z_i^p=\mu_\diamond^p-\frac{1}{2}\Delta_i^p$ (see Definition B.53);

the third inequality is by algebra; and the fourth inequality is by the fact that when $m_\diamond^p(t-1)\ge Z$, $\overline{m_\diamond^p}(t-1)\ge Z-M=\frac{640\ln T}{(\Delta_i^p)^2}\ge\frac{320\ln(T+1)}{(\Delta_i^p)^2}$ for $T>1$.

It follows that, when $\mathcal{E}_t, \overline{H_\diamond^p(t)}$ and $\left\{m_\diamond^p(t-1)\ge Z\right\}$ happen, $\psi_{i,t}^p\ge\frac{T}{T+1}$ and $\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\le\frac{1}{T}$. As a result, $(e2.2)\le\mathcal{O}(1)$. □

B.3.3 Concluding the proofs of Theorems 3.1 and 3.2

Lemma B.68. *Let a generalized ϵ-MPMAB problem instance and $\alpha>0$ be such that for all $i\in\mathcal{I}_\alpha$ and all $p\in[M]$, $\Delta_i^p\le 2\Delta_i^{\min}$. If algorithm $\mathcal{A}$ guarantees that when interacting*

with this problem instance:

1. *For any arm $i \in \mathcal{I}_\alpha$,*

$$\mathbb{E}\left[n_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right); \tag{B.29}$$

2. *For any arm $i \in \mathcal{I}_\alpha^C$ and player $p \in [M]$,*

$$\mathbb{E}\left[n_i^p(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + C\right), \tag{B.30}$$

for some $C \geq 0$, then it has the following regret bounds simultaneously:

1. *gap-dependent regret bound:*

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_\alpha}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p} + \sum_{i\in\mathcal{I}_\alpha^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p} + MK(1+C)\right), \tag{B.31}$$

2. *gap-independent regret bound:*

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_\alpha|P} + \sqrt{M\left(|\mathcal{I}_\alpha^C| - 1\right)P} + MK(1+C)\right), \tag{B.32}$$

where we recall that $P = \sum_{t=1}^T |\mathcal{P}_t|$.

Proof. We prove the two items respectively. Recall that $\Delta_i^{\min} = \min_{p\in[M]} \Delta_i^p$.

1. Note that for all $i \in \mathcal{I}_\alpha$ and all $p \in [M]$, $\Delta_i^p \leq 2\Delta_i^{\min}$, and $\sum_{p=1}^M \mathbb{E}\left[n_i^p(T)\right] = \mathbb{E}\left[n_i(T)\right]$; as a consequence,

$$\mathrm{Reg}(T) = \sum_{p=1}^M\sum_{i=1}^K \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p = \mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\Delta_i^{\min} + \sum_{i\in\mathcal{I}_\alpha^C}\sum_{\substack{p\in[M]:\\ \Delta_i^p>0}}\mathbb{E}\left[n_i^p(T)\right]\Delta_i^p\right). \tag{B.33}$$

Using Eq. (B.29), the first term can be bounded by:

$$\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\Delta_i^{\min}\leq\mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha}\frac{\ln T}{\Delta_i^{\min}}+MK\right)\leq\mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_\alpha}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}+MK\right),$$

where the second inequality follows from the assumption that for all $i\in\mathcal{I}_\alpha$ and $p\in[M]$, $\Delta_i^p\leq 2\Delta_i^{\min}$.

Using Eq. (B.30), the second term can be bounded by:

$$\sum_{i\in\mathcal{I}_\alpha^C}\sum_{p\in[M]:\Delta_i^p>0}\mathbb{E}\left[n_i^p(T)\right]\Delta_i^p\leq\mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}+MKC\right).$$

Combining the above two bounds yields Eq. (B.31).

2. As with the proof of Eq. (B.32), we continue from Eq. (B.33), but look at the two terms respectively. For the first term,

$$\begin{aligned}\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\Delta_i^{\min}\leq&\mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha}\min\left(\mathbb{E}\left[n_i(T)\right],\frac{\ln T}{(\Delta_i^{\min})^2}+M\right)\Delta_i^{\min}\right)\\ \leq&\mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha}\min\left(\mathbb{E}\left[n_i(T)\right]\Delta_i^{\min},\frac{\ln T}{\Delta_i^{\min}}\right)+MK\right)\\ \leq&\mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha}\sqrt{\mathbb{E}\left[n_i(T)\right]\ln T}+MK\right)\\ \leq&\mathcal{O}\left(\sqrt{|\mathcal{I}_\alpha|P\ln T}+MK\right)\end{aligned}\tag{B.34}$$

where the first inequality is from Eq. (B.29); the second inequality is by algebra; the third inequality is from the elementary fact that $\min(A,B)\leq\sqrt{AB}$; the last inequality is from Jensen's inequality and the concavity of function $x\mapsto\sqrt{x}$, which implies that $\sum_{i\in\mathcal{I}_\alpha}\sqrt{\mathbb{E}\left[n_i(T)\right]}\leq\sqrt{|\mathcal{I}_\alpha|\left(\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\right)}$, and the fact that $\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\leq$

$\sum_{i=1}^{M} \mathbb{E}\left[n_i(T)\right] \leq P$.

For the second term in Eq. (B.32), first observe that if $\left|\mathcal{I}_\alpha^C\right| = 1$, then let i^* be the only element in $\mathcal{I}_\alpha^C$; it must be the case that for all $p \in [M]$, i^* is the optimal arm for player p. As a consequence, $\sum_{i \in \mathcal{I}_\alpha^C} \sum_{p=1}^{M} \mathbb{E}\left[n_i^p(T)\right] \Delta_i^p = 0 = \mathcal{O}(\sqrt{M(|\mathcal{I}_\alpha^C| - 1)P})$.

Otherwise, $\left|\mathcal{I}_\alpha^C\right| \geq 2$. In this case,

$$\begin{aligned}
\sum_{p \in [M]} \sum_{i \in \mathcal{I}_\alpha^C} \mathbb{E}\left[n_i^p(T)\right] \Delta_i^p \leq & \mathcal{O}\left(\sum_{p \in [M]} \sum_{i \in \mathcal{I}_\alpha^C} \min\left(\mathbb{E}\left[n_i^p(T)\right], \frac{\ln T}{(\Delta_i^p)^2}\right) \Delta_i^p + MKC\right) \\
\leq & \mathcal{O}\left(\sum_{p \in [M]} \sum_{i \in \mathcal{I}_\alpha^C} \min\left(\mathbb{E}\left[n_i^p(T)\right] \Delta_i^p, \frac{\ln T}{\Delta_i^p}\right) + MKC\right) \\
\leq & \mathcal{O}\left(\sum_{p \in [M]} \sum_{i \in \mathcal{I}_\alpha^C} \sqrt{\mathbb{E}\left[n_i^p(T)\right] \ln T} + MKC\right) \\
\leq & \mathcal{O}\left(\sqrt{M\left|\mathcal{I}_\alpha^C\right| P \ln T} + MKC\right) \\
\leq & \mathcal{O}\left(\sqrt{M\left(|\mathcal{I}_\alpha^C| - 1\right) P \ln T} + MKC\right).
\end{aligned}$$

where the first inequality is by Eq. (B.30) and algebra; the second inequality is by algebra; the third inequality is from the elementary fact that $\min(A, B) \leq \sqrt{AB}$; the fourth inequality is from Jensen's inequality and the concavity of function $x \mapsto \sqrt{x}$, which implies that $\sum_{i \in \mathcal{I}_\alpha} \sqrt{\mathbb{E}\left[n_i(T)\right]} \leq \sqrt{|\mathcal{I}_\alpha| \left(\sum_{i \in \mathcal{I}_\alpha} \mathbb{E}\left[n_i(T)\right]\right)}$, and the fact that $\sum_{i \in \mathcal{I}_\alpha} \mathbb{E}\left[n_i(T)\right] \leq \sum_{i=1}^{M} \mathbb{E}\left[n_i(T)\right] \leq P$; the last inequality is from the simple observation that $\left|\mathcal{I}_\alpha^C\right| \leq 2(\left|\mathcal{I}_\alpha^C\right| - 1)$ when $\left|\mathcal{I}_\alpha^C\right| \geq 2$.

In summary, $\sum_{p=1}^{M} \sum_{i \in \mathcal{I}_\alpha^C} \mathbb{E}\left[n_i^p(T)\right] \Delta_i^p \leq \mathcal{O}\left(\sqrt{M\left(|\mathcal{I}_\alpha^C| - 1\right) P \ln T}\right) + MKC$. Combining this with Eq. (B.34), this concludes the proof of Eq. (B.32).

□

Proofs of Theorems 3.1 and 3.2. Combining Lemmas B.35, B.36, B.68 with $C = M$ and $\alpha = 10\epsilon$, Theorems 3.1 and 3.2 follow immediately. □

B.3.4 Auxiliary lemmas

Recall that we denote by $\overline{\Phi}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ the complementary CDF of the standard normal distribution.

Lemma B.69. *$\overline{\Phi}$ is monotonically decreasing. In addition, for $z \geq 0$,*

$$\frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} \exp\left(-\frac{z^2}{2}\right) \leq \overline{\Phi}(z) \leq \exp\left(-\frac{z^2}{2}\right),$$

where the first inequality (anti-concentration) is from [59]. In addition, for any $z \in \mathbb{R}$,

$$\overline{\Phi}(z) \leq \exp\left(-\frac{(z)_+^2}{2}\right), \quad \Phi(z) \leq \exp\left(-\frac{(-z)_+^2}{2}\right),$$

where we recall that $(z)_+ = \max(z, 0)$.

The following lemma is useful in bounding (b2.1), (e1.1), (e2.1); it can also be used to provide a simplified proof of the first case of Agrawal and Goyal [4, Lemma 2.13]. Roughly speaking, the lemma shows that a random variable X with a light lower probability tail must have a small value of $\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)}\right]$; it crucially uses the lower bound on $\overline{\Phi}$ (Gaussian anti-concentration) given in Lemma B.69.

Lemma B.70. *There exists some absolute constants $c_1, c_2 > 0$ such that the following holds. Given a random variable X, an event E and some $C_1 > 0$; if, for every $z \geq 1$, $\mathbb{P}(X \leq -z, E) \leq C_1 \exp(-2z^2)$, such that*

$$\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)} \mathbb{1}\{E\}\right] \leq c_1 C_1 + c_2.$$

Proof. Define $Y = -X$; we have $\mathbb{P}(Y \geq z, E) \leq C_1 \exp(-2z^2)$ for all $z \geq 1$.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)}\mathbb{1}\left\{E\right\}\right] \\
=&\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)}\mathbb{1}\left\{E, X \leq -1\right\}\right] + \mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)}\mathbb{1}\left\{E, X \geq -1\right\}\right] \\
\leq&\mathbb{E}\left[\frac{1}{\overline{\Phi}(Y)}\mathbb{1}\left\{E, Y \geq 1\right\}\right] + \frac{1}{\overline{\Phi}(1)} \\
\leq&8\sqrt{2\pi} \cdot \mathbb{E}\left[e^{Y^2}\mathbb{1}\left\{E, Y \geq 1\right\}\right] + \frac{1}{\overline{\Phi}(1)}.
\end{aligned}$$

where the first inequality follows due to the fact that $\frac{1}{\overline{\Phi}(z)}$ increases monotonically as z increases; and the second inequality is based on the observation that for $y \geq 1$, $\frac{1}{\overline{\Phi}(y)} \leq \sqrt{2\pi}\frac{y^2+1}{y}\exp(\frac{y^2}{2}) \leq 8\sqrt{2\pi}e^{y^2}$ (see Lemma B.69).

It suffices to show that $\mathbb{E}\left[e^{Y^2}\mathbb{1}\left\{E, Y \geq 1\right\}\right]$ is bounded by some constant, given the assumption on Y. Define $W = e^{Y^2}\mathbb{1}\left\{E, Y \geq 1\right\}$. We have that for any $w \geq e$,

$$\mathbb{P}(W \geq w) = \mathbb{P}(E, Y \geq \sqrt{\ln w}) \leq \frac{C_1}{w^2}.$$

As a result,

$$\begin{aligned}
\mathbb{E}\left[W\right] &= \int_0^\infty \mathbb{P}(W \geq w)\,\mathrm{d}w \\
&= \int_0^e \mathbb{P}(W \geq w)\,\mathrm{d}w + \int_e^\infty \mathbb{P}(W \geq w)\,\mathrm{d}w \\
&\leq e + \int_e^\infty \frac{C_1}{w^2}\,\mathrm{d}w \\
&\leq e + \frac{C_1}{e},
\end{aligned}$$

Therefore, the lemma holds by taking $c_1 = \frac{8\sqrt{2\pi}}{e}$ and $c_2 = 8\sqrt{2\pi}e + \frac{1}{\overline{\Phi}(1)}$. □

The following two lemmas are useful in bounding $(e1.1)$ (Lemma B.71), as well as

$(b2.1)$ and $(e2.1)$ (Lemma B.72), respectively.

Lemma B.71. *Fix any arm $i \in [K]$ and player $p \in [M]$. Let $N \in \mathbb{N}^+$. Suppose $\left\{g_t : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p) (Definition B.20). Then,*

$$\sum_{t:p\in\mathcal{P}_t} g_t \mathbb{1}\left\{i_t^p = i, n_i^p(t-1) < N\right\} \leq g_1 + \sum_{k=1}^{N-1} g_{\pi_k+1} \mathbb{1}\left\{\pi_k \leq T\right\},$$

where $\pi_k = \pi_k(i,p)$ denotes the round associated with the k-th pull of arm i by player p.

Proof. Let $h_t = g_t \mathbb{1}\left\{n_i^p(t-1) < N\right\}$. As seen in Example B.22, $\left\{n_i^p(t-1) : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p). This, combined with the assumption that $\left\{g_t : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p), implies that $\left\{h_t : t \in [T]\right\}$ is also invariant with respect to (i,p). Applying Lemma B.73 to the above $\left\{h_t : t \in [T]\right\}$, we have

$$\begin{aligned}
&\sum_{t:p\in\mathcal{P}_t} g_t \mathbb{1}\left\{i_t^p = i, n_i^p(t-1) < N\right\} \\
&= \sum_{t:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} \\
&\leq h_1 + \sum_{k=1}^{T} h_{\pi_k+1} \mathbb{1}\left\{\pi_k \leq T\right\} \\
&= g_1 \mathbb{1}\left\{n_i^p(0) < N\right\} + \sum_{k=1}^{T} g_{\pi_k+1} \mathbb{1}\left\{n_i^p(\pi_k) < N\right\} \mathbb{1}\left\{\pi_k \leq T\right\} \\
&= g_1 + \sum_{k=1}^{T} g_{\pi_k+1} \mathbb{1}\left\{k < N\right\} \mathbb{1}\left\{\pi_k \leq T\right\} \\
&= g_1 + \sum_{k=1}^{N-1} g_{\pi_k+1} \mathbb{1}\left\{\pi_k \leq T\right\},
\end{aligned}$$

where the first inequality is by Equation (B.36) in Lemma B.73; the second equality is by expanding the definition of h_t's; the third equality is from that $n_i^p(0) = 0$ and $n_i^p(\pi_k) = k$; and the last eqaulity is by algebra. □

Lemma B.72. *Fix any arm $i \in [K]$ and let $N \in \mathbb{N}^+$. Suppose $\left\{f_t^p : t \in [T], p \in [M]\right\}$ satisfies the invariant property with respect to arm i (Definition B.20), then,*

$$\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} f_t^p \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < N\right\} \le \sum_{p\in[M]} f_1^p + \sum_{k=1}^{N-1} f_{\tau_k+1}^{p_k} \mathbb{1}\left\{\tau_k \le T\right\},$$

where $(\tau_k, p_k) = (\tau_k(i), p_k(i))$ denote the round and player associated with the k-th pull of arm i by all players.

Proof of Lemma B.72. First, consider any fixed player p; let $h_t = f_t^p \mathbb{1}\left\{m_i^p(t-1) < N\right\}$. As seen in Example B.22, $\left\{m_i^p(t-1) : t \in [T]\right\}$ satisfies the invariant property with respect to (i, p). This, combined with the assumption that $\left\{f_t^p : t \in [T]\right\}$ satisfies the invariant property with respect to (i, p), implies that $\left\{h_t : t \in [T]\right\}$ is also invariant with respect to (i, p). Applying Lemma B.73 to the above $\left\{h_t : t \in [T]\right\}$, we have

$$\begin{aligned}
\sum_{t:p\in\mathcal{P}_t} f_t^p \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < N\right\} = &\sum_{t:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} \\
\le &h_1 + \sum_{t:p\in\mathcal{P}_t} h_{t+1} \mathbb{1}\left\{i_t^p = i\right\} \\
= &f_1^p + \sum_{t:p\in\mathcal{P}_t} f_{t+1}^p \mathbb{1}\left\{i_t^p = i, m_i^p(t) < N\right\} \\
= &f_1^p + \sum_{t:p\in\mathcal{P}_t} f_{t+1}^p \mathbb{1}\left\{i_t^p = i, n_i(t) < N\right\} \qquad \text{(B.35)}
\end{aligned}$$

where the first inequality is from Equation (B.37) of Lemma B.73; the second equality is by expanding the definition of h_t and noting that $h_1 = \mathbb{1}\left\{m_i^p(0) < N\right\} f_1^p = \mathbb{1}\left\{0 < N\right\} f_1^p = f_1^p$; the third equality is from the observation that, if $i_t^p = i$ and $u_i^p(t) = t$, then $m_i^p(t) = n_i(u_i^p(t)) = n_i(t)$.

Now, summing Equation (B.35) over all players $p \in [M]$, we have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} f_t^p \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < N\right\} \\
\leq &\sum_{p\in[M]} f_1^p + \sum_{p\in[M]}\sum_{t:p\in\mathcal{P}_t} f_{t+1}^p \mathbb{1}\left\{i_t^p = i, n_i(t) < N\right\} \\
\leq &\sum_{p\in[M]} f_1^p + \sum_{k=1}^{N-1} f_{\tau_k+1}^{p_k} \mathbb{1}\left\{\tau_k \leq T\right\},
\end{aligned}$$

where the second inequality is from the observation that for every $t \in [T]$, $p \in \mathcal{P}_t$ such that $i_t^p = i$ and $n_i(t) < N$, there must exists some unique $k \in [N-1]$ such that $\tau_k = t$ and $p_k = p$. □

The following auxiliary lemma facilitates the proofs of Lemmas B.71 and B.72.

Lemma B.73. *Fix any arm $i \in [K]$ and player $p \in [M]$. Suppose $\left\{h_t : t \in [T]\right\}$ satisfies the invariant property with respect to (i, p) (Definition B.20). Then,*

$$\begin{aligned}
\sum_{t\in[T]:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} \leq &h_1 + \sum_{k=1}^{T} h_{\pi_k+1} \mathbb{1}\left\{\pi_k \leq T\right\} &\text{(B.36)}\\
= &h_1 + \sum_{t\in[T]:p\in\mathcal{P}_t} h_{t+1} \mathbb{1}\left\{i_t^p = i\right\}, &\text{(B.37)}
\end{aligned}$$

where $\pi_k = \pi_k(i, p)$ denotes the round associated with the k-th pull of arm i by player p.

Proof.

$$\begin{aligned}
\sum_{t\in[T]:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} =& \sum_{k=1}^{T} h_{\pi_k} \mathbb{1}\left\{\pi_k \le T\right\} \\
=& \sum_{k=1}^{T} h_{\pi_{k-1}+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
\le& h_1 + \sum_{k=2}^{T} h_{\pi_{k-1}+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
=& h_1 + \sum_{k=1}^{T-1} h_{\pi_k+1} \mathbb{1}\left\{\pi_{k+1} \le T\right\} \\
\le& h_1 + \sum_{k=1}^{T-1} h_{\pi_k+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
=& h_1 + \sum_{t\in[T]:p\in\mathcal{P}_t} h_{t+1} \mathbb{1}\left\{i_t^p = i\right\},
\end{aligned}$$

where the first equality uses the definition of π_k; the second equality uses the invariant property, specifically, $h_{\pi_k} = h_{\pi_{k-1}+1}$; the first inequality uses the observation that the first term $h_{\pi_0+1}\mathbb{1}\left\{\pi_1 \le T\right\} = h_1\mathbb{1}\left\{\pi_1 \le T\right\} \le h_1$; the third equality shifts the indices in the sum by 1; the second inequality uses the observation that $\pi_{k+1} \le T \implies \pi_k \le T$; and the last equality is again by the definition of π_k. □

The following lemma is largely inspired by Agrawal and Goyal [4, Lemma 2.8]; here we generalize it to the multi-task setting, for reducing bounding (B) and (E) to bounding $(B*)$ and $(E*)$ respectively.

Lemma B.74. *For any player $p \in [M]$, time step $t \in [T]$, and arm $i \in [K]$, we have for any arm $l \in [K]$ and any threshold $z \in \mathbb{R}$:*

$$\Pr\left(i_t^p = i, \theta_i^p(t) \le z \mid \mathcal{F}_{t-1}\right) \le \frac{\Pr\left(\theta_l^p(t) \le z \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_l^p(t) > z \mid \mathcal{F}_{t-1}\right)} \cdot \Pr\left(i_t^p = l \mid \mathcal{F}_{t-1}\right).$$

Proof. First,

$$\begin{aligned}
&\Pr\left(i_t^p = i, \overline{Q_i^p(t)} \mid \mathcal{F}_{t-1}\right) \\
\leq &\Pr\left(\forall j \in [K],\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right) \\
= &\Pr\left(\theta_l^p(t) \leq z \mid \mathcal{F}_{t-1}\right) \cdot \Pr\left(\forall j \neq l,\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right),
\end{aligned}$$

where the first inequality follows because the event $\left\{i_t^p = i, \overline{Q_i^p(t)}\right\}$ happens only if $\forall j \in [K],\ \theta_j^p(t) \leq z$; and the second equality follows because conditional on $\mathcal{F}_{t-1}$, the draws $\theta_j^p(t)$'s and $\theta_l^p(t)$ are independent.

Now, observe that

$$\begin{aligned}
&\Pr\left(\forall j \neq l,\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right) \\
= &\frac{\Pr\left(\theta_l^p(t) > z, \text{ and } \forall j \neq l,\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_l^p(t) > z \mid \mathcal{F}_{t-1}\right)} \\
\leq &\frac{\Pr\left(i_t^p = l \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_l^p(t) > z \mid \mathcal{F}_{t-1}\right)}
\end{aligned}$$

where the equality follows, again, by the conditional independence of $\left\{\theta_j^p(t) : j \neq l\right\}$ and $\theta_l^p(t)$; and the inequality follows because the event $\left\{\theta_l^p(t) > z,\ \forall j \neq l,\ \theta_j^p(t) \leq y_i^p\right\}$ implies that $\left\{i_t^p = l\right\}$ happens. The lemma follows from combining the above two inequalities. □

B.4 Theoretical Guarantees of Baselines

B.4.1 IND-UCB and IND-TS in the generalized ϵ-MPMAB setting

Theorem B.75. *The expected collective regret of IND-UCB and IND-TS after T rounds satisfies the following two upper bounds simultaneously:*

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in[K]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right) \tag{B.38}$$

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{MKP}\right), \tag{B.39}$$

where we recall that $P = \sum_{t=1}^{T}|\mathcal{P}_t|$.

Proof sketch. For Eq. (B.38), we note that both IND-UCB and IND-TS guarantees that for every $p \in [M]$,

$$\mathrm{Reg}^p(T) \leq \mathcal{O}\left(\sum_{i\in[K]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right);$$

summing over p yields Eq. (B.38). For Eq. (B.39), we note that for every $p \in [M]$,

$$\mathrm{Reg}^p(T) \leq \tilde{\mathcal{O}}\left(\sqrt{K\left|\{t : p \in \mathcal{P}_t\}\right|}\right).$$

Summing over all $p \in [M]$, we have

$$\begin{aligned}\mathrm{Reg}(T) = \sum_{p=1}^{M}\mathrm{Reg}^p(T) &\leq \tilde{\mathcal{O}}\left(\sum_{p=1}^{M}\sqrt{K\left|\{t : p \in \mathcal{P}_t\}\right|}\right)\\ &\leq \tilde{\mathcal{O}}\left(\sqrt{MK\sum_{p=1}^{M}\left|\{t : p \in \mathcal{P}_t\}\right|}\right) = \tilde{\mathcal{O}}\left(\sqrt{MKP}\right).\end{aligned}$$

□

Algorithm 6: RobustAgg(ϵ) for the generalized ϵ-MPMAB setting

Input : Dissimilarity parameter $\epsilon \in [0, 1]$;

Initialization: Set $n_i^p = 0$ for all $p \in [M]$ and all $i \in [K]$.

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

  Receive active set of players $\mathcal{P}_t$;

  for $p \in \mathcal{P}_t$ **do**

    for $i \in [K]$ **do**

      Let $m_i^p = \sum_{q \in [M]: q \neq p} n_i^q$;

      Let $\overline{n_i^p} = \max(1, n_i^p)$ and $\overline{m_i^p} = \max(1, m_i^p)$;

      Let

$$\zeta_i^p(t) = \frac{1}{\overline{n_i^p}} \sum_{\substack{s<t \\ i_s^p = i}} r_s^p, \qquad \eta_i^p(t) = \frac{1}{\overline{m_i^p}} \sum_{\substack{q \in [M] \\ q \neq p}} \sum_{\substack{s<t \\ i_s^q = i}} r_s^q,$$

$$\text{and } \kappa_i^p(t, \lambda) = \lambda \zeta_i^p(t) + (1 - \lambda) \eta_i^p(t);$$

      Let $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon) = 8\sqrt{13 \ln T \left[\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}\right]} + (1 - \lambda)\epsilon$;

      Compute $\lambda^* = \text{argmin}_{\lambda \in [0,1]} F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$;

      Compute an upper confidence bound of the reward of arm i for player p:

$$\text{UCB}_i^p(t) = \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon).$$

    Let $i_t^p = \text{argmax}_{i \in [K]} \text{UCB}_i^p(t)$;

    Player p pulls arm i_t^p and observes reward r_t^p;

  for *active players* $p \in \mathcal{P}_t$ **do**

    Let $i = i_t^p$ and set $n_i^p = n_i^p + 1$.

B.4.2 RobustAgg(ϵ) and its regret analysis in the generalized ϵ-MPMAB setting

In Chapter 2, we studied a special case of ϵ-MPMAB problem, which can be viewed as ϵ-MPMAB problem defined in Section 3.2, with active sets of players $\mathcal{P}_t \equiv [M]$. In this specialized setting, they propose RobustAgg(ϵ), a UCB-based algorithm that achieves a

gap-dependent and gap-independent regret of

$$\mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_{5\epsilon}}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}+\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}+MK\right), \tag{B.40}$$

and

$$\tilde{\mathcal{O}}\left(\sqrt{M|\mathcal{I}_{5\epsilon}|T}+M\sqrt{|\mathcal{I}_{5\epsilon}^C|T}+MK\right), \tag{B.41}$$

respectively. In this section, we show that, with a few small modifications, their algorithm and analysis can be used in our (more general) ϵ-MPMAB setting, where the active sets $\mathcal{P}_t$ can change over time.

Specifically, Algorithm 6 is our modified version of RobustAgg(ϵ). Recall that RobustAgg(ϵ) performs an UCB-based exploration [8]: for every player and every arm, it constructs high-probability UCBs on the expected rewards (line 6 to 11); to this end, it makes careful use of both the player and other players' data, and construct a series of UCBs parameterized by λ (line 9), and selects the tightest one (line 10 and 11). Compared to RobustAgg(ϵ), for every round t, Algorithm 6 only computes expected reward UCBs for active players $p \in \mathcal{P}_t$ (line 4), and updates arm pull counts on active players (line 14).

We show that Algorithm 6, when applied to our ϵ-MPMAB setting, has regret guarantees that recover and generalize RobustAgg(ϵ)'s original guarantees. Specifically, in the specialized ϵ-MPMAB setting where $\mathcal{P}_t \equiv [M]$, we recover the regret guarantees of RobustAgg(ϵ) (Equations (B.40) and (B.41)).

Theorem B.76. *The expected collective regret of RobustAgg(ϵ) after T rounds satisfies*

the following two upper bounds simultaneously:

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_{5\epsilon}}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}+\sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}+MK\right), \tag{B.42}$$

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{5\epsilon}|P}+\sqrt{M\left(|\mathcal{I}_{5\epsilon}^C|-1\right)P}+MK\right), \tag{B.43}$$

where we recall that $P=\sum_{t=1}^T|\mathcal{P}_t|$.

Proof sketch. Even in the general setting where $\mathcal{P}_t$ is not necessarily $[M]$, Freedman's inquality can still be applied to establish the high-probability concentration of the empirically averaged rewards $\zeta_i^p(t)$ and $\eta_i^p(t)$; therefore, Lemma A.4 still holds in the general setting. As a result, Lemmas A.7 and A.8 carry over; hence, for all $i\in\mathcal{I}_{5\epsilon}$, Algorithm 6 still satisfies that

$$\mathbb{E}[n_i(T)] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2}+M\right), \tag{B.44}$$

and for all $i\in\mathcal{I}_{5\epsilon}^C$ and all $p\in[M]$,

$$\mathbb{E}[n_i^p(T)] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right). \tag{B.45}$$

Equations (B.42) and (B.43) now follows directly from applying Lemma B.68 with $C=0$ and $\alpha=5\epsilon$. □

B.5 Additional Experimental Results

In this section, we present the rest of the experimental results. Figures B.2, B.3, and B.4 compare the average performance of RobustAgg-TS(0.15), RobustAgg(0.15), Ind-UCB, and Ind-TS in randomly generated 0.15-MPMAB problem instances with different numbers of subpar arms.

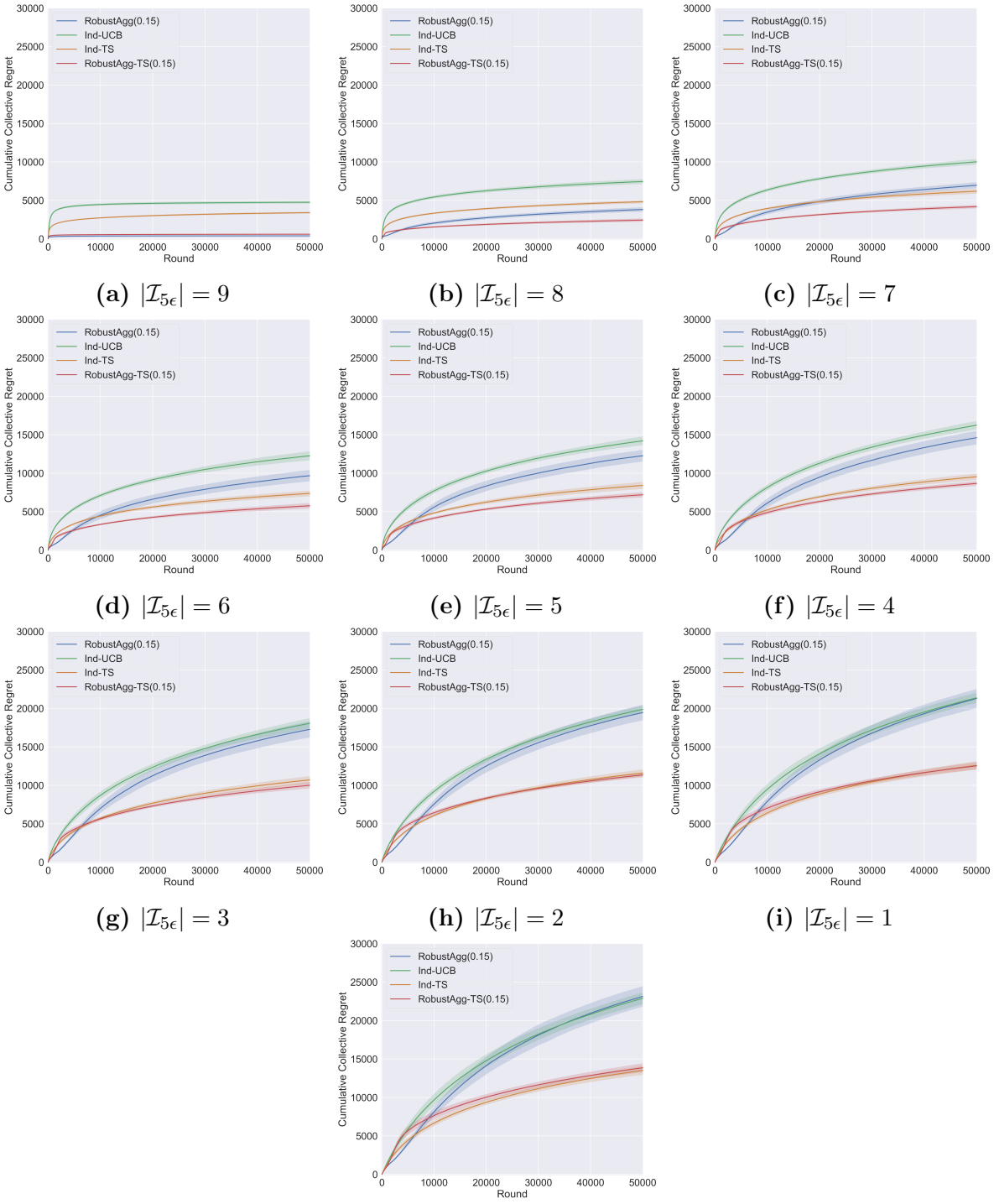
Note that, when $|\mathcal{I}_{5\epsilon}| = 9$, we have $\left|\mathcal{I}_{5\epsilon}^{C}\right| = 1$ which means that there exists one arm that is optimal to all the players and the other arms are all subpar. In this favorable special case, RobustAgg-TS(0.15) and RobustAgg(0.15) perform significantly better than the baseline algorithms without transfer, as expected.

Furthermore, when $|\mathcal{I}_{5\epsilon}| = 0$, i.e., there is no subpar arm and all the arms have relatively small suboptimality gaps. In this unfavorable special case, RobustAgg-TS(0.15)'s performance is still very competitive in comparison with Ind-TS, which demonstrates the robustness of our proposed algorithm.

B.5.1 Empirical comparison with RobustAgg-TS-V(ϵ)

We empirically evaluated a variant of Algorithm 2, named RobustAgg-TS-V(ϵ). RobustAgg-TS-V(ϵ) differs from RobustAgg-TS(ϵ) (Algorithm 2) in one way: in each round, instead of only updating the posteriors associated with each active player and its pulled arm (i.e., delayed update, line 15 of Algorithm 2), RobustAgg-TS-V(ϵ) updates the posteriors associated with every arm and player. Note that this change only affects the aggregate posteriors, as the individual posteriors associated with a player and an arm remains the same if the player does not pull the arm in this round.

Figure B.5 compares the average cumulative regret of RobustAgg-TS(0.15), RobustAgg-TS-V(0.15), RobustAgg(0.15), Ind-UCB, and Ind-TS in randomly generated 0.15-MPMAB problem instances with different numbers of subpar arms. The instances were generated following the same procedure as the other experiments. Observe that RobustAgg-TS-V(0.15)'s empirical performance is on par with that of RobustAgg-TS(0.15). However, our analysis in this chapter takes advantages of the design choice made for RobustAgg-TS(ϵ), i.e., delayed update which leads to the invariant property (Definition B.20 and Examples B.21, B.22 and B.23). It is unclear whether RobustAgg-TS-V(ϵ) enjoys similar near-optimal guarantees.



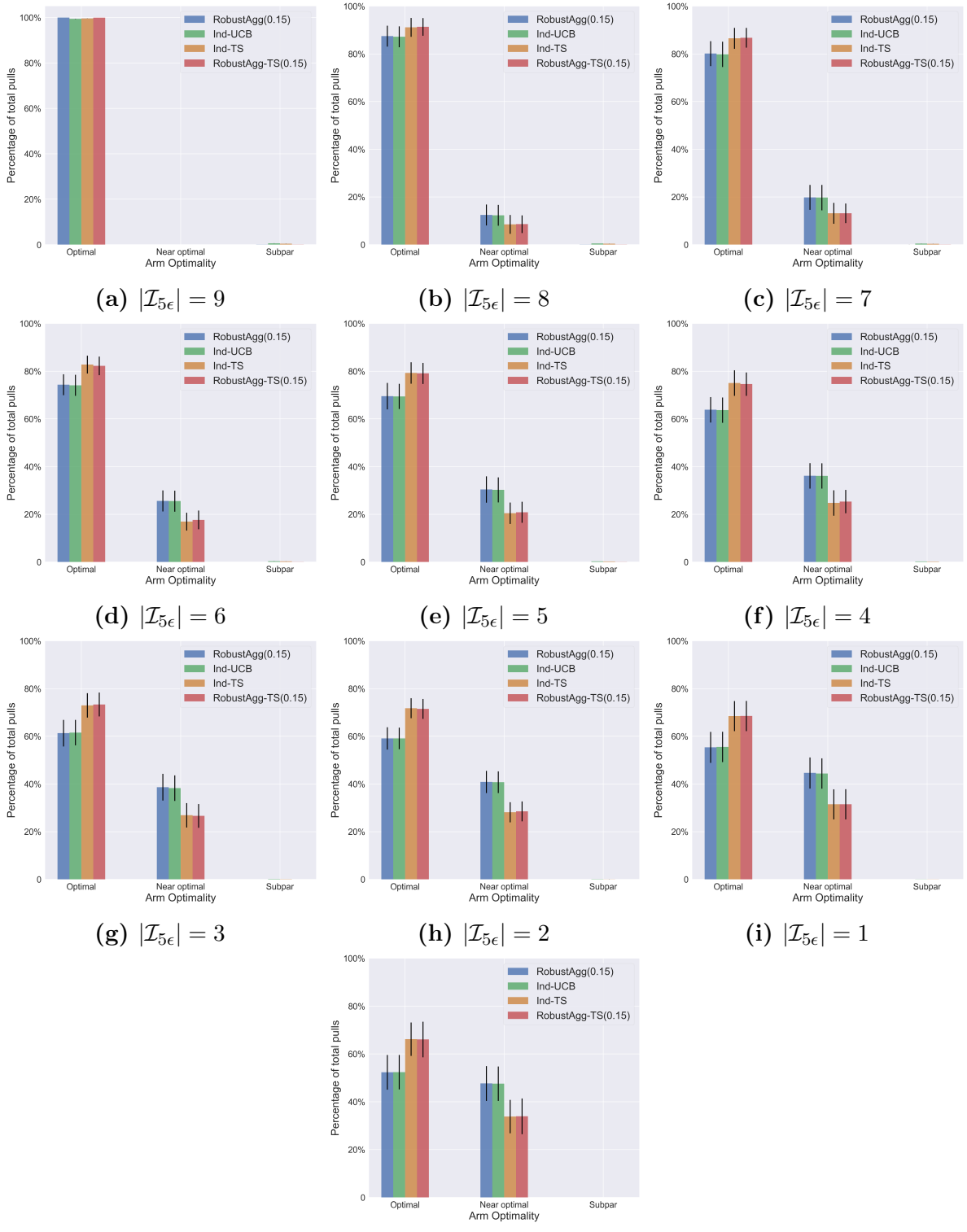
(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 8$ **(c)** $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$ **(e)** $|\mathcal{I}_{5\epsilon}| = 5$ **(f)** $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$ **(h)** $|\mathcal{I}_{5\epsilon}| = 2$ **(i)** $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure B.2. Compares the cumulative collective regret of the 4 algorithms over a horizon of $T = 50,000$ rounds.



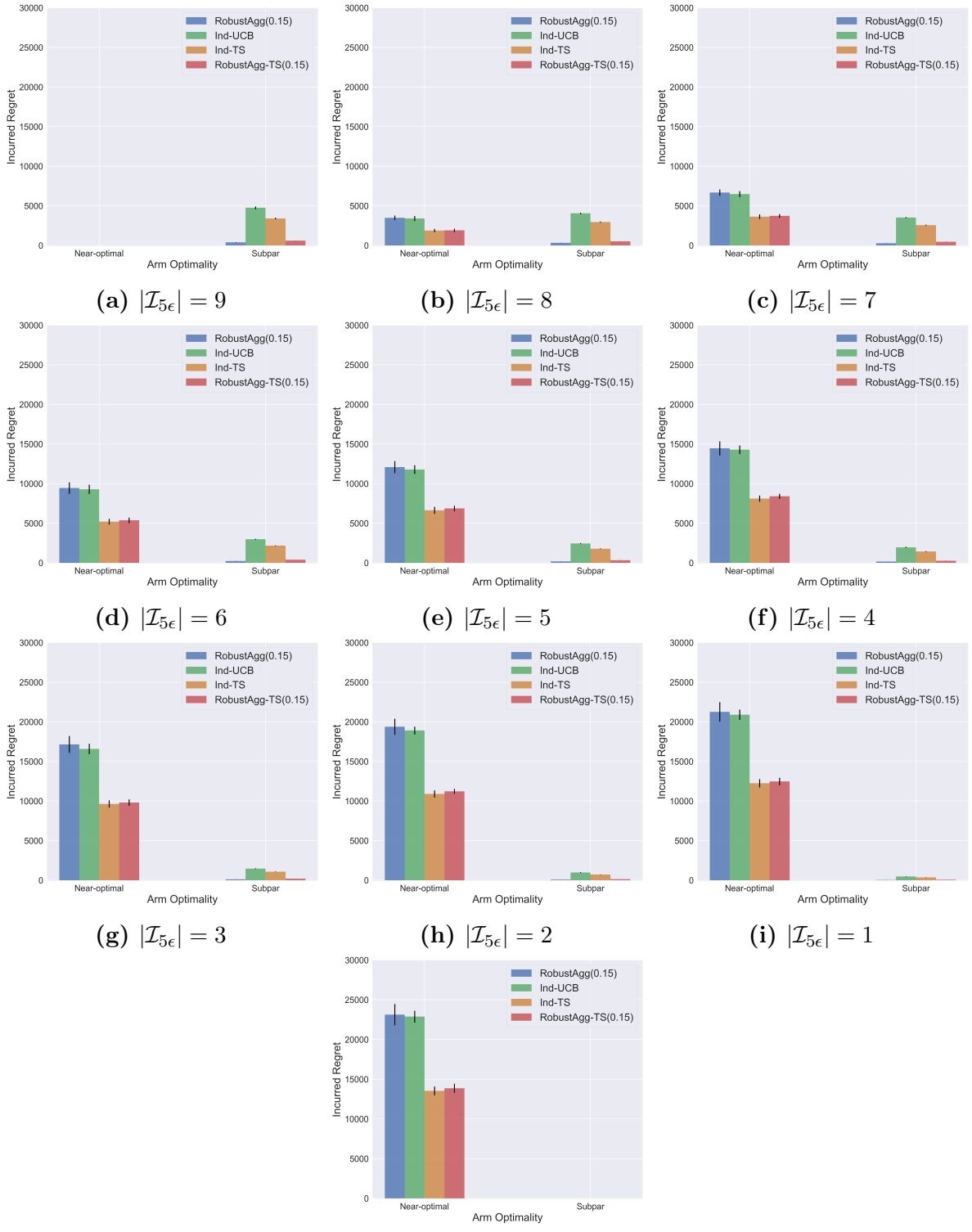
(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 8$ **(c)** $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$ **(e)** $|\mathcal{I}_{5\epsilon}| = 5$ **(f)** $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$ **(h)** $|\mathcal{I}_{5\epsilon}| = 2$ **(i)** $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure B.3. Compares the percentage of arm pulls by arm optimality for the 4 algorithms in $T = 50,000$ rounds.



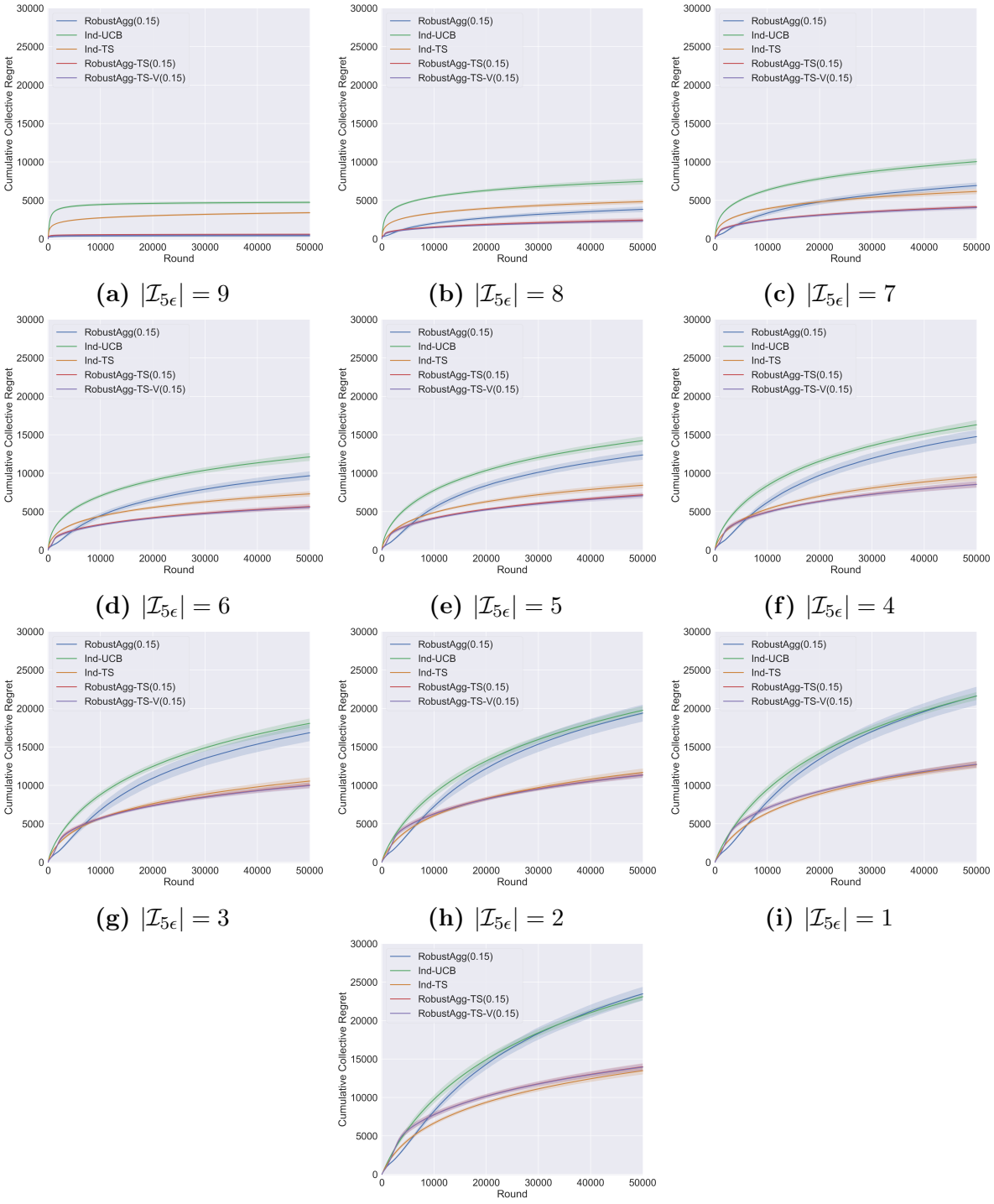
(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 8$ **(c)** $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$ **(e)** $|\mathcal{I}_{5\epsilon}| = 5$ **(f)** $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$ **(h)** $|\mathcal{I}_{5\epsilon}| = 2$ **(i)** $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure B.4. Compares the cumulative collective regret incurred by arm optimality for the 4 algorithms in $T = 50,000$ rounds.



(a) $|\mathcal{I}_{5\epsilon}| = 9$ **(b)** $|\mathcal{I}_{5\epsilon}| = 8$ **(c)** $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$ **(e)** $|\mathcal{I}_{5\epsilon}| = 5$ **(f)** $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$ **(h)** $|\mathcal{I}_{5\epsilon}| = 2$ **(i)** $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure B.5. Compares the cumulative collective regret of the 5 algorithms over a horizon of $T = 50,000$ rounds.

Appendix C

Supplementary Material for Chapter 4

C.1 Proofs of Lemmas 4.2 and 4.4

C.1.1 Proof of Lemma 4.2

Lemma 4.2. *If $(\mathcal{M}_p)_{p=1}^M$ is ϵ-dissimilar, then for every $p, q \in [M]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \leq 2H\epsilon,$$

consequently, $\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\right| \leq 4H\epsilon$.

Proof. For the first claim, we prove a stronger statement by backward induction on h, namely, for every $p, q \in [M]$, every $h \in [1, H]$, and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \leq 2(H - h + 1)\epsilon.$$

Base case:

For $h = H + 1$, we have $Q_p^\star(s,a) = 0$ for every $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, and $p \in [M]$. It follows trivially that $\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| = 0 \leq 2(H - h + 1)\epsilon$.

Inductive case:

Suppose by inductive hypothesis that for some $h \in [1, H]$ and, for every $(s,a) \in \mathcal{S}_{h+1} \times \mathcal{A}$ and $p, q \in [M]$, $\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \leq 2(H - h)\epsilon$.

We first prove the following auxiliary statement: for every $s \in \mathcal{S}_{h+1}$ and $p, q \in [M]$,

$$\left|V_p^\star(s) - V_q^\star(s)\right| \le 2(H-h)\epsilon. \tag{C.1}$$

Let $a_p = \operatorname{argmax}_{a\in\mathcal{A}} Q_p^\star(s,a)$ and $a_q = \operatorname{argmax}_{a\in\mathcal{A}} Q_q^\star(s,a)$. The above auxiliary statement can be easily proven by contradiction: without loss of generality, suppose that $V_p^\star(s) - V_q^\star(s) = Q_p^\star(s,a_p) - Q_q^\star(s,a_q) > 2(H-h)\epsilon$. Since $Q_q^\star(s,a_p) \ge Q_p^\star(s,a_p) - 2(H-h)\epsilon$, it follows that $Q_q^\star(s,a_p) > Q_q^\star(s,a_q)$, which contradicts the fact that $a_q = \operatorname{argmax}_{a\in\mathcal{A}} Q_q^\star(s,a)$.

We now return to the inductive proof, and we show that given the inductive hypothesis, for every $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ and $p, q \in [M]$,

$$\begin{aligned}
&\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \\
\le& \left|R_p(s,a) - R_q(s,a)\right| + \left|\sum_{s'\in\mathcal{S}_{h+1}} \left[\mathbb{P}_p(s' \mid s,a)V_p^\star(s') - \mathbb{P}_q(s' \mid s,a)V_q^\star(s')\right]\right| \\
\le& \epsilon + \left|\sum_{s'\in\mathcal{S}_{h+1}} \left[\mathbb{P}_p(s' \mid s,a)V_p^\star(s') - \mathbb{P}_q(s' \mid s,a)V_p^\star(s')\right]\right| + \\
&\qquad\qquad \left|\sum_{s'\in\mathcal{S}_{h+1}} \mathbb{P}_q(s' \mid s,a)\left(V_p^\star(s') - V_q^\star(s')\right)\right| \\
\le& \epsilon + \|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a))\|_1 \left(\max_{s'\in\mathcal{S}_{h+1}} \left|V_p^\star(s')\right|\right) + \\
&\qquad\qquad \|\mathbb{P}_q(\cdot \mid s,a)\|_1 \left(\max_{s'\in\mathcal{S}_{h+1}} \left|V_p^\star(s') - V_q^\star(s')\right|\right) \\
\le& \epsilon + \frac{\epsilon}{H}\cdot H + 2(H-h)\epsilon \\
=& 2(H-h+1)\epsilon,
\end{aligned}$$

where the first inequality follows from Eq. (4.1) and the triangle inequality; the second inequality follows from Definition 4.1 and the triangle inequality; the third inequality follows from Hölder's inequality; and the fourth inequality uses Definition 4.1 and Eq. (C.1).

For the second claim, we note that from the first claim, we have for any p, q, s,

$$\left|V_p^\star(s) - V_q^\star(s)\right| = \left|\max_{a\in\mathcal{A}} Q_p^\star(s,a) - \max_{a\in\mathcal{A}} Q_p^\star(s,a)\right| \le 2H\epsilon,$$

therefore, for any p, q, s, a,

$$\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\right| \le \left|V_p^\star(s) - V_q^\star(s)\right| + \left|Q_p^\star(s,a) - Q_p^\star(s,a)\right| \le 4H\epsilon.$$

□

C.1.2 Proof of Lemma 4.4

Lemma 4.4. *For any $(s,a) \in \mathcal{I}_\epsilon$, we have that: (1) for all $p \in [M]$, $(s,a) \notin Z_{p,\mathrm{opt}}$, where we recall that $Z_{p,\mathrm{opt}} = \left\{(s,a) : \mathrm{gap}_p(s,a) = 0\right\}$ is the set of optimal state-action pairs with respect to p; (2) for all $p, q \in [M]$, $\mathrm{gap}_p(s,a) \ge \frac{1}{2}\mathrm{gap}_q(s,a)$.*

Proof. For any $(s,a) \in \mathcal{I}_\epsilon$, there exists some p_0 such that $\mathrm{gap}_{p_0}(s,a) \ge 96H\epsilon$. Therefore, for every $p \in [M]$,

$$\mathrm{gap}_p(s,a) \ge \mathrm{gap}_{p_0}(s,a),$$

From Lemma 4.2 we know that $\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_{p_0}(s,a)\right| \le 4H\epsilon$. Therefore, for all p,

$$\mathrm{gap}_p(s,a) \ge \mathrm{gap}_{p_0}(s,a) - 4H\epsilon \ge 92H\epsilon > 0.$$

This proves the first item.

For the second item, for all $p, q \in [M]$,

$$\frac{\mathrm{gap}_p(s,a)}{\mathrm{gap}_q(s,a)} = \frac{\mathrm{gap}_q(s,a) - 4H\epsilon}{\mathrm{gap}_q(s,a)} \ge 1 - \frac{4H\epsilon}{\mathrm{gap}_q(s,a)} \ge 1 - \frac{4}{92} \ge \frac{1}{2}.$$

□

C.2 Additional Definitions Used in the Proofs

In this section, we define a few useful notations that will be used in our proofs. For state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, player $p \in [M]$, episode $k \in [K]$:

1. Define $n^k(s,a)$ (resp. $n_p^k(s,a)$, $\hat{\mathbb{P}}^k$, $\hat{\mathbb{P}}_p^k$, $\hat{R}^k$, $\hat{R}_p^k$) to be the value of $n(s,a)$ (resp. $n_p(s,a)$, $\hat{\mathbb{P}}$, $\hat{\mathbb{P}}_p$, $\hat{R}$, $\hat{R}_p$) at the *beginning* of episode k of Multi-task-Euler.

2. Denote by $\overline{Q}_p^k$ (resp. $\underline{Q}_p^k, \overline{V}_p^k, \underline{V}_p^k$, ind-$b_p^k(s,a)$, agg-$b_p^k(s,a)$) the values of $\overline{Q}_p$ (resp. $\underline{Q}_p, \overline{V}_p, \underline{V}_p$, ind-$b_p(s,a)$, agg-$b_p(s,a)$) right after Multi-task-Euler finishes its optimistic value iteration (line 15) at episode k.

3. Define the *surplus* [136] (also known as the Bellman error) of (s,a) at episode k and player p as:
$$E_p^k(s,a) := \overline{Q}_p^k(s,a) - R_p(s,a) - (\mathbb{P}_p \overline{V}_p^k)(s,a).$$

4. Define $w_p^k(s,a) := \frac{n_p^k(s,a)}{n^k(s,a)}$ be the proportion of player p on (s,a) at the beginning of episode k; this induces (s,a)'s *mixture expected reward*:
$$\bar{R}^k(s,a) := \sum_{q=1}^{M} w_q^k(s,a) R_q(s,a),$$
and *mixture transition probability*:
$$\bar{\mathbb{P}}^k(\cdot \mid s,a) := \sum_{q=1}^{M} w_q^k(s,a) \mathbb{P}_q(\cdot \mid s,a).$$

5. Define $\rho_p^k(s,a) := \mathbb{P}((s_h,a_h) = (s,a) \mid \pi^k(p), \mathcal{M}_p)$ to be the occupancy measure of $\pi^k(p)$ over $\mathcal{M}_p$ on (s,a), where $h \in [H]$ is the layer s is in (so that $s \in \mathcal{S}_h$). It can be seen that ρ_p^k, when restricted to $\mathcal{S}_h \times \mathcal{A}$, is a probability distribution on this set.

   Define $\rho^k(s,a) := \sum_{p=1}^{M} \rho_p^k(s,a)$; it can be seen that $\bar{\rho}^k(s,a) \in [0,M]$. Define $\bar{n}_p^k(s,a) :=$

$\sum_{j=1}^{k} \rho_p^j(s,a)$, and $\bar{n}^k(s,a) := \sum_{j=1}^{k} \rho^j(s,a)$.[1]

6. Define $N^k(s) := \sum_{a\in\mathcal{A}} n^k(s,a)$ and $N_p^k(s) := \sum_{a\in\mathcal{A}} n_p^k(s,a)$ to be the total number of encounters of state s by all players, and by player p only, respectively, at the beginning of episode k.

7. Define $N_1 \eqsim M \ln(\frac{SAK}{\delta})$, and $N_2 \eqsim \ln(\frac{MSAK}{\delta})$; define $\tau(s,a) := \min\left\{k : \bar{n}^k(s,a) \geq N_1\right\}$, and $\tau_p(s,a) := \min\left\{k : \bar{n}_p^k(s,a) \geq N_2\right\}$; With high probability, so long as $k \geq \tau(s,a)$ (resp. $k \geq \tau_p(s,a)$), $n^k(s,a)$ and $\bar{n}^k(s,a)$ (resp. $n_p^k(s,a)$ and $\bar{n}_p^k(s,a)$) are within a constant factor of each other; see Lemma C.3.

8. Define $\breve{\text{gap}}_p(s,a) := \frac{\text{gap}_p(s,a)}{4H} \vee \frac{\text{gap}_{p,\min}}{4H}$; recall the definitions of $\text{gap}_p(s,a)$ and $\text{gap}_{p,\min}$ in Section 4.2.

Define $\text{Reg}(K,p) := \sum_{k=1}^{K}\left(V_{0,p}^{\star} - V_{0,p}^{\pi^k(p)}\right)$ as player p's contribution to the collective regret; in this notation, $\text{Reg}(K) = \sum_{p=1}^{M} \text{Reg}(K,p)$.

Define the clipping function $\text{clip}(\alpha,\Delta) := \alpha\mathbf{1}(\alpha \geq \Delta)$.

We also adopt the following conventions in our proofs:

1. As ϵ-dissimilarity with $\epsilon > 2H$ does not impose any constraints on $\left\{\mathcal{M}_p\right\}_{p=1}^{M}$, throughout the proof, we only focus on the regime that $\epsilon \leq 2H$.

2. We will use $\pi^k(p)$ and π_p^k interchangeably. To avoid notational clutter, we will also sometimes slightly abuse notation, using $V_{p,h}^{\pi^k}$, $V_p^{\pi^k}$ to denote $V_{p,h}^{\pi^k(p)}$, $V_p^{\pi^k(p)}$ respectively.

C.3 Proof of the Upper Bounds

This section establishes the regret guarantees in Theorems 4.5 and 4.6. The proof follows a similar outline as STRONG-EULER's analysis [136], with important modifications tailored to the multitask setting. The proof has the following roadmap:

[1]These are the cumulative occupancy measures up to episode k, inclusively; this is in contrast with the definition of $n^k(s,a)$ and $n_p^k(s,a)$, which do not count the trajectories observed at episode k.

1. Subsection C.3.1 defines a clean event E that we show happens with probability $1-\delta$. When E happens, the observed samples are typical enough so that standard concentration inequalities apply. This will serve as the basis of our subsequent arguments.

2. Subsection C.3.2 shows that when E happens, the value function upper and lower bounds are valid; furthermore, MULTI-TASK-EULER enjoys strong optimism [136], in that all players' surpluses are always nonnegative for all state-action pairs at all time steps.

3. Subsection C.3.3 establishes a distribution-dependent upper bound on the surpluses of MULTI-TASK-EULER when E happens, which is key to our regret theorems. In comparison with STRONG-EULER [136] in the single task setting, MULTI-TASK-EULER exploits inter-task similarity, so that its surpluses on state-action pair (s,a) for player p are further controlled by a new term that depends on the dissimilarity parameter ϵ, along with $n^k(s,a)$, the total visitation counts of (s,a) *by all players*.

4. Subsection C.3.4 uses the strong optimism property and the surplus bounds established in the previous two subsections to conclude our final gap-independent and gap-dependent regret guarantees, via the clipping lemma [136] (see also Lemma C.12).

5. Finally, Subsection C.3.5 collects miscellaneous technical lemmas used in the proofs.

C.3.1 A clean event

Below we will define a "clean" event E in which all concentration bounds used in the analysis hold, which we will show happens with high probability. Specifically, we will define $E = E_{\text{ind}} \cap E_{\text{agg}} \cap E_{\text{sample}}$, where $E_{\text{ind}}, E_{\text{agg}}, E_{\text{sample}}$ are defined respectively below.

In subsequent definitions of events, we will abbreviate $\forall k \in [K], h \in [H], p \in [M], s \in \mathcal{S}_h, a \in \mathcal{A}, s' \in \mathcal{S}_{h+1}$ as $\forall k, h, p, s, a, s'$. Also, recall that $L(n) \eqsim \ln(\frac{MSAn}{\delta})$.

Define event E_{ind} as:

$$E_{\text{ind}} = E_{\text{ind,rw}} \cap E_{\text{ind,val}} \cap E_{\text{ind,prob}} \cap E_{\text{ind,var}}, \tag{C.2}$$

$$E_{\text{ind,rw}} = \left\{ \forall k, h, p, s, a \,.\, \left| \hat{R}_p^k(s,a) - R_p(s,a) \right| \leq \sqrt{\frac{L(n^k(s,a))}{2n^k(s,a)}} \right\}, \tag{C.3}$$

$$\begin{aligned} E_{\text{ind,val}} = \Bigg\{ \forall k, h, p, s, a \,.\, & \left| (\hat{\mathbb{P}}_p^k V_p^\star - \mathbb{P}_p V_p^\star)(s,a) \right| \\ & \leq 4\sqrt{\frac{\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{2HL(n_p^k(s,a))}{n_p^k(s,a)} \Bigg\}, \end{aligned} \tag{C.4}$$

$$\begin{aligned} E_{\text{ind,prob}} = \Bigg\{ \forall k, h, p, s, a, s' \,.\, & \left| (\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(s' \mid s,a) \right| \\ & \leq 4\sqrt{\frac{L(n_p^k(s,a)) \cdot \mathbb{P}_p(s' \mid s,a)}{n_p^k(s,a)}} + \frac{2L(n_p^k(s,a))}{n_p^k(s,a)} \Bigg\}, \end{aligned} \tag{C.5}$$

$$\begin{aligned} E_{\text{ind,var}} = \Bigg\{ \forall k, h, p, s, a \,.\, & \left| \frac{1}{n_p^k(s,a)} \sum_{i=1}^{n_p^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right|, \\ & \leq 4\sqrt{\frac{H^2 \text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{2H^2 L(n_p^k(s,a))}{n_p^k(s,a)} \Bigg\}, \end{aligned} \tag{C.6}$$

where in Equation (C.6), s_i' denotes the next state player p transitions to, for the i-th time it experiences (s,a). E_{ind} captures the concentration behavior of each player's individual model estimates.

Lemma C.1. $\mathbb{P}(E_{\text{ind}}) \geq 1 - \frac{\delta}{3}$.

Proof. The proof follows a similar reasoning as the proof of [e.g., 136, Proposition F.9] using Freedman's Inequality. We would like to show that each of $E_{\text{ind,rw}}, E_{\text{ind,val}}, E_{\text{ind,prob}}, E_{\text{ind,var}}$ happens with probability $1 - \frac{\delta}{12}$, which would give the lemma statement by a union bound. For brevity, we only show that $\mathbb{P}(E_{\text{ind,var}}) \geq 1 - \frac{\delta}{12}$, and the other probability statements follow from a similar reasoning.

Fix $h \in [H]$, $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, and $p \in [M]$. We will show

$$\mathbb{P}\left(\exists k \in [K]. \left| \frac{1}{n_p^k(s,a)} \sum_{i=1}^{n_p^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right| \right. \\ \left. \geq 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{2H^2 L(n_p^k(s,a))}{n_p^k(s,a)} \right) \leq \frac{\delta}{12MSA}. \tag{C.7}$$

For every $j \in \mathbb{N}_+$, define stopping time k_j as the j-th episode when (s,a) is experienced by player p, if such episode exists; otherwise, k_j is defined as ∞. it suffices to show that

$$\mathbb{P}\left(\exists j \in \mathbb{N}_+. k_j < \infty \wedge \frac{1}{j} \left| \sum_{i=1}^{j} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right| \right. \\ \left. \geq 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(j)}{j}} + \frac{2H^2 L(j)}{j} \right) \leq \frac{\delta}{12MSA}. \tag{C.8}$$

Define $\mathcal{G}_j$ as the σ-algebra generated by all observations up to time step k_j. We have that $\{\mathcal{G}_j\}_{j=0}^\infty$ is a filtration. It can be seen that the sequence

$$\left\{ X_j := (V_p^\star(s_j') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right\}_{j=1}^\infty$$

is a martingale difference sequence adapted to $\{\mathcal{G}_j\}_{j=0}^\infty$; in addition, for every $j, |X_j| \leq H^2$, and $\mathbb{E}\left[X_j^2 \mid \mathcal{G}_{j-1}\right] \leq \mathbb{E}\left[(V_p^\star(s_j') - (\mathbb{P}_p V_p^\star)(s,a))^4 \mid \mathcal{G}_{j-1}\right] \leq H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]$. Therefore, for any $\lambda \geq 0$, $\left\{ Y_j(\lambda) = \exp\left(\lambda \frac{1}{H^2} (\sum_{i=1}^j X_i) - \left((e^\lambda - \lambda - 1) \frac{j}{H^2} \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right) \right) \right\}_{j=0}^\infty$ is a nonnegative supermartingale [54]; by optional sampling theorem, $\mathbb{E}\left[Y_j(\lambda) \mathbf{1}(k_j < \infty)\right] \leq$

$\mathbb{E}\left[Y_0(\lambda)\right] = 1$. As a result, for any fixed thresholds $a, v \geq 0$ [see 54, Theorem 1.6],

$$\mathbb{P}\left(\sum_{i=1}^{j} X_i \geq a \wedge \sum_{i=1}^{j} H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \leq v \wedge k_j < \infty\right) \leq \exp\left(-\frac{a^2}{2v + 2aH^2/3}\right)$$

Now, by the doubling argument of [14, Lemma 2] (observe that $\sum_{i=1}^{j} \mathbb{E}\left[X_i^2 \mid \mathcal{G}_{i-1}\right] \in [0, H^4 j]$), we have that for all $j \in \mathbb{N}_+$:

$$\mathbb{P}\left(k_j < \infty \wedge \left|\frac{1}{j}\sum_{i=1}^{j}(V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right| \right.$$
$$\left. \geq 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(j)}{j}} + \frac{2H^2 L(j)}{j} \right) \leq \ln(4j) \cdot \frac{\delta}{48 j^2 MSA}.$$

A union bound over all $j \in \mathbb{N}_+$ yields Equation (C.8). □

Define event E_{agg} as:

$$E_{\text{agg}} = E_{\text{agg,rw}} \cap E_{\text{agg,val}} \cap E_{\text{agg,prob}} \cap E_{\text{agg,var}}, \tag{C.9}$$

$$E_{\text{agg,rw}} = \left\{ \forall k, h, p, s, a \,.\, \left| \hat{R}^k(s,a) - \bar{R}^k(s,a) \right| \leq \sqrt{\frac{L(n^k(s,a))}{2n^k(s,a)}} \right\}, \tag{C.10}$$

$$\begin{aligned} E_{\text{agg,val}} = \Bigg\{ & \forall k, h, p, s, a \,.\, \left| (\hat{PP}^k V_p^\star - \bar{\mathbb{P}}^k V_p^\star)(s,a) \right|, \\ & \leq 4\sqrt{\frac{\left(\sum_{q=1}^M w_q^k(s,a) \text{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{n^k(s,a)}} + \frac{2HL(n^k(s,a))}{n^k(s,a)} \Bigg\}, \end{aligned} \tag{C.11}$$

$$\begin{aligned} E_{\text{agg,prob}} = \Bigg\{ & \forall k, h, p, s, a, s' \,.\, \left| (\hat{PP}^k - \bar{\mathbb{P}}^k)(s' \mid s,a) \right| \\ & \leq 4\sqrt{\frac{\bar{\mathbb{P}}^k(s' \mid s,a) \cdot L(n^k(s,a))}{n^k(s,a)}} + \frac{2L(n^k(s,a))}{n^k(s,a)} \Bigg\}, \end{aligned} \tag{C.12}$$

$$\begin{aligned} E_{\text{agg,var}} = \Bigg\{ & \forall k, h, p, s, a \,. \\ & \left| \frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_{p_i} V_q^\star)(s,a))^2 - \sum_{q=1}^M w_q^k(s,a) \text{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star] \right|, \\ & \leq 4\sqrt{\frac{H^2 \left(\sum_{q=1}^M w_q^k(s,a) \text{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{n^k(s,a)}} + \frac{2H^2 L(n^k(s,a))}{n^k(s,a)} \Bigg\}, \end{aligned} \tag{C.13}$$

where in Equation (C.13), s_i' denotes the next state for the i-th time some player experiences (s,a). E_{agg} captures the concentration behavior of the aggregate model estimates.

Lemma C.2. $\mathbb{P}(E_{\text{agg}}) \geq 1 - \frac{\delta}{3}$.

Proof. The proof follows a similar reasoning as the proof of [e.g., 136, Proposition F.9] using Freedman's Inequality. We would like to show that each of $E_{\text{agg,rw}}, E_{\text{agg,val}}, E_{\text{agg,prob}}, E_{\text{agg,var}}$ happen with probability $1 - \frac{\delta}{12}$, which would give the lemma statement by a union bound.

For brevity, we show that $\mathbb{P}(E_{\text{agg,var}}) \geq 1 - \frac{\delta}{12}$, and the other probability statements follow from a similar reasoning.

Fix $h \in [H]$, $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ and $p \in [M]$; denote by p_i the identity of the player when (s,a) is experienced for the i-th time for some player. It suffices to show that

$$\mathbb{P}\left(\exists k \in [K]. \left| \frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} \left((V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - \text{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] \right) \right| \right.$$
$$\left. \geq 4\sqrt{\frac{H^2 \left(\sum_{i=1}^{n^k(s,a)} \text{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] \right) L(n^k(s,a))}{(n^k(s,a))^2}} + \frac{2H^2 L(n^k(s,a))}{n^k(s,a)} \right) \leq \frac{\delta}{12MSA}, \tag{C.14}$$

because $\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} \text{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] = \sum_{q=1}^{M} w_q^k(s,a) \text{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]$.

For every $j \in \mathbb{N}_+$, define stopping time k_j as follows: it is the index of the j-th micro-episode when (s,a) is experienced by some player, if such micro-episode exists; and k_j is defined to be ∞ otherwise. With this notation, it suffices to show:

$$\mathbb{P}\left(\exists j \in \mathbb{N}_+ . k_j < \infty \wedge \left| \frac{1}{j} \sum_{i=1}^{j} \left((V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - \text{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] \right) \right| \right.$$
$$\left. \geq 4\sqrt{\frac{H^2 \left(\sum_{i=1}^{j} \text{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] \right) L(j)}{j^2}} + \frac{2H^2 L(j)}{j} \right) \leq \frac{\delta}{12MSA}, \tag{C.15}$$

Define $\mathcal{G}_j$ as the σ-algebra generated by all observations up to micro-episode k_j. We have that $\{\mathcal{G}_j\}_{j=0}^{\infty}$ is a filtration. It can be seen that

$$\left\{ X_j := (V_p^\star(s_j') - (\mathbb{P}_{p_j} V_p^\star)(s,a))^2 - \text{var}_{\mathbb{P}_{p_j}(\cdot|s,a)}[V_p^\star] \right\}_{j=1}^{\infty}$$

is a martingale difference sequence adapted to $\{\mathcal{G}_j\}_{j=0}^{\infty}$; in addition, for every $j, \left|X_j\right| \leq H^2$, and $\mathbb{E}\left[X_j^2 \mid \mathcal{G}_{j-1}\right] \leq \mathbb{E}\left[(V_p^\star(s_j') - (\mathbb{P}_{p_j} V_p^\star)(s,a))^4 \mid \mathcal{G}_{j-1}\right] \leq H^2 \mathrm{var}_{\mathbb{P}_{p_j}(\cdot|s,a)}[V_p^\star]$. Using the same reasoning as in the proof of Lemma C.1 (and observing that $\sum_{i=1}^{j} \mathbb{E}\left[X_i^2 \mid \mathcal{G}_{i-1}\right] \in [0, H^4 j]$), we have that for all $j \in \mathbb{N}_+$:

$$\mathbb{P}\left(k_j < \infty \wedge \left| \frac{1}{j} \sum_{i=1}^{j} \left((V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right) \right| \right.$$
$$\left. \geq 4\sqrt{\frac{H \sum_{i=1}^{j} \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] L(j)}{j^2}} + \frac{2H^2 L(j)}{j} \right) \leq \ln(4j) \cdot \frac{\delta}{48 j^2 MSA}.$$

A union bound over all $j \in \mathbb{N}_+$ implies that Equation (C.15) holds. □

Define

$$E_{\text{sample}} = E_{\text{ind,sample}} \cap E_{\text{agg,sample}},$$
$$E_{\text{agg,sample}} = \left\{ \forall s, a, k \centerdot \bar{n}^k(s,a) \geq N_1 \implies n^k(s,a) \geq \frac{1}{2} \bar{n}^k(s,a) \right\},$$
$$E_{\text{ind,sample}} = \left\{ \forall s, a, k, p \centerdot \bar{n}_p^k(s,a) \geq N_2 \implies n_p^k(s,a) \geq \frac{1}{2} \bar{n}_p^k(s,a) \right\},$$

where we recall from Section C.2 that $N_1 \eqsim M \ln(\frac{SAK}{\delta})$, and $N_2 \eqsim \ln(\frac{MSAK}{\delta})$.

Lemma C.3. $\mathbb{P}(E_{\text{sample}}) \geq 1 - \frac{\delta}{3}$.

Proof. We first show $\mathbb{P}(E_{\text{agg,sample}}) \geq 1 - \frac{\delta}{6}$. Specifically, fix $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, define random variable $X_k = \sum_{p=1}^{M} \left(\mathbf{1}\left((s_{h,p}^k, a_{h,p}^k) = (s,a) \right) - \rho_p^k(s,a) \right)$. Also, define $\mathcal{G}_k$ as the σ-algebra generated by all observations up to episode k. It can be readily seen that $\{X_k\}_{k=1}^{K}$ is a martingale difference sequence adapted to filtration $\{\mathcal{G}_k\}_{k=0}^{K}$. Freedman's inequality (specifically, Lemma 2 of [14]) implies that for every fixed k, with probability

$1 - \frac{\delta}{6K}$,

$$\left|n^k(s,a) - \bar{n}^{k-1}(s,a)\right| \leq 4\sqrt{\bar{n}^{k-1}(s,a) \cdot M \ln\left(\frac{6SAK^2}{\delta}\right)} + 4M \ln\left(\frac{6SAK^2}{\delta}\right), \quad \text{(C.16)}$$

If Equation (C.16) happens, by AM-GM inequality that $\sqrt{\bar{n}^{k-1}(s,a) \cdot M \ln\left(\frac{6SAK^2}{\delta}\right)} \leq \frac{1}{4}\bar{n}^{k-1}(s,a) + 16M \ln\left(\frac{6SAK^2}{\delta}\right)$, we then have

$$\bar{n}^{k-1}(s,a) - n^k(s,a) \leq \frac{1}{4}\bar{n}^{k-1}(s,a) + 20M \ln\left(\frac{6SAK^2}{\delta}\right),$$

implying that

$$n^k(s,a) \geq \frac{3}{4}\bar{n}^{k-1}(s,a) - 20M \ln\left(\frac{6SAK^2}{\delta}\right).$$

Additionally, as $\bar{n}^{k-1}(s,a) \geq \bar{n}^k(s,a) - M$ always holds, we have

$$n^k(s,a) \geq \frac{3}{4}\bar{n}^k(s,a) - 21M \ln\left(\frac{6SAK^2}{\delta}\right).$$

In summary, for any fixed k, with probability $1 - \frac{\delta}{6K}$, if $\bar{n}^k(s,a) \geq N_1 := 84M \ln\left(\frac{6SAK^2}{\delta}\right)$,

$$n^k(s,a) \geq \frac{1}{2}\bar{n}^k(s,a).$$

Taking a union bound over all $k \in [K]$, we have $\mathbb{P}(E_{\text{agg,sample}}) \geq 1 - \frac{\delta}{6}$.

It follows similarly that $\mathbb{P}(E_{\text{ind,sample}}) \geq 1 - \frac{\delta}{6}$; the only difference in the proof is that, we need to take an extra union bound over all $p \in [M]$ - hence an additional factor M within $\ln(\cdot)$ in the definition of N_2. The lemma statement follows from a union bound over these two statements. □

Lemma C.4. $\mathbb{P}(E) \geq 1 - \delta$.

Proof. Follows from Lemmas C.1, C.2, and C.3, along with a union bound. □

C.3.2 Validity of value function bounds

In this section, we show that if the clean event E happens, then for all k and p, the value function estimates $\overline{Q}_p^k$, $\underline{Q}_p^k$, $\overline{V}_p^k$, $\underline{V}_p^k$ are valid upper and lower bounds of the optimal value functions $Q_p^\star$, $V_p^\star$ (Lemma C.7). As a by-product, we also give a general bound on the surplus (Lemma C.6) which will be refined and used in the subsequent regret bound calculations. Before going into the proof of the above two lemmas, we need a technical lemma below (Lemma C.5) that gives necessary concentration results which motivate the bonus constructions; its proof can be found at Section C.3.2.

Lemma C.5. *Fix $p \in [M]$. Suppose E happens, and suppose that for episode k and step h, we have that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \le V^\star(s') \le \overline{V}_p^k(s')$. Then, for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$:*

1.

$$\left|\hat{R}_p^k(s,a) - R_p(s,a)\right| \le b_{\mathrm{rw}}\left(n_p^k(s,a), 0\right), \tag{C.17}$$

$$\left|\hat{R}^k(s,a) - R_p(s,a)\right| \le b_{\mathrm{rw}}\left(n^k(s,a), \epsilon\right). \tag{C.18}$$

2.

$$\left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(V_p^\star)(s,a)\right| \le b_{\mathrm{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right), \tag{C.19}$$

$$\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_p^\star)(s,a)\right| \le b_{\mathrm{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right). \tag{C.20}$$

3. For any $V_1, V_2 : \mathcal{S}_{h+1} \to \mathbb{R}$ such that $\overline{V}_p^k \le V_1 \le V_2 \le \underline{V}_p^k$,

$$\left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \le b_{\mathrm{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right), \tag{C.21}$$

$$\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \le b_{\mathrm{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right). \tag{C.22}$$

Lemma C.6. *If event E happens, and suppose that for episode k and step h, we have*

that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \le V_p^\star(s') \le \overline{V}_p^k(s')$. *Then, for* $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\overline{Q}_p^k(s,a) - \left(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)\right) \in \left[0, (H-h+1) \wedge 2\textit{ind-}b_p^k(s,a) \wedge 2\textit{agg-}b_p^k(s,a)\right], \tag{C.23}$$

and

$$\left(R_p(s,a) + (\mathbb{P}_p \underline{V}_p^k)(s,a)\right) - \underline{Q}_p^k(s,a) \in \left[0, (H-h+1) \wedge 2\textit{ind-}b_p^k(s,a) \wedge 2\textit{agg-}b_p^k(s,a)\right], \tag{C.24}$$

where we recall that

$$\begin{aligned}
\textit{ind-}b_p^k(s,a) = b_{\mathrm{rw}}\left(n_p^k(s,a), 0\right) + b_{\mathrm{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right) \\
+ b_{\mathrm{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right), \\
\textit{agg-}b_p^k(s,a) = b_{\mathrm{rw}}\left(n^k(s,a), \epsilon\right) + b_{\mathrm{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right) \\
+ b_{\mathrm{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right).
\end{aligned}$$

Proof. We only show Equation (C.23) for brevity; Equation (C.24) follows from an exact symmetrical reasoning.

Recall that $\overline{Q}_p^k(s,a) = \min\left(\overline{\text{ind-}Q}_p^k(s,a), \overline{\text{agg-}Q}_p^k(s,a), H\right)$. We compare each term in the $\min(\cdot)$ operator with $(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a))$:

- For $\overline{\text{ind-}Q}_p^k(s,a)$, using Lemma C.5 and our assumption on $\overline{V}_p^k$ and $\underline{V}_p^k$ on $\mathcal{S}_{h+1}$, we have:

$$\begin{aligned}
&\overline{\text{ind-}Q}_p^k(s,a) - \left(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)\right) \\
&= (\hat{R}_p^k - R_p)(s,a) + b_{\text{rw}}\left(n_p^k(s,a), 0\right) \\
&\quad + ((\hat{\mathbb{P}}_p^k - \mathbb{P}_p)V_p^\star)(s,a) + b_{\text{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right) \\
&\quad + (\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(\overline{V}_p^k - V_p^\star))(s,a) + b_{\text{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right) \\
&\in [0, 2\text{ind-}b_p^k(s,a)].
\end{aligned}$$

- For $\overline{\text{agg-}Q}_p^k(s,a)$, using Lemma C.5 and our assumptions on $\overline{V}_p^k$ and $\underline{V}_p^k$ over $\mathcal{S}_{h+1}$, we have:

$$\begin{aligned}
&\overline{\text{agg-}Q}_p^k(s,a) - \left(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)\right) \\
&= (\hat{R}_p^k - R_p)(s,a) + b_{\text{rw}}\left(n^k(s,a), \epsilon\right) \\
&\quad + ((\hat{\mathbb{P}}^k - \mathbb{P}_p)V_p^\star)(s,a) + b_{\text{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right) \\
&\quad + ((\hat{\mathbb{P}}^k - \mathbb{P}_p)(\overline{V}_p^k - V_p^\star))(s,a) + b_{\text{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right) \\
&\in [0, 2\text{agg-}b_p^k(s,a)],
\end{aligned}$$

- For $H - h + 1$, we have:

$$(H - h + 1) - (R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)) \in [0, H - h + 1],$$

  where we use the observation that $R(s,a) \in [0,1]$, and $(\mathbb{P}_p \overline{V}_p^k)(s,a) \in [0, H-h]$, and their sum is in $[0, H]$.

Combining the above three establishes that

$$\overline{Q}_p^k(s,a) - (R(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)) \in \left[0, (H-h+1) \wedge 2\text{ind-}b_p^k(s,a) \wedge 2\text{agg-}b_p^k(s,a)\right].$$

□

Lemma C.7. *Under event E, for every $k \in [K]$, and every $p \in [M]$, and for every $h \in [H]$, For all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,*

$$\underline{Q}_p^k(s,a) \leq Q_p^{\pi^k}(s,a) \leq Q_p^\star(s,a) \leq \overline{Q}_p^k(s,a), \tag{C.25}$$

and

$$\underline{V}_p^k(s) \leq V_p^{\pi^k}(s) \leq V_p^\star(s) \leq \overline{V}_p^k(s), \tag{C.26}$$

Proof. The proof of this lemma extends [136, Proposition F.1] to our multitask setting.

For every k and p, we show the above holds for all layers $h \in [H]$ and every $(s,a) \in \mathcal{S}_h \times \mathcal{A}$; to this end, we do backward induction on layer h.

Base case:

For layer $h = H+1$, we have $\underline{V}_p^k(\perp) = V_p^{\pi^k}(\perp) = V_p^\star(\perp) = \overline{V}_p^k(\perp) = 0$.

Inductive case:

By our inductive hypothesis, for layer $h+1$ and every $s \in \mathcal{S}_{h+1}$,

$$\underline{V}_p^k(s) \leq V_p^{\pi^k}(s) \leq V_p^\star(s) \leq \overline{V}_p^k(s).$$

We will show that Equations (C.25) and (C.26) holds holds for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$.

We first show Equation (C.25). First, $Q_p^{\pi^k}(s,a) \leq Q_p^\star(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ is trivial.

To show $Q_p^\star(s,a) \leq \overline{Q}_p^k(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, by Lemma C.6 and inductive hypothesis, we have:

$$Q_p^\star(s,a) = R_p(s,a) + (\mathbb{P}_p V_p^\star)(s,a) \leq R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a) \leq \overline{Q}_p^k(s,a).$$

Likewise, we show $Q_p^{\pi^k}(s,a) \geq \underline{Q}_p^k(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, using Lemma C.6 and inductive hypothesis:

$$Q_p^{\pi^k}(s,a) = R_p(s,a) + (\mathbb{P}_p V_p^{\pi^k})(s,a) \geq R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a) \geq \underline{Q}_p^k(s,a).$$

This completes the proof of Equation (C.25) for layer h.

We now show Equation (C.26) for layer h. Again $V_p^{\pi^k}(s) \leq V_p^\star(s)$ for all $s \in \mathcal{S}_h$ is trivial.

To show $V_p^\star(s) \leq \overline{V}_p^k(s)$ for all $s \in \mathcal{S}_h$, observe that

$$V_p^\star(s) = \max_{a \in \mathcal{A}} Q_p^\star(s,a) \leq \max_{a \in \mathcal{A}} \overline{Q}_p^k(s,a) = \overline{V}_p^k(s).$$

To show $V_p^{\pi^k}(s) \geq \underline{V}_p^k(s)$ for all $s \in \mathcal{S}_h$, observe that

$$V_p^{\pi^k}(s) = Q_p^{\pi^k}(s, \pi^k(p)(s)) \geq \underline{Q}_p^k(s, \pi^k(p)(s)) = \underline{V}_p^k(s).$$

This completes the induction. □

Proof of Lemma C.5

Proof of Lemma C.5. Equations (C.17), (C.19), and (C.21) essentially follow the same reasoning as in [136]; we still include their proofs for completeness.

Equations (C.18), (C.20), and (C.22) are new, and require a more involved analysis. Our proof also relies on a technical lemma, namely Lemma C.8; we defer its statement and proof to the end of this subsection.

1. Equation (C.17) follows directly from the definition of $E_{\text{ind,rw}}$. Equation (C.18) follows from the definition of $E_{\text{agg,rw}}$, and the fact that $\left|\bar{R}^k(s,a) - R_p(s,a)\right| \leq \epsilon$.

2. We prove Equation (C.19) as follows:

$$
\begin{aligned}
&\left|(\hat{\mathbb{P}}_p^k V^\star - \mathbb{P}_p V_p^\star)(s,a)\right| \\
\leq &\mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\leq &\mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}[V^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\leq &\mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}[\overline{V}_p^k] \; L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|V_p^\star - \overline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} \; L(n_p^k(s,a))}{n_p^k(s,a)}}\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. + \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\leq &\mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}[\overline{V}_p^k] \; L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} \; L(n_p^k(s,a))}{n_p^k(s,a)}}\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. + \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\leq &b_{\mathrm{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right),
\end{aligned}
$$

where the first inequality is from the definition of $E_{\mathrm{ind,val}}$; the second inequality is from Equation (C.27) of Lemma C.8; the third inequality is from Lemma C.15; the fourth inequality is from our assumption that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \leq V^\star(s') \leq \overline{V}_p^k(s')$, and thus $\left|(V_p^\star - \underline{V}_p^k)(s')\right| \leq \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all s' in the support of $\hat{\mathbb{P}}_p^k(\cdot \mid s,a)$.

We prove Equation (C.20) as follows:

$$
\begin{aligned}
&\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_p^\star)(s,a)\right| \\
\leq &\epsilon + \left|(\hat{\mathbb{P}}^k - \bar{\mathbb{P}}_p^k)(V_p^\star)(s,a)\right| \\
\leq &\epsilon + \mathcal{O}\left(\sqrt{\frac{\left(\sum_{q=1}^M w_q^k(s,a)\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{n^k(s,a)}} + \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq &\epsilon + \mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[V_p^\star]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{L(n^k(s,a))}{n^k(s,a)} \cdot \epsilon H} + \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq &2\epsilon + \mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[\overline{V}_p^k]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - V_p^\star\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq &2\epsilon + \mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[\overline{V}_p^k]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq &b_{\mathrm{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right),
\end{aligned}
$$

where the first inequality is from the observation that $\|\bar{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_p(\cdot \mid s,a)\|_1 \leq \frac{\epsilon}{H}$ and Lemma C.16; the second inequality is from the definition of $E_{\mathrm{agg,val}}$; the third inequality is from Equation (C.28) of Lemma C.8; the fourth inequality is from Lemma C.15 and the observation that for constant $c > 0$, $c\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)} \cdot \epsilon H} \leq \epsilon + \frac{c^2}{4}\frac{L(n^k(s,a))}{n^k(s,a)}$ by AM-GM inequality; the fifth inequality is from our assumption that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \leq V^\star(s') \leq \overline{V}_p^k(s')$, and thus $\left|(V_p^\star - \underline{V}_p^k)(s')\right| \leq \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all s' in the

support of $\hat{\mathbb{P}}^k(\cdot \mid s,a)$.

3. We prove Equation (C.21) as follows:

$$
\begin{aligned}
&\left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \\
\leq &\sum_{s' \in \mathcal{S}_{h+1}} \left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(s' \mid s,a)\right| \cdot (V_2 - V_1)(s') \\
\leq &\mathcal{O}\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n_p^k(s,a)) \cdot \mathbb{P}_p(s' \mid s,a)}{n_p^k(s,a)}} + \frac{L(n_p^k(s,a))}{n_p^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\leq &\mathcal{O}\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n_p^k(s,a)) \cdot \hat{\mathbb{P}}_p^k(s' \mid s,a)}{n_p^k(s,a)}} + \frac{L(n_p^k(s,a))}{n_p^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\leq &\mathcal{O}\left(\sum_{s' \in \mathcal{S}_{h+1}} \sqrt{\hat{\mathbb{P}}_p^k(s' \mid s,a)}(\overline{V}_p^k - \underline{V}_p^k)(s') \cdot \sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}} + \sum_{s' \in \mathcal{S}_{h+1}} \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\leq &\mathcal{O}\left(\sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} \, L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\leq &b_{\mathrm{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right),
\end{aligned}
$$

where the first inequality is from the elementary fact that $\left|\sum_{i=1}^n a_i\right| \leq \sum_{i=1}^n |a_i|$; the second inequality is from the definition of $E_{\mathrm{ind,prob}}$; the third inequality is from the definition of $E_{\mathrm{ind,prob}}$ and Lemma C.17; the fourth inequality is by algebra and $0 \leq (V_2 - V_1)(s') \leq \min(H, (\overline{V}_p^k - \underline{V}_p^k)(s'))$ for all $s' \in \mathcal{S}_{h+1}$; the fifth inequality is by Cauchy-Schwarz.

We now prove Equation (C.22):

$$
\begin{aligned}
&\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \\
\leq &\left|(\bar{\mathbb{P}}^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| + \left|(\hat{\mathbb{P}}^k - \bar{\mathbb{P}}^k)(V_2 - V_1)(s,a)\right| \\
\leq &\epsilon + \sum_{s' \in \mathcal{S}_{h+1}} \left|(\hat{\mathbb{P}}^k - \bar{\mathbb{P}}^k)(s' \mid s,a)\right| \cdot (V_2 - V_1)(s') \\
\leq &\epsilon + \mathcal{O}\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n^k(s,a)) \cdot \bar{\mathbb{P}}^k(s' \mid s,a)}{n^k(s,a)}} + \frac{L(n^k(s,a))}{n^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\leq &\epsilon + \mathcal{O}\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n^k(s,a)) \cdot \hat{\mathbb{P}}^k(s' \mid s,a)}{n^k(s,a)}} + \frac{L(n^k(s,a))}{n^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\leq &\epsilon + \mathcal{O}\left(\sum_{s' \in \mathcal{S}_{h+1}} \sqrt{\hat{\mathbb{P}}^k(s' \mid s,a)}(\overline{V}_p^k - \underline{V}_p^k)(s') \cdot \sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}} + \sum_{s' \in \mathcal{S}_{h+1}} \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq &\epsilon + \mathcal{O}\left(\sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} \, L(n^k(s,a))}{n^k(s,a)}} + \frac{SHL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq &b_{\mathrm{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right),
\end{aligned}
$$

where the first inequality is triangle inequality; the second inequality is from the elementary fact that $\left|\sum_{i=1}^n a_i\right| \leq \sum_{i=1}^n |a_i|$, along with $\|\bar{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_p(\cdot \mid s,a)\|_1 \leq \frac{\epsilon}{H}$ and Lemma C.16; the third inequality is from the definition of $E_{\mathrm{agg,prob}}$; the fourth inequality is from the definition of $E_{\mathrm{agg,prob}}$ and Lemma C.17; the fifth inequality is by algebra and $0 \leq (V_2 - V_1)(s') \leq \min(H, (\overline{V}_p^k - \underline{V}_p^k)(s'))$ for all $s' \in \mathcal{S}_{h+1}$; the last inequality is by Cauchy-Schwarz.

□

Lemma C.5 relies on the following technical lemma on the concentrations of the conditional variances. Specifically, Equation (C.27) is well-known [see, e.g., 7, 109];

Equations (C.28) and (C.29) are new, and allow for heterogeneous data aggregation in the multi-task RL setting. We still include the proof of Equation (C.27) here, as it helps illustrate our ideas for proving the two new inequalities.

Lemma C.8. *If event E happens, then for any s, a, k, p, we have:*

1.

$$\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right|\lesssim H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}, \tag{C.27}$$

2.

$$\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\sum_{q=1}^M w_q^k(s,a)\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}\left[V_p^\star\right]}\right|\lesssim \sqrt{H\epsilon}+H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}, \tag{C.28}$$

and

$$\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right|\lesssim \sqrt{H\epsilon}+H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}, \tag{C.29}$$

Proof. 1. By the definition of E, we have

$$\begin{aligned}&\left|\frac{1}{n_p^k(s,a)}\sum_{i=1}^{n_p^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2-\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right|\\ \lesssim&\sqrt{\frac{H^2\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{H^2L(n_p^k(s,a))}{n_p^k(s,a)};\end{aligned}$$

this, when combined with Lemma C.17, implies that

$$\left|\sqrt{\frac{1}{n_p^k(s,a)}\sum_{i=1}^{n_p^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]}\right|\le H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}. \tag{C.30}$$

Now, observe that

$$\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right] = \frac{1}{n_p^k(s,a)} \sum_{i=1}^{n_p^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - ((\hat{\mathbb{P}}_p^k V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a))^2.$$

Recall that by the definition of event E, we have

$$\begin{aligned}\left|(\hat{\mathbb{P}}_p^k V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| &\le H \wedge \left(\sqrt{\frac{H^2 L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\ &\le 2H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}},\end{aligned}$$

where the second inequality uses Lemma C.18. Using the elementary fact that $|A - B| \le C \Rightarrow \sqrt{A} \le \sqrt{B} + \sqrt{C}$, we get that

$$\begin{aligned}&\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2}\right| \\ &\le \left|(\hat{\mathbb{P}}_p^k V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| \lesssim H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}.\end{aligned} \tag{C.31}$$

Combining Equations (C.30) and (C.31), using algebra, we get

$$\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right| \lesssim H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}},$$

establishing Equation (C.27).

2. We first show Equation (C.28). By the definition of E, we have

$$\left|\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_{p_i}V_p^\star)(s,a))^2-\sum_{p=1}^{M}w_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right|$$
$$\lesssim\sqrt{\frac{H^2\left(\sum_{p=1}^{M}w_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right)L(n^k(s,a))}{n^k(s,a)}}+\frac{H^2L(n^k(s,a))}{n^k(s,a)},$$

this, combined with Lemma C.17, implies that

$$\left|\sqrt{\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_{p_i}V_p^\star)(s,a))^2}-\sqrt{\sum_{p=1}^{M}w_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]}\right| \tag{C.32}$$
$$\lesssim H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}.$$

For the first term on the left hand side, observe that for each i, $|(\mathbb{P}_{p_i}V_p^\star)(s,a)-(\mathbb{P}_pV_p^\star)(s,a)|\le H\frac{\epsilon}{H}=\epsilon$, we therefore have

$$\left|(V_p^\star(s_i')-(\mathbb{P}_{p_i}V_p^\star)(s,a))^2-(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2\right|\le 2H\epsilon$$

by $2H$-Lipschitzness of function $f(x)=x^2$ on $[-H,H]$. By averaging over all i's and taking square root, we have

$$\left|\sqrt{\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_{p_i}V_p^\star)(s,a))^2}-\sqrt{\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2}\right|$$
$$\lesssim\sqrt{H\epsilon}. \tag{C.33}$$

Furthermore,

$$\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right] = \frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2-((\hat{\mathbb{P}}^kV_p^\star)(s,a)-(\mathbb{P}_pV_p^\star)(s,a))^2,$$

and

$$\left|(\hat{\mathbb{P}}^kV_p^\star)(s,a)-(\mathbb{P}_pV_p^\star)(s,a)\right| \lesssim \epsilon + H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}$$

Together with our assumption that $\epsilon \leq 2H$ (which implies that $\epsilon \lesssim \sqrt{H\epsilon}$), this gives

$$\begin{aligned}&\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2}\right| \\ &\lesssim \sqrt{H\epsilon}+H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}.\end{aligned} \tag{C.34}$$

Equation (C.28) is a direct consequence of Equations (C.32), (C.33) and (C.34) along with algebra.

We now show Equation (C.29) using Equation (C.28). By Lemma C.16, for every q, $\left|\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}\left[V_p^\star\right]-\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]\right| \leq 3H^2\cdot\frac{\epsilon}{H}=3H\epsilon$. Therefore,

$$\left|\sum_{q=1}^{M}w_q^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]-\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]\right| \leq 3H^2\cdot\frac{\epsilon}{H}=3H\epsilon,$$

and

$$\left|\sqrt{\sum_{q=1}^{M}w_q^k(s,a)\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right| \lesssim \sqrt{H\epsilon}$$

This, together with Equation (C.28), implies

$$\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right| \lesssim \sqrt{H\epsilon}+H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}},$$

establishing Equation (C.29).

□

C.3.3 Simplifying the surplus bounds

In this section, we show a distribution-dependent bound on the surplus terms, namely Lemma C.11, which is key to establishing our regret bound. It can be seen as an extension of Proposition B.4 of [136] to our multitask setting using the MULTI-TASK-EULER algorithm, under the ϵ-dissimilarity assumption. Before we present Lemma C.11 (Section C.3.3), we first show and prove two auxiliary lemmas, Lemma C.9 and Lemma C.10.

Lemma C.9 (Bounds on $\overline{V}_p^k - \underline{V}_p^k$, generalization of [136], Lemma F.8)**.** *If E happens, then for all $p \in [M]$, $k \in [K]$, $h \in [H+1]$ and $s \in \mathcal{S}_h$,*

$$(\overline{V}_p^k - \underline{V}_p^k)(s) \leq 4\mathbb{E}\left[\sum_{t=h}^{H}\left(H \wedge \textit{ind-}b_p^k(s_t,a_t) \wedge \textit{agg-}b_p^k(s_t,a_t)\right) \mid s_h = s, \pi^k(p), \mathcal{M}_p\right]; \tag{C.35}$$

consequently,

$$(\overline{V}_p^k - \underline{V}_p^k)(s) \lesssim H\sum_{t=h}^{H}\mathbb{E}\left[\left(1 \wedge \sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}\right) \mid s_h = s, \pi^k(p), \mathcal{M}_p\right]. \tag{C.36}$$

Proof. First, Lemmas C.7 and C.6 together imply that if E holds, Equations (C.23) and (C.24) holds for all p, k, s, a. Under this premise, we show Equation (C.35) by backward induction.

Base case:

for $h = H+1$, we have that LHS is $(\overline{V}_p^k - \underline{V}_p^k)(\perp) = 0$ which is equal to the RHS.

inductive case:

Suppose Equation (C.35) holds for all $s \in \mathcal{S}_{h+1}$. Now consider $s \in \mathcal{S}_h$. By the definitions of $\overline{V}_p^k$ and $\underline{V}_p^k$,

$$
\begin{aligned}
&(\overline{V}_p^k - \underline{V}_p^k)(s) \\
=&\overline{Q}_p^k(s, \pi_p^k(s)) - \underline{Q}_p^k(s, \pi_p^k(s)) \\
\leq&(\mathbb{P}_p(\overline{V}_p^k - \underline{V}_p^k))(s, \pi_p^k(s)) + 4(H \wedge \text{ind-}b_p^k(s, \pi_p^k(s)) \wedge \text{agg-}b_p^k(s, \pi_p^k(s))) \\
=&\mathbb{E}\left[4\min(H, \text{ind-}b_p^k(s,a), \text{agg-}b_p^k(s,a)) + (\overline{V}_p^k - \underline{V}_p^k)(s_{h+1}) \mid s_h = s, \pi_p^k, \mathcal{M}_p\right] \\
\leq&\mathbb{E}\Bigg[4(H \wedge \text{ind-}b_p^k(s,a) \wedge \text{agg-}b_p^k(s,a)) + \\
&\qquad \mathbb{E}\left[\sum_{t=h+1}^{H}\left(H \wedge 2\text{ind-}b_p^k(s_t,a_t) \wedge 2\text{agg-}b_p^k(s_t,a_t)\right) \mid s_{h+1}\right] \mid s_h = s, \pi_p^k, \mathcal{M}_p\Bigg] \\
\leq&4\mathbb{E}\left[\sum_{t=h}^{H}\left(H \wedge \text{ind-}b_p^k(s_t,a_t) \wedge \text{agg-}b_p^k(s_t,a_t)\right) \mid s_h = s, \pi_p^k, \mathcal{M}_p\right],
\end{aligned}
$$

where the first inequality is from Equations (C.23) and (C.24) for (s, a) and player p at episode k, and the second inequality is from the inductive hypothesis; the third inequality is by algebra. This completes the induction.

We now show Equation (C.36). By the definition of ind-$b_p^k(s,a)$ and algebra,

$$\begin{aligned}
&\text{ind-}b_p^k(s,a)\\
\lesssim&\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right]\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}+\\
&\qquad\qquad\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{HSL(n_p^k(s,a))}{n_p^k(s,a)}\\
\lesssim&H\sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}+\frac{HSL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)},
\end{aligned}$$

where the second inequality uses $\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right]\le H^2$ and $\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k}\le H^2$.

As a consequence, using Lemma C.18,

$$\begin{aligned}
H\wedge\text{ind-}b_p^k(s_t,a_t)\wedge\text{agg-}b_p^k(s_t,a_t)\lesssim&H\wedge\left(H\sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}+\frac{HSL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}\right)\\
\lesssim&H\left(1\wedge\sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}\right).
\end{aligned}$$

□

Lemma C.10. *If E happens, we have the following statements holding for all p,k,s,a:*

1. For two terms that appear in ind-$b_p^k(s,a)$, they are bounded respectively as:

$$\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\lesssim\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}+\frac{H^2SL(n_p^k(s,a))}{n_p^k(s,a)}\tag{C.37}$$

$$\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \lesssim \sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{H\sqrt{S}L(n_p^k(s,a))}{n_p^k(s,a)} \tag{C.38}$$

2. For two terms that appear in agg-$b_p^k(s,a)$, they are bounded respectively as:

$$\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} \lesssim 2\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{H^2 S L(n_p^k(s,a))}{n_p^k(s,a)} + H\epsilon \tag{C.39}$$

$$\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n^k(s,a))}{n^k(s,a)}} \tag{C.40}$$

$$\lesssim \sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n^k(s,a))}{n^k(s,a)}} + \frac{H\sqrt{S}L(n^k(s,a))}{n^k(s,a)} + \sqrt{\frac{H\epsilon L(n^k(s,a))}{n^k(s,a)}} \tag{C.41}$$

Proof. First, Lemmas C.7 and C.6 together imply that if E happens, the value function upper and lower bounds are valid. Conditioned on E happening, we prove the two items respectively.

1. For Equation (C.37), using the definition of $E_{\text{ind,prob}}$ and AM-GM inequality, when E happens, we have for all p, k, s, a, s',

$$\hat{\mathbb{P}}_p^k(s' \mid s,a) \lesssim \mathbb{P}_p(s' \mid s,a) + \frac{L(n_p^k(s,a))}{n_p^k(s,a)}. \tag{C.42}$$

This implies that

$$
\begin{aligned}
&\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} \\
=& \sum_{s' \in \mathcal{S}_{h+1}} \hat{\mathbb{P}}_p^k(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 \\
\lesssim& \sum_{s' \in \mathcal{S}_{h+1}} \mathbb{P}_p^k(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 + \sum_{s' \in \mathcal{S}_{h+1}} \frac{L(n_p^k(s,a))}{n_p^k(s,a)} \cdot H^2 \\
\lesssim& \|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{SH^2 L(n_p^k(s,a))}{n_p^k(s,a)},
\end{aligned}
$$

where the first inequality is from Equation (C.42), and the fact that $\overline{V}_p^k(s') - \underline{V}_p^k(s') \in [0, H]$ for any $s' \in \mathcal{S}_{h+1}$; the second inequality is by algebra.

For Equation (C.38), we have:

$$
\begin{aligned}
&\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim& \sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim& \sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)} \\
\lesssim& \sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)}
\end{aligned}
$$

where the first inequality is from Lemma C.15 and the observation that when E happens, $\left|(\overline{V}_p^k - V_p^\star)(s')\right| \le \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all $s' \in \mathcal{S}_{h+1}$; the second inequality is from Equation (C.27) of Lemma C.8 and Equation (C.37); the third inequality again uses Lemma C.15 and the observation that when E happens, $\left|(V_p^\star - V_p^{\pi^k})(s')\right| \le \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all $s' \in \mathcal{S}_{h+1}$.

2. For Equation (C.39), using the definition of $E_{\text{agg,prob}}$ and AM-GM inequality, when E

happens, we have for all p, k, s, a, s',

$$\hat{\mathbb{P}}^k(s' \mid s, a) \lesssim \bar{\mathbb{P}}^k(s' \mid s, a) + \frac{L(n^k(s,a))}{n^k(s,a)}. \tag{C.43}$$

This implies that

$$\begin{aligned}
&\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} \\
=& \sum_{s' \in \mathcal{S}_{h+1}} \hat{\mathbb{P}}^k(s' \mid s, a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 \\
\lesssim& 2 \sum_{s' \in \mathcal{S}_{h+1}} \bar{\mathbb{P}}^k(s' \mid s, a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 + \sum_{s' \in \mathcal{S}_{h+1}} \frac{L(n_p^k(s,a))}{n_p^k(s,a)} \cdot H^2 \\
\lesssim& 2 \sum_{s' \in \mathcal{S}_{h+1}} \mathbb{P}_p(s' \mid s, a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 + \epsilon H + \frac{SH^2 L(n_p^k(s,a))}{n_p^k(s,a)} \\
\lesssim& \|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{SH^2 L(n_p^k(s,a))}{n_p^k(s,a)} + \epsilon H,
\end{aligned}$$

where the first inequality is from Equation (C.43) and the fact that $\overline{V}_p^k(s') - \underline{V}_p^k(s') \in [0, H]$ for any $s' \in \mathcal{S}_{h+1}$; the second inequality is from the observation that $\|\mathbb{P}_p(\cdot \mid s, a) - \bar{\mathbb{P}}^k(\cdot \mid s, a)\|_1 \leq \frac{\epsilon}{H}$; the third inequality is by algebra.

For Equation (C.41), we have:

$$
\begin{aligned}
&\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim &\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim &\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \\
&\qquad \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)} + \sqrt{\frac{H\epsilon L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim &\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \\
&\qquad \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)} + \sqrt{\frac{H\epsilon L(n_p^k(s,a))}{n_p^k(s,a)}},
\end{aligned}
$$

where the first inequality is from Lemma C.15 and the observation that when E happens, $\left|(\overline{V}_p^k - V_p^\star)(s')\right| \le \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for $s' \in \mathcal{S}_{h+1}$; the second inequality uses Equation (C.29) of Lemma C.8 and Equation (C.39); the third inequality is from Lemma C.15 and the observation that when E happens, $\left|(\overline{V}_p^\star - V_p^{\pi^k})(s')\right| \le \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for $s' \in \mathcal{S}_{h+1}$.

□

Distribution-dependent bound on the surplus terms

Lemma C.11 (Surplus bound)**.** *If E happens, then for all p, k, s, a:*

$$
E_p^k(s,a) \lesssim B_p^{k,\mathrm{lead}}(s,a) + \mathbb{E}\left[\sum_{t=h}^{H} B_p^{k,\mathrm{fut}}(s_t,a_t) \mid (s_h,a_h) = (s,a), \pi^k(p), \mathcal{M}_p\right],
$$

where

$$B_p^{k,\text{lead}}(s,a) = H \wedge \left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{\left(1 + \text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) L(n^k(s,a))}{n^k(s,a)}} \right) \right)$$
$$\wedge \mathcal{O}\left(\sqrt{\frac{\left(1 + \text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) L(n_p^k(s,a))}{n_p^k(s,a)}} \right),$$
$$B_p^{k,\text{fut}}(s,a) = H^3 \wedge \mathcal{O}\left(\frac{H^3 S L(n_p^k(s,a))}{n_p^k(s,a)} \right).$$

Proof of Lemma C.11. First, Lemmas C.7 and C.6 together imply that if E holds, for all p, k, s, a, $E_p^k(s,a) \leq 2\left(H \wedge \text{ind-}b_p^k(s,a) \wedge \text{agg-}b_p^k(s,a)\right)$. We now bound ind-$b_p^k(s,a)$ and agg-$b_p^k(s,a)$ respectively.

Bounding ind-$b_p^k(s,a)$:

$$
\begin{aligned}
&\text{ind-}b_p^k(s,a)\\
=&\mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}[\overline{V}_p^k]\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}+\right.\\
&\qquad\left.\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)\\
\le&\mathcal{O}\left(\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}+\right.\\
&\qquad\left.\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)\\
\le&\mathcal{O}\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\right.\\
&\qquad\left.\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)\\
\le&\mathcal{O}\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n_p^k(s,a))}{n_p^k(s,a)}}+\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)
\end{aligned}
$$

where the first inequality is by expanding the definition of ind-$b_p^k(s,a)$ and algebra; the second inequality is from Equations Equation (C.37) and (C.38) of Lemma C.10, along with algebra; the third inequality is by the basic fact that $\sqrt{A}+\sqrt{B}\lesssim\sqrt{A+B}$; the fourth inequality is by AM-GM inequality.

Bounding agg-$b_p^k(s,a)$:

$$
\begin{aligned}
&\text{agg-}b_p^k(s,a)\\
\lesssim& 4\epsilon + \mathcal{O}\Bigg(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[\overline{V}_p^k]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}} + \\
&\qquad\qquad \sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \frac{SHL(n^k(s,a))}{n^k(s,a)}\Bigg)\\
\lesssim& 5\epsilon + \mathcal{O}\Bigg(\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}} + \\
&\qquad\qquad \sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \frac{SHL(n^k(s,a))}{n^k(s,a)}\Bigg)\\
\lesssim& 5\epsilon + \mathcal{O}\Bigg(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n^k(s,a))}{n^k(s,a)}} + \\
&\qquad\qquad \sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \frac{SHL(n^k(s,a))}{n^k(s,a)}\Bigg)\\
\leq& 5\epsilon + \mathcal{O}\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n^k(s,a))}{n^k(s,a)}} + \|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{SHL(n^k(s,a))}{n^k(s,a)}\right)
\end{aligned}
$$

where the first inequality is by expanding the definition of agg-$b_p^k(s,a)$ and algebra; the second inequality is from Equations (C.41) and Equation (C.39) of Lemma C.10, along with the observation that $\sqrt{\frac{S\epsilon HL(n^k(s,a))}{n^k(s,a)}} \leq \frac{SHL(n^k(s,a))}{n^k(s,a)} + \epsilon$ by AM-GM inequality; the third inequality is by the basic fact that $\sqrt{A} + \sqrt{B} \lesssim \sqrt{A+B}$; the fourth inequality is from

AM-GM inequality.

Combining the above upper bounds, and using the observation that $\frac{L(n^k(s,a))}{n^k(s,a)} \leq \frac{L(n_p^k(s,a))}{n_p^k(s,a)}$, we get

$$\begin{aligned}
&\text{ind-}b_p^k(s,a) \wedge \text{agg-}b_p^k(s,a) \wedge H \\
&\leq \mathcal{O}\left(\sqrt{\frac{\left(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n_p^k(s,a))}{n_p^k(s,a)}}\right) \wedge \\
&\left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{\left(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n^k(s,a))}{n^k(s,a)}}\right)\right) \wedge H \\
&+ \mathcal{O}\left(\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \left(\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)} \wedge H\right)\right) \\
&\leq B^{k,\text{lead}}(s,a) + \mathcal{O}\left(\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \left(\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)} \wedge H\right)\right).
\end{aligned}$$

We now show that

$$\begin{aligned}
&\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \left(\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)} \wedge H\right) \\
&\lesssim \mathbb{E}\left[\sum_{t=h}^{H} B^{k,\text{fut}}(s_t,a_t) \mid (s_h,a_h) = (s,a), \pi^k(p), \mathcal{M}_p\right],
\end{aligned} \tag{C.44}$$

which will conclude the proof. To this end, we simplify the left hand side of Equation (C.44)

using Lemma C.9:

$$
\begin{aligned}
&\|\overline{V}_p^k - \underline{V}_p^k\|_{\mathbb{P}_p(\cdot|s,a)}^2 + \left(\frac{SHL(n^k(s,a))}{n^k(s,a)} \wedge H\right) \\
\lesssim &\mathbb{E}\left[\left(H\sum_{t=h+1}^{H}\mathbb{E}\left[\left(1\wedge\sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}\right) \mid s_{h+1}\right]\right)^2 \mid (s_h,a_h)=(s,a),\pi^k(p),\mathcal{M}_p\right] + \\
&\qquad\qquad \left(\frac{SHL(n^k(s,a))}{n^k(s,a)} \wedge H\right) \\
\lesssim &H^3\mathbb{E}\left[\sum_{t=h+1}^{H}\mathbb{E}\left[\left(1\wedge\sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}\right)^2 \mid s_{h+1}\right] \mid (s_h,a_h)=(s,a),\pi^k(p),\mathcal{M}_p\right] + \\
&\qquad\qquad \left(\frac{SHL(n^k(s,a))}{n^k(s,a)} \wedge H\right) \\
\lesssim &\mathbb{E}\left[\sum_{t=h}^{H} H^3 \wedge \frac{H^3 SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)} \mid (s_h,a_h)=(s,a),\pi^k(p),\mathcal{M}_p\right] \\
\lesssim &\mathbb{E}\left[\sum_{t=h}^{H} B^{k,\mathrm{fut}}(s_t,a_t) \mid (s_h,a_h)=(s,a),\pi^k(p),\mathcal{M}_p\right],
\end{aligned}
$$

where the first inequality is from Equation (C.36) of Lemma C.9; the second inequality is by Cauchy-Schwarz and $\mathbb{E}[X]^2 \le \mathbb{E}[X^2]$ for any random variable X; and the third inequality is by the law of total expectation and algebra. □

C.3.4 Concluding the regret bounds

In this section, we present the proofs of Theorems 4.5 and 4.6. To bound the collective regret of MULTI-TASK-EULER, we first recall the following general result from [136], which is useful to establish instance-dependent regret guarantees for episodic RL.

Lemma C.12 (Clipping lemma, [136], Lemma B.6)**.** *Fix player $p \in [M]$; suppose for each episode k, it follows $\pi^k(p)$, the greedy policy with respect to $\overline{Q}_p^k$. In addition, there exists*

some event E and a collection of functions $\left\{B_p^{k,\text{lead}}, B_p^{k,\text{fut}}\right\}_{k=1}^K \subset (\mathcal{S}\times\mathcal{A}\to\mathbb{R})$, such that if E happens, then for all $k\in[K]$, $h\in[H]$ and $(s,a)\in\mathcal{S}_h\times\mathcal{A}$, the surplus of $\overline{Q}_p^k$ satisfies that

$$0 \le E_p^k(s,a) \lesssim B_p^{k,\text{lead}}(s,a) + \mathbb{E}\left[\sum_{t=h}^{H} B_p^{k,\text{fut}}(s_t,a_t) \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right],$$

then, on E:

$$\text{Reg}(K,p) \lesssim \sum_{s,a}\sum_k \rho_p^k(s,a)\,\text{clip}\left(B_p^{k,\text{lead}}(s,a), \breve{\text{gap}}_p(s,a)\right) + \sum_{s,a}\sum_k \rho_p^k(s,a)\,\text{clip}\left(B_p^{k,\text{fut}}(s,a), \frac{\text{gap}_{p,\min}}{8SAH^2}\right),$$

here, recall that $\text{clip}(\alpha,\Delta)=\alpha\mathbf{1}(\alpha\ge\Delta)$, *and* $\breve{\text{gap}}_p(s,a)=\frac{\text{gap}_p(s,a)}{4H}\vee\frac{\text{gap}_{p,\min}}{4H}$.

Remark C.13. *Our presentation of the clipping lemma is slightly different than the original one [136, Lemma B.6], in that:*

1. *We consider layered MDPs, while Simchowitz and Jamieson [136] consider general stationary MDPs where one state may be experienced at multiple different steps in $[H]$. Specifically, in a layered MDP, the occupancy distributions $\omega_{k,h}$ defined in [136] is only supported over $\mathcal{S}_h\times\mathcal{A}$. As a result, in the presentation here, we no longer need to sum over h – this is already captured in the sum over all s across all layers.*

2. *Our presentation here is in the context of multitask RL, which is with respect to a player $p\in[M]$, its corresponding MDP $\mathcal{M}_p$, and its policies used throughout the process $\left\{\pi^k(p)\right\}_{k=1}^K$. As a result, all quantities have p as subscripts.*

We are now ready to prove Theorems 4.5 and 4.6, MULTI-TASK-EULER's main regret theorems.

Proof of Theorem 4.5

Proof of Theorem 4.5. From Lemma C.12 and Lemma C.11, we have that when E happens,

$$\begin{aligned}\mathrm{Reg}(K) =& \sum_{p=1}^{M} \mathrm{Reg}(K,p) \\ \leq& \underbrace{\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right)}_{(A)} + \\ & \underbrace{\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right)}_{(B)},\end{aligned} \tag{C.45}$$

We bound each term separately. We can directly use Lemma C.14 to bound term (B) as:

$$\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \lesssim MH^3S^2A\left(\ln\left(\frac{MSAK}{\delta}\right)\right)^2. \tag{C.46}$$

For term (A), we will group the sum by $(s,a) \in \mathcal{I}_\epsilon$ and $(s,a) \notin \mathcal{I}_\epsilon$ separately.

Case 1: $(s,a) \in \mathcal{I}_\epsilon$**.**

In this case, we have that for all p, $\breve{\text{gap}}_p(s,a) = \frac{\text{gap}_p(s,a)}{4H} \geq 24\epsilon$. We simplify the corresponding term as follows:

$$
\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \,\text{clip}\left(B^{k,\text{lead}}(s,a), \breve{\text{gap}}_p(s,a)\right) \\
&\leq \sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \cdot \\
&\qquad \text{clip}\left(H \wedge \left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\right), \frac{\min_p \text{gap}_p(s,a)}{4H}\right) \\
&\leq \sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \cdot \\
&\qquad \left(H \wedge \text{clip}\left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right), \frac{\min_p \text{gap}_p(s,a)}{4H}\right)\right) \\
&\lesssim \sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \left(H \wedge \sqrt{\frac{(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)
\end{aligned}
$$

where the first inequality is from the definition of $B^{k,\text{lead}}$; the second inequality is from the basic fact that $\text{clip}(A \wedge B, C) \leq A \wedge \text{clip}(B,C)$; the third inequality uses Lemma C.19 with $a_1 = 5\epsilon$, $a_2 = \sqrt{\frac{(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}$, and $\Delta = \frac{\min_p \text{gap}_p(s,a)}{4H}$, along with the observation that $\text{clip}(5\epsilon, \frac{\min_p \text{gap}_p(s,a)}{16H}) = 0$, since for all $(s,a) \in \mathcal{I}_\epsilon$ and all $p \in [M]$, $\text{gap}_p(s,a) \geq 96\epsilon H$.

We now decompose the inner sum over k, $\sum_{k=1}^K$, to $\sum_{k=1}^{\tau(s,a)-1}$ and $\sum_{k=\tau(s,a)}^K$. The first part is bounded by:

$$
\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=1}^{\tau_p(s,a)-1} \sum_{p=1}^{M} \rho_p^k(s,a) \left(H \wedge \sqrt{\frac{(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right) \\
&\leq \sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=1}^{\tau_p(s,a)-1} \sum_{p=1}^{M} \rho_p^k(s,a) H \leq SAHN_1,
\end{aligned}
$$

which is $\lesssim MHSA \ln\left(\frac{SAK}{\delta}\right)$.

For the second part,

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\\
\lesssim&\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}}\\
\lesssim&\sqrt{\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}}\\
&\qquad\cdot\sqrt{\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=1}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)},
\end{aligned}$$

where the first inequality is by dropping the "$H\wedge$" operator; the second inequality is by Cauchy-Schwarz.

We bound each factor as follows: for the first factor,

$$\begin{aligned}
\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}=&\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\rho^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}\\
\le&L(MK)\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\frac{\rho^k(s,a)}{\bar{n}^k(s,a)}\\
\le&\sum_{(s,a)\in\mathcal{I}_\epsilon}L(MK)\cdot\int_1^{\bar{n}^K(s,a)}\frac{1}{u}du\\
\le&|\mathcal{I}_\epsilon|\,L(MK)^2\lesssim|\mathcal{I}_\epsilon|\left(\ln\left(\frac{MSAK}{\delta}\right)\right)^2,
\end{aligned}$$

where the first inequality is because L is monotonically increasing, and $\bar{n}^k(s,a)\le MK$; the second inequality is from the observation that $\rho^k(s,a)\in[0,M]$, $\bar{n}^k(s,a)\ge 2M$, and $u\mapsto\frac{1}{u}$ is monotonically decreasing; the last two inequalities are by algebra.

For the second factor,

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=1}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\\
\lesssim\ & MKH+\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\rho_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\\
\lesssim\ & MKH+\sum_{p=1}^{M}\sum_{k=1}^{K}\mathrm{Var}\left[\sum_{h=1}^{H}r_{h,p}^k\mid\pi^k(p)\right]\\
\lesssim\ & MKH^2.
\end{aligned}\tag{C.47}$$

where the first inequality is by the fact that ρ_p^k are probability distributions over every layer $h\in[H]$; the last two inequalities are by a law of total variance identity [see, e.g., 12, Equation (26)]. To summarize, the second part is at most

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\\
\lesssim\ & \sqrt{MKH^2|\mathcal{I}_\epsilon|}\ln\left(\frac{MSAK}{\delta}\right).
\end{aligned}$$

Combining the bounds for the first and the second parts, we have:

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a),\mathrm{g\breve{a}p}_p(s,a)\right)\\
\lesssim\ & \left(\sqrt{MKH^2|\mathcal{I}_\epsilon|}+MHSA\right)\ln\left(\frac{MSAK}{\delta}\right).
\end{aligned}$$

Case 2: $(s,a) \notin \mathcal{I}_\epsilon$**.**

We simplify the corresponding term as follows:

$$\begin{aligned}
&\sum_{(s,a)\notin\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right) \\
&\lesssim \sum_{(s,a)\notin\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(H \wedge \left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) L(n_p^k(s,a))}{n_p^k(s,a)}}\right), \frac{\breve{\mathrm{gap}}_p(s,a)}{4H}\right) \\
&\lesssim \sum_{(s,a)\notin\mathcal{I}_\epsilon} \sum_{k,p} \left(H \wedge \sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) L(n_p^k(s,a))}{n_p^k(s,a)}}\right)
\end{aligned}$$

For each p and (s,a), we now decompose the inner sum over k, $\sum_{k=1}^{K}$, to $\sum_{k=1}^{\tau_p(s,a)-1}$ and $\sum_{k=\tau_p(s,a)}^{K}$. The first part is bounded by:

$$\begin{aligned}
&\sum_{(s,a)\notin\mathcal{I}_\epsilon} \sum_{p=1}^{M} \sum_{k=1}^{\tau_p(s,a)-1} \rho_p^k(s,a) \left(H \wedge \sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]) L(n_p^k(s,a))}{n_p^k(s,a)}}\right) \\
&\leq \sum_{(s,a)\notin\mathcal{I}_\epsilon} \sum_{p=1}^{M} \sum_{k=1}^{\tau_p(s,a)-1} \rho_p^k(s,a) H \\
&\leq MHSAN_2,
\end{aligned}$$

which is $\lesssim MHSA \ln\left(\frac{MSAK}{\delta}\right)$.

For the second part,

$$\begin{aligned}
&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(n_p^k(s,a))}{n_p^k(s,a)}}\right)\\
\lesssim&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(\bar{n}_p^k(s,a))}{\bar{n}_p^k(s,a)}}\\
\leq&\sqrt{\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\cdot\frac{L(\bar{n}_p^k(s,a))}{\bar{n}_p^k(s,a)}}\cdot\\
&\qquad\sqrt{\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k=1}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)}
\end{aligned}$$

We bound each factor as follows: for the first factor,

$$\begin{aligned}
\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}\leq&L(K)\cdot\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\frac{\rho_p^k(s,a)}{\bar{n}^k(s,a)}\\
\leq&L(K)\cdot\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\int_1^{\bar{n}_p^K(s,a)}\frac{1}{u}du\\
\leq&\left|\mathcal{I}_\epsilon^C\right|ML(K)^2\leq\left|\mathcal{I}_\epsilon^C\right|M\left(\ln\left(\frac{MSAK}{\delta}\right)\right)^2.
\end{aligned}$$

where the first inequality is because L is monotonically increasing, and $\bar{n}_p^k(s,a)\leq K$; the second inequality is from the observation that $\rho^k(s,a)\in[0,1]$, $\bar{n}^k(s,a)\geq 2$, and $u\mapsto\frac{1}{u}$ is monotonically decreasing; the last two inequalities are by algebra.

The second factor is again bounded by (C.47). Therefore, the second part of the

sum is at most

$$\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(n_p^k(s,a))}{n_p^k(s,a)}}\right)$$
$$\leq\left(M\sqrt{KH^2|\mathcal{I}_\epsilon^C|}+MHSA\right)\ln\left(\frac{MSAK}{\delta}\right).$$

Combining the bounds for the first and the second parts, we have:

$$\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a),\mathrm{g\breve{a}p}_p(s,a)\right)$$
$$\lesssim\left(M\sqrt{KH^2|\mathcal{I}_\epsilon^C|}+MHSA\right)\ln\left(\frac{MSAK}{\delta}\right).$$

Now, combining the bounds for cases 1 and 2, we have that

$$(A)\leq\left(\sqrt{MKH^2|\mathcal{I}_\epsilon|}+M\sqrt{KH^2|\mathcal{I}_\epsilon^C|}+MHSA\right)\cdot\ln\left(\frac{MSAK}{\delta}\right).\tag{C.48}$$

In conclusion, by the regret decomposition Equation (C.45), and Equations (C.48) and (C.46), we have:

$$\mathrm{Reg}(K)\leq\left(\sqrt{MH^2|\mathcal{I}_\epsilon|\,K}+M\sqrt{H^2|\mathcal{I}_\epsilon^C|\,K}+MH^3S^2A\ln\left(\frac{MSAK}{\delta}\right)\right)\ln\left(\frac{MSAK}{\delta}\right).$$

□

Proof of Theorem 4.6

Proof of Theorem 4.6. From Lemma C.12, we have that when E happens,

$$\begin{aligned}\mathrm{Reg}(K) =& \sum_{p=1}^{M} \mathrm{Reg}(K,p) \\ \leq& \underbrace{\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right)}_{(A)} + \\ & \underbrace{\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right)}_{(B)},\end{aligned}$$

We focus on each term separately. We directly use Lemma C.14 to bound term (B) as:

$$\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \lesssim MH^3S^2A \ln\left(\frac{MSAK}{\delta}\right)\cdot \ln\frac{MHSA}{\mathrm{gap}_{\min}}. \tag{C.49}$$

For the (s,a)-th term in term (A), we will consider the cases of $(s,a) \in \mathcal{I}_\epsilon$ and $(s,a) \notin \mathcal{I}_\epsilon$ separately.

Case 1: $(s,a) \in \mathcal{I}_\epsilon$**.**

In this case, we have that for all p, $\breve{\text{gap}}_p(s,a) = \frac{\text{gap}_p(s,a)}{4H} \geq 24\epsilon$. We simplify the corresponding term as follows:

$$
\begin{aligned}
&\sum_{k,p} \rho_p^k(s,a) \,\text{clip}\left(B^{k,\text{lead}}(s,a), \breve{\text{gap}}_p(s,a)\right) \\
\leq &\sum_{k=1}^{K}\sum_{p=1}^{M} \rho_p^k(s,a) \cdot \\
&\qquad \text{clip}\left(H \wedge \left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{(1+\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\right), \frac{\min_p \text{gap}_p(s,a)}{4H}\right) \\
\leq &\sum_{k=1}^{K} \rho^k(s,a) \,\text{clip}\left(H \wedge \left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}\right)\right), \frac{\min_p \text{gap}_p(s,a)}{4H}\right) \\
\leq &\sum_{k=1}^{k} \rho^k(s,a) \left(H \wedge \text{clip}\left(5\epsilon + \mathcal{O}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}\right), \frac{\min_p \text{gap}_p(s,a)}{4H}\right)\right) \\
\lesssim &\sum_{k=1}^{K} \rho^k(s,a) \left(H \wedge \text{clip}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}, \frac{\min_p \text{gap}_p(s,a)}{16H}\right)\right)
\end{aligned}
$$

where the first inequality is by the definition of $B^{k,\text{lead}}$; the second inequality is from that $\text{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}] \leq H^2$; the third inequality is from that $\text{clip}(A \wedge B, C) \leq A \wedge \text{clip}(B,C)$; the third inequality uses Lemma C.19 with $a_1 = 5\epsilon$, $a_2 = \sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}$, and $\Delta = \frac{\min_p \text{gap}_p(s,a)}{4H}$, along with the observation that $\text{clip}(5\epsilon, \frac{\min_p \text{gap}_p(s,a)}{16H}) = 0$, since for all $(s,a) \in \mathcal{I}_\epsilon$ and all $p \in [M]$, $\text{gap}_p(s,a) \geq 96\epsilon H$.

We now decompose the inner sum over k, $\sum_{k=1}^{K}$, to $\sum_{k=1}^{\tau(s,a)-1}$ and $\sum_{k=\tau(s,a)}^{K}$. The first part's contribution is at most $N_1 \cdot H \lesssim MH \ln\left(\frac{SAK}{\delta}\right)$. For the second part, its

contribution is at most:

$$\begin{aligned}
&\sum_{k=\tau(s,a)}^{K} \rho^k(s,a) \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}, \frac{\min_p \operatorname{gap}_p(s,a)}{16H} \right) \right) \\
\lesssim & MH + \int_1^{\bar{n}^K(s,a)} \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(u)}{u}}, \frac{\min_p \operatorname{gap}_p(s,a)}{16H} \right) \right) du \\
\lesssim & MH + \frac{H^3}{\min_p \operatorname{gap}_p(s,a)} \ln\left(\frac{MSAK}{\delta}\right)
\end{aligned}$$

where the second inequality is from Lemma C.20 with $f_{\max} = H$, $C = H^2$, $\Delta = \frac{\min_p \operatorname{gap}_p(s,a)}{16H}$, $N = MSA$, $\xi = \delta$, $\Gamma = 1$, $n = \bar{n}^K(s,a) \le K$. In summary, for all $(s,a) \in \mathcal{I}_\epsilon$,

$$\sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\operatorname{gap}}_p(s,a) \right) \le \left(MH + \frac{H^3}{\min_p \operatorname{gap}_p(s,a)} \right) \ln\left(\frac{MSAK}{\delta}\right).$$

Case 2: $(s,a) \notin \mathcal{I}_\epsilon$**.**

In this case, for each $p \in [M]$, we simplify the corresponding term as follows:

$$\begin{aligned}
&\sum_{k} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\operatorname{gap}}_p(s,a) \right) \\
\lesssim & \sum_{k=1}^{K} \rho_p^k(s,a) \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(n_p^k(s,a))}{n_p^k(s,a)}}, \frac{\breve{\operatorname{gap}}_p(s,a)}{16H} \right) \right)
\end{aligned}$$

We now decompose the inner sum over k, $\sum_{k=1}^{K}$, to $\sum_{k=1}^{\tau_p(s,a)-1}$ and $\sum_{k=\tau_p(s,a)}^{K}$. The first part's contribution is at most $N_2 \cdot H \lesssim H \ln\left(\frac{MSAK}{\delta}\right)$.

For the second part, its contribution is at most:

$$\begin{aligned}
&\sum_{k=\tau_p(s,a)}^{K} \rho_p^k(s,a) \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}, \frac{\breve{\text{gap}}_p(s,a)}{16H} \right) \right) \\
\lesssim & H + \int_1^{\bar{n}_p^K(s,a)} \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(u)}{u}}, \frac{\breve{\text{gap}}_p(s,a)}{16H} \right) \right) du \\
\lesssim & H + \frac{H^3}{\breve{\text{gap}}_p(s,a)} \ln\left(\frac{MSAK}{\delta} \right)
\end{aligned}$$

where the second inequality is from Lemma C.20 with $f_{\max} = H$, $C = H^2$, $\Delta = \frac{\breve{\text{gap}}_p(s,a)}{16H}$, $N = MSA$, $\xi = \delta$, $\Gamma = 1$, $n = \bar{n}_p^K(s,a) \le K$. In summary, for any $(s,a) \in \mathcal{I}_\epsilon^C$ and $p \in [M]$,

$$\sum_k \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\text{lead}}(s,a), \breve{\text{gap}}_p(s,a) \right) \lesssim (H + \frac{H^3}{\breve{\text{gap}}_p(s,a)}) \ln\left(\frac{MSAK}{\delta} \right),$$

summing over p, we get:

$$\sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\text{lead}}(s,a), \breve{\text{gap}}_p(s,a) \right) \lesssim \left(MH + \sum_{p=1}^{M} \frac{H^3}{\breve{\text{gap}}_p(s,a)} \right) \ln\left(\frac{MSAK}{\delta} \right),$$

In summary, combining the regret bounds of cases 1 and 2 for term (A), along with Equation (C.49) for term (B), and observe that $\breve{\text{gap}}_p(s,a) = \text{gap}_{p,\min}$ if $(s,a) \in Z_{p,\text{opt}}$, and $\breve{\text{gap}}_p(s,a) = \text{gap}_p(s,a)$ otherwise, we have that on event E, MULTI-TASK-EULER satisfies:

$$\begin{aligned}
\text{Reg}(K) \lesssim \ln\left(\frac{MSAK}{\delta} \right) \Bigg(& \sum_{p \in [M]} \left(\sum_{(s,a) \in Z_{p,\text{opt}}} \frac{H^3}{\text{gap}_{p,\min}} + \sum_{(s,a) \in (\mathcal{I}_\epsilon \cup Z_{p,\text{opt}})^C} \frac{H^3}{\text{gap}_p(s,a)} \right) + \\
& \sum_{(s,a) \in \mathcal{I}_\epsilon} \frac{H^3}{\min_p \text{gap}_p(s,a)} \Bigg) + \ln\left(\frac{MSAK}{\delta} \right) \cdot MS^2 AH^3 \ln \frac{MHSA}{\text{gap}_{\min}}.
\end{aligned}$$

□

Lemma C.14 (Bounding the lower order terms)**.** *If E happens, then*

$$\begin{aligned}
&\sum_{s,a}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a),\frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right)\\
\lesssim\ & MH^3S^2A\ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right)\wedge\ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right).
\end{aligned}$$

Proof. We expand the left hand side using the definition of $B^{k,\mathrm{fut}}$, and the fact that $\mathrm{gap}_{p,\min}\geq\mathrm{gap}_{\min}$:

$$\sum_{k=1}^{K}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a),\frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \tag{C.50}$$

$$\lesssim\sum_{k=1}^{K}\rho_p^k(s,a)\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(n_p^k(s,a))}{n_p^k(s,a)},\frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right) \tag{C.51}$$

We now decompose the sum $\sum_{k=1}^{K}$ to $\sum_{k=1}^{\tau_p(s,a)-1}$ and $\sum_{k=\tau_p(s,a)}^{K}$. The first part can be bounded by

$$\sum_{k=1}^{\tau_p(s,a)-1}\rho_p^k(s,a)\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(n_p^k(s,a))}{n_p^k(s,a)},\frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right)\leq\sum_{k=1}^{\tau_p(s,a)-1}H^3\rho_p^k(s,a)\leq H^3N_2,$$

which is at most $\mathcal{O}\left(H^3\cdot\ln\left(\frac{MSAK}{\delta}\right)\right)$. For the second part, it can be bounded by:

$$\begin{aligned}
&\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(n_p^k(s,a))}{n_p^k(s,a)},\frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right)\\
\leq\ & H^3\cdot1+\int_1^{\bar{n}_p^K(s,a)}\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(u)}{u},\frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right)du\\
\lesssim\ & H^3+H^3\ln\left(\frac{MSA}{\delta}\right)+H^3S\ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right)\wedge\ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right),
\end{aligned}$$

where the second inequality is from Lemma C.20 with $f_{\max} = H^3$, $C = H^3 S$, $\Delta = \frac{\mathrm{gap}_{\min}}{8SAH^2}$, $N = MSA$, $\xi = \delta$, $\Gamma = 1$, $n = \bar{n}_p^K(s,a) \le K$. In summary,

$$\begin{aligned}
&\sum_k \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \frac{\mathrm{gap}_{\min}}{8SAH^2}\right) \\
&\lesssim H^3 S \ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right) \wedge \ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right)
\end{aligned}$$

Summing over $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $p \in [M]$, we get

$$\begin{aligned}
&\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \frac{\mathrm{gap}_{\min}}{8SAH^2}\right) \\
&\lesssim MH^3S^2A \ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right) \wedge \ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right).
\end{aligned}$$

□

C.3.5 Miscellaneous lemmas

This subsection collects a few miscellaneous lemmas used throughout the upper bound proofs.

Lemma C.15 (136, Lemma F.5)**.** *For random variables X and Y,*

$$\left|\sqrt{\mathrm{var}[X]} - \sqrt{\mathrm{var}[Y]}\right| \le \sqrt{\mathbb{E}\left[(X-Y)^2\right]}.$$

Lemma C.16. *Suppose distributions P and Q are supported over $[0, B]$, and $\|P - Q\|_1 \le \epsilon \le 2$. Then:*

$$\left|\mathbb{E}_{X\sim P}[X] - \mathbb{E}_{X\sim Q}[X]\right| \le B\epsilon,$$

$$\left|\mathrm{var}_{X\sim P}[X] - \mathrm{var}_{X\sim Q}[X]\right| \le 3B^2\epsilon.$$

Proof. First,

$$\begin{aligned}\left|\mathbb{E}_{X\sim P}[X]-\mathbb{E}_{X\sim Q}[X]\right| &= \left|\int_0^B x(p_X(x)-q_X(x))dx\right| \\ &\le \int_0^B |x|\left|p_X(x)-q_X(x)\right| dx \le B\|P-Q\|_1 \le B\epsilon.\end{aligned}$$

Second, observe that

$$\left|\mathbb{E}_{X\sim P}[X^2]-\mathbb{E}_{X\sim Q}[X^2]\right| \le B^2\epsilon.$$

Meanwhile,

$$\begin{aligned}\left|(\mathbb{E}_{X\sim P}[X])^2-(\mathbb{E}_{X\sim Q}[X])^2\right| &\le \left|\mathbb{E}_{X\sim P}[X]-\mathbb{E}_{X\sim Q}[X]\right|\cdot\left|\mathbb{E}_{X\sim P}[X]+\mathbb{E}_{X\sim Q}[X]\right| \\ &\le 2B\cdot B\epsilon \\ &= 2B^2\epsilon.\end{aligned}$$

Combining the above, we have

$$\left|\mathrm{var}_{X\sim P}[X]-\mathrm{var}_{X\sim Q}[X]\right| \le 3B^2\epsilon.$$

□

Lemma C.17. *For $A, B, C, D, E, F \ge 0$:*

1. *If $|A-B| \le \sqrt{BC}+C$, then we have $\left|\sqrt{A}-\sqrt{B}\right| \le 2\sqrt{C}$.*
2. *If $D \le E + F\sqrt{D}$, then $\sqrt{D} \le \sqrt{E}+F$.*

Proof. 1. First, $A-B \le |A-B| \le \sqrt{BC}+C$; this implies that $A \le B+2\sqrt{BC}+C$, and therefore $\sqrt{A} \le \sqrt{B}+\sqrt{C}$.

On the other hand, $B \le A + C + \sqrt{BC}$; therefore, applying item 1 with $D = B$, $E = A + C$, and $F = \sqrt{C}$, we have $\sqrt{B} \le \sqrt{A + C} + \sqrt{C} \le \sqrt{A} + 2\sqrt{C}$.

2. The roots of $x^2 - Fx - E = 0$ are $\frac{F \pm \sqrt{F^2+4E}}{2}$, and therefore D must satisfy $\sqrt{D} \le \frac{F+\sqrt{F^2+4E}}{2} \le \frac{F+F+2\sqrt{E}}{2} = F + \sqrt{E}$.

□

Lemma C.18. *For $a \ge 0$, $1 \wedge (a + \sqrt{a}) \le 1 \wedge 2\sqrt{a}$.*

Proof. We consider the cases of $a \ge 1$ and $a < 1$ respectively. If $a \ge 1$, LHS $= 1 =$ RHS. Otherwise, $a \le 1$; in this case, LHS $= 1 \wedge (a + \sqrt{a}) \le 1 \wedge (\sqrt{a} + \sqrt{a}) =$ RHS. □

Lemma C.19 (Special case of [136], Lemma B.5)**.** *For $a_1, a_2, \Delta \ge 0$, $\operatorname{clip}(a_1 + a_2, \Delta) \le 2\operatorname{clip}(a_1, \Delta/4) + 2\operatorname{clip}(a_2, \Delta/4)$.*

Lemma C.20 (Integral calculation, [136], Lemma B.9 therein)**.** *Let*

$$f(u) \le \min(f_{\max}, \operatorname{clip}(g(u), \Delta)),$$

where $\Delta \in [0, \Gamma]$, and $g(u)$ is nonincreasing. Let $N \ge 1$ and $\xi \in (0, \frac{1}{2})$. Then:

1. *If $g(u) \lesssim \sqrt{\frac{C \log \frac{Nu}{\xi}}{u}}$ for some $C > 0$ such that $\ln C \lesssim \ln N$, then*

$$\int_\Gamma^n f(u/4)du \lesssim \sqrt{Cn \ln \frac{Nn}{\xi}} \wedge \frac{C}{\Delta} \ln\left(\frac{Nn}{\xi}\right).$$

2. *If $g(u) \lesssim \frac{C \ln \frac{Nu}{\xi}}{u}$ for some $C > 0$ such that $\ln C \lesssim \ln N$, then*

$$\int_\Gamma^n f(u/4)du \lesssim f_{\max} \ln \frac{N}{\xi} + C \ln \frac{Nn}{\xi} \cdot \left(\ln \frac{Nn}{\xi} \wedge \ln \frac{N\Gamma}{\Delta}\right).$$

C.4 Proof of the Lower Bounds

C.4.1 Auxiliary lemmas

Lemma C.21 (Regret decomposition, [136], Section H.2)**.** *For any MPERL problem instance and any algorithm, we have*

$$\mathbb{E}\left[\mathrm{Reg}(K)\right] \geq \sum_{p=1}^{M} \sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbb{E}\left[n_p^{K+1}(s,a)\right] \mathrm{gap}_p(s,a), \tag{C.52}$$

where we recall that $n_p^{K+1}(s,a)$ is the number of visits of (s,a) by player p at the beginning of the $(K+1)$-th episode (after the first K episodes). Furthermore, for any $(s,a)\in\mathcal{S}_1\times\mathcal{A}$, we have

$$\sum_{p=1}^{M} \mathbb{E}\left[n_p^{K+1}(s,a)\right] \mathrm{gap}_p(s,a) \geq \mathbb{E}\left[n^{K+1}(s,a)\right] \left(\min_{p\in[M]} \mathrm{gap}_p(s,a)\right), \tag{C.53}$$

where we recall that $n^{K+1}(s,a) = \sum_{p=1}^{M} n_p^{K+1}(s,a)$.

Proof. Eq. (C.53) follows straightforwardly from the fact that for every $(s,a,p)\in\mathcal{S}_1\times\mathcal{A}\times[M]$, $\min_{p'\in[M]}\mathrm{gap}_{p'}(s,a) \leq \mathrm{gap}_p(s,a)$.

We now prove Eq. (C.52). Let π_p^k denote $\pi^k(p)$. We have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}(K)\right] &= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} p_0(s_{1,p}^k = s)\left(V_p^\star(s) - V_p^{\pi_p^k}(s)\right)\right] \\
&\geq \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} p_0(s_{1,p}^k = s)\left(V_p^\star(s) - Q_p^\star(s,\pi_p^k(s))\right)\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} p_0(s)\mathrm{gap}_p(s,\pi_p^k(s))\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} \mathbf{1}\left(s_{1,p}^k = s\right)\mathrm{gap}_p(s,\pi_p^k(s))\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbf{1}\left(s_{1,p}^k, \pi_p^k(s) = (s,a)\right)\mathrm{gap}_p(s,a)\right] \\
&= \sum_{p=1}^{M}\sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbb{E}\left[n_p^K(s,a)\right]\mathrm{gap}_p(s,a)
\end{aligned} \tag{C.54}$$

where the first equality is from the definition of collective regret; the first inequality is from the simple fact that $V_p^\pi(s) = Q_p^\pi(s,\pi(s)) \leq Q_p^\star(s,\pi(s))$ for any policy π; the second equality is from the definition of suboptimality gaps; and the third equality is from the basic observation that $s_{1,p}^k \sim p_0$. □

Lemma C.22 (Divergence decomposition [90, 168])**.** *For two MPERL problem instances, $\mathfrak{M}$ and $\mathfrak{M}'$, which only differ in the transition probabilities $\left\{\mathbb{P}_p(\cdot \mid s,a)\right\}_{p\in[M],(s,a)\in\mathcal{S}\times\mathcal{A}}$, and for a fixed algorithm, let $\mathbb{P}_{\mathfrak{M}}$ and $\mathbb{P}_{\mathfrak{M}'}$ be the probability measures on the outcomes of running the algorithm on $\mathfrak{M}$ and $\mathfrak{M}'$, respectively. Then,*

$$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) = \sum_{p=1}^{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}_{\mathfrak{M}}\left[n_p^{K+1}(s,a)\right]\mathrm{KL}\left(\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s,a), \mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s,a)\right),$$

where $\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s,a)$ *and* $\mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s,a)$ *are the transition probabilities of the problem instance* $\mathfrak{M}$ *and* $\mathfrak{M}'$*, respectively.*

Lemma C.23 (Bretagnolle-Huber inequality, [90], Theorem 14.2)**.** *Let* $\mathbb{P}$ *and* $\mathbb{Q}$ *be two distributions on the same measurable space, and* A *be an event. Then,*

$$\mathbb{P}(A) + \mathbb{Q}(A^C) \geq \frac{1}{2}\exp\left(-\operatorname{KL}(\mathbb{P},\mathbb{Q})\right).$$

Lemma C.24 (see Lemma A.12)**.** *For any* $x, y \in [\frac{1}{4}, \frac{3}{4}]$, $\operatorname{KL}\left(\operatorname{Ber}(x), \operatorname{Ber}(y)\right) \leq 3(x-y)^2$.

Lemma C.25. *Let* X *be a Binomial random variable and* $X \sim Bin(n,p)$*, where* $n \geq \frac{1}{p}$*. Then,*

$$\mathbb{E}\left[X^{\frac{3}{2}}\right] \leq 2(np)^{\frac{3}{2}}.$$

Proof. Let $Y = X^2$, and $f(Y) = Y^{\frac{3}{4}}$. We have $\mathbb{E}[Y] = \mathbb{E}\left[X^2\right] = \operatorname{var}[X] + \mathbb{E}[X]^2 = (np)^2 + np(1-p) \leq (np)^2 + np \leq 2(np)^2$, where the last inequality follows from the assumption that $n \geq \frac{1}{p}$. By Jensen's inequality, we have $\mathbb{E}\left[X^{\frac{3}{2}}\right] = \mathbb{E}\left[f(Y)\right] \leq f\left(\mathbb{E}[Y]\right) \leq \left(2n^2p^2\right)^{\frac{3}{4}} \leq 2(np)^{\frac{3}{2}}$. □

C.4.2 Gap independent lower bounds

Theorem C.26 (Restatement of Theorem 4.7)**.** *For any* $A \geq 2$*,* $H \geq 2$*,* $S \geq 4H$*,* $K \geq SA$*,* $M \in \mathbb{N}$*, and* $l, l^C \in \mathbb{N}$ *such that* $l + l^C = SA$ *and* $l \leq SA - 4(S + HA)$*, there exists some* ϵ *such that for any algorithm* Alg*, there exists an* ϵ*-MPERL problem instance with* S *states,* A *actions,* M *players and an episode length of* H *such that* $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \geq l$*, and*

$$\mathbb{E}\left[\operatorname{Reg}_{\text{Alg}}(K)\right] \geq \Omega\left(M\sqrt{H^2 l^C K} + \sqrt{MH^2 l K}\right).$$

Proof. The construction and techniques in this proof are inspired by Appendix A.5 and [136].

Fix any algorithm Alg; we consider two cases:

1. $l > Ml^C$;

2. $Ml^C \geq l$.

Case 1: $l > Ml^C$**.**

Let $S_1 = S - 2(H-1)$, and $b = \lceil \frac{l}{S_1} \rceil \geq 1$. Let $\Delta = \sqrt{\frac{l+1}{384MK}}$, and let $\epsilon = \frac{1}{2}H\Delta$. We note that under the assumption that $K \geq SA$, and the observation that $l \leq SA$, we have $\Delta \leq \frac{1}{4}$. We define $(b+1)^{S_1}$ ϵ-MPERL problem instances, each indexed by an element in $[b+1]^{S_1}$. It suffices to show that, on at least one of the problem instances, $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(\sqrt{MH^2lK}\right)$.

Construction. For $\mathbf{a} = (a_1, \ldots, a_{S_1}) \in [b+1]^{S_1}$, we define the following ϵ-MPERL problem instance, $\mathfrak{M}(\mathbf{a}) = \left\{\mathcal{M}_p\right\}_{p=1}^{M}$, with S states, A actions, and an episode length of H, such that for each $p \in [M]$, $\mathcal{M}_p$ is constructed as follows:

- $\mathcal{S}_1 = [S_1]$, and p_0 is a uniform distribution over the states in $\mathcal{S}_1$.
- For $h \in [2, H]$, $\mathcal{S}_h = \{S_1 + 2h - 3, S_1 + 2h - 2\}$.
- $\mathcal{A} = [A]$.
- For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution $r_p(s, a)$ is a Bernoulli distribution, $\mathrm{Ber}(R_p(s, a))$, and we will specify $R_p(s, a)$ subsequently.
- For each state $s \in [S_1]$,

$$\mathbb{P}_p(S_1 + 1 \mid s, a) = \begin{cases} \frac{1}{2} + \Delta, & \text{if } a = a_s; \\ \frac{1}{2}, & \text{if } a \in [b+1] \setminus a_s; \\ 0, & \text{if } a \notin [b+1]; \end{cases}$$

and for each $a \in \mathcal{A}$, $\mathbb{P}_p(S_1 + 2 \mid s, a) = 1 - \mathbb{P}_p(S_1 + 1 \mid s, a)$, and $R_p(s, a) = 0$.

- For $h \in [2, H]$, and $a \in \mathcal{A}$, let
  - $\mathbb{P}_p\left(S_1 + 2h - 1 \mid S_1 + 2h - 3, a\right) = 1$, $\mathbb{P}_p\left(S_1 + 2h \mid S_1 + 2h - 3, a\right) = 0$, and $R_p(S_1 + 2h - 3, a) = 1$.
  - $\mathbb{P}_p\left(S_1 + 2h \mid S_1 + 2h - 2, a\right) = 0$, $\mathbb{P}_p\left(S_1 + 2h - 1 \mid S_1 + 2h - 2, a\right) = 1$, and $R_p(S_1 + 2h - 2, a) = 0$.

It can be easily verified that $\mathfrak{M}(\mathbf{a}) = \left\{\mathcal{M}_p\right\}_{p=1}^{M}$ is a 0-MPERL problem instance, and hence an ϵ-MPERL problem instance—the reward distributions and the transition probabilities are the same for all players, i.e., for every $p, q \in [M]$, and every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\left|R_p(s,a) - R_q(s,a)\right| = 0 \leq \epsilon, \quad \left|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a)\right| = 0 \leq \frac{\epsilon}{H}.$$

Suboptimality gaps. We now calculate the suboptimality gaps of the state-action pairs in the above MDPs. For each $p \in [M]$ and each $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathrm{gap}_p(s,a) = V_p^\star(s) - Q_p^\star(s,a) = \max_{a'} Q_p^\star(s,a') - Q_p^\star(s,a).$$

In $\mathfrak{M}(\mathbf{a})$, it can be easily observed that for every $p \in [M]$, and every $(s, a) \in \left(\mathcal{S} \setminus \mathcal{S}_1\right) \times \mathcal{A}$, $\mathrm{gap}_p(s, a) = 0$. Now, for every $p \in [M]$, $(s, a) \in \mathcal{S}_1 \times \mathcal{A}$, we have

$$\begin{aligned}\mathrm{gap}_p(s,a) &= \max_{a'} Q_p^\star(s,a') - Q_p^\star(s,a) \\ &= (H-1)\left(\max_{a'} \mathbb{P}_p(S_1 + 1 \mid s, a') - \mathbb{P}_p(S_1 + 1 \mid s, a)\right).\end{aligned}$$

It follows that, for every $p \in [M]$ and every state $s \in [S_1]$,

$$\mathrm{gap}_p(s,a) = \begin{cases} 0, & \text{if } a = a_s; \\ (H-1)\Delta, & \text{if } a \in [b+1] \setminus a_s; \\ (H-1)\left(\frac{1}{2} + \Delta\right), & \text{if } a \notin [b+1]. \end{cases}$$

Subpar state-action pairs. It can be verified that in $\mathfrak{M}(\mathbf{a})$, $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \geq l$. Specifically, since $(H-1)\Delta = (H-1)\frac{2\epsilon}{H} \geq \epsilon \geq \frac{\epsilon}{2} = 96H\frac{\epsilon}{192H}$, we have that the number of subpar state-action pairs is at least $S_1 b = S_1 \lceil \frac{l}{S_1} \rceil \geq l$.

It suffices to prove that

$$\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{1}{640}\sqrt{MH^2lK},$$

where we recall that $\mathbf{a} = (a_1, \ldots, a_{S_1})$; furthermore, it suffices to show that, for any $s' \in [S_1]$,

$$\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N^{K+1}(s') - n^{K+1}(s', a_{s'})\right] \geq \frac{MK}{4S_1}, \tag{C.55}$$

where $N^{K+1}(s') = \sum_{a\in\mathcal{A}} n^{K+1}(s',a)$; this is because it follows from Eq. (C.55) that

$$\begin{aligned} & \mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\ \geq & \sum_{s'\in\mathcal{S}_1}(H-1)\frac{\Delta}{4}\cdot\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N^{K+1}(s') - n^{K+1}(s', a_{s'})\right] \\ \geq & \sum_{s'\in\mathcal{S}_1}\frac{H}{2}\cdot\frac{\Delta}{4}\cdot\frac{MK}{4S_1} \\ \geq & \frac{1}{640}\sqrt{MH^2lK}, \end{aligned}$$

where the first inequality uses Lemma C.21 (the regret decomposition lemma).

Without loss of generality, let us choose $s' = 1$. To prove Eq. (C.55), we use a standard technique and define a set of helper problem instances. Specifically, for any $(a_2, a_3, \ldots, a_{S_1}) \in [b+1]^{S_1-1}$, we define a problem instance $\mathfrak{M}(0, a_2, \ldots, a_{s_1})$ such that it agrees with $\mathfrak{M}(a_1, a_2, \ldots, a_{s_1})$ on everything but $\mathbb{P}_p(\cdot \mid 1, a_1)$'s, i.e., in $\mathfrak{M}(0, a_2, \ldots, a_{s_1})$, for every $p \in [M]$,

$$\mathbb{P}_p(S_1 + 1 \mid 1, a_1) = \frac{1}{2}.$$

Now, for each $(j, a_2, \ldots, a_{s_1}) \in \big([0] \cup [b+1]\big) \times [b+1]^{S_1-1}$, let $\mathbb{P}_{j,a_2,\ldots,a_{S_1}}$ denote the probability measure on the outcomes of running Alg on the problem instance $\mathfrak{M}(j, a_2, \ldots, a_{s_1})$. Further, for each $j \in \{0\} \cup [b+1]$, we define

$$\mathbb{P}_j = \frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\ldots,a_{S_1} \in [b+1]^{S_1-1}} \mathbb{P}_{j,a_2,\ldots,a_{S_1}};$$

and we use $\mathbb{E}_j$ to denote the expectation with respect to $\mathbb{P}_j$.

In subsequent calculations, for any index $m \in \big([0] \cup [b+1]\big) \times [b+1]^{S_1-1}$, we also denote by $\mathbb{P}_m\big(\cdot \mid N^{K+1}(1)\big)$ and $\mathbb{E}_m\big[\cdot \mid N^{K+1}(1)\big]$ the probability and expectation, respectively, conditional on a realization of $N^{K+1}(1)$ under $\mathbb{P}_m$. Observe that, for any $j \in \{0\} \cup [b+1]$,

$$\begin{aligned}
\mathbb{P}_j(\cdot \mid N^{K+1}(1)) &= \frac{\mathbb{P}_j(\cdot, N^{K+1}(1))}{\mathbb{P}_j(N^{K+1}(1))} \\
&= \frac{\frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\ldots,a_{S_1} \in [b+1]^{S_1-1}} \mathbb{P}_{j,a_2,\ldots,a_{S_1}}(\cdot, N^{K+1}(1))}{\mathbb{P}_j(N^{K+1}(1))} \\
&= \frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\ldots,a_{S_1} \in [b+1]^{S_1-1}} \frac{\mathbb{P}_{j,a_2,\ldots,a_{S_1}}(\cdot, N^{K+1}(1))}{\mathbb{P}_{j,a_2,\ldots,a_{S_1}}(N^{K+1}(1))} \\
&= \frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\ldots,a_{S_1} \in [b+1]^{S_1-1}} \mathbb{P}_{j,a_2,\ldots,a_{S_1}}(\cdot \mid N^{K+1}(1)), \qquad \text{(C.56)}
\end{aligned}$$

where the first equality is from the definition of conditional probability; the second equality is from the definition of $\mathbb{P}_j$; the third equality uses the fact that $\mathbb{P}_j(N^{K+1}(1)) =$

$\mathbb{P}_{j,a_2,\ldots,a_{S_1}}(N^{K+1}(1))$ for any $a_2,\ldots,a_{S_1}$, which is true because $N^{K+1}(1)$ is independent of $a_2,\ldots,a_{S_1}$ conditional on j; and the last equality, again, is from the definition of conditional probability.

We have, for each $j \in [b+1]$,

$$
\begin{aligned}
&\mathbb{E}_j\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right] - \mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right] \\
\leq & N^{K+1}(1)\left\|\mathbb{P}_j\left(\cdot \mid N^{K+1}(1)\right) - \mathbb{P}_0\left(\cdot \mid N^{K+1}(1)\right)\right\|_1 \\
\leq & N^{K+1}(1)\cdot\frac{1}{(b+1)^{S_1-1}}\cdot \\
& \sum_{a_2,\ldots,a_{S_1}\in[b+1]^{S_1-1}}\left\|\mathbb{P}_{j,a_2,\ldots,a_{S_1}}\left(\cdot \mid N^{K+1}(1)\right) - \mathbb{P}_{0,a_2,\ldots,a_{S_1}}\left(\cdot \mid N^{K+1}(1)\right)\right\|_1 \\
\leq & N^{K+1}(1)\cdot\frac{1}{(b+1)^{S_1-1}}\cdot \\
& \sum_{a_2,\ldots,a_{S_1}\in[b+1]^{S_1-1}}\sqrt{2\,\mathrm{KL}\left(\mathrm{Ber}(\frac{1}{2}+\Delta),\mathrm{Ber}(\frac{1}{2})\right)\mathbb{E}_{0,a_2,\ldots,a_{S_1}}\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]} \\
\leq & N^{K+1}(1)\cdot\frac{1}{(b+1)^{S_1-1}}\sum_{a_2,\ldots,a_{S_1}\in[b+1]^{S_1-1}}\sqrt{6\Delta^2\mathbb{E}_{0,a_2,\ldots,a_{S_1}}\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]} \\
\leq & N^{K+1}(1)\sqrt{(6)\frac{l+1}{384MK}\cdot\mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]} \\
= & \frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK}\cdot\mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]}. \qquad \text{(C.57)}
\end{aligned}
$$

where the first inequality is based on Lemma C.16 and the fact that, conditional on $N^{K+1}(1)$, $n^{K+1}(1,j)$ has distribution supported on $[0, N^{K+1}(1)]$; the second inequality follows from Equation (C.56) and the triangle inequality; the third inequality uses Pinsker's inequality and Lemma C.22 (the divergence decomposition lemma); the fourth inequality uses Lemma C.24 and the fact that $\Delta \leq \frac{1}{4}$; and the last inequality follows from Jensen's inequality.

Since $N^{K+1}(1)$ has the same distribution under both $\mathbb{P}_0$ and any $\mathbb{P}_j$ (which is

$\mathrm{Bin}(K, \frac{1}{S_1})$), taking expectation with respect to $N^{K+1}(1)$, we have that, for any $j \in [b+1]$,

$$\mathbb{E}_j\left[n^{K+1}(1,j)\right] - \mathbb{E}_0\left[n^{K+1}(1,j)\right] \leq \mathbb{E}_0\left[\frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK}\cdot\mathbb{E}_0\left[n^{K+1}(1,j)\mid N^{K+1}(1)\right]}\right].$$

In subsequent derivations, we can now avoid bounding the conditional expectation. Specifically, we have

$$\begin{aligned}
&\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_j\left[n^{K+1}(1,j)\right]\\
\leq&\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_0\left[n^{K+1}(1,j)\right]+\\
&\qquad\qquad\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_0\left[\frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK}\cdot\mathbb{E}_0\left[n^{K+1}(1,j)\mid N^{K+1}(1)\right]}\right]\\
\leq&\frac{1}{b+1}\mathbb{E}_0\left[\sum_{j\in[b+1]}n^{K+1}(1,j)\right]+\\
&\qquad\qquad\mathbb{E}_0\left[\frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK}\cdot\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_0\left[n^{K+1}(1,j)\mid N^{K+1}(1)\right]}\right]\\
\leq&\frac{1}{b+1}\mathbb{E}_0\left[N^{K+1}(1)\right]+\mathbb{E}_0\left[\frac{1}{8}\sqrt{\frac{l+1}{MK}\cdot\frac{1}{b+1}}\left(N^{K+1}(1)\right)^{\frac{3}{2}}\right]\\
\leq&\frac{1}{b+1}\mathbb{E}_0\left[N^{K+1}(1)\right]+\frac{1}{8}\sqrt{\frac{S_1}{MK}}\cdot\mathbb{E}_0\left[\left(N^{K+1}(1)\right)^{\frac{3}{2}}\right],
\end{aligned}\tag{C.58}$$

where the first inequality follows from Eq. (C.57) and algebra; the second inequality uses linearity of expectation and Jensen's inequality; the third inequality uses the facts that $\sum_{j\in[b+1]} n^{K+1}(1,j) \leq N^{K+1}(1)$ and, for every $z \in [0]\cup[b+1]$,

$$\sum_{j\in[b+1]}\mathbb{E}_z\left[n^{K+1}(1,j)\mid N^{K+1}(1)\right] \leq \sum_{j\in\mathcal{A}}\mathbb{E}_z\left[n^{K+1}(1,j)\mid N^{K+1}(1)\right] = N^{K+1}(1);$$

and the last inequality uses the linearity of expectation and the construction that $b = \lceil\frac{l}{S_1}\rceil$,

which implies that $l \leq bS_1$ and therefore $l+1 \leq bS_1+1 \leq bS_1+S_1 = (b+1)S_1$.

It follows from Equation (C.58) that

$$\begin{aligned}\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_j\left[n^{K+1}(1,j)\right] &\leq \frac{1}{b+1}\cdot\frac{MK}{S_1}+\frac{1}{8}\sqrt{\frac{S_1}{MK}}\cdot\mathbb{E}_0\left[\left(N^{K+1}(1)\right)^{\frac{3}{2}}\right]\\ &\leq \frac{MK}{2S_1}+\frac{1}{4}\sqrt{\frac{S_1}{MK}\left(\frac{MK}{S_1}\right)^3}\\ &\leq \frac{3MK}{4S_1},\end{aligned}$$

where the second inequality uses the fact that $\frac{1}{b+1} \leq \frac{1}{2}$ and Lemma C.25 under the assumption that $K \geq S_1$.

It then follows that

$$\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_j\left[N^{K+1}(1)-n^{K+1}(1,j)\right] \geq \frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_j\left[N^{K+1}(1)\right]-\frac{3MK}{4S_1} = \frac{MK}{4S_1},$$

and we have

$$\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N^{K+1}(1)-n^{K+1}(1,a_1)\right] \geq \frac{MK}{4S_1}.$$

Case 2: $Ml^C \geq l$**.**

Again, let $S_1 = S-2(H-1)$. Let $u = \lceil\frac{l}{S_1}\rceil$ and $v = A-u = A-\lceil\frac{l}{S_1}\rceil$. Furthermore, let $\Delta = \sqrt{\frac{vS_1}{384K}}$, and $\epsilon = 2H\Delta$. We note that under the assumption that $K \geq SA$ and the fact that $vS_1 \leq SA$, we have $\Delta \leq \frac{1}{4}$. We will define $v^{S_1\times M}$ ϵ-MPERL problem instances, each indexed by an element in $[v]^{S_1\times M}$. It suffices to show that, on at least one of the instances, $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(M\sqrt{H^2l^CK}\right)$.

Facts about v**.** There are two helpful facts about v that can be easily verified:

- $vS_1 \geq \frac{1}{2}l^C$. This is true because, by definition, $vS_1 \geq S_1A-l-S_1 = S_1A-(SA-l^C)-S_1 = l^C-(SA-S_1A)-S_1 = l^C-\big(2(H-1)A+S_1\big)$; since, by assumption,

$l \leq SA - 4(S + HA)$, we have $l^C \geq 4(HA + S) \geq 2\big(2(H-1)A + S_1\big)$; it then follows that $vS_1 \geq l^C - \big(2(H-1)A + S_1\big) \geq \frac{1}{2} l^C$.

- $v \geq 2$. This is true because, as shown above, $vS_1 \geq \frac{1}{2} l^C$ and $l^C \geq 4(HA + S)$, which imply that $v \geq \frac{2(HA+S)}{S_1} \geq \frac{2S_1}{S_1} = 2$.

Construction. For $\mathbf{a} = (a_{1,1}, \ldots, a_{1,M}, a_{2,1}, \ldots, a_{S_1,M}) \in [v]^{S_1 \times M}$, we define the following ϵ-MPERL problem instance, $\mathfrak{M}(\mathbf{a}) = \big\{\mathcal{M}_p\big\}_{p=1}^{M}$, with S states, A actions, and an episode length of H, such that for each $p \in [M]$, $\mathcal{M}_p$ is constructed in the same way as it is for case 1, except for the transition probabilities of $(s, a) \in \mathcal{S}_1 \times \mathcal{A}$:

- For each state $s \in [S_1]$,

$$\mathbb{P}_p(S_1 + 1 \mid s, a) = \begin{cases} \frac{1}{2} + \Delta, & \text{if } a = a_{s,p}; \\ \frac{1}{2}, & \text{if } a \in [v] \setminus a_{s,p}; \\ 0, & \text{if } a \notin [v]; \end{cases}$$

and for each $a \in \mathcal{A}$, $\mathbb{P}_p(S_1 + 2 \mid s, a) = 1 - \mathbb{P}_p(S_1 + 1 \mid s, a)$, and $R_p(s, a) = 0$.

We now verify that $\mathfrak{M}(\mathbf{a})$ is an ϵ-MPMAB problem instance. It can be easily observed that the reward distributions are the same for all players, i.e., for every $p, q \in [M]$ and every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\big|R_p(s, a) - R_q(s, a)\big| = 0 \leq \epsilon.$$

Regarding the transition probabilities, $\forall (s, a) \in \Big(\big(\mathcal{S}_1 \times \big(\mathcal{A} \setminus [v]\big)\big)\Big) \cup \Big(\big(\mathcal{S} \setminus \mathcal{S}_1\big) \times \mathcal{A}\Big)$, we observe that the transition probabilities are the same for all players. Furthermore, for every $p, q \in [M]$ and every $(s, a) \in \mathcal{S}_1 \times [v]$,

$$\Big\|\mathbb{P}_p\big(\cdot \mid s, a\big) - \mathbb{P}_q\big(\cdot \mid s, a\big)\Big\|_1 \leq 2\Delta = \frac{\epsilon}{H}.$$

Therefore, $\mathfrak{M}(\mathbf{a})$ is an ϵ-MPMAB problem instance.

Suboptimality gaps. Similar to the arguments in Case 1, it can be shown that for every $p \in [M]$, and every $(s,a) \in \big(\mathcal{S} \setminus \mathcal{S}_1\big) \times \mathcal{A}$, $\mathrm{gap}_p(s,a) = 0$. And, for every $p \in [M]$, and every $s \in \mathcal{S}_1$,

$$\mathrm{gap}_p(s,a) = \begin{cases} 0, & \text{if } a = a_{s,p}; \\ (H-1)\Delta, & \text{if } a \in [v] \setminus a_{s,p}; \\ (H-1)\left(\frac{1}{2} + \Delta\right), & \text{if } a \notin [v]. \end{cases}$$

Subpar state-action pairs. Based on the above construction, for every $(s,a) \in \mathcal{S}_1 \times \big(\mathcal{A} \setminus [v]\big)$ and every $p \in [M]$, $\mathrm{gap}_p(s,a) = (H-1)\left(\frac{1}{2} + \Delta\right) \geq 3(H-1)\Delta = \frac{3(H-1)}{2H}\epsilon \geq \frac{3}{4}\epsilon \geq 96H\left(\frac{\epsilon}{192H}\right)$, where the first inequality uses the fact that $\Delta \leq \frac{1}{4}$. Therefore, there are at least $(A-v)S_1 = uS_1 \geq l$ state-action pairs in $\mathcal{I}_{\frac{\epsilon}{192H}}$, i.e., $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \geq l$.

Now, it suffices to prove that

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{1}{240} M \sqrt{H^2 l^C K},$$

where we recall that $\mathbf{a} = (a_{1,1}, \ldots, a_{1,M}, a_{2,1}, \ldots, a_{S_1,M})$. It suffices to show, for any $s' \in [S_1]$ and any $p' \in [M]$,

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N_{p'}^{K+1}(s') - n_{p'}^{K+1}(s', a_{s'})\right] \geq \frac{K}{4S_1}, \tag{C.59}$$

where $N_{p'}^{K+1}(s') = \sum_{a\in\mathcal{A}} n_{p'}^{K+1}(s',a)$. To see this, by Lemma C.21, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([v]^{S_1\times M}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\
\geq\ & \sum_{p=1}^{M}\sum_{s'\in\mathcal{S}_1}(H-1)\Delta\cdot\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([v]^{S_1\times M}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N_p^{K+1}(s') - n_p^{K+1}(s',a_{s'})\right] \\
\geq\ & \frac{H-1}{4}MK\sqrt{\frac{vS_1}{384K}} \\
\geq\ & \frac{1}{160}M\sqrt{H^2(vS_1)K} \\
\geq\ & \frac{1}{240}M\sqrt{H^2l^CK},
\end{aligned}$$

where the last inequality uses the fact that $vS_1 \geq \frac{1}{2}l^C$.

Without loss of generality, let us choose $s'=1$ and $p'=1$. Similar to case 1, we define a set of helper problem instances: for any $(a_{1,2},\ldots,a_{S_1,M})\in[v]^{S_1\times M-1}$, we define a problem instance $\mathfrak{M}(0,a_{1,2},\ldots,a_{S_1,M})$ such that it agrees with $\mathfrak{M}(a_{1,1},a_{1,2},\ldots,a_{S_1,M})$ on everything but $\mathbb{P}_1(\cdot\mid 1,a_1)$, namely, in $\mathfrak{M}(0,a_{1,2},\ldots,a_{S_1,M})$, $\mathbb{P}_1(S_1+1\mid 1,a_1)=\frac{1}{2}$.

For each $(j,a_{1,2},\ldots,a_{S_1,M})\in\left([0]\cup[v]\right)\times[v]^{S_1\times M-1}$, let $\mathbb{P}_{j,a_{1,2},\ldots,a_{S_1,M}}$ denote the probability measure on the outcomes of running Alg on the instance $\mathfrak{M}(j,a_{1,2},\ldots,a_{S_1,M})$. Further, for each $j\in\{0\}\cup[v]$, we define

$$\mathbb{P}_j = \frac{1}{v^{S_1\times M-1}}\sum_{a_{1,2},\ldots,a_{S_1,M}\in[v]^{S_1\times M-1}}\mathbb{P}_{j,a_{1,2},\ldots,a_{S_1,M}};$$

and we use $\mathbb{E}_j$ to denote the expectation with respect to $\mathbb{P}_j$. In subsequent calculations, for any $m\in\left([0]\cup[v]\right)\times[v]^{S_1\times M-1}$, we also denote by $\mathbb{P}_m\left(\cdot\mid N_1^{K+1}(1)\right)$ and $\mathbb{E}_m\left[\cdot\mid N_1^{K+1}(1)\right]$ the probability and expectation conditional on a realization of $N_1^{K+1}(1)$ under $\mathbb{P}_m$. Similar to case 1, it can be shown that, for any $j\in\{0\}\cup[v]$,

$$\mathbb{P}_j(\cdot \mid N^{K+1}(1)) = \frac{1}{v^{S_1 \times M-1}} \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \mathbb{P}_{j,a_{1,2},\ldots,a_{S_1,M}} \left(\cdot \mid N^{K+1}(1)\right). \tag{C.60}$$

Now, for each $j \in [v]$, we have

$$\begin{aligned}
&\mathbb{E}_j \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right] - \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right] \\
\leq & N_1^{K+1}(1) \left\| \mathbb{P}_j \left(\cdot \mid N_1^{K+1}(1)\right) - \mathbb{P}_0 \left(\cdot \mid N_1^{K+1}(1)\right) \right\|_1 \\
\leq & N_1^{K+1}(1) \cdot \frac{1}{v^{S_1 \times M-1}} \cdot \qquad (C.61) \\
& \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \left\| \mathbb{P}_{j,a_{1,2},\ldots,a_{S_1,M}} \left(\cdot \mid N_1^{K+1}(1)\right) - \mathbb{P}_{0,a_{1,2},\ldots,a_{S_1,M}} \left(\cdot \mid N_1^{K+1}(1)\right) \right\|_1 \\
\leq & N_1^{K+1}(1) \cdot \frac{1}{v^{S_1 \times M-1}} \cdot \qquad (C.62) \\
& \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \sqrt{2\,\mathrm{KL}\left(\mathrm{Ber}(\frac{1}{2}+\Delta), \mathrm{Ber}(\frac{1}{2})\right) \mathbb{E}_{0,a_2,\ldots,a_{S_1}} \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]} \\
\leq & N_1^{K+1}(1) \cdot \frac{1}{v^{S_1 \times M-1}} \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \sqrt{6\Delta^2 \mathbb{E}_{0,a_2,\ldots,a_{S_1}} \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]} \\
\leq & N_1^{K+1}(1) \cdot \sqrt{\frac{6vS_1}{384MK} \cdot \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]} \\
= & \frac{1}{8} N_1^{K+1}(1) \sqrt{\frac{vS_1}{MK} \cdot \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]}. \qquad (C.63)
\end{aligned}$$

where the first inequality is based on Lemma C.16 and the fact that, conditional on $N_1^{K+1}(1)$, $n_1^{K+1}(1,j)$ has distribution supported on $[0, N_1^{K+1}(1)]$; the second inequality follows from Equation (C.60) and the triangle inequality; the third inequality uses Pinsker's inequality and Lemma C.22 (the divergence decomposition lemma); the fourth inequality uses Lemma C.24 and the fact that $\Delta \leq \frac{1}{4}$; and the last inequality follows from Jensen's inequality.

Using arguments similar to the ones shown for case 1, we have that

$$\begin{aligned}
&\frac{1}{v}\sum_{j\in[v]}\mathbb{E}_j\left[n_1^{K+1}(1,j)\right]\\
\leq&\frac{1}{v}\mathbb{E}_0\left[n_1^{K+1}(1,j)\right]+\mathbb{E}_0\left[\frac{1}{8}N_1^{K+1}(1)\sqrt{\frac{vS_1}{K}\cdot\frac{1}{v}\sum_{j\in[v]}\mathbb{E}_0\left[n_1^{K+1}(1,j)\mid N_1^{K+1}(1)\right]}\right]\\
\leq&\frac{1}{v}\mathbb{E}_0\left[N^{K+1}(1)\right]+\frac{1}{8}\sqrt{\frac{S_1}{K}}\cdot\mathbb{E}_0\left[\left(N_1^{K+1}(1)\right)^{\frac{3}{2}}\right]\\
\leq&\frac{1}{v}\cdot\frac{K}{S_1}+\frac{1}{4}\sqrt{\frac{S_1}{K}\left(\frac{K}{S_1}\right)^3}\\
\leq&\frac{3K}{4S_1},
\end{aligned}$$

where the second to last inequality is from Lemma C.25 under the assumption that $K \geq S_1$, and the last inequality uses the fact that $v \geq 2$.

It then follows that

$$\frac{1}{v}\sum_{j\in[v]}\mathbb{E}_j\left[N_1^{K+1}(1)-n_1^{K+1}(1,j)\right]\geq\frac{1}{v}\sum_{j\in[v]}\mathbb{E}_j\left[N_1^{K+1}(1)\right]-\frac{K}{4S_1}=\frac{K}{4S_1},$$

and we thereby have shown that

$$\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([v]^{S_1\times M}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N_1^{K+1}(1)-n_1^{K+1}(1,a_1)\right]\geq\frac{K}{4S_1}.$$

□

C.4.3 Gap dependent lower bound

Theorem C.27 (Restatement of Theorem 4.8)**.** *Fix $\epsilon \geq 0$. For any $S \in \mathbb{N}$, $A \geq 2$, $H \geq 2$, $M \in \mathbb{N}$, such that $S \geq 2(H-1)$, let $S_1 = S - 2(H-1)$; and let $\left\{\Delta_{s,a,p}\right\}_{(s,a,p)\in[S_1]\times[A]\times[M]}$ be any set of values such that*

- *for every* $(s, a, p) \in [S_1] \times [A] \times [M]$, $\Delta_{s,a,p} \in [0, H/48]$*;*
- *for every* $(s, p) \in [S_1] \times [M]$*, there exists at least one action* $a \in [A]$ *such that* $\Delta_{s,a,p} = 0$*;*
- *and, for every* $(s, a) \in [S_1] \times [A]$ *and* $p, q \in [M]$, $\left|\Delta_{s,a,p} - \Delta_{s,a,q}\right| \leq \epsilon/4$.

There exists an ϵ*-MPERL problem instance with* S *states,* A *actions,* M *players and an episode length of* H*, such that* $\mathcal{S}_1 = [S_1]$, $|\mathcal{S}_h| = 2$ *for all* $h \geq 2$*, and*

$$\mathrm{gap}_p(s, a) = \Delta_{s,a,p}, \quad \forall (s, a, p) \in [S_1] \times [A] \times [M].$$

For this problem instance, any sublinear regret algorithm Alg *for the* ϵ*-MPERL problem must have regret at least*

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(\ln K \left(\sum_{p \in [M]} \sum_{\substack{(s,a) \in \mathcal{I}^C_{(\epsilon/192H)}: \\ \mathrm{gap}_p(s,a) > 0}} \frac{H^2}{\mathrm{gap}_p(s, a)} + \sum_{(s,a) \in \mathcal{I}_{(\epsilon/192H)}} \frac{H^2}{\min_p \mathrm{gap}_p(s, a)}\right)\right).$$

Proof. The construction and techniques in this proof are inspired by [136] and Appendix A.5.

Proof outline.

We will construct an ϵ-MPERL problem instance, $\mathfrak{M}$, and show that, for any sublinear regret algorithm and sufficiently large K, the following two claims are true:

1. for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ such that for all p, $\mathrm{gap}_p(s, a) > 0$,

$$\mathbb{E}_{\mathfrak{M}}\left[n^K(s, a)\right] \geq \Omega\left(\frac{H^2}{\left(\min_p \mathrm{gap}_p(s, a)\right)^2} \ln K\right); \tag{C.64}$$

2. for any $(s,a) \in \mathcal{I}^{C_\epsilon}_{\frac{\epsilon}{192H}}$ and any $p \in [M]$ such that $\mathrm{gap}_p(s,a) > 0$,

$$\mathbb{E}_{\mathfrak{M}}\left[n_p^K(s,a)\right] \geq \Omega\left(\frac{H^2}{\left(\mathrm{gap}_p(s,a)\right)^2}\ln K\right). \tag{C.65}$$

The rest then follows from Lemma C.21 (the regret decomposition lemma).

Construction of $\mathfrak{M}$.

Given any set of values $\left\{\Delta_{s,a,p}\right\}_{(s,a,p)\in[S_1]\times[A]\times[M]}$ that satisfies the assumptions in the theorem statement, we can construct a collection of MDPs $\left\{\mathcal{M}_p\right\}_{p=1}^M$, such that for each $p \in [M]$, $\mathcal{M}_p$ is as follows, and $\mathfrak{M} = \left\{\mathcal{M}_p\right\}_{p=1}^M$ is an ϵ-MPERL problem instance:

- $\mathcal{S}_1 = [S_1]$, and p_0 is a uniform distribution over the states in $\mathcal{S}_1$.
- For $h \in [2,H]$, $\mathcal{S}_h = \{S_1 + 2h - 3, S_1 + 2h - 2\}$.
- $\mathcal{A} = [A]$.
- For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution $r_p(s,a)$ is a Bernoulli distribution, $\mathrm{Ber}(R_p(s,a))$, and we specify $R_p(s,a)$ subsequently.
- For every $(s,a) \in \mathcal{S}_1 \times [A]$, set $\bar{\Delta}^p_{s,a} = \frac{\Delta_{s,a,p}}{H-1}$. Then, let

  $$\mathbb{P}_p\left(S_1 + 1 \mid s,a\right) = \frac{1}{2} - \bar{\Delta}^p_{s,a}, \quad \mathbb{P}_p\left(S_1 + 2 \mid s,a\right) = \frac{1}{2} + \bar{\Delta}^p_{s,a},$$

  and $R_p(s,a) = 0$. Since $\Delta_{s,a,p} \in [0, H/48]$, $\bar{\Delta}^p_{s,a} \leq \frac{H}{48(H-1)} \leq \frac{1}{24}$, where the last inequality follows from the assumption that $H \geq 2$. Therefore, $\mathbb{P}_p\left(S_1 + 1 \mid s,a\right) \in [0,1]$, and $\mathbb{P}_p\left(S_1 + 2 \mid s,a\right) \in [0,1]$.
- For $h \in [2,H]$, and $a \in [A]$, let

- $\mathbb{P}_p\left(S_1+2h-1 \mid S_1+2h-3, a\right)=1$, $\mathbb{P}_p\left(S_1+2h \mid S_1+2h-3, a\right)=0$, and $R_p(S_1+2h-3, a)=1$.
- $\mathbb{P}_p\left(S_1+2h \mid S_1+2h-2, a\right)=0$, $\mathbb{P}_p\left(S_1+2h-1 \mid S_1+2h-2, a\right)=1$, and $R_p(S_1+2h-2, a)=0$.

By the assumption that for every $(s,p) \in [S_1] \times [M]$, there exists at least one action $a \in [A]$ such that $\Delta_{s,a,p}=0$, we have that there is at least one action a such that $\bar{\Delta}^p_{s,a}=0$. We verify that for every $(s,a,p) \in [S_1] \times [A] \times [M]$,

$$\begin{aligned}
\operatorname{gap}_p(s,a) &= V_p^\star(s) - Q_p^\star(s,a) \\
&= \max_{a'} Q_p^\star(s,a') - Q_p^\star(s,a) \\
&= (H-1)\bar{\Delta}^p_{s,a} \\
&= \Delta_{s,a,p}.
\end{aligned}$$

We now verify that the above MPERL problem instance $\mathfrak{M}=\left\{\mathcal{M}_p\right\}_{p=1}^M$ is an ϵ-MPERL problem instance:

1. The reward distributions are the same for all players, namely, for all p, q,

$$\left|R_p(s,a)-R_q(s,a)\right|=0 \leq \epsilon, \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

2. Further, by the assumption that for every $(s,a) \in [S_1] \times [A]$ and $p, q \in [M]$, $\left|\Delta_{s,a,p}-\Delta_{s,a,q}\right| \leq \epsilon/4$, we have that

$$\left|\bar{\Delta}^p_{s,a}-\bar{\Delta}^q_{s,a}\right|=\frac{\left|\Delta_{s,a,p}-\Delta_{s,a,q}\right|}{H-1} \leq \frac{\epsilon}{4(H-1)} \leq \frac{\epsilon}{2H}.$$

It then follows that

$$\|\mathbb{P}_p\left(\cdot \mid s,a\right) - \mathbb{P}_q\left(\cdot \mid s,a\right)\|_1 = 2\left|\bar{\Delta}^p_{s,a} - \bar{\Delta}^q_{s,a}\right| \leq \frac{\epsilon}{H}.$$

Meanwhile, for every $(s,a) \in \left(\mathcal{S} \setminus \mathcal{S}_1\right) \times \mathcal{A}$

$$\|\mathbb{P}_p\left(\cdot \mid s,a\right) - \mathbb{P}_q\left(\cdot \mid s,a\right)\|_1 = 0 \leq \frac{\epsilon}{H}.$$

In summary, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\|\mathbb{P}_p\left(\cdot \mid s,a\right) - \mathbb{P}_q\left(\cdot \mid s,a\right)\|_1 \leq \frac{\epsilon}{H}.$$

We are now ready to prove the two claims.

1. **Proving claim 1 (Equation** (C.64)**):**

   Fix any $(s_0,a_0) \in [S_1] \times [A]$ such that $\bar{\Delta}^{\min}_{s_0,a_0} = \min_p \bar{\Delta}^p_{s_0,a_0} > 0$. It can be easily observed that $\text{gap}_p(s_0,a_0) > 0$ for all p. Define $p_0 = \text{argmin}_p \bar{\Delta}^p_{s_0,a_0}$. We can construct a new problem instance, $\mathfrak{M}'$, which agrees with $\mathfrak{M}$, except that, $\forall p \in [M]$,

   $$\mathbb{P}_p\left(S_1+1 \mid s_0,a_0\right) = \frac{1}{2} - \bar{\Delta}^p_{s_0,a_0} + 2\bar{\Delta}^{\min}_{s_0,a_0}, \mathbb{P}_p\left(S_1+2 \mid s_0,a_0\right) = \frac{1}{2} + \bar{\Delta}^p_{s_0,a_0} - 2\bar{\Delta}^{\min}_{s_0,a_0}.$$

   $\mathfrak{M}'$ is an ϵ-MPERL problem instance. To see this, we note that the only change is in $\mathbb{P}_p\left(\cdot \mid s_0,a_0\right)$ for all $p \in [M]$. In this new instance, it is still true that for every $p,q \in [M]$,

   $$\|\mathbb{P}_p\left(\cdot \mid s_0,a_0\right) - \mathbb{P}_q\left(\cdot \mid s_0,a_0\right)\|_1 = 2\left|\bar{\Delta}^p_{s_0,a_0} - \bar{\Delta}^q_{s_0,a_0}\right| \leq \frac{\epsilon}{H}.$$

   Fix any sublinear regret algorithm Alg for the ϵ-MPERL problem. By Lemma C.22

(the divergence decomposition lemma), we have

$$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) = \sum_{p=1}^{M} \mathbb{E}_{\mathfrak{M}}\left[n_p^K(s_0, a_0)\right] \mathrm{KL}\left(\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s_0, a_0), \mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s_0, a_0)\right),$$

where $\mathbb{P}_{\mathfrak{M}}$ and $\mathbb{P}_{\mathfrak{M}'}$ are the probability measures on the outcomes of running Alg on $\mathfrak{M}$ and $\mathfrak{M}'$, respectively; $\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s_0, a_0)$, $\mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s_0, a_0)$ are the transition probabilities for (s_0, a_0) and player p in $\mathfrak{M}$ and $\mathfrak{M}'$, respectively.

We observe that, for any $p \in [M]$,

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s_0, a_0), \mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s_0, a_0)\right) \\
=&\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2} - \bar{\Delta}_{s_0,a_0}^{p}\right), \mathrm{Ber}\left(\frac{1}{2} - \bar{\Delta}_{s_0,a_0}^{p} + 2\bar{\Delta}_{s_0,a_0}^{\min}\right)\right) \\
\leq& 12(\bar{\Delta}_{s_0,a_0}^{\min})^2,
\end{aligned}$$

where the last inequality follows from Lemma C.24 and the assumption that $\Delta_{s,a,p} \leq \frac{H}{48}$. In addition, $\sum_{p=1}^{M} \mathbb{E}_{\mathfrak{M}}\left[n_p^K(s_0, a_0)\right] = \mathbb{E}_{\mathfrak{M}}\left[n^K(s_0, a_0)\right]$. It then follows that

$$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) \leq 12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0, a_0)\right](\bar{\Delta}_{s_0,a_0}^{\min})^2. \tag{C.66}$$

Now, in the original ϵ-MPERL problem instance, $\mathfrak{M}$, by Equation (C.52) and Markov's Inequality, we have

$$\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}_{s_0,a_0}^{\min}\right)\mathbb{P}_{\mathfrak{M}}\left(n_{p_0}^K(s_0, a_0) \geq \frac{K}{4S_1}\right);$$

where we note that $\bar{\Delta}_{s_0,a_0}^{p_0} = \bar{\Delta}_{s_0,a_0}^{\min}$. In $\mathfrak{M}'$, the new ϵ-MPERL problem instance, we

have

$$\begin{aligned}
\mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] &\geq \left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{E}_{\mathfrak{M}'}\left[\sum_{a\neq a_0} n_{p_0}(s_0,a)\right] \\
&= \left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{E}_{\mathfrak{M}'}\left[N^K_{p_0}(s_0) - n_{p_0}(s_0,a_0)\right] \\
&\geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}'}\left(N^K_{p_0}(s_0) - n_{p_0}(s_0,a_0) \geq \frac{K}{4S_1}\right) \\
&\geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}'}\left(N^K_{p_0}(s_0) \geq \frac{K}{2S_1}, n_{p_0}(s_0,a_0) \leq \frac{K}{4S_1}\right) \\
&\geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\left(\mathbb{P}_{\mathfrak{M}'}\left(n_{p_0}(s_0,a_0) \leq \frac{K}{4S_1}\right) - \exp(-\frac{K}{8S_1})\right),
\end{aligned}$$

where the first inequality is by Equation (C.52); the second inequality is by Markov's Inequality; the third inequality is by simple algebra; and the last inequality is by Chernoff bound that $\mathbb{P}_{\mathfrak{M}'}\left(N^K_{p_0}(s_0) < \frac{K}{2S_1}\right) \leq \exp(-\frac{K}{8S_1})$, and $\mathbb{P}(A\cap B) \geq \mathbb{P}(B) - \mathbb{P}(A^C)$ for events A, B.

It then follows that

$$\begin{aligned}
&\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] + \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\
=&\frac{K}{2}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\cdot \\
&\qquad\left(\mathbb{P}_{\mathfrak{M}}\left(n^K_{p_0}(s_0,a_0) \geq \frac{K}{2}\right) + \mathbb{P}_{\mathfrak{M}'}\left(n^K_{p_0}(s_0,a_0) < \frac{K}{2}\right) - \exp(-\frac{K}{8S_1})\right) \\
\geq&\frac{K}{2}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\left(\frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathfrak{M}},\mathbb{P}_{\mathfrak{M}'})\right) - \exp(-\frac{K}{8S_1})\right) \\
\geq&\frac{K}{2}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\left(\frac{1}{2}\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{\min}_{s_0,a_0})^2\right) - \exp(-\frac{K}{8S_1})\right),
\end{aligned}$$

where the first inequality follows from Lemma C.23 (the Bretagnolle-Huber inequality), and the second inequality follows from Eq. (C.66). Observe that $\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right] \leq \frac{K}{S_1}$; in addition, by our assumption that $\Delta_{s,a,p} \leq \frac{H}{48}$ for every (s,a,p), we have $\bar{\Delta}^{\min}_{s_0,a_0} \leq$

$\frac{1}{24}$. These together implies that $\frac{1}{4}\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{\min}_{s_0,a_0})^2\right) \geq \exp(-\frac{K}{8S_1})$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] + \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\
&\geq \frac{K}{2}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\cdot\frac{1}{4}\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{\min}_{s_0,a_0})^2\right).
\end{aligned}$$

Now, under the assumption that Alg is a sublinear regret algorithm, we have

$$\frac{K}{8}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{\min}_{s_0,a_0})^2\right) \leq 2CK^{\alpha}.$$

It follows that

$$\begin{aligned}
\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right] &\geq \frac{1}{12\left(\bar{\Delta}^{\min}_{s_0,a_0}\right)^2}\ln\left(\frac{(H-1)\bar{\Delta}^{\min}_{s_0,a_0}K^{1-\alpha}}{16C}\right) \\
&= \frac{(H-1)^2}{12\left(\min_p \mathrm{gap}_p(s_0,a_0)\right)^2}\ln\left(\frac{\min_p \mathrm{gap}_p(s_0,a_0)K^{1-\alpha}}{16C}\right) \\
&\geq \frac{H^2}{24\left(\min_p \mathrm{gap}_p(s_0,a_0)\right)^2}\ln\left(\frac{\min_p \mathrm{gap}_p(s_0,a_0)K^{1-\alpha}}{16C}\right).
\end{aligned}$$

We then have

$$\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right] \geq \Omega\left(\frac{H^2}{\left(\min_p \mathrm{gap}_p(s_0,a_0)\right)^2}\ln K\right).$$

2. **Proving Claim 2 (Equation** (C.65)**):**

Fix any $(s_0,a_0) \in \mathcal{I}^{C_\epsilon}_{\frac{\epsilon}{192H}}$ and $p_0 \in [M]$ such that $\bar{\Delta}^{p_0}_{(s_0,a_0)} > 0$, which means that

$\text{gap}_{p_0}(s_0, a_0) > 0$. We have that for all $p \in [M]$,

$$\bar{\Delta}^p_{s_0,a_0} = \frac{\Delta^p_{s_0,a_0}}{H-1} = \frac{\text{gap}_p(s_0,a_0)}{H-1} \leq \frac{24H(\epsilon/(192H))}{(H-1)} \leq \frac{\epsilon}{8(H-1)} \leq \frac{\epsilon}{4H}. \tag{C.67}$$

We can construct a new problem instance, $\mathfrak{M}'$, which agrees with $\mathfrak{M}$ except that

$$\begin{aligned}
\mathbb{P}_{p_0}\left(S_1 + 1 \mid s_0, a_0\right) &= \frac{1}{2} - \bar{\Delta}^{p_0}_{s_0,a_0} + 2\bar{\Delta}^{p_0}_{s_0,a_0} = \frac{1}{2} + \bar{\Delta}^{p_0}_{s_0,a_0}, \\
\mathbb{P}_{p_0}\left(S_1 + 2 \mid s_0, a_0\right) &= \frac{1}{2} + \bar{\Delta}^{p_0}_{s_0,a_0} - 2\bar{\Delta}^{p_0}_{s_0,a_0} = \frac{1}{2} - \bar{\Delta}^{p_0}_{s_0,a_0}.
\end{aligned}$$

$\mathfrak{M}'$ is an ϵ-MPERL problem instance. To see this, we note that the only change is in $\mathbb{P}_{p_0}\left(\cdot \mid s_0, a_0\right)$. In this new instance, it is still true that for any $q \neq p_0$,

$$\|\mathbb{P}_{p_0}(\cdot \mid s_0, a_0) - \mathbb{P}_q(\cdot \mid s_0, a_0)\|_1 \leq 2\left|\bar{\Delta}^{p_0}_{s_0,a_0} + \bar{\Delta}^{q}_{s_0,a_0}\right| \leq \frac{\epsilon}{H}.$$

where the last inequality uses Equation (C.67) that $\bar{\Delta}^p_{s_0,a_0} \leq \frac{\epsilon}{4H}$ for every $p \in [M]$.

Fix any sublinear regret algorithm Alg. By Lemma C.22 (the divergence decomposition lemma), we have

$$\text{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) = \mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] \text{KL}\left(\mathbb{P}^{\mathfrak{M}}_{p_0}(\cdot \mid s_0, a_0), \mathbb{P}^{\mathfrak{M}'}_{p_0}(\cdot \mid s_0, a_0)\right).$$

Using a similar reasoning as before, we can show that

$$\text{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) \leq 12\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] (\bar{\Delta}^{p_0}_{s_0,a_0})^2. \tag{C.68}$$

Similar to case 1, we have the following argument. In the original ϵ-MPERL problem instance, $\mathfrak{M}$, we have $\mathbb{E}_{\mathfrak{M}}\left[\text{Reg}_{\text{Alg}}(K)\right] \geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}}\left(n^K_{p_0}(s_0,a_0) \geq \frac{K}{4S_1}\right)$;

and in $\mathfrak{M}'$, the new ϵ-MPERL problem instance, we have

$$\mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\left(\mathbb{P}_{\mathfrak{M}'}\left(n^K_{p_0}(s_0,a_0) < \frac{K}{4S_1}\right) - \exp(-\frac{K}{8S_1})\right).$$

It then follows that

$$\begin{aligned}
&\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] + \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\
\geq &\frac{K}{2}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\left(\frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathfrak{M}},\mathbb{P}_{\mathfrak{M}'})\right) - \exp(-\frac{K}{8S_1})\right) \\
\geq &\frac{K}{8}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{p_0}_{s_0,a_0})^2\right).
\end{aligned}$$

Now, under the assumption that Alg is a sublinear regret algorithm, we have

$$\frac{K}{8}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right](\bar{\Delta}^{p_0}_{s_0,a_0})^2\right) \leq 2CK^{\alpha}.$$

It follows that

$$\begin{aligned}
\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] &\geq \frac{1}{12\left(\bar{\Delta}^{p_0}_{s_0,a_0}\right)^2}\ln\left(\frac{(H-1)\bar{\Delta}^{p_0}_{s_0,a_0}K^{1-\alpha}}{16C}\right) \\
&\geq \frac{H^2}{24\left(\mathrm{gap}_{p_0}(s_0,a_0)\right)^2}\ln\left(\frac{\mathrm{gap}_{p_0}(s_0,a_0)K^{1-\alpha}}{16C}\right).
\end{aligned}$$

We then have that

$$\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] \geq \Omega\left(\frac{H^2}{\left(\mathrm{gap}_{p_0}(s_0,a_0)\right)^2}\ln K\right).$$

Combing the two claims:

We note that in $\mathfrak{M}$, for any $(s,a,p) \in \left(\mathcal{S} \setminus \mathcal{S}_1\right) \times \mathcal{A} \times [M]$, $\text{gap}_p(s,a) = 0$. It then follows from Lemma C.21 (the regret decomposition lemma) and the fact that for any $(s,a,p) \in \mathcal{I}_{\epsilon/192H} \times [M], \text{gap}_p(s,a) > 0$, that

$$\begin{aligned}
&\mathbb{E}\left[\text{Reg}_{\text{Alg}}(K)\right] \\
&\geq \sum_{p=1}^{M} \sum_{(s,a) \in \mathcal{S}_1 \times \mathcal{A}} \mathbb{E}\left[n_p^K(s,a)\right] \text{gap}_p(s,a) \\
&\geq \Omega\left(\ln K \left(\sum_{p \in [M]} \sum_{\substack{(s,a) \in \mathcal{I}^C_{\epsilon/192H}: \\ \text{gap}_p(s,a) > 0}} \frac{H^2}{\text{gap}_p(s,a)} + \sum_{(s,a) \in \mathcal{I}_{\epsilon/192H}} \frac{H^2}{\min_p \text{gap}_p(s,a)}\right)\right).
\end{aligned}$$

□

Appendix D

Supplementary Material for Chapter 5

D.1 Related Work

There is a rich literature on metric learning; see [84] for a survey. In this chapter, we focus on learning Mahalanobis distances from relative comparisons that involve triplets of items, in the form of "is u closer to x or x'?" [128, 152, 106]. In particular, we study metric learning from preference comparisons in the ideal point model [38]: a preference comparison is a special type of triplet comparison, where the comparator u is latent and represents a user's ideal item. If the ideal points are known beforehand, one can simply treat this as a problem of metric learning from triplet comparisons. Conversely, if the metric is known, one can also localize user ideal points using techniques from [69, 107, 144].

This chapter builds upon recent research that studies simultaneous metric and preference learning [167, 28]. In a single user setting, [167] developed an algorithm that iteratively alternate between estimating the metric and the user ideal point. [28] generalized the setting to involve multiple users. They established identifiability guarantees when users provide unquantized measurements, and presented generalization bounds and recovery guarantees when users provide binary responses. While [28] showed that it is possible to jointly recover a metric and user ideal points when each user answers $\Theta(d)$ queries, we address the fundamental question of learning Mahalanobis distances when we have a much limited budget of $o(d)$ preference comparisons per user. The $o(d)$ budget is more realistic

especially when items are embedded in higher dimensions, but also poses interesting new challenges as learning user ideal points is no longer possible.

Several other works in the broader literature are related. For example, learning ordinal embeddings or kernel functions from triplet comparisons has been well studied. [142] developed an active multi-dimensional scaling algorithm to learn item embeddings, with the goal of capturing item similarities perceived by humans. See also [151, 67, 77], among other works. [64] introduced a collaborative metric learning algorithm, which uses matrix factorization to learn user and item embeddings such that the Euclidean distance reflects user preferences and item/user similarities. A divide-and-conquer approach for deep metric learning has been studied by [126], who use k-means to cluster items and learn separate metrics for each cluster before concatenating them together; they performed an extensive empirical study based on image data. In this chapter, we consider the ideal point model to study the fundamental problem of metric learning from limited preference comparisons.

D.2 Additional Algorithms from Existing Work

Algorithm 7 and Algorithm 8 describe the procedures for learning an unknown Mahalanobis distance using unquantized measurements from a single user and a large pool of users, respectively. See Section 2 of [28].

Algorithm 9 describes the convex optimization problem introduced in Section 3 of [28] for simultaneous metric and preference learning using quantized measurements from multiple users. Here, $\ell : \mathbb{R} \to \mathbb{R}_{0+}$ can be any convex loss function that is L-Lipschitz-continuous. In particular, to achieve the recovery guarantee in Proposition 5.18, we assume the probabilistic model in Assumption 5.17 with link function f and use the loss function $\ell(z) = -\log f(z)$.

Algorithm 7: Metric learning using unquantized measurements from a single user [28]

Input: A set $\mathcal{D} = \left\{(x_{i_0}, x_{i_1}, \psi_i)\right\}_{i=1}^{m}$ of unquantized measurements from a single user.

1 Solve the system of linear equations over symmetric matrices $A \in \mathbb{R}^{d\times d}$ and vectors $w \in \mathbb{R}^d$:

$$\left\langle x_{i_0}x_{i_0}^\top - x_{i_1}x_{i_1}^\top,\, A\right\rangle + \left\langle x_{i_0} - x_{i_1}, w\right\rangle = \psi_i.$$

Output: $\hat{A}$, the solution to the above linear equations.

Algorithm 8: Metric learning using unquantized measurements from multiple users [28]

Input: A family of $\mathcal{D}_k = \left\{(x_{i_0;k}, x_{i_1;k}, \psi_{i;k})\right\}_{i=1}^{m_k}$ of unquantized measurements from users $k \in [K]$.

1 Solve the system of linear equations over symmetric matrices $A \in \mathbb{R}^{d\times d}$ and vectors $w_1, w_2, \ldots, w_K \in \mathbb{R}^d$:

$$\left\langle x_{i_0;k}x_{i_0;k}^\top - x_{i_1;k}x_{i_1;k}^\top,\, A\right\rangle + \left\langle x_{i_0;k} - x_{i_1;k}, w_k\right\rangle = \psi_{i;k}.$$

Output: $\hat{A}$, the solution to the above linear equations.

D.3 Direct Sums of Inner Product Spaces

In the paper, we've liberally made use of direct sums of inner product spaces, for example, $\mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d$, which we treat as an inner product space. It allows us ready access to well-established machinery including inner products, norms, singular values, and pseudoinverses. The direct sum of inner product spaces is defined:

Definition D.1. *Let* $\big(V, \langle\cdot,\cdot\rangle_V\big)$ *and* $\big(W, \langle\cdot,\cdot\rangle_W\big)$ *be two inner product spaces. Their* direct sum *is the vector space* $V \oplus W$ *equipped with the inner product:*

$$\langle v_1 \oplus w_1, v_2 \oplus w_2\rangle_{V\oplus W} = \langle v_1, v_2\rangle_V + \langle w_1, w_2\rangle_W.$$

In particular, this induces the norm on $V \oplus W$ *satisfying* $\|v \oplus w\|_{V\oplus W}^2 = \|v\|_V^2 + \|w\|_W^2$.

Algorithm 9: Metric learning using quantized measurements from multiple users [28]

Input: A family of $\mathcal{D}_k = \left\{(x_{i_0;k}, x_{i_1;k}, y_{i;k})\right\}_{i=1}^{m_k}$ of quantized measurements from users $k \in [K]$; hyperparameters $\zeta_M, \zeta_v > 0$.

1 Solve the convex optimization problem over symmetric matrices $A \in \mathbb{R}^{d\times d}$ and vectors $w_1, w_2, \dots w_K \in \mathbb{R}^d$:

$$\hat{M}, \{\hat{v}_k\}_k \leftarrow \min_{A,\{w_k\}_k} \sum_k \sum_{\mathcal{D}_k} \ell\left(y_{i;k}\left(\left\langle x_{i_0}x_{i_0}^\top - x_{i_1;k}x_{i_1;k}^\top, A\right\rangle + \left\langle x_{i_0;k} - x_{i_1;k}, w_k\right\rangle\right)\right) \tag{D.1}$$

$$\text{s.t.} \quad A \succeq 0,\ \|A\|_{\mathrm{F}} \le \zeta_M,\ \|w_k\|_2 \le \zeta_v\ \forall k$$

Output: $\hat{M}$.

Moore-Penrose pseudoinverse.

The pseudoinverse can be defined for any map between inner product spaces:

Definition D.2. *Let $A : V \to W$ be a linear map between inner product spaces V and W. Let $K = \ker(A)$ and let $K^\perp$ be its orthogonal complement. Let $A_{K^\perp} : K^\perp \to \mathrm{Im}(A)$ be the restriction of A to $K^\perp$ and let $\Pi_{\mathrm{Im}(A)} : W \to \mathrm{Im}(A)$ be the orthogonal projection onto* $\mathrm{Im}(A)$. *The* Moore-Penrose pseudoinverse *of A is the map $A^+ : W \to V$ given by:*

$$A^+ = A_{K^\perp}^{-1} \circ \Pi_{\mathrm{Im}(A)}.$$

Note that $A_{K^\perp}^{-1}$ exists by the first isomorphism theorem of algebra.

Universal property.

The following property of direct sum allows us to decompose a linear map $A : V_1 \oplus V_2 \to V$, which we use in the proof of Theorem 5.15, when decomposing $\Pi^+ : \bigoplus_\lambda \mathrm{Sym}(V_\lambda) \to \mathrm{Sym}(\mathbb{R}^d)$.

Proposition D.3 (Universal property of the direct sum, [104])**.** *Let $A : V_1 \oplus V_2 \to V$ be a*

linear map. Then, there exists $A_i : V_i \to V$ *for* $i = 1, 2$ *such that for all* $v_1 \oplus v_2 \in V_1 \oplus V_2$,

$$A(v_1 \oplus v_2) = A_1(v_1) + A_2(v_2).$$

Schatten norm.

The Frobenius norm over matrices can be generalized to linear maps between inner product spaces:

Definition D.4. *Let* $A : V \to W$ *be a linear map between finite-dimensional inner product spaces of rank* r. *Let* $\sigma_1 \geq \cdots \geq \sigma_r$ *be its nonzero singular values. The* 2-Schatten norm $\|A\|_2$ *is given by:*

$$\|A\|_2^2 = \sum_{i=1}^{r} \sigma_i^2.$$

In particular, this implies $\|A\|_2 \leq \sigma_{\max}(A) \cdot \sqrt{\operatorname{rank}(A)}$.

Proposition D.5. *Let* $A : V_1 \oplus V_2 \to V$ *be a linear map between finite-dimensional inner product spaces. Let* $A_i : V_i \to V$ *for* $i = 1, 2$ *be given as in Proposition D.3. Then:*

$$\|A\|_2^2 = \|A_1\|_2^2 + \|A_2\|_2^2,$$

where $\|\cdot\|_2$ *denotes the 2-Schatten norm.*

D.4 Proofs and Additional Results for Section 5.3

For the proof of Theorem 5.3, we will make use of the notion of a *comparison graph* over a set of items. Given preference comparisons from a user, the induced comparison graph is simply the directed graph over items where two items are connected by an edge if the user has compared them:

Definition D.6. *A* comparison graph $G = (V, E)$ *is a graph whose vertices* $V = \{x_1, \ldots, x_N\}$ *is a set of items and whose edges* $E = \{(x_{i_0}, x_{i_1})\}_{i=1}^{m}$ *is a set of item pairs.*

Its edge-vertex incidence matrix $S \in \{-1, 0, +1\}^{m \times N}$ *is defined by:*

$$S_{ij} = \begin{cases} 1 & j = i_0 \\ -1 & j = i_1 \\ 0 & o.w. \end{cases}$$

Theorem 5.3. *Fix $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ and $v_k \in \mathbb{R}^d$ for each $k \in \mathbb{N}$. Let $(D_k)_{k\in\mathbb{N}}$ be a collection of design matrices, each for a set of $m \leq d$ pairwise comparisons. If each set of compared items has generic pairwise relations, then for all $M' \in \mathrm{Sym}^+(\mathbb{R}^d)$, there exists $(v'_k)_{k\in\mathbb{N}} \subset \mathbb{R}^d$ such that:*

$$D_k(M, v_k) = D_k(M', v'_k), \qquad \forall k \in \mathbb{N}.$$

Proof of Theorem 5.3. Fix $M' \in \mathrm{Sym}^+(\mathbb{R}^d)$. It suffices to prove the result for a single user, since the covariates v_k's impose no constraints on each other. Fix a pseudo-ideal point $v \in \mathbb{R}^d$. Let D be a design matrix induced by the collection of pairs $\{(x_{i_0}, x_{i_1})\}_{i=1}^m$ from a set of items $\mathcal{X} = \{x_1, \ldots, x_N\}$. We show that when $\mathcal{X}$ has generic pairwise relations, then there exists $v' \in \mathbb{R}^d$ such $D(M, v) = D(M', v')$. By expanding and rearranging this equation, we obtain a linear system of equations $Av' = b$, where $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$, and where the ith set of equations is given by:

$$\underbrace{(x_{i_0} - x_{i_1})^\top}_{i\text{th row of } A} v' = \underbrace{\left\langle x_{i_0}x_{i_0}^\top - x_{i_1}x_{i_1}^\top, M - M' \right\rangle + \left\langle x_{i_0} - x_{i_1}, v \right\rangle}_{i\text{th entry of } b}.$$

The Rouché–Capelli theorem states that the system $A\hat{u} = b$ has a solution if the rank of the augmented matrix $[A|b]$ is equal to the rank of the design matrix A. If this is the case, then, there is a solution v' for any choice of $M' \in \mathrm{Sym}^+(\mathbb{R}^d)$.

To finish the proof, we show that the ranks of A and $[A|b]$ are equal. To this end,

let S be the edge-vertex incidence matrix induced by $\{(x_{i_0}, x_{i_1})\}_{i=1}^m$. Define the matrix $X \in \mathbb{R}^{N \times d}$ and vector $b' \in \mathbb{R}^N$ so that the jth row of each is:

$$X_j = x_j^\top \qquad \text{and} \qquad b'_j = \langle x_j x_j^\top, M - M' \rangle + \langle x_j - v \rangle,$$

so that $A = SX$ and $b = Sb'$. The items have generic pairwise relations, and so $\text{rank}(S) = \text{rank}(SX)$ by Lemma D.7. The augmented matrix $[A|b]$ has the decomposition $S[X|b']$, so its rank is upper bounded by $\text{rank}(S)$. And because the $\text{rank}([A|b])$ is at least $\text{rank}(A)$, we obtain equality, as claimed. □

Lemma D.7. *Let $V = \{x_1, \ldots, x_N\}$ be a set of items in $\mathbb{R}^d$, and let $X \in \mathbb{R}^{N \times d}$ be its matrix representation, so that the jth row is $X_j = x_j^\top$. Let $G = (V, E)$ be a comparison graph with $|E| \leq d$. Let S be its edge-vertex incidence matrix. If the items have generic pairwise relations, then:*

$$\text{rank}(SX) = \text{rank}(S).$$

Proof. Let $G' = (V, E')$ be a maximal acyclic subgraph $G' \subset G$, say with m' edges, and let $S' \in \mathbb{R}^{m' \times N}$ be its corresponding edge-vertex incidence matrix. On the one hand, we have:

$$\text{rank}(S') \geq \text{rank}(S'X).$$

On the other, because $\mathcal{X}$ has pairwise generic relations and $m' \leq d$, we have:

$$\text{rank}(S'X) = \dim\big(\text{span}\big(\{x - x' : (x, x') \in E'\}\big)\big) = m' \geq \text{rank}(S').$$

The first equality is obtained by the definition of rank applied to $S'X$. The second equality follows from pairwise genericity. Thus, we have $\text{rank}(S') = \text{rank}(S'X) \leq \text{rank}(SX)$. Furthermore, we claim that:

$$\text{rank}(S) = \text{rank}(S').$$

It would follow that $\text{rank}(S) \leq \text{rank}(SX) \leq \text{rank}(S)$, which implies the result.

We prove the claim by showing that for any $e \in E \setminus E'$, the row S_e is a linear combination of rows $S_{e'}$ where $e' \in E'$. Let $e = (x, x')$. By the maximality of G', a cycle containing e is created by including e into G'. Thus, there is an undirected path P from x to x' in G', where $P = (x_0, \ldots, x_k)$ satisfies:

- $x_0 = x$ and $x_k = x'$,
- either (x_{i-1}, x_i) or its reversal (x_i, x_{i-1}) is contained in E'.

For each i, let $e_i \in E'$ be one of these edges (x_{i-1}, x_i) or (x_i, x_{i-1}) and let $r_i \in \{-1, +1\}$ indicate whether e_i was the reversal of (x_{i-1}, x_i). It follows that indeed S_e is a linear combination of the rows of S',

$$S_e = \sum_{i=1}^{k} r_i S_{e_i}.$$

□

D.4.1 Generic pairwise relations

In the next proposition, we show that our notion of *generic pairwise relations* is a notion of points being in general position [108]; almost all finite subsets of $\mathbb{R}^d$ are have pairwise generic relations. Recall:

Definition 5.2. *A set $\mathcal{X} \subset \mathbb{R}^d$ has* generic pairwise relations *if for any acyclic graph $G = (\mathcal{X}, E)$ with at most d edges, the set $\{x - x' : (x, x') \in E\}$ is linearly independent.*

Proposition D.8. *Fix $N \in \mathbb{N}$. We say that $X \in \mathbb{R}^{N \times d}$ has generic pairwise relations if its rows have generic pairwise relations. The following set has Lesbegue measure zero:*

$$\big\{X \in \mathbb{R}^{N \times d} : X \text{ is not pairwise generic}\big\}.$$

Proof. Let $\mathcal{S}$ be the finite collection of all edge-vertex incidence matrices S for acyclic comparison graphs with at most d edges on N items. Notice that if $X \in \mathbb{R}^{N\times d}$ is not pairwise generic, then there exists some $S \in \mathcal{S}$ such that $SX \in \mathbb{R}^{m\times d}$ is not full rank. It follows that:

$$\big\{X \text{ not pairwise generic}\big\} = \bigcup_{S\in\mathcal{S}} \big\{ \det(SXX^\top S^\top) = 0\big\}.$$

The zero set $\{\det(SXX^\top S^\top) = 0\}$ of a non-zero polynomial has Lebesgue measure zero, by Sard's theorem. The finite union of measure zero sets also has measure zero. □

The concept of *general linear position* is a standard notion of general position. We present the definition in a way to highlight its relationship to pairwise genericity. Recall that a star graph is a tree with a root vertex connected to all other vertices.

Definition D.9. *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$. We say that $\mathcal{X}$ is in* general linear position *if for any star graph $G = (V, E)$ with at most d edges on $V \subset \mathcal{X}$, the set $\{x - x' : (x, x') \in E\}$ is linearly independent.*

Because star graphs are acyclic graphs, the following is immediate:

Proposition D.10. *If $\mathcal{X}$ has generic pairwise relations, then $\mathcal{X}$ is in general linear position.*

On the other hand, the converse is not necessarily true. As we can see from the following example, having pairwise generic relations is a strictly stronger condition than being in general linear position.

Example D.11. *Consider the following points in $\mathbb{R}^2$:*

$$x_1 = \begin{bmatrix}0\\0\end{bmatrix} \qquad x_2 = \begin{bmatrix}1\\0\end{bmatrix} \qquad x_3 = \begin{bmatrix}1\\1\end{bmatrix} \qquad x_4 = \begin{bmatrix}2\\1\end{bmatrix}.$$

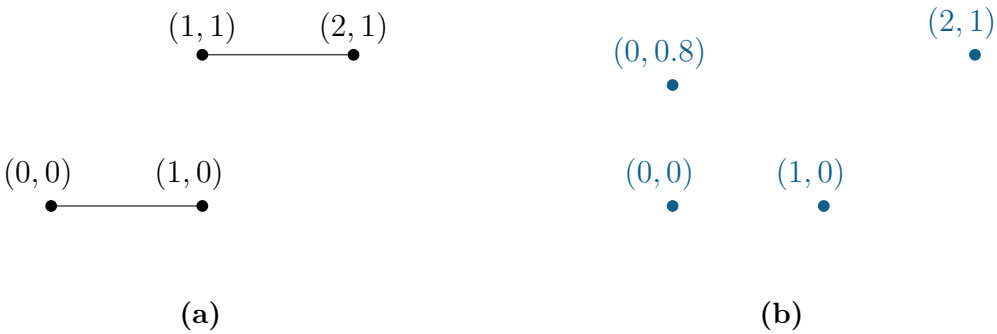


Figure D.1. (a) Illustration of Example D.11. The set of four points is in general linear position, but does not have generic pairwise relations. (b) A set of four points that has generic pairwise relations; it must also be in general linear position.

This collection of points is in general linear position, since no three points are collinear. However, these points do not have generic pairwise relations. We have:

$$x_2 - x_1 = x_4 - x_3.$$

D.5 Proofs and Additional Results for Section 5.4

D.5.1 An additional result for Section 5.4.1

Proposition D.12. *There is a one-to-one correspondence between* $\mathrm{Sym}^+(V)$ *and Mahalanobis distances on* V*. In particular,* $\rho_V : V \times V \to \mathbb{R}$ *is a Mahalanobis distance if and only if there exists some* $Q \in \mathrm{Sym}^+(V)$ *such that:*

$$\rho_V(x, x') = \sqrt{(x - x')BQB^\top(x - x')}.$$

Moreover, Q *is unique. We say that* Q *is the* matrix representation *of the Mahalanobis distance* ρ_V*. If* ρ_V *is the subspace metric on* V *of a Mahalanobis distance* ρ *on* $\mathbb{R}^d$ *with representation* $M \in \mathrm{Sym}^+(\mathbb{R}^d)$*, then:*

$$Q = \Pi_V(M).$$

Proof. ($\Longrightarrow$). Suppose that ρ_V is a Mahalanobis distance on V. We show that it has a representation in $\mathrm{Sym}^+(V)$. By definition, there exists a Mahalanobis distance ρ on $\mathbb{R}^d$ such that:

$$\rho_V = \rho\big|_V.$$

Let M be the matrix representation of ρ and let $Q = \Pi_V(M) \in \mathrm{Sym}^+(V)$. Then:

$$\begin{aligned}\rho_V(x, x') &= \sqrt{(x - x')^\top M (x - x')} \\ &= \sqrt{(x - x')^\top B B^\top M B B^\top (x - x')} \\ &= \sqrt{(x - x')^\top B Q B^\top (x - x')},\end{aligned}$$

where the first equality expands the equality $\rho_V(x, x') = \rho(x, x')$, the second uses the fact that $BB^\top x = x$ for all $x \in V$ since $B \in \mathbb{R}^{d \times r}$ is an orthonormal basis, and the third equality is uses the definition of Π_V.

To prove uniqueness, suppose that $Q, Q' \in \mathrm{Sym}^+(V)$ represent ρ_V. We claim that $Q = Q'$. To show this, it suffices to prove that for all $z \in \mathbb{R}^r$,

$$\left\langle Q - Q', zz^\top \right\rangle = 0.$$

This is because the collection $\{zz^\top : z \in \mathbb{R}^r\}$ spans all $(r \times r)$-symmetric matrices. To this end, fix $z \in \mathbb{R}^r$. We take $x, x' \in V \subset$ by setting $x = Bz$ and $x' = 0$. We have:

$$\rho_V(x, x') = \sqrt{(x - x')^\top B Q B^\top (x - x')} = \sqrt{z^\top Q z}.$$

The same equation holds for Q' since both represent ρ_V. Squaring both equations and taking their difference shows that $\langle Q - Q', zz^\top \rangle = 0$, as desired. Thus, $Q = Q'$ and the matrix representation of ρ_V is unique.

($\Longleftarrow$). Let $Q \in \mathrm{Sym}^+(V)$. We can extend the orthonormal basis B of V to an

orthonormal basis of $\mathbb{R}^d$. In particular, let $B_\perp \in \mathbb{R}^{d\times(d-r)}$ be an orthonormal basis of the orthogonal complement of V. Set:

$$M = B_\perp B_\perp^\top + BQB^\top,$$

so that $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ is positive-definite. Let ρ be the Mahalanobis distance on $\mathbb{R}^d$ represented by M. Then, the Mahalanobis distance $\rho\big|_V$ on V has representation:

$$\Pi_V(M) = B^\top MB = B^\top B_\perp B_\perp^\top B + B^\top BQB^\top B = Q,$$

which shows that each $Q \in \mathrm{Sym}^+(\mathbb{R}^d)$ corresponds to a Mahalanobis distance on V. $\square$

D.5.2 Proofs for Section 5.4.2

Lemma 5.8. *Let V be an r-dimensional subspace of $\mathbb{R}^d$ with a canonical representation given by $B \in \mathbb{R}^{d\times r}$. Fix any Mahalanobis distance $M \in \mathrm{Sym}^+(\mathbb{R}^d)$, any pair of items $x, x' \in \mathbb{R}^d$, and ideal point $u \in \mathbb{R}^d$. Suppose that x and x' are contained in V with canonical representation $x_V = B^\top x$ and $x'_V = B^\top x'$ in $\mathbb{R}^r$. Then:*

$$\psi_M\big(x, x'; u\big) = \psi_Q\big(x_V, x'_V; u_V\big),$$

where the phantom ideal point u_V of u on V satisfies $(B^\top MB)u_V = B^\top Mu$, and $Q = \Pi_V(M)$ is the matrix representation in $\mathrm{Sym}^+(V)$ of the subspace metric $\rho\big|_V$.

Proof. Let $v = -2Mu$ and $v_V = -2Qu_V$ be the pseudo-ideal user points for u and u_V, respectively. The following shows that v_V is given by the canonical representation of the orthogonal projection of v to V,

$$v_V = -2\underbrace{B^\top MB}_{Q}\underbrace{(B^\top MB)^{-1}B^\top Mu}_{u_V} = -2B^\top Mu = B^\top v.$$

We now expand the definitions of ψ_M and ψ_Q,

$$\begin{aligned}
\psi_M(x, x'; u) &\overset{(i)}{=} \left\langle xx^\top - x'x'^\top, M \right\rangle + \left\langle x - x', v \right\rangle \\
&\overset{(ii)}{=} \left\langle BB^\top(xx^\top - x'x'^\top)BB^\top, M \right\rangle + \left\langle BB^\top(x - x'), v \right\rangle \\
&\overset{(iii)}{=} \left\langle B^\top xx^\top B - B^\top x'x'^\top B, B^\top MB \right\rangle + \left\langle B^\top x - B^\top x', B^\top v \right\rangle \\
&\overset{(iv)}{=} \left\langle x_V {x_V}^\top - x'_V {x'_V}^\top, Q \right\rangle + \left\langle x_V - x'_V, v_V \right\rangle \\
&\overset{(v)}{=} \psi_Q(x_V, x'_V; u_V),
\end{aligned}$$

where (i) and (v) follow by definition, (ii) uses the fact that as $B \in \mathbb{R}^{d \times r}$ is an orthonormal basis, $BB^\top v = v$ for all $v \in V$, (iii) applies the following property for the trace inner product $\langle BA, C \rangle = \mathrm{tr}(C^\top BA) = \langle A, B^\top C \rangle$, and (iv) rewrites the equation in terms of the canonical representations. □

Proposition 5.10. *Let $\mathcal{X}$ quadratically span a subspace V of dimension r. There exists a collection $D_1, \ldots, D_K$ of design matrices, each over m pairs of items in $\mathcal{X}$, such that given a (distinct) user's response to each design, $\rho\big|_V$ can be identified when $m \geq r+1$ and $K \geq r(r+1)/2$.*

Proof. By Lemma 5.8, it suffices to prove the result for $V = \mathbb{R}^d$. We show that if $\mathcal{X}$ quadratically spans $\mathbb{R}^d$, then we can construct an (m, K)-experimental design where $m = d + 1$ and $K = d(d+1)/2$ such that there is a unique matrix consistent with all user responses. Let $D = d + \frac{d(d+1)}{2}$ be the dimension of $\mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d$.

Since $\mathcal{X}$ quadratically spans V, there exists a collection of pairs $\{(x_{i_0}, x_{i_1})\}_{i=1}^D$ such that:

$$\mathrm{span}\big(\{\Delta_i \oplus \delta_i : i \in [D]\}\big) \;=\; \mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d, \tag{D.2}$$

where we let $\Delta_i = x_{i_0} x_{i_0}^\top - x_{i_1} x_{i_1}^\top$ and $\delta_i = x_{i_0} - x_{i_1}$. In particular, the collection $\{\Delta_i \oplus \delta_i\}_{i \in [D]}$ is linearly independent. Without loss of generality, we may select these so that the first d

pairwise differences δ_i are also linearly independent:

$$\operatorname{span}\big(\{\delta_i : i \in [d]\}\big) = \mathbb{R}^d.$$

We will ask all users to compare the first d pairs and one additional pair, unique to the user. In particular, set the kth collection of preference comparison queries by:

$$\mathcal{D}_k = \big\{(x_{i_0}, x_{i_1}) : i \in \mathcal{I}_k\big\}, \qquad \text{where } \mathcal{I}_k = [d] \cup \{d+k\}.$$

First, we show that the responses from a single user must reveal at least one dimension of $\mathrm{Sym}(\mathbb{R}^d)$. To see this, let's fix a user $k \in [K]$. From Equation (D.2), we can define the vector $(\alpha_{i,k} : i \in \mathcal{I}_k)$ so that:

$$\alpha_{d+k;k} = 1 \qquad \text{and} \qquad \sum_{i \in \mathcal{I}_k} \alpha_{i;k}\delta_i = 0.$$

Therefore, from the preference measurements, we deduce that at least one degree of freedom of M is revealed:

$$\sum_{i\in\mathcal{I}_k} \alpha_{i;k}\psi_{i;k} = \sum_{i\in\mathcal{I}_k} \alpha_{i;k}\big\langle \Delta_i, M\big\rangle + \underbrace{\sum_{i\in\mathcal{I}_k} \alpha_{i;k}\big\langle \delta_i, v_k\big\rangle}_{\langle 0, v_k\rangle} = \left\langle \sum_{i\in\mathcal{I}_k} \alpha_{i;k}\Delta_i, M \right\rangle. \tag{D.3}$$

We now claim that each user reveals a different degree of freedom of M. In particular, it suffices to show that the following collection of matrices spans $\mathrm{Sym}(\mathbb{R}^d)$,

$$\left\{ \sum_{i\in\mathcal{I}_k} \alpha_{i;k}\Delta_i : k \in [K] \right\}.$$

Suppose otherwise. Since $K = \frac{d(d+1)}{2}$, this means that this collection of matrices are linearly dependent, and that there exists a non-zero vector $(\mu_k : k \in [K])$ such that

$0 \in \mathrm{Sym}(\mathbb{R}^d)$ is the linear combination:

$$\sum_{k\in[K]} \mu_k \sum_{i\in\mathcal{I}_k} \alpha_{i;k}\Delta_i = 0.$$

Because we chose $\alpha_{d+k;k} = 1$ for each user $k \in [K]$, this implies that zero in $\mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d$ is also a non-trivial linear combination of the collection $\Delta_i \oplus \delta_i$, where:

$$\sum_{i=1}^{d} \left(\sum_{k\in[K]} \mu_k \alpha_{i;k} \right) \Delta_i \oplus \delta_i + \sum_{i=d+1}^{D} \mu_{i-d} \cdot \Delta_i \oplus \delta_i = 0.$$

But then this collection is not full rank and cannot span $\mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d$, as assumed in Equation (D.2). It follows that M is the unique solution to the system of linear equations corresponding to Equation (D.3). □

Proposition 5.11. *Let $(D_k)_{k\in\mathbb{N}}$ be a set of design matrices over items in $\mathcal{X} \subset V$. If $\mathcal{X}$ does not quadratically span V, then infinitely many Mahalanobis distances on V are consistent with any set of user responses to the design matrices.*

Proof. Because $\mathcal{X}_V$ does not quadratically span V, there exists an element $Q_\perp \oplus v_\perp \in \mathrm{Sym}(V) \oplus V$ such that:

$$\left\langle \left(x_V x_V^\top - x'_V x_V'^\top\right) \oplus (x - x'), Q_\perp \oplus v_\perp \right\rangle = 0,$$

for all $x, x' \in \mathcal{X}_V$, where $x_V = B^\top x$ and $x'_V = B^\top x'$. Let $M_\perp = BQB^\top$, so that:

$$\left\langle \left(xx^\top - x'x'^\top\right) \oplus (x - x'), M_\perp \oplus v_\perp \right\rangle = 0, \qquad \text{for all } x, x' \in \mathcal{X}_V. \tag{D.4}$$

We claim that if $M \in \mathrm{Sym}^+(\mathbb{R}^d)$ is consistent with the kth user's responses $\mathcal{D}_k = \{(x_{i_0;k}, x_{i_1;k}, \psi_{i;k})\}_{i=1}^m$, then the matrix $M + \lambda M_\perp$ is also consistent, provided that $M + \lambda M_\perp$ remains in $\mathrm{Sym}^+(\mathbb{R}^d)$. In particular, if M is consistent, there exists an ideal point u_k so

that for all $i \in [m]$:

$$\begin{aligned}\psi_{i;k} = \psi_M(x_{i_0}, x_{i_1}; u_k) &\overset{(i)}{=} \left\langle \left(x_{i_0}x_{i_0}^\top - x_{i_1}x_{i_1}^\top\right) \oplus (x_{i_0} - x_{i_1}), M \oplus v_k \right\rangle \\ &\overset{(ii)}{=} \left\langle \left(x_{i_0}x_{i_0}^\top - x_{i_1}x_{i_1}^\top\right) \oplus (x_{i_0} - x_{i_1}), M \oplus v_k + \lambda M_\perp \oplus v_\perp \right\rangle \\ &\overset{(iii)}{=} \psi_{M+\lambda M_\perp}(x_{i_0}, x_{i_1}; \lambda \tilde{u}_k),\end{aligned}$$

where (i) expands the definition of ψ_M while setting the pseudo-ideal point to $v_k = -2Mu_k$, (ii) applies Equation (D.4), and (iii) applies the definition of $\psi_{M+M_\perp}$ while setting $\tilde{u}_k = -\frac{1}{2}M^{-1}(v_k + v_\perp)$.

Thus, if M is the matrix representation of the underlying Mahalanobis distance, the following matrices are also consistent:

$$\left\{ M + \lambda M_\perp : 0 \leq \lambda < \frac{\sigma_{\min}(M)}{\sigma_{\max}(M_\perp)} \right\},$$

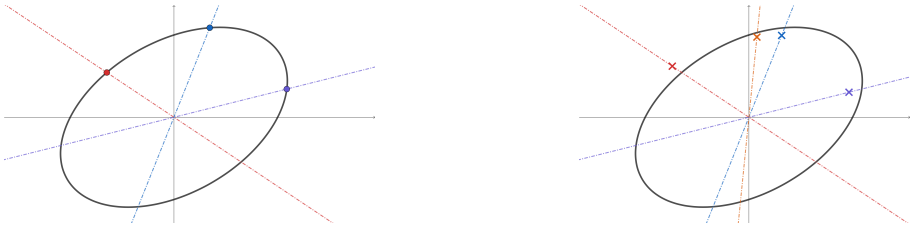
where $\sigma_{\max}(M_\perp)$ is the maximum singular value of $M_\perp$ while $\sigma_{\min}(M)$ is the minimum singular value of M; this implies that $M + \lambda M_\perp$ is positive-definite. Infinitely such λ's exist because (a) $\sigma_{\max}(M_\perp) < \infty$ is finite and (b) $\sigma_{\min}(M) > 0$ is bounded away from zero because M is positive-definite. □

D.5.3 Proof of Proposition 5.13 from Section 5.4.3

Proposition 5.13. *Let ρ be a Mahalanobis distance on $\mathbb{R}^d$. Let $(V_\lambda)_{\lambda\in\Lambda}$ be a collection of subspaces with canonical representations given by the orthonormal bases $(B_\lambda)_{\lambda\in\Lambda}$. The following are equivalent:*

1. $\left\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\right\}$ *spans* $\mathrm{Sym}(\mathbb{R}^d)$.

2. *Let Π_{V_λ} be given by Equation* (5.3). *The linear map* $\Pi : \mathrm{Sym}(\mathbb{R}^d) \to \bigoplus_{\lambda\in\Lambda} \mathrm{Sym}(V_\lambda)$ *is*



(a) **(b)**

Figure D.2. (a) Illustrates the number of subspaces needed to reconstruct a high-dimensional ellipsoid from its intersections with low-dimensional subspaces. In $\mathbb{R}^2$, we need 3 points on distinct 1-dimensional subspaces to possibly recover an ellipse centered at the origin. (b) When we cannot exactly identify where the high-dimensional ellipsoid intersects with each subspace, we may still fit an ellipsoid from approximate estimations using least squares [55].

injective, where:

$$\Pi(A) = \bigoplus_{\lambda \in \Lambda} \Pi_{V_\lambda}(A).$$

3. If $\hat{\rho}$ is a Mahalanobis distance such that $\hat{\rho}\big|_{V_\lambda} = \rho\big|_{V_\lambda}$ for all $\lambda \in \Lambda$, then $\hat{\rho} = \rho$.

Proof. (1 $\implies$ 2). Suppose $\operatorname{span}\left\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\right\} = \operatorname{Sym}(\mathbb{R}^d)$. To show that Π is injective, it suffices to show that its kernel is trivial. Let $M \in \ker(\Pi)$. We claim that for any $\lambda \in \Lambda$ and $x \in V_\lambda$, we have:

$$\left\langle xx^\top, M \right\rangle = 0. \tag{D.5}$$

Assume this for now. Then, $M \in \operatorname{Sym}(\mathbb{R}^d) = \operatorname{span}\left\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\right\}$, so that $\langle M, M \rangle = 0$. This implies that $M = 0$, so the kernel is trivial. We now show Eq. (D.5). Using the definition of Π_{V_λ}, when $M \in \ker(\Pi)$, we have:

$$\Pi_{V_\lambda}(M) = B_\lambda^\top M B_\lambda = 0. \tag{D.6}$$

Say that $\dim(V_\lambda) = r_\lambda$ and $x \in V_\lambda$. As $B_\lambda \in \mathbb{R}^{d \times r_\lambda}$ is a basis of V_λ, there exists $z \in \mathbb{R}^{r_\lambda}$

such that $x = B_\lambda z$. By Eq. (D.6),

$$\left\langle xx^\top, M \right\rangle = z^\top B_\lambda^\top M B_\lambda z = 0.$$

$(2 \implies 1)$. We prove the contrapositive. Suppose $S = \operatorname{span}\left\{xx^\top : x \in V_\lambda, \lambda \in \Lambda\right\}$ does not span $\operatorname{Sym}(\mathbb{R}^d)$. Then, there exists some nonzero $A \in S^\perp$ in its orthogonal complement. To show that Π is not injective, we show that $A \in \ker(\Pi)$. That is, for all $\lambda \in \Lambda$, that $B_\lambda^\top A B_\lambda = 0$. We do this by proving that all eigenvalues of $B_\lambda^\top A B_\lambda$ are zero.

Let $v \in \mathbb{R}^{r_\lambda}$ be any unit eigenvector of $B_\lambda^\top A B_\lambda$ and α be the corresponding eigenvalue, so that:

$$\alpha = v^\top B_\lambda^\top A B_\lambda v = \left\langle xx^\top, A \right\rangle,$$

where $x = B_\lambda v$ is an element of V_λ. But because $A \in S^\perp$, this implies that the eigenvalue is zero, $\alpha = 0$.

$(2 \implies 3)$. Let M and $\hat{M}$ be the matrix representations of ρ and $\hat{\rho}$, respectively. By assumption, their subspace metrics coincide over $(V_\lambda)_\lambda$, so Proposition D.12 implies:

$$\Pi_{V_\lambda}(M) = \Pi_{V_\lambda}(\hat{M}).$$

And as Π is injective, we must have $M = \hat{M}$, so that $\rho = \hat{\rho}$.

$(3 \implies 2)$. We prove that Π is injective by showing that its kernel is trivial. Let $A \in \ker(\Pi)$. Then, let $c, \hat{c} > \|A\|_{\text{op}}$ and define $M = c^{-1}A + I$ and $\hat{M} = \hat{c}^{-1}A + I$, which are positive-definite by construction. Let ρ and $\hat{\rho}$ be their corresponding Mahalanobis distances. Their subspace metrics on all V_λ's coincide, since $A \in \ker(\Pi)$,

$$\Pi(M) = \Pi(c^{-1}A + I) = \Pi(I) = \Pi(\hat{c}^{-1}A + I) = \Pi(\hat{M}).$$

And so, by assumption $\rho = \hat{\rho}$. But as the matrix representation of a Mahalanobis distance

is unique (Proposition D.12), this implies that $M = \hat{M}$, proving that $A = 0$. $\square$

D.6 Proofs and Additional Results for Section 5.5

D.6.1 Proofs and additional remarks for Theorem 5.15

Theorem 5.15. *Let $\mathbb{R}^d$ have a Mahalanobis distance with matrix representation $M \in \mathrm{Sym}^+(\mathbb{R}^d)$. Let $\mathcal{X} \subset \mathbb{R}^d$ be subspace-clusterable over subspaces V_λ indexed by $\lambda \in \Lambda$, where $|\Lambda| = n$. Let $\hat{M}$ be the estimator of M and let $\hat{Q}_\lambda$ be the estimator of the subspace metric Q_λ for each λ learned from Algorithm 5. Suppose there exist $\gamma \leq \varepsilon$ such that $\left\|\mathbb{E}\left[\hat{Q}_\lambda\right] - Q_\lambda\right\|_{\mathrm{F}} \leq \gamma$ and $\left\|\hat{Q}_\lambda - Q_\lambda\right\|_{\mathrm{F}} \leq \varepsilon$ for each λ. Fix $p \in (0, 1]$. Then, there is a universal constant $c > 0$ such that with probability at least $1 - p$,*

$$\left\|\hat{M} - M\right\|_{\mathrm{F}} \leq c \cdot \frac{1}{\sigma_{\min}(\Pi)} \left(\gamma\sqrt{n} + \varepsilon d\sqrt{\log \frac{2d}{p}}\right),$$

where $\sigma_{\min} > 0$ is the least singular value of Π.

Proof of Theorem 5.15. Let $c = 2c_0$ where c_0 is a universal constant to be defined later. Recall from Eq. (5.5) that $\hat{M}$ minimizes $\|A - \hat{M}_{\mathrm{LS}}\|_{\mathrm{F}}$ over all $A \in \mathrm{Sym}^+(\mathbb{R}^d)$. Since M is also contained in $\mathrm{Sym}^+(\mathbb{R}^d)$, we have:

$$\left\|\hat{M} - \hat{M}_{\mathrm{LS}}\right\|_{\mathrm{F}} \leq \left\|M - \hat{M}_{\mathrm{LS}}\right\|_{\mathrm{F}}.$$

By the triangle inequality,

$$\left\|\hat{M} - M\right\|_{\mathrm{F}} \leq \left\|\hat{M} - \hat{M}_{\mathrm{LS}}\right\|_{\mathrm{F}} + \left\|\hat{M}_{\mathrm{LS}} - M\right\|_{\mathrm{F}} \leq 2\left\|\hat{M}_{\mathrm{LS}} - M\right\|_{\mathrm{F}}.$$

Therefore, it suffices to show that, with probability $1 - \delta$,

$$\left\|\hat{M}_{\mathrm{LS}} - M\right\|_{\mathrm{F}} \leq c_0 \cdot \frac{1}{\sigma_{\min}(\Pi)} \left(\gamma\sqrt{m} + \varepsilon d\sqrt{\log \frac{2d}{\delta}}\right). \tag{D.7}$$

Before proving Eq. (D.7), we introduce some notation.

Notation and facts.

For each subspace V_λ, we denote the recovery error by:

$$E_\lambda = \hat{Q}_\lambda - Q_\lambda = \Big(\underbrace{\mathbb{E}[\hat{Q}_\lambda] - Q_\lambda}_{H_\lambda \text{ (bias)}} \Big) + \Big(\underbrace{\hat{Q}_\lambda - \mathbb{E}[\hat{Q}_\lambda]}_{\xi_\lambda \text{ (noise)}} \Big),$$

which we decompose into a bias term $H_\lambda := \mathbb{E}[\hat{Q}_\lambda] - Q_\lambda$ and a noise term $\xi_\lambda := \hat{Q}_\lambda - \mathbb{E}[\hat{Q}_\lambda]$. By assumption,

$$\|H_\lambda\|_{\mathrm{F}} \le \gamma \qquad \text{and} \qquad \mathbb{E}[\xi_\lambda] = 0, \quad \|\xi_\lambda\|_{\mathrm{F}} \le \|E_\lambda\|_{\mathrm{F}} \le \varepsilon.$$

Let $H := \bigoplus_{\lambda\in\Lambda} H_\lambda$, $\xi := \bigoplus_{\lambda\in\Lambda} \xi_\lambda$, and $E := H + \xi$. Thus, $E = \bigoplus_{\lambda\in\Lambda} \Big(\hat{Q}_\lambda - Q_\lambda\Big)$, by the above bias/noise decomposition. In addition, since $\|H_\lambda\|_{\mathrm{F}} \le \gamma$, we have $\|H\| = \sqrt{\sum_{\lambda\in\Lambda} \|H_\lambda\|_{\mathrm{F}}^2} \le \sqrt{m}\gamma$.

We now prove Eq. (D.7). Recall from Eq. (5.4) that $\hat{M}_{\mathrm{LS}}$ is the least-squares solution, so that:

$$\begin{aligned} \hat{M}_{\mathrm{LS}} - M &= \Pi^+(E) \\ &= \Pi^+(H + \xi), \end{aligned} \tag{D.8}$$

where $\Pi^+ : \bigoplus_{\lambda\in\Lambda} \mathrm{Sym}(V_\lambda) \to \mathrm{Sym}(\mathbb{R}^d)$ denotes the Moore–Penrose inverse of Π (see Definition D.2). It then follows from Eq. (D.8) and the triangle inequality that

$$\big\|\hat{M}_{\mathrm{LS}} - M\big\|_{\mathrm{F}} \le \big\|\Pi^+(H)\big\|_{\mathrm{F}} + \big\|\Pi^+(\xi)\big\|_{\mathrm{F}}. \tag{D.9}$$

By Proposition 5.13, the map Π is injective since $\mathcal{X}$ is subspace-clusterable. Thus,

$\sigma_{\min}(\Pi) > 0$, and:

$$\left\|\Pi^+(H)\right\|_{\mathrm{F}} \leq \frac{1}{\sigma_{\min}(\Pi)}\|H\|_{\mathrm{F}} \leq \frac{1}{\sigma_{\min}(\Pi)}\gamma\sqrt{m}. \tag{D.10}$$

It then follows from Eq. (D.9) and Eq. (D.10) that, to prove Eq. (D.7), it suffices to show that, with probability at least $1-\delta$,

$$\left\|\Pi^+(\xi)\right\|_{\mathrm{F}} \leq c_0 \cdot \frac{1}{\sigma_{\min}(\Pi)}\left(\varepsilon d\sqrt{\log\frac{2d}{\delta}}\right). \tag{D.11}$$

By the universal property of the direct sum (see Proposition D.3), there exist $\Pi_\lambda^+ : \mathrm{Sym}(V_\lambda) \to \mathrm{Sym}(\mathbb{R}^d)$ for each $\lambda \in \Lambda$, such that

$$\Pi^+(\xi) = \sum_{\lambda\in\Lambda} \Pi_\lambda^+(\xi_\lambda).$$

Observe that

1. Each ξ_λ is from subspace V_λ; and thus, ξ_λ's and $\Pi_\lambda^+(\xi_\lambda)$'s across subspaces are independent,
2. $\mathbb{E}\left[\Pi_\lambda^+(\xi_\lambda)\right] = \Pi_\lambda^+\left(\mathbb{E}\left[\xi_\lambda\right]\right) = 0$; and,
3. $\left\|\Pi_\lambda^+(\xi_\lambda)\right\|_{\mathrm{F}} \leq \left\|\Pi_\lambda^+\right\|_{\mathrm{op}} \cdot \left\|\xi_\lambda\right\|_{\mathrm{F}} \leq \left\|\Pi_\lambda^+\right\|_2 \cdot \varepsilon$,

where $\|\cdot\|_2$ denotes the 2-Schatten norm (see Definition D.4).

Corollary D.16 gives a Hoeffding-style concentration inequality for independent sub-Gaussian random matrices. Applied here, it states that there exists a universal constant

c_0 such that, with probability $1-\delta$,

$$\begin{aligned}\left\|\Pi^+(\xi)\right\|_{\mathrm{F}} &= \left\|\sum_{\lambda\in\Lambda}\Pi_\lambda^+(\xi_\lambda)\right\|_{\mathrm{F}} \\ &\overset{(i)}{\leq} c_0\cdot\sqrt{\sum_{\lambda\in\Lambda}\left\|\Pi_\lambda^+\right\|_2^2\cdot\varepsilon^2\log\frac{2d}{\delta}} \\ &\overset{(ii)}{=} c_0\cdot\left\|\Pi^+\right\|_2\cdot\varepsilon\sqrt{\log\frac{2d}{\delta}} \\ &\overset{(iii)}{\leq} c_0\cdot\frac{1}{\sigma_{\min}(\Pi)}\cdot\varepsilon d\sqrt{\log\frac{2d}{\delta}},\end{aligned} \tag{D.12}$$

where (i) applies the third observation from above, (ii) applies Proposition D.5 about 2-Schatten norms, and (iii) uses the following facts:

- $\|\Pi^+\|_2 \leq \sigma_{\max}(\Pi^+)\cdot\sqrt{\operatorname{rank}(\Pi^+)}$, (see Definition D.4),
- $\sigma_{\max}(\Pi^+) = \dfrac{1}{\sigma_{\min}(\Pi)}$,
- $\operatorname{rank}(\Pi^+) \leq \frac{d(d+1)}{2} \leq d^2$.

□

D.6.2 Proofs and additional remarks for Proposition 5.18

Proposition 5.18 (Theorem 4.1, [28])**.** *Suppose that $\mathbb{R}^r$ has a Mahalanobis distance with representation $Q\in\mathrm{Sym}^+(\mathbb{R}^r)$ where $\|Q\|_{\mathrm{F}}\leq\zeta_M$. Let each user $k\in[K]$ have pseudo-ideal point $v_k\in\mathbb{R}^r$ where $v_k\leq\zeta_v$. Let $\mathcal{P}_m$ be a distribution over designs of size m over $\mathbb{R}^r$ (Definition 5.1). For each user, let $D_k\sim\mathcal{P}_m$ be an i.i.d. random design, and let $\mathcal{D}_k=\{(x_{i_0},x_{i_1},y_{i;k})\}_{i\in[m]}$ be the user's responses under Assumption 5.17. Fix $p\in(0,1]$. Given loss function $\ell(z)=-\log f(z)$, Algorithm 9 returns $\hat{Q}\in\mathrm{Sym}^+(\mathbb{R}^r)$, where with probability at least $1-p$,*

$$\|\hat{Q}-Q\|_{\mathrm{F}}^2 \leq \frac{16L}{c_f^2\cdot\sigma_{\min}^2(\mathcal{P}_m)}\sqrt{\frac{(\zeta_M^2+K\zeta_v^2)\log\frac{4}{p}}{mK}}.$$

Proof. The objective over which the parameters $(A, w_1, \ldots, w_K)$ is optimized in Eq. (D.1) of Algorithm 9 can be written as:

$$\hat{R}(A, w_1, \ldots, w_K) = \sum_{k\in[K]} \sum_{i\in[m]} -\log f\big(y_{i;k} \cdot D_{i;k}(A, w_k)\big).$$

Let $(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_K)$ be the solution recovered in this step of Algorithm 9. The excess risk of these parameters is defined to be how much worse in expectation the parameters are at explaining observed data compared to the true parameters $(Q, v_1, \ldots, v_K)$ that generated the data. The excess risk leads to a bound on $\|\hat{Q} - Q\|_{\mathrm{F}}^2$,

$$\begin{aligned}
&\mathbb{E}\left[\hat{R}(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_K)\right] - \mathbb{E}\left[\hat{R}(Q, v_1, \ldots, v_K)\right] \\
&\qquad \overset{(a)}{=} \sum_{k\in[K]} \mathop{\mathbb{E}}_{D_k\sim\mathcal{P}_m} \left[\sum_{i\in[m]} \mathrm{KL}\left(f\big(D_{i;k}(Q, v_k)\big) \,\middle\|\, f\big(D_{i;k}(\hat{Q}, \hat{v}_k)\big)\right)\right] \\
&\qquad \overset{(b)}{\geq} 2c_f^2 \sum_{k\in[K]} \mathop{\mathbb{E}}_{D_k\sim\mathcal{P}_m} \left\|D_k\big(\hat{Q} - Q, \hat{v}_k - v_k\big)\right\|^2 \\
&\qquad \overset{(c)}{\geq} 2c_f^2 \sum_{k\in[K]} m \cdot \sigma_{\min}^2(\mathcal{P}_m) \cdot \left(\|\hat{Q} - Q\|_{\mathrm{F}}^2 + \|\hat{v}_k - v_k\|^2\right) \\
&\qquad \geq 2mKc_f^2 \cdot \sigma_{\min}^2(\mathcal{P}_m) \cdot \|\hat{Q} - Q\|_{\mathrm{F}}^2,
\end{aligned} \tag{D.13}$$

where each inequality is justified below. We just need to show that the excess risk of $\hat{Q}$ returned by the algorithm has small excess risk. Lemma D.13 approaches this via a standard generalization argument, showing that with probability at least $1 - \delta$,

$$\begin{aligned}
&\mathbb{E}\left[\hat{R}(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_K)\right] - \mathbb{E}\left[\hat{R}(Q, v_1, \ldots, v_K)\right] \\
&\qquad \leq \underbrace{\hat{R}(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_K) - \hat{R}(Q, v_1, \ldots, v_K)}_{\leq 0} + 32L\sqrt{mK(\zeta_M^2 + K\zeta_v^2)\log\frac{4}{\delta}},
\end{aligned} \tag{D.14}$$

where the indicated difference is less than zero because $(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_k)$ is the minimizer of $\hat{R}$. The result is obtained by combining Eqs. (D.13) and (D.14). To finish the prove, we justify the above inequalities:

(a) Recall that $\Pr[Y_{i;k} = +1] = f(D_{i;k}(Q, v_k))$. Because $f(z) = 1 - f(-z)$, we also have that:

$$\Pr[Y_{i;k} = -1] = 1 - f(D_{i;k}(Q, v_k)) = f(-D_{i;k}(Q, v_k)).$$

Therefore, $\Pr[Y_{i;k} = y] = f(y \cdot D_{i;k}(Q, v_k))$. It follows that the excess risk is equal to:

$$\begin{aligned}
&\mathbb{E}\left[\hat{R}(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_K)\right] - \mathbb{E}\left[\hat{R}(Q, v_1, \ldots, v_K)\right] \\
&= \sum_{k\in[K]} \mathop{\mathbb{E}}_{D_k, Y} \left[\sum_{i\in[m]} -\log\left(\frac{f(Y_{i;k} \cdot D_{i;k}(Q, v_k))}{f(Y_{i;k} \cdot D_{i;k}(\hat{Q}, \hat{v}_k))}\right)\right] \\
&= \sum_{k\in[K]} \mathop{\mathbb{E}}_{D_k} \left[\sum_{i\in[m]} \sum_{y\in\{-1,+1\}} -f(y \cdot D_{i;k}(Q, v_k)) \log\left(\frac{f(y \cdot D_{i;k}(Q, v_k))}{f(y \cdot D_{i;k}(\hat{Q}, \hat{v}_k))}\right)\right],
\end{aligned}$$

where we obtain the equality (a) by applying the definition $\mathrm{KL}(p\|q)$,

$$\mathrm{KL}(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

(b) The following is the same argument used in [28, Proposition E.3].

$$\begin{aligned}
\sum_{i\in[m]} \mathrm{KL}\left(f\big(D_{i;k}(Q, v_k)\big) \,\middle\|\, f\big(D_{i;k}(\hat{Q}, \hat{v}_k)\big)\right) &\geq 2 \sum_{i\in[m]} \Big(f\big(D_{i;k}(Q, v_k)\big) - f\big(D_{i;k}(\hat{Q}, \hat{v}_k)\big)\Big)^2 \\
&\geq 2c_f^2 \sum_{i\in[m]} \Big(D_{i;k}(Q, v_k) - D_{i;k}(\hat{Q}, \hat{v}_k)\Big)^2 \\
&= 2c_f^2 \sum_{i\in[m]} \Big(D_{i;k}(\hat{Q} - Q, \hat{v}_k - v_k)\Big)^2 \\
&= 2c_f^2 \Big\|D_k(\hat{Q} - Q, \hat{v}_k - v_k)\Big\|^2,
\end{aligned}$$

where the first inequality comes from $\mathrm{KL}(p\|q) \geq 2(p-q)^2$, see [106, Lemma 5.2], the second uses the monotonicity of f and the lower bound of f', the third applies linearity of $D_{i;k}$, and the fourth just rewrites the sum in terms of the squared ℓ_2-norm over $\mathbb{R}^m$.

(c) Recall that $\sigma^2(\mathcal{P}_m) = \frac{1}{m} \cdot \sigma_{\min}(\mathbb{E}[D^*D])$ when $D \sim \mathcal{P}_m$. Let $X = (\hat{Q} - Q) \oplus (\hat{v}_k - v_k)$ for short. Then,

$$\begin{aligned}
\mathbb{E}\left\|D_k(\hat{Q} - Q, \hat{v}_k - v_k)\right\|^2 &= \mathbb{E}\left\langle D_k X, D_k X\right\rangle \\
&= X^\top \mathbb{E}\left[D_k^* D_k\right] X \\
&\geq \sigma_{\min}(\mathbb{E}\left[D_k^* D_k\right]) \cdot \|X\|^2 \\
&= m \cdot \sigma_{\min}^2(\mathcal{P}_m) \cdot \left(\|\hat{Q} - Q\|_{\mathrm{F}}^2 + \|\hat{v}_k - v_k\|^2\right),
\end{aligned}$$

where the inequality applies the variational characterization of the minimum singular value.

□

Lemma D.13. *Let $\delta \in (0,1)$. Given the assumptions of Proposition 5.18, Eq.* (D.14) *holds with probability at least $1-\delta$.*

Proof. Let $\Theta \subset \mathrm{Sym}^+(\mathbb{R}^r) \oplus \mathbb{R}^{r\times K}$ denote the set of parameters $\theta \equiv (A, w_1, \ldots, w_K)$ such that $A \in \mathrm{Sym}^+(\mathbb{R}^r)$ with $\|A\|_{\mathrm{F}} \leq \zeta_M$ and $w_k \in \mathbb{R}^r$ with $\|w_k\| \leq \zeta_v$. We claim that with probability at least $1-\delta$, we have uniform convergence:

$$\sup_{\theta\in\Theta} \left|\hat{R}(\theta) - \mathbb{E}\left[\hat{R}(\theta)\right]\right| \leq 16L\sqrt{mK(\zeta_M^2 + K\zeta_v^2)\log\frac{4}{\delta}}. \tag{D.15}$$

Before proving this, notice that this implies Eq. D.14. In particular, let $\hat{\theta}$ correspond to the parameters $(\hat{Q}, \hat{v}_1, \ldots, \hat{v}_K)$ and let θ correspond to $(Q, v_1, \ldots, v_K)$. Then we have that with probability at least $1-\delta$, both $\hat{R}(\hat{\theta})$ and $\hat{R}(\theta)$ are close to their expected values, each

contributing at most the right-hand side of Eq. (D.15):

$$\mathbb{E}\left[\hat{R}(\hat{\theta})\right] - \mathbb{E}\left[\hat{R}(\theta)\right] \leq \hat{R}(\hat{\theta}) - \hat{R}(\theta) + 32L\sqrt{mK(\zeta_M^2 + K\zeta_v^2)\log\frac{4}{\delta}}.$$

In the remainder of the proof, we show Eq. (D.15). For any $\theta \in \Theta$, consider the empirical risk $\hat{R}(\theta)$. We claim that the risk contribution by the ith comparison by the kth user is a bounded random variable,

$$\left| -\log\left(f\left(Y_{i;k} \cdot D_{i;k}(A, w_k)\right)\right) + \log\frac{1}{2} \right| \overset{(a)}{\leq} 2L(\zeta_M + \zeta_v).$$

Let us verify this claim later. For now, the bounded difference inequality (reproduced below as Lemma D.17) implies that with probability at least $1 - \delta$,

$$\sup_{\theta\in\Theta} \left|\hat{R}(\theta) - \mathbb{E}\left[\hat{R}(\theta)\right]\right| \leq \mathbb{E}\left[\sup_{\theta\in\Theta} \left|\hat{R}(\theta) - \mathbb{E}\left[\hat{R}(\theta)\right]\right|\right] + 4L(\zeta_M + \zeta_v)\sqrt{2mK\log\frac{2}{\delta}}. \quad \text{(D.16)}$$

To bound the expectation term, let us combine each user's random design matrix D_k into a single (m, K)-experimental design matrix $D : \mathrm{Sym}(\mathbb{R}^r) \oplus \mathbb{R}^{r\times K} \to \mathbb{R}^{m\times K}$, so that it is the following linear map:

$$D(A, w_1, \ldots, w_K)_{i;k} = D_{i;k}(A, w_k).$$

Let $D^* : \mathbb{R}^{m\times K} \to \mathrm{Sym}(\mathbb{R}^r) \oplus \mathbb{R}^{m\times K}$ be its adjoint. Let $\epsilon \in_R \{-1, +1\}^{m\times K}$ be an array of independent Rademacher random variables, so that $\epsilon_{i;k}$ is equal to -1 or $+1$ uniformly at

random. Then:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\hat{R}(\theta)-\mathbb{E}\left[\hat{R}(\theta)\right]\right|\right] &\overset{(b)}{\leq} 2\,\mathbb{E}\left[\sup_{A,w_1,\ldots,w_K}\left|\sum_{k\in[K]}\sum_{i\in[m]}\epsilon_{i;k}\left(-\log f\big(Y_{i;k}\cdot D_{i;k}(A,w_k)\big)\right)\right|\right] \\
&\overset{(c)}{\leq} 2L\cdot\mathbb{E}\left[\sup_{A,w_1,\ldots,w_K}\left|\sum_{k\in[K]}\sum_{i\in[m]}\epsilon_{i;k}\big(Y_{i;k}\cdot D_{i;k}(A,w_k)\big)\right|\right] \\
&\overset{(d)}{=} 2L\cdot\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\big\langle\epsilon,D(\theta)\big\rangle\right|\right] \\
&\overset{(e)}{\leq} 2L\cdot\mathbb{E}\left\|D^*\epsilon\right\|\cdot\sup_{\theta\in\Theta}\|\theta\| \\
&\overset{(f)}{\leq} 4L\sqrt{2mK(\zeta_M^2+K\zeta_v^2)}, \qquad \text{(D.17)}
\end{aligned}$$

where we justify each step below. We obtain Eq. (D.15) by combining Eqs. (D.16) and (D.17),

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|\hat{R}(\theta)-\mathbb{E}\left[\hat{R}(\theta)\right]\right| &\leq 4L\sqrt{2mK(\zeta_M^2+K\zeta_v^2)}+4L(\zeta_M+\zeta_v)\sqrt{2mK\log\frac{2}{\delta}} \\
&\overset{(i)}{\leq} 4L\sqrt{2\left(2mK(\zeta_M^2+K\zeta_v^2)+2mK(\zeta_m+\zeta_v)^2\log\frac{2}{\delta}\right)} \\
&\overset{(ii)}{\leq} 8L\sqrt{mK\cdot(\zeta_m^2+K\zeta_v^2)\cdot\left(1+3\log\frac{2}{\delta}\right)} \\
&\overset{(iii)}{\leq} 16L\sqrt{mK\cdot(\zeta_M^2+K\zeta_v^2)\log\frac{4}{\delta}},
\end{aligned}$$

where (i) applies a variant of the AM-GM inequality $\sqrt{a}+\sqrt{b}\leq\sqrt{2(a+b)}$, (ii) uses the following upper bound $(\zeta_M+\zeta_v)^2\leq 3(\zeta_M^2+K\zeta_v^2)$, which holds whenever $\zeta_M,\zeta_v\geq 0$ and $K\geq 1$, and (iii) uses $1<3\log 2$ and $8\sqrt{3}<16$. Finally, we prove the remaining inequalities:

(a) Because we have assumed that items lie in the unit ball and that the parameters satisfy

$\|A\|_{\mathrm{F}} \le \zeta_M$ and $\|w_i\| \le \zeta_v$, the unquantized measurements are bounded:

$$\begin{aligned}\left|D_{i;k}(A, v_k)\right| &\le \sup_{x,x' \in B(0,1)} \left|\left\langle xx^\top - x'x'^\top, A\right\rangle + \left\langle x - x', v_k\right\rangle\right| \\ &\le 2\|A\|_{\mathrm{F}} + 2\|v_k\| \\ &\le 2(\zeta_M + \zeta_v),\end{aligned}$$

where we have used triangle inequality for $\left\|xx^\top - x'x'^\top\right\|_{\mathrm{F}} \le 2$ and $\|x - x'\| \le 2$. Because $-\log f(\cdot)$ is L-Lipschitz on this domain, whenever $|z| \le 2(\zeta_M + \zeta_v)$, we have:

$$\left|-\log f(z) + \log\frac{1}{2}\right| = \left|-\log f(z) + \log f(0)\right| \le L|z|.$$

(b) This inequality follows from a standard symmetrization argument. Let $\mathcal{H}$ be a set of N-tuples of functions, where $h \equiv (h_1, \ldots, h_N)$. Given a set of i.i.d. random variables $Z_1, \ldots, Z_N, Z'_1, \ldots, Z'_N$ and a set of Rademacher random variables $\epsilon_1, \ldots, \epsilon_N \in \{-1, +1\}$, we have:

$$\begin{aligned}&\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^N h_i(Z_i) - \mathbb{E}\left[\sum_{i=1}^N h_i(Z_i)\right]\right|\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^N \epsilon_i h_i(Z_i) - \sum_{i=1}^N \epsilon_i h_i(Z'_i)\right|\right] \\ &\le \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^N \epsilon_i h(Z_i)\right|\right] + \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^N \epsilon_i h_i(Z'_i)\right|\right] \\ &= 2\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^N \epsilon_i h_i(Z_i)\right|\right].\end{aligned}$$

In our setting, we have an index set $(i,k) \in [m] \times [K]$ and $h_{i;k} : Z \mapsto -\log f\big(Z \cdot$

$D_{i;k}(A, w_k)\big)$.

(c) We use the fact that the function $-\log f(z)$ is L-Lipschitz over the domain $|z| \leq 2(\zeta_M + \zeta_v)$. We can move the Lipschitz constant out of the expectation by applying [178, Theorem 6.28], reproduced below.

(d) This step first makes use of the fact that the random variables $\epsilon_{i;k} Y_{i;k} \stackrel{d}{=} \epsilon_{i;k}$ are equal in distribution. Then, it consolidates everything using the trace inner product on $\mathbb{R}^{m \times K}$.

(e) This step uses the property of the adjoint $\langle \epsilon, D(\theta) \rangle = \langle D^*(\epsilon), \theta \rangle \leq \|D^*(\epsilon)\| \cdot \|\theta\|$. The first inner product is over $\mathbb{R}^{m \times K}$, the second inner product and norm are over $\mathrm{Sym}(\mathbb{R}^r) \otimes \mathbb{R}^{r \times K}$.

(f) We apply the bound on the parameters $\sup_{\theta \in \Theta} \|\theta\| \leq \sqrt{\zeta_M^2 + K\zeta_v^2}$ along with the following:

$$\begin{aligned}
\mathbb{E} \left\| D^* \epsilon \right\| &\stackrel{(i)}{\leq} \sqrt{\mathbb{E} \left\langle DD^*, \epsilon\epsilon^\top \right\rangle} \\
&\stackrel{(ii)}{=} \sqrt{\left\langle \mathbb{E}[DD^*], \mathbb{E}[\epsilon\epsilon^\top] \right\rangle} \\
&\stackrel{(iii)}{=} \sqrt{\sum_{i,k} \left\| \Delta_{i;k} \oplus \delta_{i;k} \right\|^2} \stackrel{(iv)}{\leq} 2\sqrt{2mK}.
\end{aligned}$$

The (i) uses Jensen's inequality, (ii) uses the independence of the randomness over the design matrices and the Rademacher random variables, (iii) uses the fact that $\mathbb{E}[\epsilon\epsilon^\top]$ is the identity on $\mathbb{R}^{m \times K}$, and (iv) uses the fact that items are contained in the unit Euclidean ball, so that:

$$\|\Delta_{i;k} \oplus \delta_{i;k}\|^2 = \|\Delta_{i;k}\|^2 + \|\delta_{i;k}\|^2 \leq 2^2 + 2^2.$$

□

Remark D.14. *To show that there exists $\mathcal{P}_m$ such that $\sigma^2_{\min}(\mathcal{P}_m) = \Omega(1)$, assume that the space $\mathbb{R}^r$ is quadratically spanned by $\mathcal{X}$. In particular, there exists a collection of items*

$(x_{i_0}, x_{i_1})_{i=1}^n$ *such that its design matrix* D *is full rank. Define* $X_i \in \mathrm{Sym}(\mathbb{R}^r) \oplus \mathbb{R}^r$ *for* $i = 1, \ldots, n$ *by* $X_i = \Delta_i \oplus \delta_i$*. Then,* D^*D *corresponds to:*

$$D^*D = \sum_{i=1}^n X_i X_i^\top,$$

where $\sigma_{\min}(D^*D) > 0$*. Let* $\mathcal{P}_m$ *be constructed by drawing* m *pairs uniformly at random. Let* D_m *be the random design matrix. Let* $I_j \sim \mathrm{Unif}([n])$ *for* $j = 1, \ldots, m$ *be the index of the* j*th random pair, so that we obtain:*

$$\begin{aligned}
\mathbb{E}[D_m^* D_m] &= \mathbb{E}\left[\sum_{i=1}^m X_{I_j} X_{I_j}^\top\right] \\
&= \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \\
&= \frac{m}{n} D^*D.
\end{aligned}$$

It follows that for this choice of random design, we have $\sigma_{\min}^2(\mathcal{P}_m) = \sigma_{\min}(D^*D)$*, which is a constant.*

D.6.3 Auxiliary lemmas

Lemma D.15 (Hoeffding-style inequality for independent bounded random vectors, [72], Corollary 7)**.** *There exists a universal constant* c *such that for any random vectors* $X_1, X_2, \ldots, X_m \in \mathbb{R}^d$ *that are independent and satisfy* $\mathbb{E}[X_i] = 0$ *and* $\|X_i\|_2 \leq \kappa_i$ *for* $i \in [m]$*, we have, for any* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$*,*

$$\left\| \sum_{i=1}^m X_i \right\|_2 \leq c \cdot \sqrt{\sum_{i=1}^m \kappa_i^2 \log \frac{2d}{\delta}}.$$

Corollary D.16 (Matrix version, [72], Corollary 7)**.** *There exists a universal constant* c *such that for any random matrices* $X_1, X_2, \ldots, X_m \in \mathbb{R}^{d \times d}$ *that are independent and*

satisfy $\mathbb{E}[X_i] = 0$ *and* $\|X_i\|_{\mathrm{F}} \leq \kappa_i$ *for* $i \in [m]$*, we have, for any* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$*,*

$$\left\| \sum_{i=1}^{m} X_i \right\|_{\mathrm{F}} \leq c \cdot \sqrt{\sum_{i=1}^{m} \kappa_i^2 \log \frac{2d}{\delta}}.$$

Proof. Since $\log\left(\frac{2d^2}{\delta}\right) \leq 2\log\left(\frac{2d}{\delta}\right)$ for $\delta \leq 1$, the corollary follows directly from Lemma D.15. □

Lemma D.17 (Bounded difference inequality)**.** *Let* $f : \mathcal{X}^N \to \mathbb{R}$ *satisfy the bounded difference property,*

$$\sup_{x_1,\ldots,x_N,x_i'} \left| f(x_1, \ldots, x_N) - f(x_1, \ldots, x_i', \ldots, x_N) \right| \leq C, \qquad \forall i \in [N].$$

Let $X_1, \ldots, X_N$ *be i.i.d. random variables. Then, with probability at least* $1 - \delta$*,*

$$\left| f(X_1, \ldots, X_N) - \mathbb{E}[f(X_1, \ldots, X_N) \right| \leq C\sqrt{2N \log \frac{2}{\delta}}.$$

This theorem is also known as McDiarmid's inequality; as reference, see for example [178, Theorem 6.16].

Lemma D.18 (Theorem 6.28, [178])**.** *Let* h *be an* L*-Lipschitz function* $h : \mathbb{R} \to \mathbb{R}$*. Let* $\mathcal{F}$ *be a function class with functions* $f : \mathcal{Z} \to \mathbb{R}$*. Let* $z_1, \ldots, z_N \in \mathcal{Z}$ *and let* $\epsilon_1, \ldots, \epsilon_N$ *be independent Rademacher random variables. Then:*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{N} \epsilon_i h\big(f(z_i)\big) \right| \right] \leq L \cdot \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{N} \epsilon_i f(z_i) \right| \right].$$

D.7 Details and Additional Results for Section 5.6

Our experimental setup and implementation are inspired by and adapted from [28]. In Section D.7.1, we provide further details to our experimental setup. In Section D.7.2,

we present additional experimental results.

D.7.1 Experimental details

Each simulation run is defined by several parameters: the ambient dimension d, the number of subspaces n, the dimension of each subspace r, the number of users per subspace K, and the number of preference comparisons per user m.

Data generation.

In each simulation run, we generate a symmetric positive definite matrix M from the Wishart distribution $W(d, I_d)$ and normalize it so that $\|M\|_{\mathrm{F}} = d$, following the same procedure in [28, Section F.3]. We generate n r-dimensional subspaces uniformly at random [140]: for each subspace, we independently draw r isotropic random vectors from the normal distribution $\mathcal{N}(0, \frac{1}{d} I_d)$ and use QR decomposition to find an orthonormal basis.

For each subspace V equipped with orthonormal basis B, we randomly generate K user ideal points by sampling independently from $\mathcal{N}(0, \frac{1}{d} I_d)$, and for each user, we generate independently $2m$ items (m pairs). To generate an item $x \in V$, we first draw a vector z from $\mathcal{N}(0, \frac{1}{r} I_d)$, and then compute $x = Pz$, where P is the orthogonoal projection matrix onto V given by $P = BB^\top$. Note that

$$\mathbb{E}\left[\|x\|_2^2\right] = \mathbb{E}\left[\operatorname{tr}\left(xx^\top\right)\right] = \mathbb{E}\left[\operatorname{tr}\left(Pzz^\top P^\top\right)\right] = \operatorname{tr}\left(P\mathbb{E}\left[zz^\top\right]P\right) = \frac{1}{r}\operatorname{tr}(P) = 1.$$

Equivalently, we have $x \sim \mathcal{N}(0, \frac{1}{r} BB^\top)$.

Given subspace noise level $\sigma > 0$, we now describe the procedure for generating items that that lie near V with $\dim(V) = r$. Let $t = 2Km$. We first sample t items in V using the procedure above. We perturb each item by adding an independent noise ξ: To obtain ξ, we first draw $z \sim \mathcal{N}(0, \frac{\sigma^2}{d-r} I_d)$, and then project it onto, $V^\perp$, the orthogonal complement of V, $\xi = \left(I_d - BB^\top\right) z$. Note that $\mathbb{E}\left[\|\xi\|_2^2\right] = \sigma^2$. This is equivalent to independently drawing t vectors from $\mathcal{N}(0, \frac{1}{r} BB^\top + \frac{\sigma^2}{d-r} B_\perp B_\perp^\top)$, where $B_\perp$ is an orthonormal basis of $V^\perp$.

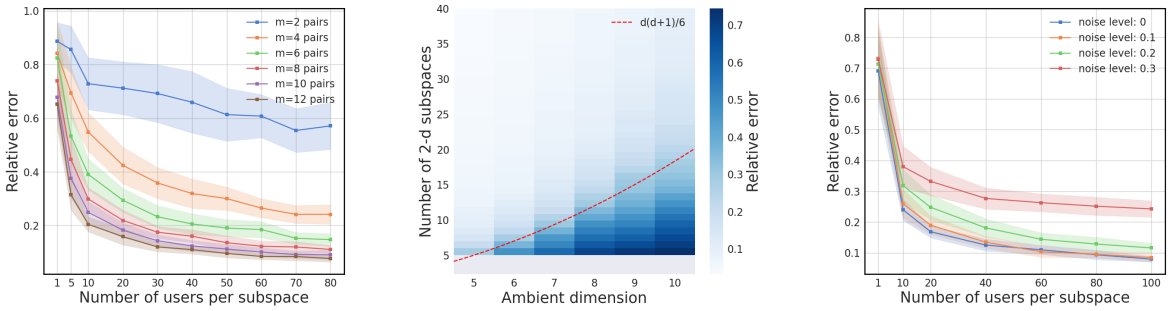
Let $X \in \mathbb{R}^{d \times t}$ be the matrix who columns consist of these perturbed items. We use singular value decomposition on X to compute the rank-r approximation, $\hat{X}$, that minimizes $\|\hat{X} - X\|_{\mathrm{F}}$ (Eckart–Young–Mirsky theorem, see, e.g., [58]). We then use QR decomposition to find an orthonormal basis $\hat{B}$ for the vector space $\hat{V}$ spanned by columns in $\hat{X}$. For each perturbed item $\hat{x}$, its canonical representation in $\hat{V}$ is given by $\hat{x}_{\hat{V}} = \hat{B}^\top \hat{x} \in \mathbb{R}^r$, and these canonical representations are used in Stage 1 of Algorithm 5.

Binary responses are generated using user ideal points and (possibly perturbed) item embeddings in $\mathbb{R}^d$ under the probabilistic model in Assumption 5.17 with link function $f(x; \beta) = \frac{1}{1+\exp(-\beta x)}$. Unless otherwise specified, we set $\beta = 4$; this is the "medium" noise setting considered in [28].

Algorithm implementation.

We provide additional details on the implementation of Algorithm 5. In Stage 1 (learning subspace metrics), we use Algorithm 9 and set constraints based on oracle knowledge of optimal hyperparameters ζ_M and ζ_v (also called the best-case hyperparameter setting in [28]). We use $\ell(x; \beta) = \log(1 + \exp(-\beta x))$ as the loss function, where β is assumed known and given by the logistic link function above. We use the Splitting Conic Solver (SCS) in CVXPy with hyperparameters `eps = 1e4` and `max_iters = 1e5` to solve the convex optimization problem.

In Stage 2 of our practical implementation (reconstruction from subspace metrics), we note that least squares can be sensitive to outliers, and therefore we use the Huber loss for robust regression [66]. In particular, we use the HuberRegressor from scikit-learn [118] with default hyperparameters, except for setting `max_iters = 1e4`. To reconstruct a full metric, we use subspace metrics learned in Stage 1. We note that we do not include a subspace (and the learned subspace metric) into our reconstruction step if CVXPy/SCS does not solve the corresponding optimization problem in Stage 1 successfully, that is, `prob.status != OPTIMAL`. Nevertheless, given n subspaces, if CVXPy/SCS does not



(a) **(b)** **(c)**

Figure D.3. (a) shows the average relative errors over items that lie in a union of 40 2-dimensional subspaces. (b) shows the average relative errors for reconstructing $\hat{M}$ from increasing numbers of 2-dimensional subspaces; for each subspace, 80 users each provides 10 preference comparisons. The dotted red curve illustrates the counting argument in Remark 5.14; here, each 2-dimensional subspace can contribute at most 3 degrees of freedom. (c) shows the average relative errors for varying subspace noise levels; here, items lie approximately in a union of 40 2-dimensional subspaces and each user provides 10 preference comparisons.

successfully solve any of them, we use the n-th subspace alone for reconstruction.

D.7.2 Additional experimental results

We ran the same experiments in Section 5.6 for subspace dimension $r = 2$, with slightly different parameters. The response noise was again set to $\beta = 4$, and each experiment was run 30 times. Figure D.3a compares the average relative errors for varying K and m, where items lie in a union of 40 subspaces. Figure D.3b shows the average errors given increasing numbers of subspaces, where $K = 80$ and $m = 10$. Note that by the dimension-counting argument in Remark 5.14, each 2-dimensional subspace contributes at most $\frac{2(2+1)}{2} = 3$ degrees of freedom, and therefore a minimum of $\left\lceil \frac{d(d+1)}{6} \right\rceil$ subspaces are needed. Figure D.3b shows the average recovery errors for varying subspace noise levels, $\sigma \in \{0, 0.1, 0.2, 0.3\}$, and varying K, where items lie in a union of 40 subspaces and we set $m = 10$. The behaviors demonstrated in these experiments are analogous to those observed for $r = 1$ discussed in Section 5.6.

Bibliography

[1] D. Abel, Y. Jinnai, S. Y. Guo, G. Konidaris, and M. Littman. Policy and value transfer in lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 20–29. PMLR, 2018.

[2] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

[3] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[4] S. Agrawal and N. Goyal. Near-optimal regret bounds for thompson sampling. *J. ACM*, 64(5), sep 2017. ISSN 0004-5411. doi: 10.1145/3088510. URL https://doi.org/10.1145/3088510.

[5] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

[6] R. Arora, T. V. Marinov, and M. Mohri. Corralling stochastic bandit algorithms. *arXiv preprint arXiv:2006.09255*, 2020.

[7] J.-Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, pages 150–165. Springer, 2007.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[9] B. Awerbuch and R. D. Kleinberg. Competitive collaborative learning. In *International Conference on Computational Learning Theory*, pages 233–248. Springer, 2005.

[10] M. G. Azar, A. Lazaric, and E. Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2220–2228. Curran Associates, Inc., 2013.

[11] M. G. Azar, A. Lazaric, and E. Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2220–2228. Curran Associates, Inc., 2013.

[12] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[13] Y. Bar-On and Y. Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3110–3120, 2019.

[14] P. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In T. Zhang and R. A. Sevedio, editors, *Proceedings of the 21st Annual Conference on Learning Theory - COLT 2008*, pages 335–342, United States, 2008. Omnipress.

[15] P. L. Bartlett and A. Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42, 2009.

[16] H. Bastani, D. Simchi-Levi, and R. Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 2021.

[17] S. Basu, B. Kveton, M. Zaheer, and C. Szepesvári. No regrets for learning the prior in bandits. *arXiv preprint arXiv:2107.06196*, 2021.

[18] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[19] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[20] I. Bistritz, T. Baharav, A. Leshem, and N. Bambos. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11–21, 2020.

[21] E. Boursier and V. Perchet. Selfish robustness and equilibria in multi-player bandits. In *Conference on Learning Theory*, pages 530–581. PMLR, 2020.

[22] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[23] G. Bresler, G. H. Chen, and D. Shah. A latent source model for online collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 3347–3355, 2014.

[24] E. Brunskill and L. Li. Sample complexity of multi-task reinforcement learning. *UAI*, 2013.

[25] S. Bubeck and T. Budzinski. Coordination without communication: optimal regret in two players multi-armed bandits. In *Conference on Learning Theory*, pages 916–939. PMLR, 2020.

[26] S. Bubeck, Y. Li, Y. Peres, and M. Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, pages 961–987, 2020.

[27] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff. Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 289–300, 2014.

[28] G. Canal, B. Mason, R. Korlakai Vinayak, and R. Nowak. One for all: Simultaneous metric and preference learning over multiple users. *Advances in Neural Information Processing Systems*, 35:4943–4956, 2022.

[29] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 142–151, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

[30] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[31] N. Cesa-Bianchi, C. Gentile, and G. Zappella. A gang of bandits. In *Advances in Neural Information Processing Systems*, pages 737–745, 2013.

[32] N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

[33] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

[34] R. Chawla, A. Sankararaman, A. Ganesh, and S. Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International conference on artificial intelligence and statistics*, pages 3471–3481. PMLR, 2020.

[35] K. Christakopoulou and A. Banerjee. Learning to interact with users: A collaborative-bandit approach. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 612–620. SIAM, 2018.

[36] R. Cohen, M. Y. Cheng, and M. W. Fleming. Why bother about bother: Is it worth

it to ask the user. In *Proceedings of AAAI Fall Symposium*, 2005.

[37] L. Colucci, P. Doshi, K.-L. Lee, J. Liang, Y. Lin, I. Vashishtha, J. Zhang, and A. Jude. Evaluating item-item similarity algorithms for movies. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 2141–2147, 2016.

[38] C. H. Coombs. Psychological scaling without a unit of measurement. *Psychological review*, 57(3):145, 1950.

[39] C. Dann and E. Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2818–2826, 2015.

[40] C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5717–5727, 2017.

[41] C. Dann, L. Li, W. Wei, and E. Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

[42] C. D'Eramo, D. Tateo, A. Bonarini, M. Restelli, and J. Peters. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgpv2VFvr.

[43] A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. In *Advances in neural information processing systems*, pages 4848–4856, 2017.

[44] M. Dimakopoulou and B. Van Roy. Coordinated exploration in concurrent reinforcement learning. In *International Conference on Machine Learning*, pages 1271–1279. PMLR, 2018.

[45] D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52 (4):1289–1306, 2006.

[46] A. Dubey and A. Pentland. Cooperative multi-agent bandits with heavy tails. In *Proceedings of the 37th International Conference on Machine Learning*, pages 451–460, 2020.

[47] A. Dubey and A. Pentland. Kernel methods for cooperative contextual bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pages 428–450, 2020.

[48] A. Dubey and A. Pentland. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

[49] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[50] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35 (11):2765–2781, 2013.

[51] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[52] R. Feraud, R. Alami, and R. Laroche. Decentralized exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1901–1909, 2019.

[53] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[54] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[55] W. Gander, G. H. Golub, and R. Strebel. Least-squares fitting of circles and ellipses. *BIT Numerical Mathematics*, 34:558–578, 1994.

[56] F. M. Garcia and P. S. Thomas. A meta-mdp approach to exploration for lifelong reinforcement learning. *arXiv preprint arXiv:1902.00843*, 2019.

[57] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765, 2014.

[58] G. H. Golub, A. Hoffman, and G. W. Stewart. A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88: 317–327, 1987.

[59] R. D. Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

[60] Z. Guo and E. Brunskill. Concurrent pac rl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[61] H. Hadiji. Polynomial cost of adaptation for x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1029–1038, 2019.

[62] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[63] J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh. Hierarchical bayesian bandits. *arXiv preprint arXiv:2111.06929*, 2021.

[64] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin. Collaborative metric learning. In *Proceedings of the 26th international conference on world wide web*, pages 193–201, 2017.

[65] J. Hu, X. Chen, C. Jin, L. Li, and L. Wang. Near-optimal representation learning for linear bandits and linear rl. *arXiv preprint arXiv:2102.04132*, 2021.

[66] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[67] L. Jain, K. G. Jamieson, and R. Nowak. Finite sample prediction and recovery bounds for ordinal embedding. *Advances in neural information processing systems*, 29, 2016.

[68] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[69] K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

[70] A. J. Jeckmans, M. Beye, Z. Erkin, P. Hartel, R. L. Lagendijk, and Q. Tang. Privacy in recommender systems. *Social media retrieval*, pages 263–281, 2013.

[71] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

[72] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[73] T. Jin, P. Xu, J. Shi, X. Xiao, and Q. Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.

[74] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multi-

armed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[75] S. Kar, H. V. Poor, and S. Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 1771–1778. IEEE, 2011.

[76] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[77] M. Kleindessner and U. von Luxburg. Kernel functions based on triplet comparisons. *Advances in neural information processing systems*, 30, 2017.

[78] R. K. Kolla, K. Jagannathan, and A. Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4): 1782–1795, 2018.

[79] N. Korda, B. Szorenyi, and S. Li. Distributed clustering of linear bandits in peer to peer networks. In *International Conference on Machine Learning*, pages 1301–1309, 2016.

[80] R. M. Kretchmar. Parallel reinforcement learning. In *The 6th World Conference on Systemics, Cybernetics, and Informatics*. Citeseer, 2002.

[81] A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Conference on Learning Theory*, pages 2025–2027, 2019.

[82] A. Kubota, E. I. Peterson, V. Rajendren, H. Kress-Gazit, and L. D. Riek. Jessie: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 121–130, 2020.

[83] A. Kubota, E. I. Peterson, V. Rajendren, H. Kress-Gazit, and L. D. Riek. Jessie: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 121–130, 2020.

[84] B. Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

[85] B. Kveton, M. Konobeev, M. Zaheer, C.-W. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning*

Research, pages 5884–5893. PMLR, 18–24 Jul 2021.

[86] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[87] P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

[88] P. C. Landgren. *Distributed Multi-Agent Multi-Armed Bandits*. PhD thesis, Princeton University, 2019.

[89] N. C. Landolfi, G. Thomas, and T. Ma. A model-based approach for sample-efficient multi-task reinforcement learning. *arXiv preprint arXiv:1907.04964*, 2019.

[90] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[91] A. Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.

[92] A. Lazaric and M. Restelli. Transfer from multiple mdps. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[93] A. Lazaric, M. Restelli, and A. Bonarini. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 544–551, 2008.

[94] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

[95] S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

[96] S. Li, W. Chen, S. Li, and K.-S. Leung. Improved algorithm on online clustering of bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2923–2929. AAAI Press, 2019.

[97] X. Liang, Y. Liu, T. Chen, M. Liu, and Q. Yang. Federated transfer reinforcement learning for autonomous driving. *arXiv preprint arXiv:1910.06001*, 2019.

[98] B. Liu, L. Wang, and M. Liu. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019.

[99] K. Liu and Q. Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

[100] Y. Liu, Z. Guo, and E. Brunskill. Pac continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, page 438–446, 2016.

[101] A. Locatelli and A. Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492, 2018.

[102] Y. M. Lu and M. N. Do. A theory for sampling signals from a union of subspaces. *IEEE transactions on signal processing*, 56(6):2334–2345, 2008.

[103] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM review*, 50(3): 413–458, 2008.

[104] S. Mac Lane and G. Birkhoff. *Algebra*, volume 330. American Mathematical Soc., 1999.

[105] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

[106] B. Mason, L. Jain, and R. Nowak. Learning low-dimensional metrics. *Advances in neural information processing systems*, 30, 2017.

[107] A. K. Massimino and M. A. Davenport. As you like it: Localization via paired comparisons. *The Journal of Machine Learning Research*, 22(1):8357–8395, 2021.

[108] J. Matousek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2013.

[109] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *COLT*, 2009.

[110] A. Mehrabian, E. Boursier, E. Kaufmann, and V. Perchet. A practical algorithm for multiplayer bandits when arm means vary among players. In *International Conference on Artificial Intelligence and Statistics*, pages 1211–1221. PMLR, 2020.

[111] T. T. Nguyen and H. W. Lauw. Dynamic clustering of contextual multi-armed

bandits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1959–1962, 2014.

[112] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

[113] A. OroojlooyJadid and D. Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

[114] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

[115] A. Pacchiano, M. Phan, Y. Abbasi-Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. *arXiv preprint arXiv:2003.01704*, 2020.

[116] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, 6(1):90–105, 2004.

[117] J. Pazis and R. Parr. Efficient pac-optimal exploration in concurrent, continuous state mdps with delayed updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[118] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[119] A. Peleg, N. Pearl, and R. Meir. Metalearning linear bandits by prior update. *arXiv preprint arXiv:2107.05320*, 2021.

[120] X. Qian, H. Feng, G. Zhao, and T. Mei. Personalized recommendation combining user interest and social circle. *IEEE transactions on knowledge and data engineering*, 26(7):1763–1777, 2013.

[121] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[122] L. D. Riek. Healthcare robotics. *Communications of the ACM*, 60(11):68–78, 2017.

[123] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, 2005.

[124] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[125] Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

[126] A. Sanakoyeu, V. Tschernezki, U. Buchler, and B. Ommer. Divide and conquer the embedding space for metric learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 471–480, 2019.

[127] A. Sankararaman, A. Ganesh, and S. Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

[128] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

[129] S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790. IEEE, 2017.

[130] N. Sharma, S. Basu, K. Shanmugam, and S. Shakkottai. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

[131] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha. Portfolio choices with orthogonal bandit learning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 974–980. AAAI Press, 2015. ISBN 9781577357384.

[132] C. Shi, W. Xiong, C. Shen, and J. Yang. Decentralized multi-player multi-armed bandits with no collision information. In *International Conference on Artificial Intelligence and Statistics*, pages 1519–1528. PMLR, 2020.

[133] P. Shivaswamy and T. Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054, 2012.

[134] D. Silver, L. Newnham, D. Barker, S. Weller, and J. McFall. Concurrent reinforcement learning from customer interactions. In *International conference on machine learning*, pages 924–932. PMLR, 2013.

[135] D. Silver, L. Newnham, D. Barker, S. Weller, and J. McFall. Concurrent reinforcement

learning from customer interactions. In *International conference on machine learning*, pages 924–932. PMLR, 2013.

[136] M. Simchowitz and K. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*, 2019.

[137] A. Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

[138] M. Soare, O. Alsharif, A. Lazaric, and J. Pineau. Multi-task linear bandits. *NIPS2014 Workshop on Transfer and Multi-task Learning : Theory meets Practice*, 2014.

[139] L. Song, C. Tekin, and M. Van Der Schaar. Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing*, 9(3):433–445, 2014.

[140] G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3): 403–409, 1980.

[141] B. Szörényi, R. Busa-Fekete, I. Hegedűs, R. Ormándi, M. Jelasity, and B. Kégl. Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pages 1056–1064. International Machine Learning Society, 2013.

[142] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on Machine Learning*, pages 673–680, 2011.

[143] F. Tanaka and M. Yamamura. Multitask reinforcement learning on the distribution of mdps. In *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, volume 3, pages 1108–1113. IEEE, 2003.

[144] G. Tatli, Y. Chen, and R. K. Vinayak. Learning populations of preferences via pairwise comparison queries. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

[145] M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

[146] M. E. Taylor, N. K. Jong, and P. Stone. Transferring instances for model-based reinforcement learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 488–505. Springer, 2008.

[147] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[148] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[149] A. Tirinzoni, M. Salvini, and M. Restelli. Transfer of samples in policy search via multiple importance sampling. In *International Conference on Machine Learning*, pages 6264–6274. PMLR, 2019.

[150] K. Tsiakas, C. Abellanoza, and F. Makedon. Interactive learning and adaptation for robot assisted therapy for people with dementia. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–4, 2016.

[151] L. Van Der Maaten and K. Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

[152] N. Verma and K. Branson. Sample complexity of learning mahalanobis distance metrics. *Advances in neural information processing systems*, 28, 2015.

[153] D. Vial, S. Shakkottai, and R. Srikant. Robust multi-agent multi-armed bandits. *arXiv preprint arXiv:2007.03812*, 2020.

[154] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[155] R. Wan, L. Ge, and R. Song. Metadata-based multi-task bandits with bayesian hierarchical models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[156] H. Wang, Q. Wu, and H. Wang. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[157] P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

[158] Q. Wang, C. Zeng, W. Zhou, T. Li, S. S. Iyengar, L. Shwartz, and G. Y. Grabarnik. Online interactive collaborative filtering using multi-armed bandit with dependent arms. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1569–1580, 2018.

[159] X. Wang, S. C. Hoi, C. Liu, and M. Ester. Interactive social recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 357–366, 2017.

[160] Y. Wang, J. Hu, X. Chen, and L. Wang. Distributed bandit learning: How much communication is needed to achieve (near) optimal regret. *arXiv preprint arXiv:1904.06309*, 2019.

[161] Z. Wang, C. Zhang, M. K. Singh, L. Riek, and K. Chaudhuri. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 1531–1539. PMLR, 2021.

[162] Z. Wang, C. Zhang, and K. Chaudhuri. Thompson sampling for robust transfer in multi-task bandits. In *International Conference on Machine Learning*, pages 23363–23416. PMLR, 2022.

[163] Z. Wang, G. So, and R. K. Vinayak. Metric learning from limited pairwise preference comparisons. *Manuscript*, 2024.

[164] C.-Y. Wei, H. Luo, and A. Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pages 3583–3634. PMLR, 2020.

[165] Q. Wu, H. Wang, Q. Gu, and H. Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

[166] Y. Wu, A. György, and C. Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.

[167] A. Xu and M. Davenport. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33:454–465, 2020.

[168] H. Xu, T. Ma, and S. S. Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

[169] X. Xu, S. Vakili, Q. Zhao, and A. Swami. Online learning with side information. In *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*, pages 303–308, 2017.

[170] K. Yang, L. Yang, and S. Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

[171] J. Yu, D. Tao, J. Li, and J. Cheng. Semantic preserving distance metric learning and applications. *Information Sciences*, 281:674–686, 2014.

[172] A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

[173] C. Zhang and Z. Wang. Provably efficient multi-task reinforcement learning with model transfer. *Advances in Neural Information Processing Systems*, 34:19771–19783, 2021.

[174] C. Zhang, A. Agarwal, H. Daumé III, J. Langford, and S. Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *International Conference on Machine Learning*, pages 7335–7344, 2019.

[175] J. Zhang and E. Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.

[176] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881. PMLR, 2018.

[177] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[178] T. Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

[179] Z. Zhang, Y. Zhou, and X. Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020.

[180] Z. Zhu, L. Huang, and H. Xu. Collaborative thompson sampling. *Mobile Networks and Applications*, 2020. doi: 10.1007/s11036-019-01453-x.

[181] Z. Zhu, K. Lin, and J. Zhou. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020.

[182] H. H. Zhuo, W. Feng, Q. Xu, Q. Yang, and Y. Lin. Federated reinforcement learning. *arXiv preprint arXiv:1901.08277*, 2019.