On Interpolating Experts and Multi-Armed Bandits

Houshuang Chen[1] **Yuchen He**[1] **Chihao Zhang**[2]

Abstract

Learning with expert advice and multi-armed bandit are two classic online decision problems which differ on how the information is observed in each round of the game. We study a family of problems interpolating the two. For a vector $\mathbf{m} = (m_1, \dots, m_K) \in \mathbb{N}^K$, an instance of $\mathbf{m}$-MAB indicates that the arms are partitioned into K groups and the i-th group contains m_i arms. Once an arm is pulled, the losses of all arms in the same group are observed. We prove tight minimax regret bounds for $\mathbf{m}$-MAB and design an optimal PAC algorithm for its pure exploration version, $\mathbf{m}$-BAI, where the goal is to identify the arm with minimum loss with as few rounds as possible. We show that the minimax regret of $\mathbf{m}$-MAB is $\Theta\left(\sqrt{T \sum_{k=1}^{K} \log(m_k + 1)}\right)$ and the minimum number of pulls for an $(\varepsilon, 0.05)$-PAC algorithm of $\mathbf{m}$-BAI is $\Theta\left(\frac{1}{\varepsilon^2} \cdot \sum_{k=1}^{K} \log(m_k + 1)\right)$. Both our upper bounds and lower bounds for $\mathbf{m}$-MAB can be extended to a more general setting, namely the bandit with graph feedback, in terms of the *clique cover* and related graph parameters. As consequences, we obtained tight minimax regret bounds for several families of feedback graphs.

1. Introduction

A typical family of online decision problems is as follows: In each round of the game, the player chooses one of N arms to pull. At the same time, the player will incur a loss of the pulled arm. The objective is to minimize the expected regret defined as the difference between the cumulative losses of the player and that of the single best arm over T rounds. The minimax regret, denoted as $R^*(T)$, represents the minimum expected regret achievable by any algorithm against the worst loss sequence.

There are variants of the problem according to amount of information the player can observe in each round. In the problem of multi-armed bandit (MAB), the player can only observe the loss of the arm just pulled. The minimax regret is $\Theta\left(\sqrt{NT}\right)$ (Audibert & Bubeck, 2009). Another important problem is when the player can observe the losses of all arms in each round, often refered to as learning with expert advice. The minimax regret is $\Theta\left(\sqrt{T \log N}\right)$ (Freund & Schapire, 1997; Haussler et al., 1995). Bandit with graph feedback generalizes and interpolates both models. In this model, a directed graph G, called the feedback graph, is given. The vertex set of G is the set of arms and a directed edge from i to j indicates that pulling the arm i can observe the loss of arm j. As a result, the MAB corresponds to when G consists of singletons with self-loop, and learning with expert advice corresponds to when G is a clique. A number of recent works devote to understanding how the structure of G affects the minimax regret (Alon et al., 2015; Chen et al., 2021; He & Zhang, 2023; Eldowa et al., 2024; Kocák & Carpentier, 2023; Rouyer et al., 2022; Dann et al., 2023).

In this paper, we consider a natural interpolation between learning with expert advice and multi-armed bandit. Let $\mathbf{m} = (m_1, m_2, \dots, m_K) \in \mathbb{N}^K$ be a vector with each $m_i \geq 1$. An instance of $\mathbf{m}$-MAB is that the all N arms are partitioned into K groups and the pull of each arm can observe the losses of all arms in the same group. In the language of bandit with graph feedback, the feedback graph G is the disjoint union of K cliques with size $m_1, m_2, \dots, m_k$ respectively. We show that the minimax regret for $\mathbf{m}$-MAB is $\Theta\left(\sqrt{T \cdot \sum_{k \in [K]} \log(m_k + 1)}\right)$. As a result, this generalizes the optimal regret bounds for both MAB and learning with expert advice.

A closely related problem is the so-called "pure exploration" version of bandit, often referred to as the *best arm identification* (BAI) problem where the loss of each arm follows some (unknown) distribution. The goal of the problem is to identify the arm with minimum mean loss with as few rounds as possible. Similarly, we introduced the problem of $\mathbf{m}$-BAI with the same feedback pattern as $\mathbf{m}$-MAB. We design an $(\varepsilon, 0.05)$-PAC algorithm for $\mathbf{m}$-BAI which ter-

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China [2] John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Chihao Zhang <chihao@sjtu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

minates in $T = O\left(\frac{1}{\varepsilon^2}\sum_{k\in[K]}\log(m_k+1)\right)$ rounds for every $\varepsilon < \frac{1}{8}$. This means that after T rounds of the game, with probability at least 0.95, the algorithm can output an arm whose mean loss is less than ε plus the mean of the best one. We show that our algorithm is optimal by proving a matching lower bound $\Omega\left(\frac{1}{\varepsilon^2}\sum_{k\in[K]}\log(m_k+1)\right)$ for any $(\varepsilon, 0.05)$-PAC algorithm.

Both our upper bounds and lower bounds for the minimax regret of $\mathbf{m}$-MAB can be generalized to bandit with graph feedback. To capture the underlying structure necessary for our proofs, we introduce some new graph parameters which yield optimal bound for several families of feedback graphs. The main results are summarized in Section 1.1.

Our algorithm deviates from the standard *online stochastic mirror descent* (OSMD) algorithm for bandit problems. We employ the two-stage OSMD developed in (He & Zhang, 2023) and give a novel analysis which yields the optimal regret bound. For the lower bound, we prove certain new "instance-specific" lower bounds for the best arm identification problem. These lower bounds may find applications in other problems. We will give an overview of our techniques in Section 1.2.

1.1. Main Results

We summarize our main results in this section. Formal definitions of $\mathbf{m}$-MAB, $\mathbf{m}$-BAI and bandit with graph feedback are in Section 2. All the proof details can be found in the appendix.

Theorem 1.1. *There exists an algorithm such that for any instance of $(m_1,\dots,m_K)$-MAB, any $T > 0$ and any loss sequence $\ell^{(0)},\ell^{(1)},\dots,\ell^{(T-1)} \in [0,1]^N$, its regret is at most*

$$c\cdot\sqrt{T\cdot\sum_{k=1}^{K}\log(m_k+1)},$$

where $c > 0$ is a universal constant.

Given an instance of $\mathbf{m}$-BAI, for $\varepsilon,\delta \in (0,1)$, an (ε,δ)-PAC algorithm can output an arm whose mean loss is less than ε plus the mean of the optimal one with probability at least $1-\delta$. Using a reduction from $\mathbf{m}$-BAI to $\mathbf{m}$-MAB (Lemma A.1), we obtain a PAC algorithm for $\mathbf{m}$-BAI:

Theorem 1.2. *There exists an $(\varepsilon, 0.05)$-PAC algorithm for $(m_1,\dots,m_K)$-BAI which pulls*

$$T \le c\cdot\sum_{k=1}^{K}\frac{\log(m_k+1)}{\varepsilon^2}$$

arms where $c > 0$ is a universal constant.

Let $\mathtt{Ber}(p)$ denote the Bernoulli distribution with mean p. We complement the above algorithm with the following lower bound:

Theorem 1.3. *There exists an instance $\mathscr{H}$ such that for every $(\varepsilon, 0.05)$-PAC algorithm $\mathcal{A}$ of $(m_1,\dots,m_K)$-BAI with $\varepsilon \in \left(0,\frac{1}{8}\right)$, the expected number of pulls T of $\mathcal{A}$ on $\mathscr{H}$ satisfies*

$$\mathbf{E}[T] \ge c'\cdot\sum_{k=1}^{T}\frac{\log(m_k+1)}{\varepsilon^2},$$

where $c' > 0$ is a universal constant. Moreover, we can pick $\mathscr{H}$ as the one in which each arm follows $\mathtt{Ber}(\frac{1}{2})$.

Using the reduction from $\mathbf{m}$-BAI to $\mathbf{m}$-MAB (Lemma A.1) again, we obtain the lower bound for $\mathbf{m}$-MAB.

Theorem 1.4. *For any algorithm $\mathcal{A}$ of $(m_1,\dots,m_k)$-MAB, for any sufficiently large $T > 0$, there exists a loss sequence $\ell^{(0)},\ell^{(1)},\dots,\ell^{(T-1)}$ such that the regret of $\mathcal{A}$ in T rounds is at least*

$$c'\cdot\sqrt{T\cdot\sum_{k=1}^{K}\log(m_k+1)},$$

where $c' > 0$ is a universal constant.

Our results generalize to the setting of *bandit with graph feedback*. Let $G = (V,E)$ be a directed graph with self-loop on each vertex. Let $V_1,\dots,V_K \subseteq V$ be subsets of vertices. We say that they form a $(V_1,\dots,V_K)$-*clique cover* of G if each induced subgraph $G[V_k]$ for $k\in[K]$ is a clique and $\bigcup_{k\in[K]}V_k = V$.

Corollary 1.5. *Let G be a feedback graph with a self-loop on each vertex. If G contains a $(V_1,\dots,V_K)$-clique cover where $|V_k| = m_k$ for every $k\in[K]$, then the minimax regret of bandit with graph feedback G is at most*

$$c\cdot\sqrt{T\cdot\sum_{k=1}^{K}\log(m_k+1)}$$

for some universal constant $c > 0$.

Our lower bounds generalize to bandit with graph feedback as well. The terms "strongly observable feedback graphs" and "weakly observable feedback graphs" are defined in Section 2.

Theorem 1.6. *Let $G = (V,E)$ be the feedback graph. Assume that there exist K* disjoint *sets $S_1,\dots,S_K \subseteq V$ such that*

- *each $G[S_k]$ is a strongly observable graph with a self-loop on each vertex;*
- *there is no edge between S_i and S_j for any $i \ne j$.*

Then for any algorithm $\mathcal{A}$ and any sufficiently large time horizon $T > 0$, there exists some loss sequence on which the regret of $\mathcal{A}$ is at least $c' \cdot \sqrt{T \cdot \sum_{k=1}^{K} \log\left(|S_k| + 1\right)}$ for some universal constant $c' > 0$.

The following lower bound for weakly observable feedback graphs confirms a conjecture in (He & Zhang, 2023) and implies the optimality of several regret bounds established there, e.g., when the feedback graph is the disjoint union of loopless complete bipartite graphs. The notion of t-packing independent set is defined in Section 2.

Theorem 1.7. *Let $G = (V, E)$ be the feedback graph. Assume that V can be partitioned into K disjoint sets $V = V_1 \cup V_2 \cup \cdots \cup V_K$ such that*

- *for every $k \in [K]$, each $G[V_k]$ is observable;*
- *for every $k \in [K]$, there exists a t_k-packing independent set S_k in $G[V_k]$ such that every vertex in S_k does not have a self-loop;*
- *there is no edge from V_i to S_j for any $i \neq j$ in G.*

Then for any algorithm $\mathcal{A}$ and any sufficiently large time horizon $T > 0$, there exists some loss sequence on which the regret of $\mathcal{A}$ with feedback graph G is at least $c' \cdot T^{\frac{2}{3}} \cdot \left(\sum_{k=1}^{K} \max\left\{\log|S_k|, \frac{|S_k|}{t_k}\right\}\right)^{\frac{1}{3}}$ for some universal constant $c' > 0$.

Theorem 1.7 implies tight regret lower bounds for several weakly observable graphs. We summarize the minimax regret for some feedback graphs, weakly or strongly observable, in Table 1.

1.2. Overview of Technique & Contribution

We note that a simple reduction (Lemma A.1) implies that any algorithm for $\mathbf{m}$-MAB can be turned into a PAC algorithm for $\mathbf{m}$-BAI. As a result, Theorems 1.1 to 1.4 follow from a minimax regret upper bound for $\mathbf{m}$-MAB and a lower bound for $\mathbf{m}$-BAI.

1.2.1. Upper bounds for $\mathbf{m}$-MAB

We design a new two-stage algorithm (Algorithm 1) to establish an upper bound for $\mathbf{m}$-MAB. The algorithm is similar to the one used in (He & Zhang, 2023) to study weakly observable graphs with a few tweaks to accommodate our new analysis.

The algorithm maintains a distribution $Y^{(t)}$ over K groups and for each group $k \in [K]$, it maintains a distribution $X_k^{(t)}$ for arms in that group. In each round of the game, the algorithm pulls an arm in a two-stage manner: First pick the group according to the distribution over groups and then pick the arm in that group following the distribution in the group. At the end of each round, all distributions are updated in the manner similar to *online stochastic mirror descent* (OSMD) with carefully designed loss vectors and various potential functions. Each group can be viewed as a super arm, and $Y^{(t)}$ updates with the corresponding loss sequence $\hat{L}^{(t)}(k)$ for $k \in [K]$, while $X_k^{(t)}$ updates with the corresponding loss estimator $\hat{\ell}_k^{(t)}$ in group k.

Our main technical contribution is a novel analysis of this two-stage algorithm. We design auxiliary two-stage *piecewise continuous processes* whose regret is relatively easy to analyze. Then we view our algorithm as a discretization of the process and bound the accumulated discretization errors.

Our new analysis is the key to the tight regret bound. If we apply the classical analysis for OSMD to the two-stage algorithm, as done in (He & Zhang, 2023), the regret decomposes into two parts: (1) the regret due to choosing the group and (2) the regret due to running in the optimal group k^*. Let $R^{(t)}(\ell)$ denote the t-th instant regret for loss sequence ℓ. That is

$$R^{(t)}(\ell) \leq O\left(R^{(t)}(\hat{L}) + R^{(t)}(\hat{\ell}_{k^*})\right).$$

The first part is easy to bound, while the second part is challenging because it contains a factor $\frac{1}{Y^{(t)}(k^*)}$ (the inverse of the probability to choose the optimal group at each round) which is usually hard to bound. However, with our new analysis for OSMD, the regret in the second part can be improved to the expectation of the regret for each group. That is

$$R^{(t)}(\ell) \leq O\left(R^{(t)}(\hat{L}) + \sum_{k \in [K]} Y^{(t)}(k) \cdot R^{(t)}(\hat{\ell}_k)\right).$$

Technically, the $Y^{(t)}(k)$ term will eliminate the $\frac{1}{Y^{(t)}(k)}$ factor, making it possible for the optimal bound (see Lemma B.4 for details).

Since the notion of $\mathbf{m}$-MAB generalizes both learning with expert advice and multi-armed bandit, we remark that our analysis of Algorithm 1 can specialize to an analysis of both ordinary mirror descent (MD) algorithm and OSMD algorithm. We believe that the viewpoint of discretizing a piecewise continuous process is more intuitive than the textbook analysis of OSMD and may be of independent pedagogical interest.

1.2.2. Lower bounds for $\mathbf{m}$-BAI

Our lower bound for the number of rounds in an $(\varepsilon, 0.05)$-PAC algorithm for $\mathbf{m}$-BAI where $\mathbf{m} = (m_1, \ldots, m_K)$ is

$$\Omega\Big(\sum_{k=1}^{K} \frac{\log(m_k + 1)}{\varepsilon^2}\Big),$$

Table 1. Minimax Regret Bound on Various Feedback Graphs

Graph Type	Previous Result	This Work
General strongly observable graphs with self-loops	$O\left(\sqrt{\alpha T}\log NT\right)$ $\Omega\left(\sqrt{\alpha T}\right)$ [1]	$O\left(\sqrt{T\sum_{k=1}^{K}\log m_k}\right)$ [2] See Theorem 1.6 for the lower bound
Disjoint union of K cliques	$O\left(\sqrt{KT}\log NT\right)$ $\Omega\left(\sqrt{KT}\right)$	$\Theta\left(\sqrt{T\sum_{k=1}^{K}\log m_k}\right)$
General weakly observable graphs	$\Omega\left(T^{\frac{2}{3}}\max\left\{\frac{\lvert S\rvert}{k},\log\lvert S\rvert\right\}^{\frac{1}{3}}\right)$ [3]	$\Omega\left(T^{\frac{2}{3}}\left(\sum_{k=1}^{K}\max\left\{\log\lvert S_k\rvert,\frac{\lvert S_k\rvert}{t_k}\right\}\right)^{\frac{1}{3}}\right)$
Disjoint union of K loopless bipartite graphs	$\Omega\left(T^{\frac{2}{3}}\left(\log N\right)^{\frac{1}{3}}\right)$	$\Omega\left(T^{\frac{2}{3}}\left(\sum_{k=1}^{K}\log m_k\right)^{\frac{1}{3}}\right)$

[1] Here α is the independence number of the graph.
[2] Here K is the clique cover number of the graph and $m_1, m_2, \dots m_K$ are the size of the K cliques respectively.
[3] Here S is a t-packing independent set of the graph. S_k and t_k are defined in Theorem 1.7.
[4] Previous results are from (Alon et al., 2015), (Alon et al., 2017) and (Chen et al., 2021).

which is the sum of lower bounds on each (m_k)-$\mathtt{BAI}$ instance. To achieve this, we show that the instance where all arms are $\mathtt{Ber}(\frac{1}{2})$ is in fact a universal hard instance in the sense that every $(\varepsilon, 0.05)$-PAC algorithm requires $\Omega\left(\sum_{k=1}^{K}\frac{\log(m_k+1)}{\varepsilon^2}\right)$ to identify. Via a reduction of "direct-sum" flavor, we show that every $(\varepsilon, 0.05)$-PAC algorithm, when applied to this instance, must successfully identify that each group consists of $\mathtt{Ber}(\frac{1}{2})$ arms. As a result, the lower bound is the sum of the lower bounds for each "all $\mathtt{Ber}(\frac{1}{2})$" (m_k)-$\mathtt{BAI}$ instance.

We then prove the lower bound for "all $\mathtt{Ber}(\frac{1}{2})$" (m)-$\mathtt{BAI}$ instance for every $m \geq 2$. We use $\mathscr{H}_0^{(m)}$ to denote this instance. The $\mathscr{H}_0^{(m)}$ specified lower bound is obtained by constructing another m instances $\mathscr{H}_1^{(m)}, \dots, \mathscr{H}_m^{(m)}$ and compare the distribution of losses generated by $\mathscr{H}_0^{(m)}$ and the distribution of losses generated by a *mixture* of $\mathscr{H}_1^{(m)}, \dots, \mathscr{H}_m^{(m)}$. For technical reasons, we first prove the lower bound when all arms are Gaussian and reduce the Gaussian arms to Bernoulli arms.

1.3. Related Works

The bandit feedback setting as an online decision problem has received considerable attention. The work of (Audibert & Bubeck, 2009) first provided a tight bound for the bandit feedback setting, while the full information feedback case has been well studied in (Freund & Schapire, 1997; Haussler et al., 1995). Building upon these works, (Mannor & Shamir, 2011) introduced an interpolation between these two extremes and generalized the feedback of the classic bandit problem to a graph structure. Several prior studies, such as (Alon et al., 2015; Zimmert & Lattimore, 2019; Chen et al., 2021; He & Zhang, 2023), have proposed various graph parameters to characterize the factors that influence regret. However, the algorithms proposed in these works for more general graphs do not yield a tight bound in our specific setting.

The pure exploration version of the bandit problem, known as the *best arm identification* ($\mathtt{BAI}$) problem, has also received significant attention in the literature (Even-Dar et al., 2002; Mannor & Tsitsiklis, 2004; Bubeck et al., 2009; Audibert et al., 2010; Karnin et al., 2013; Chen et al., 2017). While the $\mathtt{BAI}$ problem may appear deceptively simple, determining the precise bound for $\mathtt{BAI}$ under the bandit feedback setting remains an open question. However, for the problem of identifying an ε-optimal arm with high probability, (Even-Dar et al., 2002) established a tight bound for the bandit feedback setting, while the bound for the full feedback model is relatively straightforward (see e.g. Chen et al. (2021)).

1.3.1. Comparison with Previous Work

The very recent work of (Eldowa et al., 2024) studied interpolation of learning with experts and multi-armed bandit as well from a different perspective. They proved an $O\left(\sqrt{T\alpha(1+\log\left(N/\alpha\right))}\right)$ upper bound for the minimax regret of bandit with strongly feedback graph G where α is the *independence number* of G. The parameter is in general *not* comparable with clique covers used in this work for feedback graphs. Particularly on an $\mathbf{m}$-$\mathtt{MAB}$ instance where $\mathbf{m} = (m_1, \dots, m_K)$, the independence number is K and therefore their upper bound becomes to $O\left(\sqrt{TK\log(N/K)}\right)$ while our results showed that the minimax regret is indeed $\Theta\left(\sqrt{T\sum_{k=1}^{K}\log(m_k+1)}\right)$.

Another work (Foster et al., 2020) using the idea of the two-stage algorithm bears similarity to ours. But applying their algorithm and analysis to our problem can only derive an upper bound of $O\left(\sqrt{TK\max_{k\in[K]}\log m_k}\right)$, which is also suboptimal. To see the difference, assume $K=\lfloor\log N\rfloor$ and $\mathbf{m}=(1,1,\ldots,1,N-K+1)$, then the minimax regret is $\Theta\left(\sqrt{T\log N}\right)$ while the upper bounds in both (Eldowa et al., 2024) and (Foster et al., 2020) are $O\left(\sqrt{T}\log N\right)$. We believe that the algorithms and the analysis of previous work cannot achieve the same bound in our setting without significant extra effort.

2. Preliminaries

In this section, we formally define the notations used and introduce some preparatory knowledge that will help in understanding this work.

2.1. Mathematical Notations

Let n be a non-negative integer. We use $[n]$ to denote the set $\{1,2,\ldots,n\}$ and $\Delta_{n-1}=\left\{\mathbf{x}\in\mathbb{R}^n_{\geq 0}:\sum_{i=1}^n\mathbf{x}(i)=1\right\}$ to denote the $n-1$ dimensional standard simplex where $\mathbb{R}_{\geq 0}$ is the set of all non-negative real numbers. For a real vector $\mathbf{x}\in\mathbb{R}^n$, the i-th entry of $\mathbf{x}$ is denoted as $\mathbf{x}(i)$ for every $i\in[n]$. We define $\mathbf{e}_i^{[n]}$ as the indicator vector of the i-th coordinate such that $\mathbf{e}_i^{[n]}(i)=1$ and $\mathbf{e}_i^{[n]}(j)=0$ for all $j\neq i$ and $j\in[n]$. We may write $\mathbf{e}_i^{[n]}$ as $\mathbf{e}_i$ if the information on n is clear from the context.

Given two vectors $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, we define their inner product as $\langle\mathbf{x},\mathbf{y}\rangle=\sum_{i=1}^n\mathbf{x}(i)\mathbf{y}(i)$. For any $a,b\in\mathbb{R}$, let $[a,b]=\{c\in\mathbb{R}\mid\min\{a,b\}\leq c\leq\max\{a,b\}\}$ be the interval between a and b. For any $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, we say $\mathbf{y}\geq\mathbf{x}$ if $\mathbf{y}(i)\geq\mathbf{x}(i)$ for every $i\in[n]$. Then we can define the rectangle formed by $\mathbf{x}$ and $\mathbf{y}$: $\mathtt{Rect}(\mathbf{x},\mathbf{y})=\{\mathbf{z}\in\mathbb{R}^n:\mathbf{y}\geq\mathbf{z}\geq\mathbf{x}\}$.

For any positive semi-definite matrix $M\in\mathbb{R}^{n\times n}$, let $\|\mathbf{x}\|_M=\sqrt{\mathbf{x}^{\mathrm{T}}M\mathbf{x}}$ be the norm of $\mathbf{x}$ with respect to M. Specifically, we abbreviate $\|\mathbf{x}\|_{(\nabla^2\psi)^{-1}}$ as $\|\mathbf{x}\|_{\nabla^{-2}\psi}$ where $\nabla^2\psi$ is the Hessian matrix of a convex function ψ.

Let $F:\mathbb{R}^n\to\mathbb{R}$ be a convex function which is differentiable in its domain $\mathtt{dom}(F)$. Given $\mathbf{x},\mathbf{y}\in\mathtt{dom}(F)$, the Bregman divergence with respect to F is defined as $B_F(\mathbf{x},\mathbf{y})=F(\mathbf{x})-F(\mathbf{y})-\langle\mathbf{x}-\mathbf{y},\nabla F(\mathbf{y})\rangle$. Given two measures $\mathbf{P}_1$ and $\mathbf{P}_2$ on the same measurable space $(\Omega,\mathcal{F})$, the KL-divergence between $\mathbf{P}_1$ and $\mathbf{P}_2$ is defined as $D_{\mathrm{KL}}(\mathbf{P}_1,\mathbf{P}_2)=\sum_{\omega\in\Omega}\mathbf{P}_1[\omega]\log\frac{\mathbf{P}_1[\omega]}{\mathbf{P}_2[\omega]}$ if Ω is discrete or $D_{\mathrm{KL}}(\mathbf{P}_1,\mathbf{P}_2)=\int_\Omega\log\frac{\mathbf{P}_1[\omega]}{\mathbf{P}_2[\omega]}\,\mathrm{d}\mathbf{P}_1[\omega]$ if Ω is continuous provided $\mathbf{P}_1$ is absolutely continuous with respect to $\mathbf{P}_2$.

2.2. Graph Theory

Let $G=(V,E)$ be a directed graph where $|V|=N$. We use (u,v) to denote the directed edge from vertex u to vertex v. For any $U\subseteq V$, we denote the subgraph induced by U as $G[U]$. For $v\in V$, let $N_{\mathtt{in}}(v):=\{u\in V:(u,v)\in E\}$ be the set of in-neighbors of v and $N_{\mathtt{out}}(v):=\{u\in V:(v,u)\in E\}$ be the set of out-neighbors. If the graph is undirected, we have $N_{\mathtt{in}}(v)=N_{\mathtt{out}}(v)$, and we use $\mathrm{N}(v)$ to denote the neighbors for brevity. We say $S\subseteq V$ is an independent set of G if for every $v\in S$, $\{u\in S\mid u\neq v,u\in N_{\mathtt{in}}(v)\cup N_{\mathtt{out}}(v)\}=\varnothing$. The maximum independence number of G is denoted as $\alpha(G)$ and abbreviated as α when G is clear from the context. Furthermore, we say an independent set S is a t-packing independent set if and only if for any $v\in V$, there are at most t out-neighbors of v in S, i.e., $|N_{\mathtt{out}}(v)\cap S|\leq t$. We say the subsets $V_1,\ldots,V_K\subseteq V$ form a $(V_1,\ldots,V_K)$-*clique cover* of G if each induced subgraph $G[V_k]$ for $k\in[K]$ is a clique and $\bigcup_{k\in[K]}V_k=V$.

2.3. $\mathbf{m}$-MAB and $\mathbf{m}$-BAI

Let $K>0$ be an integer. Given a vector $\mathbf{m}=(m_1,m_2,\ldots,m_K)\in\mathbb{Z}^K_{\geq 1}$ with $\sum_{k\in[K]}m_k=N$, we now define problems $\mathbf{m}$-`MAB` and $\mathbf{m}$-`BAI` respectively.

2.3.1. $\mathbf{m}$-MAB

In the problem of $\mathbf{m}$-`MAB`, there are N arms. The arms are partitioned into K groups and the k-th group contains m_k arms. Let $T\in\mathbb{N}$ be the time horizon. Then $\mathbf{m}$-`MAB` is the following online decision game. The game proceeds in T rounds. At round $t=0,1,\ldots,T-1$:

- The player pulls an arm $A_t\in[N]$;
- The adversary chooses a loss function $\ell^{(t)}\in[0,1]^N$;
- The player incurs loss $\ell^{(t)}(A_t)$ and observes the losses of all arms in the group containing A_t.

Clearly the vector $\mathbf{m}$ encodes the amount of information the player can observe in each round. Two extremes are the problem of learning with expert advice and multi-armed bandit, which correspond to (N)-`MAB` and $(1,\ldots,1)$-`MAB` respectively.

We assume the player knows $\mathbf{m}$ and T in advance and use $\mathcal{A}$ to denote the player's algorithm (which can be viewed as a function from previous observed information and the value of its own random seeds to the arm pulled at each round).

The performance of the algorithm $\mathcal{A}$ is measured by the notion of *regret*. Fix a loss sequence $\vec{L}=\left\{\ell^{(0)},\ldots,\ell^{(T-1)}\right\}$. Let $a^*=\arg\min_{a\in[N]}\sum_{t=1}^T\ell^{(t)}(a)$ be the arm with minimum accumulated losses. The regret of the algorithm $\mathcal{A}$ and time horizon T on $\vec{L}$ with respect to the arm a is defined as $R_a(T,\mathcal{A},\vec{L}) = \mathbf{E}\left[\sum_{t=0}^{T-1}\ell^{(t)}(A_t)\right] - \sum_{t=0}^{T-1}\ell^{(t)}(a)$. If there is no ambiguity, we abbreviate $R_a(T,\mathcal{A},\vec{L})$ as $R_a(T)$. We also use $R(T)$ to denote $R_{a^*}(T)$.

We are interested in the regret of the best algorithm against the worst adversary, namely the quantity

$$R_a^*(T) = \inf_{\mathcal{A}}\sup_{\vec{L}} R_a(T,\mathcal{A},\vec{L}).$$

We call $R_{a^*}^*(T)$ the *minimax regret* of $\mathbf{m}$-MAB and usually write it as $R^*(T)$.

We may use the following two ways to name an arm in $\mathbf{m}$-MAB:

- use the pair (k,j) where $k\in[K]$ and $j\in[m_k]$ to denote "the j-th arm in the k-th group";
- use a global index $i\in[N]$ to denote the i-th arm.

Following this convention, we use $\ell^{(t)}(i)$ and $\ell_k^{(t)}(j)$ to denote the loss of arm i and arm (k,j) at round t respectively.

2.3.2. Best Arm Identification and $\mathbf{m}$-BAI

The *best arm identification* (BAI) problem asks the player to identify the best arm among N given arms with as few pulls as possible. To be specific, each arm i is associated with a parameter p_i and each pull of arm i gives an observation of its random loss, which is drawn from a fixed distribution with mean p_i independently. The loss of each arm is restricted to be in $[0,1]$. The one with smallest p_i, indexed by i^*, is regarded as the best arm. An arm j is called an *ε-optimal arm* if its mean is less than the mean of the best arm plus ε for some $\varepsilon\in(0,1)$, namely $p_j < p_{i^*}+\varepsilon$. With fixed $\varepsilon,\delta>0$, an (ε,δ)-*probably approximately correct* algorithm, or (ε,δ)-PAC algorithm for short, can find an ε-optimal arm with probability at least $1-\delta$. In most parts of this paper, we choose $\delta = 0.05$. For an algorithm $\mathcal{A}$ of BAI, we usually use T to denote the number of arms $\mathcal{A}$ pulled before termination. Similarly for any arm i, we use T_i to denote the number of times that the arm i has been pulled by $\mathcal{A}$ before its termination. We also use N_i to denote the number of times that the arm i has been *observed* by $\mathcal{A}$.

Let $\mathbf{m} = (m_1,m_2,\cdots,m_K)\in\mathbb{Z}_{\geq 1}^K$ be a vector. Similar to $\mathbf{m}$-MAB, the arms are partitioned into K groups and the k-th group consists of m_k arms. Each pull of an arm can observe the losses of all arms in the group. As usual, the goal is to identify the best arm (the one with minimum p_i) with as few rounds as possible.

Similar to $\mathbf{m}$-MAB, we use $i\in[N]$ or (k,j) where $k\in[K]$ and $j\in[m_k]$ to name an arm. For a fixed algorithm, we use T_i or $T_{(k,j)}$ to denote the number of times the respective arm has been pulled and use N_i or $N_{(k,j)}$ to denote the number of times it has been observed. For every $k\in[K]$ we use $T^{(k)}$ to denote the number of times the arms in the k-th group have been pulled, namely $T^{(k)} = \sum_{j\in[m_k]} T_{(k,j)}$. By definition, it holds that $T = \sum_{k\in[K]} T^{(k)}$ and $N_{(k,j)} = T^{(k)}$ for every $j\in[m_k]$.

2.4. Bandit with Graph Feedback

A more general way to encode the observability of arms is to use feedback graphs. In this problem, a directed graph $G=(V,E)$ is given. The vertex set $V=[N]$ is the collection of all arms.

The game proceeds in the way similar to $\mathbf{m}$-MAB. The only difference is that when an arm A_t is pulled by the player at a certain round, all arms in $N_{\mathtt{out}}(A_t)$ can be observed. As a result, given a vector $\mathbf{m}=(m_1,m_2,\cdots,m_K)\in\mathbb{Z}_{\geq 1}^K$, the $\mathbf{m}$-MAB problem is identical to bandit with graph feedback $G=(V,E)$ where G is the disjoint union of K cliques $G_1=(V_1,E_1), G_2=(V_2,E_2),\dots,G_K=(V_K,E_K)$ with $m_k=|V_k|$ and $E_k=V_k^2$ for every $k\in[K]$.

According to (Alon et al., 2015), we measure the observability of each vertex in terms of its in-neighbors. If a vertex has no in-neighbor, we call it a *non-observable* vertex, otherwise it is *observable*. If a vertex v has a self-loop *or* $N_{\mathtt{in}}(v)$ exactly equals to $V\setminus\{v\}$, then v is *strongly observable*. If an observable vertex is not strongly observable, then it is *weakly observable*. In this work, we assume each vertex is observable. If all the vertices are strongly observable, the graph G is called a strongly observable graph. If G contains weakly observable vertices (and does not have non-observable ones), we say G is a weakly observable graph.

We can also define the notion of regret for bandit with graph feedback. Assume notations before, the regret of an algorithm $\mathcal{A}$ with feedback graph G and time horizon T on a loss sequence $\vec{L}$ with respect to the arm a is defined as $R_a(G,T,\mathcal{A},\vec{L}) = \mathbf{E}\left[\sum_{t=0}^{T-1}\ell^{(t)}(A_t)\right] - \sum_{t=0}^{T-1}\ell^{(t)}(a)$. If there is no ambiguity, we abbreviate $R_a(G,T,\mathcal{A},\vec{L})$ as $R_a(G,T)$ or $R_a(T)$. We also use $R(T)$ to denote $R_{a^*}(T)$. Then minimax regret is again

$$R_{a^*}^*(G,T) = \inf_{\mathcal{A}}\sup_{\vec{L}} R_{a^*}(G,T,\mathcal{A},\vec{L}).$$

When G is clear from the context, we write it as $R^*(T)$.

3. The Upper Bounds

In this section, we prove Theorem 1.1 and Theorem 1.2. We describe the algorithm for $\mathbf{m}$-MAB in Section 3.1 and analyze it in Section 3.2. The algorithm for $\mathbf{m}$-BAI is obtained by a reduction to $\mathbf{m}$-$\texttt{MAB}$ described in Appendix A. Finally we discuss how to extend the algorithm to bandit with strongly observable feedback graphs and prove Corollary 1.5 in Appendix B.3.

3.1. The Algorithm

As discussed in the introduction, our algorithm basically follows the framework of the two-stage online stochastic mirror descent developed in (He & Zhang, 2023). However, our updating rules is slightly different from the one in (He & Zhang, 2023) in order to incorporate with our new analysis.

Given a K-dimensional vector $\mathbf{m} = (m_1, \dots, m_K)$ as input, in each round t, the algorithm proceeds in the following two-stage manner:

- A distribution $Y^{(t)}$ over $[K]$ is maintained, indicating which group of arms the algorithm is going to pick.
- For each $k \in [K]$, a distribution $X_k^{(t)}$ is maintained, indicating which arm in the k-th group the algorithm will pick conditioned on that the k-th group is picked in the first stage.
- The algorithm then picks the j-th arm in the k-group with probability $Y^{(t)}(k) \cdot X_k^{(t)}(j)$.

The algorithm is described in Algorithm 1 and we give an explanation for each step below. Assuming $Y^{(0)}$ and $X_k^{(0)}$ for all $k \in [K]$ are well initialized, in each time step $t = 0, 1, \dots, T-1$, the player will repeat the following operations:

Sampling: For each arm (k, j), the algorithm pulls it with probability

$$Z^{(t)}(k, j) = Y^{(t)}(k) \cdot X_k^{(t)}(j).$$

The arm pulled at this round is denoted by $A_t = (k_t, j_t)$. Our algorithm can guarantee that $Z^{(t)}$ is a distribution over all arms.

Observing: Observe partial losses $\ell_{k_t}^{(t)}(j)$ for all $j \in [m_{k_t}]$.

Estimating: For each arm (k, j), define the unbiased estimator $\hat{\ell}_k^{(t)}(j) = \frac{\mathbb{1}[k=k_t]}{\mathbf{Pr}[k=k_t]} \cdot \ell_k^{(t)}(j)$. It is clear that $\mathbf{E}\left[\hat{\ell}_k^{(t)}(j)\right] = \ell_k^{(t)}(j)$.

Updating:

- For each $k \in [K]$, update $X_k^{(t)}$ in the manner of standard OSMD:
$$\nabla\phi_k(\overline{X}_k^{(t+1)}) = \nabla\phi_k(X_k^{(t)}) - \hat{\ell}_k^{(t)}; \quad X_k^{(t+1)} = \operatorname*{arg\,min}_{\mathbf{x}\in\Delta_{m_k-1}} B_{\phi_k}(\mathbf{x}, \overline{X}_k^{(t+1)}),$$
where $\phi_k(\mathbf{x}) = \eta_k^{-1}\sum_i x(i)\log x(i)$ is the negative entropy scaled by the learning rate η_k.
- Define $\overline{Y}^{(t)}$ in the way that, for any $k \in [K]$
$$\frac{1}{\sqrt{\overline{Y}^{(t+1)}(k)}} = \frac{1}{\sqrt{Y^{(t)}(k)}} + \sum_{j\in[m_k]} \frac{\eta}{\eta_k} X_k^{(t)}(j)\left(1 - \exp\left(-\eta_k \cdot \hat{\ell}_k^{(t)}(j)\right)\right). \tag{1}$$
where η is the learning rate. Then let $Y^{(t+1)}$ be the projection of $\overline{Y}^{(t+1)}$ on Δ_{K-1}:
$$Y^{(t+1)} = \operatorname*{arg\,min}_{\mathbf{y}\in\Delta_{K-1}} B_\psi(\mathbf{y}, \overline{Y}^{(t+1)}),$$
where $\psi(\mathbf{y}) = -2\sum_i \sqrt{y(i)}$ for any $\mathbf{y} = (y(1), \dots, y(K)) \in \mathbb{R}^K$, referred to as Tsallis entropy in literature. Note that when x is small, $1 - \exp(-x) \approx x$. So when η_k is small (and it is so), the updating rule is approximately for any $k \in [K]$
$$\frac{1}{\sqrt{\overline{Y}^{(t+1)}(k)}} = \frac{1}{\sqrt{Y^{(t)}(k)}} + \eta \sum_{j\in[m_k]} X_k^{(t)}(j) \cdot \hat{\ell}_k^{(t)}(j),$$
which is equivalent to
$$\boldsymbol{\nabla}\psi(\overline{Y}^{(t+1)}) = \boldsymbol{\nabla}\psi(\overline{Y}^{(t)}) - \eta \cdot \widehat{L}^{(t)},$$
where $\widehat{L}^{(t)} = (\widehat{L}^{(t)}(1), \dots, \widehat{L}^{(t)}(K)) \in \mathbb{R}^K$ satisfying $\widehat{L}^{(t)}(k) = \sum_{j\in[m_k]} X_k^{(t)}(j) \cdot \hat{\ell}_k^{(t)}(j)$. One can think of $\widehat{L}^{(t)}(k)$ as the "average loss" of the arms in the k-th group at round t. Nevertheless, we use rule Equation (1) in the algorithm to guarantee the result in Lemma B.1 since it is convenient for our analysis later.

In the realization of Algorithm 1, we will choose $\eta = \frac{1}{\sqrt{T}}$ and $\eta_k = \frac{\log(m_k+1)}{\sqrt{T\sum_{k=1}^K \log(m_k+1)}}$.

3.2. Regret Bound for MAB

We prove the following theorem, which implies Theorem 1.1.

Theorem 3.1. *For every $T > 0$ and every loss sequence $\ell^{(0)}, \dots, \ell^{(T-1)} \in [0,1]^N$, the regret of Algorithm 1 satisfies*

$$R(T) \le O\left(\sqrt{T\sum_{k=1}^K \log(m_k+1)}\right).$$

Algorithm 1 Two-Stage Algorithm for $\mathbf{m}$-$\texttt{MAB}$

```
**Input:** An $(m_1, \dots, m_K)$-$\texttt{MAB}$ instance
$X_k^{(0)} \leftarrow \operatorname*{arg\,min}_{a \in \Delta_{m_k - 1}} \phi_k(a)$, for all $k \in [K]$;
$Y^{(0)} \leftarrow \operatorname*{arg\,min}_{b \in \Delta_{K-1}} \psi(b)$;
**for** $t \leftarrow 0$ **to** $T - 1$ **do**
    $Z^{(t)}(k, j) \leftarrow Y^{(t)}(k) \cdot X_k^{(t)}(j)$, for all $k \in [K]$ and $j \in [m_k]$;
    Pull $A_t = (k_t, j_t) \sim Z^{(t)}$ and observe $\ell_{k_t}^{(t)}(j)$ for all $j \in [m_k]$;
    Update $\nabla \phi_k(\overline{X}_k^{(t+1)}) = \nabla \phi_k(X_k^{(t)}) - \hat{\ell}_k^{(t)}$;
        $X_k^{(t+1)} = \arg\min_{\mathbf{x} \in \Delta_{m_k - 1}} B_{\phi_k}(\mathbf{x}, \overline{X}_k^{(t+1)})$;
    Update $\forall k \in [K]$, $\frac{1}{\sqrt{\overline{Y}^{(t+1)}(k)}} = \frac{1}{\sqrt{Y^{(t)}(k)}} +$
        $\frac{\eta}{\eta_k} \sum_{j \in [m_k]} X_k^{(t)}(j) \left(1 - \exp\left(-\eta_k \cdot \hat{\ell}_k^{(t)}(j)\right)\right))$;
        $Y^{(t+1)} = \arg\min_{\mathbf{y} \in \Delta_{K-1}} B_\psi(\mathbf{y}, \overline{Y}^{(t+1)})$;
**end for**
```

Instead of directly bounding the regret of the sequence of the action distributions $\left\{Z^{(t)}\right\}_{0 \le t \le T-1}$, we study an auxiliary *piecewise continuous* process $\left\{\mathcal{Z}^{(s)}\right\}_{s \in [0,T)}$. We define and bound the *regret* of $\left\{\mathcal{Z}^{(s)}\right\}_{s \in [0,T)}$ in Appendix B.1.1, and compare it with the regret of $\left\{Z^{(t)}\right\}_{0 \le t \le T-1}$ in Appendix B.1.2. Finally, we prove Theorem 3.1 in Appendix B.1.3

4. Lower Bound

In this section, the main work is to prove a lower bound for $\mathbf{m}$-$\texttt{BAI}$. A natural way is to prove a lower bound for each group and then sum up all in a "direct sum " flavor. A conventional method is to design a family of hard instances and claim that there exists some instance requiring a sufficient number of pulls. However, different groups may have different hard instances, preventing a direct summation of the lower bound for each group. Instead, we prove that the instance with all m arms following a $\texttt{Ber}(1/2)$ distribution, denoted by $\mathscr{H}_0^{(m)}$, is always the most challenging one.

Lemma 4.1. *Let $\mathcal{A}$ be an $(\varepsilon, 0.05)$-PAC algorithm. Assume $m \ge 2$. There exists a universal constant $c_1 > 0$ such that $\mathcal{A}$ terminates on $\mathscr{H}_0^{(m)}$ after at least $\frac{c_1}{\varepsilon^2} \log(m+1)$ rounds in expectation.*

Armed with above lemma, we can establish the lower bound for $\mathbf{m}$-$\texttt{BAI}$.

Lemma 4.2. *Let ε be a number in $\left(0, \frac{1}{8}\right)$. For every $(\varepsilon, 0.05)$-PAC algorithm of $\mathbf{m}$-$\texttt{BAI}$, we have $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}\left[T^{(k)}\right] \ge \frac{c_1 \log(m_k+1)}{\varepsilon^2}$ for every $k \in [K]$ with $m_k \ge 2$ and $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}[T] \ge \sum_{k=1}^{K} \frac{c_1 \log(m_k+1)}{2\varepsilon^2}$ if the total number of arms $\sum_{k=1}^{K} m_k \ge 2$, where c_1 is the constant in Lemma 4.1, $T^{(k)}$ is the number of rounds that arms in group k is played and T is the total pull times.*

Moreover, these lower bounds still hold even the algorithm can identify the ε-optimal arm with probability 0.95 only when the input arms have losses drawn from either $\texttt{Ber}\left(\frac{1}{2}\right)$ or $\texttt{Ber}\left(\frac{1}{2} - \varepsilon\right)$.

Now let us fix $\mathbf{m} = (m_1, \dots, m_K)$. We then derive a regret lower bound for $\mathbf{m}$-MAB and thus prove Theorem 1.4 using Lemmas 4.1 and A.1.

Lemma 4.3. *For any algorithm $\mathcal{A}$ of $(m_1, \dots, m_k)$-MAB, for any sufficiently large $T > 0$, there exists $\mathscr{H} \in \mathcal{H}$ such that the expected regret of $\mathcal{A}$ satisfies*

$$\mathbf{E}_{\mathscr{H}}\left[R(T)\right] \ge c' \cdot \sqrt{T \cdot \sum_{k=1}^{K} \log(m_k + 1)}$$

where $c' > 0$ is a universal constant. Here the expectation is taken over the randomness of losses which are drawn from $\mathscr{H}$ independently in each round.

We can,of course, apply our more powerful Lemma 4.1 to a broader class of graphs, thus obtaining improved lower bounds for both strongly and weakly observable graphs.

5. Conclusion

In this study, we delve into the $\mathbf{m}$-$\texttt{MAB}$ and $\mathbf{m}$-$\texttt{BAI}$, and reduce the the latter to the former. We propose a two stage algorithm for $\mathbf{m}$-$\texttt{MAB}$ and prove a lower bound for $\mathbf{m}$-$\texttt{BAI}$, thereby providing both problems with tight bounds. Furthermore, we utilize the bound proven for $\texttt{BAI}$ to more general graphs and yield some improved lower bounds. The technique developed in the upper bound is more intuitive than standard methods. The proof of the lower bound for $\texttt{BAI}$ reveals that to address the failure error of m arms, calculations must consider them together, such as using mixture distribution, instead of assuming m failures to distinguish one distribution from all others one by one to construct a contradiction. We believe our approach may inspire others to enhance the logarithmic factor in other problems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

This work was supported by CCF-Huawei Populus Grove Fund CCF-HuaweiLK2022007, and the NSFC No.

62372289.

References

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pp. 23–35. PMLR, 2015.

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pp. 1–122, 2009.

Audibert, J.-Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *COLT*, pp. 41–53, 2010.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pp. 23–37. Springer, 2009.

Chen, H., Huang, Z., Li, S., and Zhang, C. Understanding bandits with graph feedback. *Advances in Neural Information Processing Systems*, 34:24659–24669, 2021.

Chen, L., Li, J., and Qiao, M. Towards instance optimal bounds for best arm identification. In *Conference on Learning Theory*, pp. 535–592. PMLR, 2017.

Dann, C., Wei, C.-Y., and Zimmert, J. A blackbox approach to best of both worlds in bandits and beyond. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5503–5570. PMLR, 2023.

Eldowa, K., Esposito, E., Cesari, T., and Cesa-Bianchi, N. On the minimax regret for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 36, 2024.

Erez, L. and Koren, T. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 34:28511–28521, 2021.

Even-Dar, E., Mannor, S., and Mansour, Y. Pac bounds for multi-armed bandit and markov decision processes. In *COLT*, volume 2, pp. 255–270. Springer, 2002.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Haussler, D., Kivinen, J., and Warmuth, M. K. Tight worst-case loss bounds for predicting with expert advice. In *European Conference on Computational Learning Theory*, pp. 69–83. Springer, 1995.

He, Y. and Zhang, C. Improved algorithms for bandit with graph feedback via regret decomposition. *Theoretical Computer Science*, 979:114200, 2023.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246. PMLR, 2013.

Kocák, T. and Carpentier, A. Online learning with feedback graphs: The true shape of regret. In *International Conference on Machine Learning*, pp. 17260–17282. PMLR, 2023.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24, 2011.

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and Seldin, Y. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 35:35035–35048, 2022.

Zimmert, J. and Lattimore, T. Connections between mirror descent, thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems*, pp. 11973–11982, 2019.

A. A Reduction from BAI to MAB

In this section, we construct a PAC algorithm for $\mathbf{m}$-BAI by leveraging an algorithm designed for $\mathbf{m}$-MAB, employing Lemma A.1. We then reduce BAI to MAB. Consequently, the upper bound of an algorithm for MAB is also applicable to the constructed algorithm for BAI, and a lower bound for BAI implies one for MAB.

Let $r(T, \vec{L})$ be a real valued function with the time horizon T and loss sequence $\vec{L} = \left(\ell^{(1)}, \dots, \ell^{(T)}\right)$ as its input. Let $\mathscr{H}$ be a BAI instance. With fixed $T > 0$, we use $\mathbf{E}_{\mathscr{H}}\left[r(T, \vec{L})\right]$ to denote the expectation of $r(T, \vec{L})$ where $\ell^{(t)}$ in $\vec{L}$ is drawn from $\mathscr{H}$ independently for every $t \in [T]$. Let $\mathcal{H}$ be a set of BAI instances.

Lemma A.1. *Let $\mathcal{A}$ be an algorithm for $\mathbf{m}$-MAB with regret $R_{a^*}(T, \mathcal{A}, \vec{L}) \le r(T, \vec{L})$ for every time horizon T and every loss sequence $\vec{L}$. Then there exists an $(\varepsilon, 0.05)$-PAC algorithm $\mathcal{A}'$ for $\mathbf{m}$-BAI that terminates in T^* rounds where T^* is the solution of the equation*

$$T^* = \frac{2500 \cdot \max_{\vec{L}} r(T^*, \vec{L})}{\varepsilon}.$$

Moreover, if we only care about identifying an ε-optimal arm with probability 0.95 when the input is chosen from a known family $\mathcal{H}$, we can construct an algorithm solving this problem that terminates in $T^*_{\mathcal{H}}$ rounds where $T^*_{\mathcal{H}}$ is the solution of the equation

$$T^*_{\mathcal{H}} = \frac{2500 \cdot \max_{\mathscr{H} \in \mathcal{H}} \mathbf{E}_{\mathscr{H}}\left[r(T^*_{\mathcal{H}}, \vec{L})\right]}{\varepsilon}.$$

Proof. Given an instance $\mathscr{H}$ of $\mathbf{m}$-BAI, we run $\mathcal{A}$ for T^* rounds. Let T_i be the number of times that the arm i has been pulled, i.e., $T_i = \sum_{t=0}^{T^*-1} \nVdash[A_t = i]$. Let $\overline{Z} = \left(\overline{Z}_1, \overline{Z}_2, \dots, \overline{Z}_N\right) = \left(\frac{T_1}{T^*}, \frac{T_2}{T^*}, \dots, \frac{T_N}{T^*}\right)$ be a distribution on N arms. We construct $\mathcal{A}'$ by simply sampling from $\overline{Z} = \left(\frac{T_1}{T^*}, \frac{T_2}{T^*}, \dots, \frac{T_N}{T^*}\right)$ and outputting the result.

Recall that p_i is the mean of the i-th arm in $\mathscr{H}$ and arm a^* is the one with the minimum mean. Define the gap vector $\Delta = (p_1 - p_{a^*}, \cdots, p_N - p_{a^*})$. Note that $\overline{Z}$ is a random vector and define conditional expected regret $R(\overline{Z}) = \langle \Delta, \overline{Z} \rangle \cdot T^*$ given $\overline{Z}$. Thus the expected regret $\mathbf{E}_{\overline{Z}}\left[R(\overline{Z})\right] \le \max_{\vec{L}} r(T^*, \vec{L})$. By Markov's inequality, $R(\overline{Z}) \le 100 \max_{\vec{L}} r(T^*, \vec{L})$ with probability at least 0.99. Now we only consider $\overline{Z}$ conditioned on $R(\overline{Z}) \le 100 \max_{\vec{L}} r(T^*, \vec{L})$. Let $B \subseteq [N]$ denote the "bad set" which contains arms that are not ε-optimal. Then $\varepsilon T^* \sum_{i \in B} \overline{Z}_i \le 100 \max_{\vec{L}} r(T^*, \vec{L})$. Note that $T^* = \frac{2500 \cdot \max_{\vec{L}} r(T^*, \vec{L})}{\varepsilon}$. Therefore $\sum_{i \in B} \overline{Z}_i \le 0.04$. In total, this algorithm will make a mistake with probability no more than 0.05 by the union bound.

When we only care about the input instances chosen from $\mathcal{H}$, we run $\mathcal{A}$ for $T^*_{\mathcal{H}}$ rounds and similarly, we output an arm drawn from $\left(\frac{T_1}{T^*_{\mathcal{H}}}, \frac{T_2}{T^*_{\mathcal{H}}}, \dots, \frac{T_N}{T^*_{\mathcal{H}}}\right)$. It is easy to verify via the same arguments that this algorithm can output an ε-optimal arm with probability 0.95 when the input is chosen from $\mathcal{H}$. □

B. Upper bound

In this section, we first prove the regret bound for MAB, and then reduce BAI to MAB to give a bound for BAI with the help of Lemma A.1. Finally we can easily apply the algorithm to more general graph.

B.1. Regret Upper Bound for MAB

B.1.1. The piecewise continuous process

Assuming notations in Algorithm 1, the process $\left\{\mathcal{Z}^{(s)}\right\}_{s \in [0,T)}$ is defined as

$$\mathcal{Z}^{(s)}(k, j) = \mathcal{Y}^{(s)}(k) \cdot \mathcal{X}_k^{(s)}(j), \quad \forall k \in [K], j \in [m_k],$$

where $\left\{\mathcal{Y}^{(s)}\right\}_{s \in [0,T)}$ and $\left\{\mathcal{X}_k^{(s)}\right\}_{s \in [0,T)}$ for every $k \in [K]$ are piecewise continuous processes defined in the following way.

- For every integer $t \in \{0, 1, \dots, T-1\}$, we let $\mathcal{Y}^{(t)} = Y^{(t)}$ and $\mathcal{X}_k^{(t)} = X_k^{(t)}$ for every $k \in [K]$.

- For every integer $t \in \{0, 1, \dots, T-1\}$ and every $k \in [K]$, the trajectory of $\left\{\mathcal{X}_k^{(s)}\right\}_{s\in[t,t+1)}$ is a continuous path in $\mathbb{R}^{m_k}$ governed by the ordinary differential equation

$$\frac{\mathrm{d}\boldsymbol{\nabla}\phi_k(\mathcal{X}_k^{(s)})}{\mathrm{d}s} = -\hat{\ell}_k^{(t)}. \tag{2}$$

- For every integer $t \in \{0, 1, \dots, T-1\}$, the trajectory of $\left\{\mathcal{Y}^{(s)}\right\}_{s\in[t,t+1)}$ is a continuous path in $\mathbb{R}^K$ governed by the ordinary differential equation

$$\frac{\mathrm{d}\boldsymbol{\nabla}\psi(\mathcal{Y}^{(s)})}{\mathrm{d}s} = -\widehat{L}^{(s)}, \tag{3}$$

  where $\widehat{L}^{(s)} = \left(\widehat{L}^{(s)}(1), \dots, \widehat{L}^{(s)}(K)\right) \in \mathbb{R}^K$ satisfies $\widehat{L}^{(s)}(k) = \sum_{j\in[m_k]} \mathcal{X}_k^{(s)}(j) \cdot \hat{\ell}_k^{(t)}(j)$.

Clearly the trajectories of $\mathcal{Z}^{(s)}$, $\mathcal{Y}^{(s)}$ and $\mathcal{X}_k^{(s)}$ for every $k \in [K]$ are piecewise continuous paths in the time interval $s \in [0, T)$. An important property is that the end of each piece of the trajectories of $\mathcal{Y}^{(s)}$ and $\mathcal{X}_k^{(s)}$ coincides with its discrete counterpart *before* performing projection to the probability simplex.

Formally, for every $t \in [T]$ and $k \in [K]$, define $\mathcal{X}_k^{(t)^-} := \lim_{s\to t^-} \mathcal{X}_k^{(s)}$ and $\mathcal{Y}^{(t)^-} := \lim_{s\to t^-} \mathcal{Y}^{(s)}$. We have the following lemma.

Lemma B.1. *For every $t \in [T]$ and $k \in [K]$, it holds that $\mathcal{X}_k^{(t)^-} = \overline{X}_k^{(t)}$ and $\mathcal{Y}^{(t)^-} = \overline{Y}^{(t)}$.*

Proof. To ease the notation, for any fixed $t \in \{0, 1, \dots, T-1\}$ and fixed $k \in [K]$, we now prove that $\mathcal{X}_k^{(t+1)^-} = \overline{X}_k^{(t+1)}$ and $\mathcal{Y}^{(t+1)^-} = \overline{Y}^{(t+1)}$ respectively.

In fact, $\mathcal{X}_k^{(t+1)^-} = \overline{X}_k^{(t+1)}$ immediately follows by integrating both sides of (2) from t to $t+1$ and noting that $\mathcal{X}_k^{(t)} = X_k^{(t)}$. More efforts are needed to prove the identity for $\mathcal{Y}^{(t)}$. Recall $\phi_k(\mathbf{x}) = \eta_k^{-1} \sum_j x(j) \log x(j)$ for every $\mathbf{x} = \big(x(1), \dots, x(m_k)\big)$. It follows from (2) that for every $s \in [t, t+1)$ every $k \in [K]$ and every $j \in [m_k]$,

$$\mathcal{X}_k^{(s)}(j) = \mathcal{X}_k^{(t)}(j) \cdot \exp\left(-(s-t)\eta_k \hat{\ell}_k^{(t)}(j)\right).$$

As a result, we know that

$$\widehat{L}^{(s)}(k) = \sum_{j\in[m_k]} \mathcal{X}_k^{(t)}(j) \cdot \exp\left(-(s-t)\eta_k \hat{\ell}_k^{(t)}(j)\right) \cdot \hat{\ell}_k^{(t)}(j).$$

Integrating (3) from t to s, plugging in above and noting that $\mathcal{Y}^{(t)} = Y^{(t)}$, we obtain

$$\frac{1}{\sqrt{\mathcal{Y}^{(s)}(k)}} = \frac{1}{\sqrt{Y^{(t)}(k)}} + \frac{\eta}{\eta_k} \sum_{j\in[m_k]} X_k^{(t)}(j) \left(1 - \exp\left(-\eta_k \cdot (s-t) \cdot \hat{\ell}_k^{(t)}(j)\right)\right),$$

which is exactly our rule to define $\overline{Y}^{(t+1)}$ in Line 1 of Algorithm 1 (take $s = t+1$). □

We define the regret for the piecewise continuous process as follows.

Definition B.2. The *continuous regret* contributed by the process $\left\{\mathcal{Z}^{(s)}\right\}_{s\in[0,T)}$ with respect to a fixed arm $a \in [N]$ is defined as

$$\mathscr{R}_a(T) := \sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} \langle \mathcal{Z}^{(s)} - \mathbf{e}_a^{[N]}, \ell^{(t)} \rangle \, \mathrm{d}s\right].$$

Then we are ready to bound $\mathscr{R}_a(T)$. Recall that we may write $\mathbf{e}_a^{[N]}$ as $\mathbf{e}_a$ if the information on N is clear from the context.

Lemma B.3. *For any time horizon $T > 0$, any loss sequence $\ell^{(0)}, \ell^{(1)}, \dots, \ell^{(T-1)} \in [0,1]^N$, and any arm $a = (k, j)$, it holds that*

$$\mathscr{R}_a(T) \le B_\psi(\mathbf{e}_k^{[K]}, Y^{(0)}) + B_{\phi_k}(\mathbf{e}_j^{[m_k]}, X_k^{(0)}).$$

Proof. Assume $a = (k, j)$. For every $t \in \{0, 1, \dots, T-1\}$, we compute the decreasing rate of the Bregman divergence caused by the evolution of $\mathcal{Y}^{(s)}$ and $\mathcal{X}_k^{(s)}$ respectively.

First consider the change of $B_\psi(\mathbf{e}_k, \mathcal{Y}^{(s)})$ over time:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s} B_\psi(\mathbf{e}_k, \mathcal{Y}^{(s)}) &= \frac{\mathrm{d}}{\mathrm{d}s}\Big(\psi(\mathbf{e}_k) - \psi(\mathcal{Y}^{(s)}) - \langle \mathbf{e}_k - \mathcal{Y}^{(s)}, \boldsymbol{\nabla}\psi(\mathcal{Y}^{(s)})\rangle\Big) \\ &= \langle \frac{\mathrm{d}\boldsymbol{\nabla}\psi(\mathcal{Y}^{(s)})}{\mathrm{d}s}, \mathcal{Y}^{(s)} - \mathbf{e}_k \rangle \\ &= -\langle \widehat{L}^{(s)}, \mathcal{Y}^{(s)} - \mathbf{e}_k \rangle.\end{aligned}$$

Integrating above from t to $t+1$, we have

$$\int_t^{t+1} \langle \widehat{L}^{(s)}, \mathcal{Y}^{(s)} - \mathbf{e}_k \rangle \,\mathrm{d}s = B_\psi(\mathbf{e}_k, \mathcal{Y}^{(t)}) - B_\psi(\mathbf{e}_k, \mathcal{Y}^{(t+1)^-}) = B_\psi(\mathbf{e}_k, Y^{(t)}) - B_\psi(\mathbf{e}_k, \overline{Y}^{(t+1)}), \tag{4}$$

where the last equality follows from Lemma B.1.

Note that *projection never increases Bregman divergence*; that is, we have

$$\begin{aligned}&B_\psi(\mathbf{e}_k, \overline{Y}^{(t+1)}) - B_\psi(\mathbf{e}_k, Y^{(t+1)}) \\ =\;& \psi(Y^{(t+1)}) - \psi(\overline{Y}^{(t+1)}) + \langle \boldsymbol{\nabla}\psi(Y^{(t+1)}), \mathbf{e}_k - Y^{(t+1)} \rangle - \langle \boldsymbol{\nabla}\psi(\overline{Y}^{(t+1)}), \mathbf{e}_k - \overline{Y}^{(t+1)} \rangle \\ =\;& \underbrace{\psi(Y^{(t+1)}) - \psi(\overline{Y}^{(t+1)}) - \langle \boldsymbol{\nabla}\psi(\overline{Y}^{(t+1)}), Y^{(t+1)} - \overline{Y}^{(t+1)} \rangle}_{A} + \underbrace{\langle \boldsymbol{\nabla}\psi(\overline{Y}^{(t+1)}) - \boldsymbol{\nabla}\psi(Y^{(t+1)}), Y^{(t+1)} - \mathbf{e}_k \rangle}_{B}.\end{aligned}$$

Since ψ is convex, we have $A \ge 0$. By the definition of $Y^{(t+1)}$,

$$Y^{(t+1)} = \operatorname*{arg\,min}_{\mathbf{y} \in \Delta_{K-1}} B_\psi(\mathbf{y}, \overline{Y}^{(t+1)}) = \operatorname*{arg\,min}_{\mathbf{y} \in \Delta_{K-1}} \psi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\nabla}\psi(\overline{Y}^{(t+1)}) \rangle.$$

The first-order optimality condition (see Section 26.5 in (Lattimore & Szepesvári, 2020)) implies that $B \ge 0$. As a result, $B_\psi(\mathbf{e}_k, \overline{Y}^{(t+1)}) \ge B_\psi(\mathbf{e}_k, Y^{(t+1)})$ and it follows from Equation (4) that

$$\int_t^{t+1} \langle \widehat{L}^{(s)}, \mathcal{Y}^{(s)} - \mathbf{e}_k \rangle \,\mathrm{d}s \le B_\psi(\mathbf{e}_k, Y^{(t)}) - B_\psi(\mathbf{e}_k, Y^{(t+1)}). \tag{5}$$

Then we consider the change of $B_{\phi_k}(\mathbf{e}_j, \mathcal{X}_k^{(s)})$ over time. Likewise we have

$$\frac{\mathrm{d}}{\mathrm{d}s} B_{\phi_k}(\mathbf{e}_j, \mathcal{X}_k^{(s)}) = \langle \frac{\mathrm{d}\boldsymbol{\nabla}\phi_k(\mathcal{X}_k^{(s)})}{\mathrm{d}s}, \mathcal{X}_k^{(s)} - \mathbf{e}_j \rangle = -\langle \hat{\ell}_k^{(t)}, \mathcal{X}_k^{(s)} - \mathbf{e}_j \rangle.$$

By an argument similar to the one for $\mathcal{Y}^{(s)}$ above, we can obtain

$$\int_t^{t+1} \langle \hat{\ell}_k^{(t)}, \mathcal{X}_k^{(s)} - \mathbf{e}_j \rangle \,\mathrm{d}s \le B_{\phi_k}(\mathbf{e}_j, X_k^{(t)}) - B_{\phi_k}(\mathbf{e}_j, X_k^{(t+1)}). \tag{6}$$

On the other hand, we have for every $s \in [t, t+1)$ and any arm $a^* = (k^*, j^*)$,

$$\mathbf{E}\left[\langle \mathcal{Z}^{(s)} - \mathbf{e}_{a^*}, \ell^{(t)} \rangle\right] = \mathbf{E}\left[\langle \mathcal{Z}^{(s)} - \mathbf{e}_{a^*}, \hat{\ell}^{(t)} \rangle\right] = \mathbf{E}\left[\sum_{k \in [K]} \sum_{j \in [m_k]} \mathcal{Y}^{(s)}(k) \cdot \mathcal{X}_k^{(s)}(j) \cdot \hat{\ell}_k^{(t)}(j) - \hat{\ell}^{(t)}(a^*)\right].$$

Recall that for every $k \in [K]$, it holds that $\widehat{L}^{(s)}(k) = \sum_{j \in [m_k]} \mathcal{X}_k^{(s)}(j) \cdot \hat{\ell}_k^{(t)}(j)$. Rearranging above yields

$$\begin{aligned}
\mathbf{E}\left[\langle \mathcal{Z}^{(s)} - \mathbf{e}_{a^*}, \ell^{(t)} \rangle\right] &= \mathbf{E}\left[\sum_{k \in [K]} \mathcal{Y}^{(s)}(k) \cdot \widehat{L}^{(s)}(k) - \hat{\ell}^{(t)}(a^*)\right] \\
&= \mathbf{E}\left[\langle \mathcal{Y}^{(s)}, \widehat{L}^{(s)} \rangle - \hat{\ell}^{(t)}(a^*)\right] \\
&= \mathbf{E}\left[\langle \mathcal{Y}^{(s)} - \mathbf{e}_{k^*}, \widehat{L}^{(s)} \rangle + \widehat{L}^{(s)}(k^*) - \hat{\ell}_{k^*}^{(t)}(j^*)\right] \\
&= \mathbf{E}\left[\langle \mathcal{Y}^{(s)} - \mathbf{e}_{k^*}, \widehat{L}^{(s)} \rangle\right] + \mathbf{E}\left[\langle \mathcal{X}_{k^*}^{(s)} - \mathbf{e}_{j^*}, \hat{\ell}_{k^*}^{(t)} \rangle\right].
\end{aligned}$$

Integrating above from t to $t+1$ and plugging in Equations (5) and (6), we obtain

$$\begin{aligned}
\int_t^{t+1} \mathbf{E}\left[\langle \mathcal{Z}^{(s)} - \mathbf{e}_{a^*}, \ell^{(t)} \rangle\right] \mathrm{d}s &= \int_t^{t+1} \mathbf{E}\left[\langle \mathcal{Y}^{(s)} - \mathbf{e}_{k^*}, \widehat{L}^{(s)} \rangle\right] \mathrm{d}s + \int_t^{t+1} \mathbf{E}\left[\langle \mathcal{X}_k^{(s)} - \mathbf{e}_{j^*}, \hat{\ell}_{k^*}^{(t)} \rangle\right] \mathrm{d}s \\
&\le B_\psi(\mathbf{e}_k, Y^{(t)}) - B_\psi(\mathbf{e}_k, Y^{(t+1)}) + B_{\phi_k}(\mathbf{e}_j, X_k^{(t)}) - B_{\phi_k}(\mathbf{e}_j, X_k^{(t+1)}).
\end{aligned}$$

Summing above over t from 0 to $T-1$ finishes the proof. □

B.1.2. Comparison of $R_a(T)$ and $\mathscr{R}_a(T)$

For any fixed loss sequence $\ell^{(0)}, \ell^{(1)}, \ldots, \ell^{(T-1)}$, we bound the difference between the regret $R_a(T)$ of Algorithm 1 and the continuous regret $\mathscr{R}_a(T)$ for any arm a. Formally, we establish the following lemma:

Lemma B.4.

$$R_a(T) - \mathscr{R}_a(T) \le \frac{1}{2} \sum_{t=0}^{T-1} \mathbf{E}\left[\sup_{\xi \in \mathtt{Rect}(Y^{(t)}, \overline{Y}^{(t+1)})} \|\widehat{L}^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\psi(\xi)} + \sum_{k \in [K]} Y^{(t)}(k) \cdot \sup_{\zeta_k \in \mathtt{Rect}(X_k^{(t)}, \overline{X}_k^{(t+1)})} \|\hat{\ell}_k^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k)}\right].$$

Proof. By the definition of the regret, we have

$$\begin{aligned}
R_a(T) &= \mathbf{E}\left[\sum_{t=0}^{T-1} \langle Z^{(t)} - \mathbf{e}_a, \hat{\ell}^{(t)} \rangle\right] \\
&= \sum_{t=0}^{T-1} \mathbf{E}\left[\langle Z^{(t)} - \mathbf{e}_a, \hat{\ell}^{(t)} \rangle\right] \\
&= \sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} \langle \mathcal{Z}^{(s)} - \mathbf{e}_a, \hat{\ell}^{(t)} \rangle \,\mathrm{d}s + \int_t^{t+1} \langle Z^{(t)} - \mathcal{Z}^{(s)}, \hat{\ell}^{(t)} \rangle \,\mathrm{d}s\right] \\
&= \mathscr{R}_a(T) + \sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} \langle Z^{(t)} - \mathcal{Z}^{(s)}, \hat{\ell}^{(t)} \rangle \,\mathrm{d}s\right],
\end{aligned}$$

where the first equality holds due to Fubini's theorem. Therefore, we only need to bound the term $\sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} \langle Z^{(t)} - \mathcal{Z}^{(s)}, \hat{\ell}^{(t)} \rangle \,\mathrm{d}s\right]$.

Fix $t \in \{0, 1, \ldots, T-1\}$. We have shown in the proof of Lemma B.1 that

$$\mathcal{X}_k^{(s)}(j) = X_k^{(t)}(j) \cdot \exp\left(-(s-t)\eta_k \hat{\ell}_k^{(t)}(j)\right) \le X_k^{(t)}(j)$$

for any $s \in [t, t+1)$ and any $j \in [m_k]$.

Recall that $\widehat{L}^{(s)}(k) = \sum_{j \in [m_k]} \mathcal{X}_k^{(s)}(j) \cdot \hat{\ell}_k^{(t)}(j)$ for every $k \in [K]$. Then by the discussion above, we have $\widehat{L}^{(s)} \le \widehat{L}^{(t)}$ for any $s \in [t, t+1)$. As a result, it follows from (3) that for any $s \in [t, t+1)$,

$$\boldsymbol{\nabla}\psi(\mathcal{Y}^{(s)}) - \boldsymbol{\nabla}\psi(Y^{(t)}) = \int_t^s -\widehat{L}^{(w)} \,\mathrm{d}w \ge -(s-t) \cdot \widehat{L}^{(t)}. \tag{7}$$

Recall that for any two vectors $\mathbf{x}, \mathbf{y}$ of the same dimension, $\texttt{Rect}(\mathbf{x}, \mathbf{y})$ is the rectangle between $\mathbf{x}$ and $\mathbf{y}$. Since our ψ is a *separable function* (and therefore $\boldsymbol{\nabla}^2\psi$ is diagonal), we can apply the *mean value theorem* entrywise and obtain

$$\boldsymbol{\nabla}\psi(\mathcal{Y}^{(s)}) - \boldsymbol{\nabla}\psi(Y^{(t)}) = \boldsymbol{\nabla}^2\psi(\xi^{(s)})(\mathcal{Y}^{(s)} - Y^{(t)}) \tag{8}$$

for some $\xi^{(s)} \in \texttt{Rect}(\mathcal{Y}^{(s)}, Y^{(t)})$.

By our choice of ψ, it holds that $\boldsymbol{\nabla}^2\psi(\xi^{(s)}) \succ 0$ for any $\xi^{(s)} \in \texttt{Rect}(\mathcal{Y}^{(s)}, Y^{(t)})$. Therefore, combining Equations (7) and (8), we have

$$\mathcal{Y}^{(s)} \geq Y^{(t)} - (s-t) \cdot \boldsymbol{\nabla}^{-2}\psi(\xi^{(s)}) \cdot \widehat{L}^{(t)}.$$

Similar argument yields that

$$\mathcal{X}_k^{(s)} \geq X_k^{(t)} - (s-t) \cdot \boldsymbol{\nabla}^{-2}\phi_k(\zeta_k^{(s)}) \cdot \hat{\ell}_k^{(t)}$$

for some $\zeta_k^{(s)} \in \texttt{Rect}(\mathcal{X}_k^{(s)}, X_k^{(t)})$.

Therefore for any $k \in [K], j \in [m_k]$ and any $s \in [t, t+1)$, we can bound the difference between $Z^{(t)}(k,j)$ and $\mathcal{Z}^{(s)}(k,j)$:

$$\begin{aligned}
&Z^{(t)}(k,j) - \mathcal{Z}^{(s)}(k,j) = Y^{(t)}(k) \cdot X_k^{(t)}(j) - \mathcal{Y}^{(s)}(k) \cdot \mathcal{X}_k^{(s)}(j) \\
&\leq Y^{(t)}(k) \cdot X_k^{(t)}(j) - \left(Y^{(t)}(k) - (s-t) \cdot \left[\boldsymbol{\nabla}^{-2}\psi(\xi^{(s)}) \cdot \widehat{L}^{(t)}\right](k)\right) \cdot \left(X_k^{(t)}(j) - (s-t) \cdot \left[\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k^{(s)}) \cdot \hat{\ell}_k^{(t)}\right](j)\right) \\
&= -(s-t)^2 \cdot \left[\boldsymbol{\nabla}^{-2}\psi(\xi^{(s)}) \cdot \widehat{L}^{(t)}\right](k) \cdot \left[\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k^{(s)}) \cdot \hat{\ell}_k^{(t)}\right](j) + (s-t) \cdot X_k^{(t)}(j) \cdot \left[\boldsymbol{\nabla}^{-2}\psi(\xi^{(s)}) \cdot \widehat{L}^{(t)}\right](k) \\
&\quad + (s-t) \cdot Y^{(t)}(k) \cdot \left[\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k^{(s)}) \cdot \hat{\ell}_k^{(t)}\right](j) \\
&\leq (s-t) \cdot X_k^{(t)}(j) \cdot \left[\boldsymbol{\nabla}^{-2}\psi(\xi^{(s)}) \cdot \widehat{L}^{(t)}\right](k) + (s-t) \cdot Y^{(t)}(k) \cdot \left[\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k^{(s)}) \cdot \hat{\ell}_k^{(t)}\right](j)
\end{aligned}$$

for some $\xi^{(s)} \in \texttt{Rect}(\mathcal{Y}^{(s)}, Y^{(t)})$ and $\zeta_k^{(s)} \in \texttt{Rect}(\mathcal{X}_k^{(s)}, X_k^{(t)})$.

We are now ready to bound the gap between $R_a(T)$ and $\mathscr{R}_a(T)$:

$$\begin{aligned}
&R_a(T) - \mathscr{R}_a(T) \\
&= \sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} \langle Z^{(t)} - \mathscr{Z}^{(s)}, \hat{\ell}^{(t)} \rangle\right] \\
&\leq \underbrace{\sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} (s-t) \left(\sum_{k\in[K]} \sum_{j\in[m_k]} X_k^{(t)}(j) \cdot \sup_{\xi \in \texttt{Rect}(Y^{(t)}, \overline{Y}^{(t+1)})} \left[\boldsymbol{\nabla}^{-2}\psi(\xi) \cdot \widehat{L}^{(t)}\right](k)\right) \cdot \hat{\ell}_k^{(t)}(j)\, \mathrm{d}s\right]}_{(A)} \\
&+ \underbrace{\sum_{t=0}^{T-1} \mathbf{E}\left[\int_t^{t+1} (s-t) \left(\sum_{k\in[K]} \sum_{j\in[m_k]} Y^{(t)}(k) \cdot \sup_{\zeta_k \in \texttt{Rect}(X_k^{(t)}, \overline{X}_k^{(t+1)})} \left[\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k) \cdot \hat{\ell}_k^{(t)}\right](j)\right) \cdot \hat{\ell}_k^{(t)}(j)\, \mathrm{d}s\right]}_{(B)}.
\end{aligned}$$

Note that in both expressions (A) and (B) above, only the term $(s-t)$ depend on s. So we can integrate and obtain:

$$\begin{aligned}
(A) &= \frac{1}{2}\sum_{t=0}^{T-1} \mathbf{E}\left[\left(\sum_{k\in[K]} \sum_{j\in[m_k]} X_k^{(t)}(j) \cdot \sup_{\xi \in \texttt{Rect}(Y^{(t)}, \overline{Y}^{(t+1)})} \left[\boldsymbol{\nabla}^{-2}\psi(\xi) \cdot \widehat{L}^{(t)}\right](k)\right) \cdot \hat{\ell}_k^{(t)}(j)\right] \\
&= \frac{1}{2}\sum_{t=0}^{T-1} \mathbf{E}\left[\sum_{k\in[K]} \sup_{\xi \in \texttt{Rect}(Y^{(t)}, \overline{Y}^{(t+1)})} \left[\boldsymbol{\nabla}^{-2}\psi(\xi) \cdot \widehat{L}^{(t)}\right](k) \cdot \left(\sum_{j\in[m_k]} X_k^{(t)}(j) \cdot \hat{\ell}_k^{(t)}(j)\right)\right] \\
&= \frac{1}{2}\sum_{t=0}^{T-1} \mathbf{E}\left[\sum_{k\in[K]} \sup_{\xi \in \texttt{Rect}(Y^{(t)}, \overline{Y}^{(t+1)})} \left[\boldsymbol{\nabla}^{-2}\psi(\xi) \cdot \widehat{L}^{(t)}\right](k) \cdot \widehat{L}^{(t)}(k)\right]
\end{aligned} \tag{9}$$

$$= \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\left[\sup_{\xi\in\mathtt{Rect}(Y^{(t)},\overline{Y}^{(t+1)})}\|\widehat{L}^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\psi(\xi)}\right].$$

Similarly,

$$(B) = \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\left[\left(\sum_{k\in[K]}\sum_{j\in[m_k]}Y^{(t)}(k)\cdot\sup_{\zeta_k\in\mathtt{Rect}(X_k^{(t)},\overline{X}_k^{(t+1)})}\left[\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k)\cdot\hat{\ell}_k^{(t)}\right](j)\right)\cdot\hat{\ell}_k^{(t)}(j)\right] \tag{10}$$

$$= \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\left[\sum_{k\in[K]}Y^{(t)}(k)\cdot\sup_{\zeta_k\in\mathtt{Rect}(X_k^{(t)},\overline{X}_k^{(t+1)})}\|\hat{\ell}_k^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k)}\right].$$

Combining Equations (9) and (10), we have

$$R_a(T) - \mathscr{R}_a(T) \le \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\left[\sup_{\xi\in\mathtt{Rect}(Y^{(t)},\overline{Y}^{(t+1)})}\|\widehat{L}^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\psi(\xi)} + \sum_{k\in[K]}Y^{(t)}(k)\cdot\sup_{\zeta_k\in\mathtt{Rect}(X_k^{(t)},\overline{X}_k^{(t+1)})}\|\hat{\ell}_k^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k)}\right]. \tag{11}$$

□

If we apply the "regret decomposition theorem" in (He & Zhang, 2023) and use the standard OSMD bound for each stage, we will get the term

$$\sup_{\zeta_{k^*}\in\mathtt{Rect}(X_{k^*}^{(t)},\overline{X}_{k^*}^{(t+1)})}\|\hat{\ell}_{k^*}^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\phi_{k^*}(\zeta_{k^*})} \tag{12}$$

where k^* is the index of the group containing the optimal arm instead of the term

$$\sum_{k\in[K]}Y^{(t)}(k)\cdot\sup_{\zeta_k\in\mathtt{Rect}(X_k^{(t)},\overline{X}_k^{(t+1)})}\|\hat{\ell}_k^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k)}$$

in Equation (11). The new $Y^{(t)}(k)$ term is crucial to our optimal regret bound since it cancels a $Y^{(t)}(k)$ term hidden in the denominator of $\|\hat{\ell}_k^{(t)}\|^2_{\boldsymbol{\nabla}^{-2}\phi_k(\zeta_k)}$. This will be clear in Appendix B.1.3.

B.1.3. The Regret of Algorithm 1

Note that the regret of Algorithm 1 is composed of the two parts in Lemma B.3 and Lemma B.4. In this section, we will prove Theorem 3.1 by providing more specific bounds for the terms in these two lemmas.

Proof of Theorem 3.1. By definition of Bregman divergence,

$$B_\psi(\mathbf{e}_k, Y^{(0)}) = \psi(\mathbf{e}_k) - \psi(Y^{(0)}) - \langle\nabla\psi(Y^{(0)}), \mathbf{e}_k - Y^{(0)}\rangle.$$

Since we initialize $Y^{(0)} = \arg\min_{b\in\Delta_{K-1}}\psi(b)$, $Y^{(0)}(k) = \frac{1}{K}$ for $k\in[K]$ and $\langle\nabla\psi(Y^{(0)}), \mathbf{e}_k - Y^{(0)}\rangle \ge 0$ follows the first-order optimality condition for $Y^{(0)}$. Thus

$$B_\psi(\mathbf{e}_k, Y^{(0)}) \le \psi(\mathbf{e}_k) - \psi(Y^{(0)}) = \frac{-2+2\sqrt{K}}{\eta} \le \frac{2\sqrt{K}}{\eta}.$$

Similarly we have $X_k^{(0)}(j) = \frac{1}{m_k}$ for $j\in[m_k]$ and

$$B_{\phi_k}(\mathbf{e}_j, X_k^{(0)}) \le \phi_k(\mathbf{e}_j) - \phi_k(X_k^{(0)}) = \frac{\log m_k}{\eta_k}.$$

Therefore

$$\mathscr{R}_a(T) \le \frac{2\sqrt{K}}{\eta} + \frac{\log m_k}{\eta_k}. \tag{13}$$

Recall that $A_t = (k_t, j_t)$ is the arm pulled by the algorithm at round t. Now we plug our estimator $\hat{\ell}_k^{(t)}(j) = \frac{\mathbb{1}[k_t=k]}{Y^{(t)}(k)}\ell_k^{(t)}(j)$ and $\nabla^2\psi(\xi) = \operatorname{diag}\left(\frac{1}{2\eta\xi(1)^{3/2}}, \frac{1}{2\eta\xi(2)^{3/2}}, \cdots, \frac{1}{2\eta\xi(K)^{3/2}}\right)$ into the first term on the RHS of Lemma B.4.

$$\begin{aligned}
\mathbf{E}\left[\sup_{\xi\in\mathtt{Rect}(Y^{(t)},\overline{Y}^{(t+1)})}\|\widehat{L}^{(t)}\|^2_{\nabla^{-2}\psi(\xi)}\right] &= 2\eta\mathbf{E}\left[\sup_{\xi\in\mathtt{Rect}(Y^{(t)},\overline{Y}^{(t+1)})}\sum_{k\in[K]}\xi(k)^{3/2}\cdot\left(\frac{\mathbb{1}[k_t=k]}{Y^{(t)}(k)}\sum_{j\in[m_k]}\ell_k^{(t)}(j)X_k^{(t)}(j)\right)^2\right]\\
&\overset{(a)}{\le} 2\eta\mathbf{E}\left[\sum_{k\in[K]}\left(Y^{(t)}(k)\right)^{3/2}\cdot\left(\frac{\mathbb{1}[k_t=k]}{Y^{(t)}(k)}\sum_{j\in[m_k]}\ell_k^{(t)}(j)X_k^{(t)}(j)\right)^2\right]\\
&\overset{(b)}{\le} 2\eta\mathbf{E}\left[\mathbf{E}\left[\sum_{k\in[K]}\frac{\mathbb{1}[k_t=k]}{\sqrt{Y^{(t)}(k)}}\right]\,\middle|\, Y^{(t)}\right]\\
&= 2\eta\sum_{k=1}^{K}\mathbf{E}\left[\sqrt{Y^{(t)}(k)}\right] \overset{(c)}{\le} 2\eta\sum_{k=1}^{K}\sqrt{\mathbf{E}\left[Y^{(t)}(k)\right]} \le 2\eta\sqrt{K}.
\end{aligned}$$

In the calculation above: (a) follows from $\overline{Y}^{(t+1)}(k) \le Y^{(t)}(k)$, (b) is due to $\sum_{j\in[m_k]}\ell_k^{(t)}(j)X_k^{(t)}(j) \in [0,1]$, and (c) is due to Jensen's inequality.

Similarly we have for the second term with $\nabla^2\phi_k(\zeta_k) = \operatorname{diag}\left(\frac{1}{\eta_k\zeta_k(1)}, \frac{1}{\eta_k\zeta_k(2)}, \cdots, \frac{1}{\eta_k\zeta_k(m_k)}\right)$

$$\begin{aligned}
&\mathbf{E}\left[\sum_{k\in[K]}Y^{(t)}(k)\cdot\sup_{\zeta_k\in\mathtt{Rect}(X_k^{(t)},\overline{X}_k^{(t+1)})}\|\hat{\ell}_k^{(t)}\|^2_{\nabla^{-2}\phi_k(\zeta_k)}\right]\\
&= \mathbf{E}\left[\sum_{k\in[K]}\eta_k Y^{(t)}(k)\cdot\sup_{\zeta_k\in\mathtt{Rect}(X_k^{(t)},\overline{X}_k^{(t+1)})}\sum_{j\in[m_k]}\zeta_k(j)\cdot\left(\frac{\mathbb{1}[k_t=k]}{Y^{(t)}(k)}\ell_k^{(t)}(j)\right)^2\right]\\
&\overset{(d)}{\le} \mathbf{E}\left[\sum_{k\in[K]}\eta_k Y^{(t)}(k)\cdot\sum_{j\in[m_k]}X_k^{(t)}(j)\cdot\left(\frac{\mathbb{1}[k_t=k]}{Y^{(t)}(k)}\ell_k^{(t)}(j)\right)^2\right]\\
&\overset{(e)}{\le} \mathbf{E}\left[\mathbf{E}\left[\sum_{k\in[K]}\eta_k\cdot\sum_{j\in[m_k]}X_k^{(t)}(j)\cdot\frac{\mathbb{1}[k_t=k]}{Y^{(t)}(k)}\right]\,\middle|\, Y^{(t)}(k)\right]\\
&= \sum_{k\in[K]}\eta_k\sum_{j\in[m_k]}X_k^{(t)}(j) = \sum_{k\in[K]}\eta_k.
\end{aligned}$$

In the calculation above: (d) follows from $\overline{X}_k^{(t+1)}(j) \le X_k^{(t)}(j)$ and (e) is due to $\ell_k^{(t)}(j) \in [0,1]$.

Hence, summing up above two terms from 0 to $T-1$, we obtain

$$R_a(T) - \mathscr{R}_a(T) \le \eta\sqrt{K}T + \frac{1}{2}T\sum_{k\in[K]}\eta_k. \tag{14}$$

Combining Equations (13) and (14) and choosing $\eta = \frac{1}{\sqrt{T}}$ and $\eta_k = \frac{\log(m_k+1)}{\sqrt{T\sum_{k=1}^{K}\log(m_k+1)}}$, we obtain for any fixed arm a,

$$R_a(T) \le \frac{2\sqrt{K}}{\eta} + \frac{\log m_k}{\eta_k} + \frac{T}{2}\sum_{k\in[K]}\eta_k + \eta T\sqrt{K} \le O\left(\sqrt{T\sum_{k=1}^{K}\log(m_k+1)}\right).$$

□

B.2. Upper Bound for `BAI`

We can use the Algorithm 1 and Theorem 3.1 to give an upper bound for $\mathbf{m}$-`BAI` through Lemma A.1.

Proof of Theorem 1.2. We use Algorithm 1 to construct an $(\varepsilon, 0.05)$-PAC algorithm for $\mathbf{m}$-`BAI` as described in Lemma A.1. Since the regret satisfies $R(T) \le c\sqrt{T\sum_{k=1}^{K}\log(1+m_k)}$ for some constant c on every loss sequence by Theorem 1.1, running Algorithm 1 with $T^* = \frac{(2500c)^2\sum_{k=1}^{K}\log(1+m_k)}{\varepsilon^2}$, we can get an $(\varepsilon, 0.05)$-PAC algorithm which always terminates in $O\left(\sum_{k=1}^{K}\frac{\log(m_k+1)}{\varepsilon^2}\right)$ rounds. □

B.3. The Strongly Observable Graph with Self-loops

We can generalize our results to any strongly observable graph $G = (V, E)$ with each vertex owning a self-loop. Assume G contains a $(V_1, \dots, V_K)$-clique cover. We construct a new graph $G' = (V, E')$ by ignoring the edges between any two distinct cliques. It is clear that $R^*(G, T) \le R^*(G', T)$. Then we can prove Corollary 1.5 by directly applying Algorithm 1 with feedback graph G'. This proves Corollary 1.5, which asserts that

$$R^*(G, T) = O\left(\sqrt{T \cdot \sum_{k=1}^{K}\log(m_k+1)}\right).$$

Although we assume that each vertex contains a self-loop for the sake of simplicity, we note that our algorithm can still be applied to strongly observable graphs that have some vertices without self-loops. In such cases, we can incorporate an additional exploration term into our algorithm, and a similar analysis to that in Section 3.2 still works.

There have been several works using the clique cover as the parameter to bound the minimax regret of graph bandit. For example, (Erez & Koren, 2021) applies FTRL algorithm with a carefully designed potential function which combines the Tsallis entropy with negative entropy. It achieves a regret of $(\log T)^{O(1)} \cdot O\left(\sqrt{KT}\right)$. Our new bound takes into account the size of each clique and is always superior.

C. Lower Bounds for $\mathbf{m}$-`BAI`

Let $\mathcal{A}$ be an algorithm for $\mathbf{m}$-`BAI` where $\mathbf{m} = (m_1, \dots, m_K)$ is a vector. Given an instance of $\mathbf{m}$-`BAI`, we use T to denote the number of rounds the algorithm $\mathcal{A}$ proceeds. Recall that for every group $k \in [K]$ and $j \in [m_k]$, we use $T_{(k,j)}$ to denote the number of times that the arm (k, j) has been pulled. For every $k \in [K]$, let $T^{(k)} = \sum_{j\in[m_k]} T_{(k,j)}$ be the number of rounds the arms in the k-th group have been pulled. We also use $N_{(k,j)}$ to denote the number of times the arm (k, j) has been observed. Clearly $N_{(k,j)} = T^{(k)}$.

In the following part, we only consider stochastic environment. That is, $\ell^{(t)}$ is independently drawn from the same distribution for each $t \in \mathbb{N}$. Therefore, we omit the superscript (t) and only use $\ell(i)$ or $\ell_k(j)$ to denote the one-round loss of arm i or arm (k, j) respectively when the information is clear from the context.

In Appendix C.1, we lower bound the number of rounds for a PAC algorithm on a specific $\mathbf{m}$-`BAI` instance with $\mathbf{m} = (m)$ and then prove the result for $\mathbf{m}$-`BAI` in Appendix C.4. We then use these results to prove a regret lower bound for $\mathbf{m}$-`MAB` and bandit problems with general feedback graphs in Appendix E.

C.1. An Instance-Specific Lower Bound for (m)-`BAI`

In this section, we study the number of rounds required for (m)-`BAI` in an $(\varepsilon, 0.05)$-PAC algorithm. In this setting, the pull of any arm can observe the losses of all arms. We will establish a lower bound for a specified instance, namely the one where all arms follow `Ber`$(\frac{1}{2})$. This is key to our lower bound later.

We focus on instances of (m)-`BAI` where each arm is Bernoulli. As a result, each instance can be specified by a vector $(p_1, \dots, p_{m-1}, p_m) \in \mathbb{R}^m$ meaning the loss of arm i follows `Ber`(p_i) in each round *independently*.

Let $\varepsilon \in \left(0, \frac{1}{2}\right)$. In the following context, when we denote an instance as $\mathscr{H}^{\mathbf{m}}$, the superscript $\mathbf{m}$ indicates that it is an

$\mathbf{m}$-$\mathtt{BAI}$ instance. Consider the following $m+1$ (m)-$\mathtt{BAI}$ instances $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$:

- The instance $\mathscr{H}_0^{(m)}$ is $\left(\frac{1}{2},\frac{1}{2},\frac{1}{2},\cdots,\frac{1}{2}\right)$. That is, $p_i=\frac{1}{2}$ for every $i\in[m]$ in $\mathscr{H}_0^{(m)}$;
- For $j\in[m]$,

$$\mathscr{H}_j^{(m)}=\left(\frac{1}{2},\frac{1}{2},\cdots,\frac{1}{2},\quad \underset{\substack{\uparrow\\ \text{the } j\text{-th arm}}}{\frac{1}{2}-\varepsilon}\quad,\frac{1}{2},\cdots,\frac{1}{2}\right);$$

  that is, the instance satisfies $p_j=\frac{1}{2}-\varepsilon$ and $p_i=\frac{1}{2}$ for every $i\neq j$.

We say an algorithm $\mathcal{A}$ distinguishes $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$ with probability p if

$$\mathbf{Pr}\left[\mathcal{A}\text{ outputs } j\;\middle|\;\text{the input instance is }\mathscr{H}_j^{(m)}\right]\geq p,$$

and the output can be arbitrary among $\{0,1,\ldots m\}$ when the input is not in $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$. We refer it as a *distinguishing algorithm*, which is different from the $(\varepsilon,0.05)$-PAC algorithm.

The main result of this section is

Lemma C.1 (restate Lemma 4.1)**.** *Let $\mathcal{A}$ be an $(\varepsilon,0.05)$-PAC algorithm. Assume $m\geq 2$. There exists a universal constant $c_1>0$ such that $\mathcal{A}$ terminates on $\mathscr{H}_0^{(m)}$ after at least $\frac{c_1}{\varepsilon^2}\log(m+1)$ rounds in expectation.*

We will demonstrate in Lemma C.6 that an $(\varepsilon,0.05)$-PAC algorithm can be adapted into an algorithm to distinguish $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$ with few extra samples. Thus, we first establish a lower bound for distinguishing algorithms. For technical reasons, we begin by proving a lower bound for distinguishing algorithms for *Gaussian arms* in Appendix C.2 and subsequently reduce the Bernoulli arms to Gaussian arms in Appendix C.3.

C.2. The Gaussian Arms

In this section, we relax the constraint on the range of each arm's loss and allow the losses to be arbitrary real numbers. Let $\varepsilon\in\left(0,\frac{1}{2}\right)$ and $\sigma\in\left(\frac{1}{2\sqrt{2\pi}},\frac{1}{\sqrt{2\pi}}\right)$. We construct $m+1$ instances $\{\mathscr{N}_j\}_{j\in\{0\}\cup[m]}$ with Gaussian distributions:

- In the instance $\mathscr{N}_0$, for each $i\in[m]$, $\ell(i)$ is independently drawn from a Gaussian distribution $\mathcal{N}(0,\sigma^2)$;
- In the instance $\mathscr{N}_j$ for $j\in[m]$, $\ell(j)\sim\mathcal{N}(-\varepsilon,\sigma^2)$ and $\ell(i)\sim\mathcal{N}(0,\sigma^2)$ for each $i\neq j$ and $i\in[m]$ independently.

Lemma C.2 (Bretagnolle-Huber inequality, see e.g. (Lattimore & Szepesvári, 2020))**.** *Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be two probability measures on the same measurable space $(\Omega,\mathcal{F})$, and let $E\in\mathcal{F}$ be an arbitrary event. Then*

$$\mathbf{P}_1[E]+\mathbf{P}_2[\overline{E}]\geq\frac{1}{2}e^{-D_{\mathrm{KL}}(\mathbf{P}_1,\mathbf{P}_2)}$$

Let $\mathscr{N}_{\mathtt{mix}}$ be the mixture of $\{\mathscr{N}_j\}_{j\in[m]}$ meaning that the environment chooses k from $[m]$ uniformly at random and generates losses according to $\mathscr{N}_k$ in the following $\mathtt{BAI}$ game. Let $\mathcal{A}$ be an algorithm distinguishing $\{\mathscr{N}_j\}_{j\in[m]\cup\{0\}}$. Let Ω be the set of all possible outcomes during the first t^* rounds, including the samples according to the input distribution and the output of $\mathcal{A}$ (if $\mathcal{A}$ does not terminate after the t^*-th round, we assume its output is -1). Note that if the algorithm terminates in $t'<t^*$ rounds, we can always add t^*-t' virtual rounds so that it still produces a certain loss sequence in $\mathbb{R}^{m\times t^*}$.

As a result, each outcome $\omega\in\Omega$ can be viewed as a pair $\omega=(w,x)$ where $w\in\mathbb{R}^{m\times t^*}$ is the loss sequence and $x\in\{-1,0,1,\ldots,m\}$ indicates the output of $\mathcal{A}$. Thus $\Omega=W\times\{-1,0,1,\ldots,m\}$ where $W=\mathbb{R}^{m\times t^*}$.

To ease the proof below, we slightly change $\mathcal{A}$'s output: if the original output is $x\in\{-1,0,\ldots,m\}$, we instead output a uniform real in $[x,x+1)$. Therefore, we can let $\Omega=W\times X$ where $W=\mathbb{R}^{m\times t^*}$ and $X=\mathbb{R}$. The benefit of doing so is that we can let $\mathcal{F}$ be the Borel sets in Ω which is convenient to work with. Clearly it is sufficient to establish lower bounds for the algorithms after the change.

For any instance $\mathscr{H}^{(m)}$, let $\mathbf{P}_{\mathscr{H}^{(m)}}$ be the measure of outcomes of $\mathcal{A}$ in t^* rounds with input instance $\mathscr{H}^{(m)}$ and $\mathbf{p}_{\mathscr{H}^{(m)}}$ be the corresponding probability density function (PDF). Then $\mathbf{P}_{\mathscr{N}_0}$ and $\mathbf{P}_{\mathscr{N}_{\mathtt{mix}}}$ are two probability measures on $(\Omega, \mathcal{F})$ and $\mathbf{p}_{\mathscr{N}_{\mathtt{mix}}}(\omega) = \frac{1}{m}\sum_{j\in[m]} \mathbf{p}_{\mathscr{N}_j}(\omega)$ for any $\omega = (w, x) \in \Omega = \mathbb{R}^{m\times t^*+1}$. We also let $\mathbf{p}^W_{\mathscr{H}^{(m)}}$ be the PDF of the samples during the first t^* rounds according to the input $\mathscr{H}^{(m)}$ and $\mathbf{p}^X_{\mathscr{H}^{(m)}}$ be the PDF of $\mathcal{A}$'s output. Furthermore, we let $\mathbf{p}^{X|W}_{\mathscr{H}^{(m)}}$ to be the conditional density function of X given W. By definition, we have $\mathbf{p}^{X|W}_{\mathscr{H}^{(m)}}(x|w) = \frac{\mathbf{p}_{\mathscr{H}^{(m)}}(\omega)}{\mathbf{p}^W_{\mathscr{H}^{(m)}}(w)}$.

Lemma C.3.

$$D_{\mathtt{KL}}\left(\mathbf{P}_{\mathscr{N}_{\mathtt{mix}}}, \mathbf{P}_{\mathscr{N}_0}\right) \le \log \frac{m - 1 + \exp\left(\frac{\varepsilon^2 t^*}{\sigma^2}\right)}{m}.$$

Proof. For any $\omega = (w, x) \in \Omega$, let $w_{j,t}$ denote the $(j, t)^{\text{th}}$ entry of the matrix w for every $j \in [m]$ and $t \in [t^*]$. That is, $w_{j,t} = \ell^{(t)}(j)$, which is the loss of arm j in the t-th round. Then for each $i \in [m]$,

$$\mathbf{p}^W_{\mathscr{N}_i}(w) = \left(2\pi\sigma^2\right)^{-\frac{mt^*}{2}} \exp\left(-\frac{\sum_{t\in[t^*]}\left((w_{i,t}+\varepsilon)^2 + \sum_{j\neq i} w_{j,t}^2\right)}{2\sigma^2}\right)$$

and

$$\mathbf{p}^W_{\mathscr{N}_0}(w) = \left(2\pi\sigma^2\right)^{-\frac{mt^*}{2}} \exp\left(-\frac{\sum_{t\in[t^*], j\in[m]} w_{j,t}^2}{2\sigma^2}\right).$$

Therefore we have

$$\begin{aligned}\frac{\mathbf{p}_{\mathscr{N}_i}(\omega)}{\mathbf{p}_{\mathscr{N}_0}(\omega)} = \frac{\mathbf{p}^W_{\mathscr{N}_i}(w)}{\mathbf{p}^W_{\mathscr{N}_0}(w)} &= \frac{\left(2\pi\sigma^2\right)^{-\frac{mt^*}{2}} \exp\left(-\frac{\sum_{t\in[t^*]}\left((w_{i,t}+\varepsilon)^2 + \sum_{j\neq i} w_{j,t}^2\right)}{2\sigma^2}\right)}{\left(2\pi\sigma^2\right)^{-\frac{mt^*}{2}} \exp\left(-\frac{\sum_{t\in[t^*], j\in[m]} w_{j,t}^2}{2\sigma^2}\right)} \\ &= \exp\left(-\frac{\varepsilon^2 t^* + 2\varepsilon\sum_{t\in[t^*]} w_{i,t}}{2\sigma^2}\right).\end{aligned}$$

From Jensen's inequality, we have

$$\begin{aligned}D_{\mathtt{KL}}\left(\mathbf{P}_{\mathscr{N}_{\mathtt{mix}}}, \mathbf{P}_{\mathscr{N}_0}\right) &= \int_\Omega \log\frac{\mathbf{p}_{\mathscr{N}_{\mathtt{mix}}}(\omega)}{\mathbf{p}_{\mathscr{N}_0}(\omega)}\,\mathrm{d}\mathbf{P}_{\mathscr{N}_{\mathtt{mix}}}(\omega) \le \log\int_\Omega \frac{\mathbf{p}_{\mathscr{N}_{\mathtt{mix}}}(\omega)}{\mathbf{p}_{\mathscr{N}_0}(\omega)}\,\mathrm{d}\mathbf{P}_{\mathscr{N}_{\mathtt{mix}}}(\omega) \\ &= \log\int_\Omega \frac{1}{m}\sum_{j\in[m]} \mathbf{p}_{\mathscr{N}_j}(\omega)\frac{\frac{1}{m}\sum_{i\in[m]}\mathbf{p}_{\mathscr{N}_j}(\omega)}{\mathbf{p}_{\mathscr{N}_0}(\omega)}\,\mathrm{d}\omega.\end{aligned}$$

Note that for $\omega = (w, x)$, For $i, j \in [m]$ and $i \neq j$,

$$\begin{aligned}\int_\Omega \mathbf{p}_{\mathscr{N}_i}(\omega)\frac{\mathbf{p}_{\mathscr{N}_j}(\omega)}{\mathbf{p}_{\mathscr{N}_0}(\omega)}\,\mathrm{d}\omega &= \int_W\int_X \mathbf{p}^W_{\mathscr{N}_i}(w)\cdot\mathbf{p}^{X|W}_{\mathscr{N}_i}(x|w)\frac{\mathbf{p}^W_{\mathscr{N}_j}(w)}{\mathbf{p}^W_{\mathscr{N}_0}(w)}\mathrm{d}x\mathrm{d}w \\ &= \int_W \mathbf{p}^W_{\mathscr{N}_i}(w)\frac{\mathbf{p}^W_{\mathscr{N}_j}(w)}{\mathbf{p}^W_{\mathscr{N}_0}(w)}\mathrm{d}w \\ &= \left(2\pi\sigma^2\right)^{-\frac{mt^*}{2}}\cdot\int_\Omega \exp\left(-\frac{\sum_{t\in[t^*]}\left((w_{i,t}+\varepsilon)^2 + (w_{j,t}+\varepsilon)^2\right) + \sum_{\substack{j'\neq i \\ j'\neq j}} w_{j',t}^2}{2\sigma^2}\right)\mathrm{d}w = 1.\end{aligned}$$

For $i \in [m]$,

$$\int_\Omega \mathbf{p}_{\mathscr{N}_i}(\omega)\frac{\mathbf{p}_{\mathscr{N}_i}(\omega)}{\mathbf{p}_{\mathscr{N}_0}(\omega)}\,\mathrm{d}\omega = \int_W\int_X \mathbf{p}^W_{\mathscr{N}_i}(w)\cdot\mathbf{p}^{X|W}_{\mathscr{N}_i}(x|w)\frac{\mathbf{p}^W_{\mathscr{N}_i}(w)}{\mathbf{p}^W_{\mathscr{N}_0}(w)}\mathrm{d}x\mathrm{d}w$$

$$
\begin{aligned}
&= \int_W \mathbf{p}^W_{\mathcal{N}_i}(w)\frac{\mathbf{p}^W_{\mathcal{N}_i}(w)}{\mathbf{p}^W_{\mathcal{N}_0}(w)}\mathrm{d}w \\
&= \left(2\pi\sigma^2\right)^{-\frac{mt^*}{2}} \cdot \int_\Omega \exp\left(-\frac{\sum_{t\in[t^*]}\left((w_{i,t}+2\varepsilon)^2+\sum_{j'\neq i} w_{j',t}^2\right)-2\varepsilon^2 t^*}{2\sigma^2}\right)\mathrm{d}w \\
&= \exp\left(\frac{\varepsilon^2 t^*}{\sigma^2}\right).
\end{aligned}
$$

Therefore, combining the equations above, we get

$$
\begin{aligned}
\int_\Omega \frac{1}{m}\sum_{j\in[m]} \mathbf{p}_{\mathcal{N}_j}(\omega)\frac{\frac{1}{m}\sum_{i\in[m]}\mathbf{p}_{\mathcal{N}_i}(\omega)}{\mathbf{p}_{\mathcal{N}_0}(\omega)}\mathrm{d}\omega &= \frac{1}{m^2}\sum_{i,j\in[m]}\int_\Omega \mathbf{p}_{\mathcal{N}_i}(\omega)\frac{\mathbf{p}_{\mathcal{N}_j}(\omega)}{\mathbf{p}_{\mathcal{N}_0}(\omega)}\mathrm{d}\omega \\
&= \frac{m(m-1)+m\cdot\exp\left(\frac{\varepsilon^2 t^*}{\sigma^2}\right)}{m^2} = \frac{m-1+\exp\left(\frac{\varepsilon^2 t^*}{\sigma^2}\right)}{m},
\end{aligned}
$$

where the first equality follows from Fubini's theorem. This indicates that $D_{\mathtt{KL}}\left(\mathbf{P}_{\mathcal{N}_{\mathtt{mix}}},\mathbf{P}_{\mathcal{N}_0}\right) \le \log\frac{m-1+\exp\left(\frac{\varepsilon^2 t^*}{\sigma^2}\right)}{m}$. □

Let $t^* = \frac{c_0\log(m+1)}{\varepsilon^2}$, where $c_0 \le \sigma^2$ is a universal constant. We have the following lemma to bound $\mathbf{Pr}_{\mathcal{N}_0}\left[T \ge t^*\right]$. Here the randomness comes from the algorithm and environment when the input instance is $\mathcal{N}_0$.

Lemma C.4. *For any algorithm distinguishing* $\{\mathcal{N}_j\}_{j\in[m]\cup\{0\}}$ *with probability* 0.925*, we have* $\mathbf{Pr}_{\mathcal{N}_0}\left[T\ge t^*\right]\ge 0.1$.

Proof. Let $\mathcal{A}$ be an algorithm that can distinguish $\{\mathcal{N}_j\}_{j\in[m]\cup\{0\}}$ with probability 0.925. Let E be the event that $\mathcal{A}$ terminates within t^* rounds and gives answer $\mathcal{N}_0$. Recall that T is a random variable which represents the rounds that $\mathcal{A}$ runs. Assume $\mathbf{Pr}_{\mathcal{N}_0}\left[T\ge t^*\right] < 0.1$. Then we have $\mathbf{Pr}_{\mathcal{N}_0}\left[\overline{E}\right] < 0.075+0.1$ from the union bound. Combining Lemma C.2 and Lemma C.3, we get

$$
\mathbf{Pr}_{\mathcal{N}_{\mathtt{mix}}}\left[E\right] \ge \frac{m}{2\left(m-1+\exp\left(\frac{\varepsilon^2 t^*}{\sigma^2}\right)\right)} - \mathbf{Pr}_{\mathcal{N}_0}\left[\overline{E}\right] > \frac{m}{2\left(m-1+m+1\right)} - 0.1 - 0.075 \ge 0.075
$$

for every $m\ge 1$. This indicates the existence of some $j\in[m]$ such that $\mathbf{Pr}_{\mathcal{N}_j}\left[E\right] > 0.075$, which is in contradiction to the promised success probability of $\mathcal{A}$. Therefore $\mathcal{A}$ satisfies

$$
\mathbf{Pr}_{\mathcal{N}_0}\left[T\ge t^*\right]\ge 0.1.
$$

□

C.3. From Gaussian to Bernoulli

We then show a reduction from Gaussian arms to Bernoulli arms which implies lower bounds for instances $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$.

Given an input instance from $\{\mathcal{N}_j\}_{j\in[m]\cup\{0\}}$, we can map it to a corresponding instance among $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$ by the following rules.

In each round, if an arm receives a loss $\ell\in\mathbb{R}$, let

$$
\widehat{\ell} = \begin{cases} 0, & \text{if} \quad \ell < 0; \\ 1, & \text{if} \quad \ell \ge 0. \end{cases} \tag{15}
$$

Obviously, losses drawn from Gaussian distribution $\mathcal{N}(0,\sigma^2)$ are mapped to $\mathtt{Ber}\left(\frac{1}{2}\right)$ losses. For a biased Gaussian $\mathcal{N}\left(-\varepsilon,\sigma^2\right)$, as Figure 1 shows, it holds that

$$
\mathbf{Pr}\left[\widehat{\ell}<0\right] = \int_{-\infty}^{-\varepsilon}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x+\varepsilon)^2}{2\sigma^2}}\,\mathrm{d}x + \int_{-\varepsilon}^{0}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x+\varepsilon)^2}{2\sigma^2}}\,\mathrm{d}x
$$

$$= \frac{1}{2} + \int_{-\varepsilon}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x+\varepsilon)^2}{2\sigma^2}}\, \mathrm{d}x\,.$$

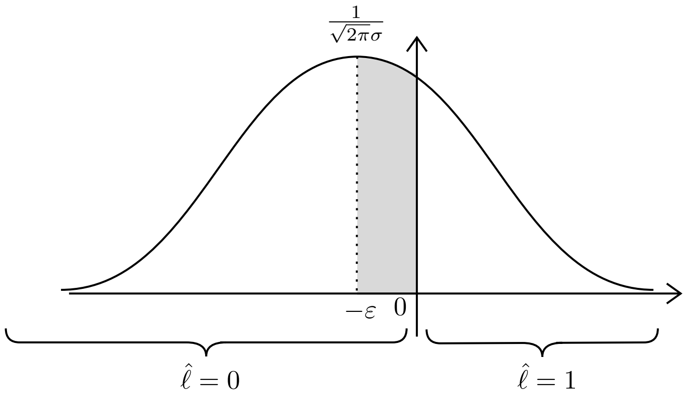

Figure 1. From Gaussian to Bernoulli

Let $f(\sigma) = \int_{-\varepsilon}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x+\varepsilon)^2}{2\sigma^2}}\, \mathrm{d}x$ denote the shadowed area in Figure 1. Note that f is continuous with regard to σ and

$$f(\sigma) \in \left(\frac{\varepsilon}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}}, \frac{\varepsilon}{\sqrt{2\pi}\sigma} \right).$$

Assume that $\varepsilon < \frac{1}{8}$. Therefore, there exists $\sigma_0 \in \left(\frac{1}{2\sqrt{2\pi}}, \frac{1}{\sqrt{2\pi}} \right)$ such that $f(\sigma_0) = \varepsilon$. Choose $\sigma = \sigma_0$. Then we map $\mathcal{N}(-\varepsilon, \sigma^2)$ to $\texttt{Ber}\left(\frac{1}{2} - \varepsilon\right)$ and transform the sample space from $\mathbb{R}^{m \times t^*}$ to $\{0,1\}^{m \times t^*}$.

Lemma C.5. *Let ε be a number in $\left(0, \frac{1}{8}\right)$. For any algorithm distinguishing $\left\{\mathscr{H}_j^{(m)}\right\}_{j \in [m] \cup \{0\}}$ with probability 0.925, we have $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq t^*\right] \geq 0.1$.*

Proof. Assume that there exists such an algorithm $\mathcal{A}$ with $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq t^*\right] < 0.1$. We then construct an algorithm $\mathcal{A}'$ to distinguish $\{\mathcal{N}_j\}_{j \in [m] \cup \{0\}}$.

The algorithm $\mathcal{A}'$ proceeds as follows: When $\mathcal{A}'$ receives a loss ℓ, it first calculates $\widehat{\ell}$ as Equation (15) and treats $\widehat{\ell}$ as the loss to apply $\mathcal{A}$. If $\mathcal{A}$ outputs $\mathscr{H}_j^{(m)}$, $\mathcal{A}'$ output $\mathcal{N}_j$. Therefore, $\mathcal{A}'$ also succeeds with probability 0.925 while satisfying $\mathbf{Pr}_{\mathcal{N}_0}\left[T \geq t^*\right] < 0.1$. This violates Lemma C.4. □

We remark that we cannot replace $\mathscr{H}_0^{(m)}$ by $\mathscr{H}_j^{(m)}$ for any $j \in [m]$ in Lemma C.5, since an "$\mathscr{H}_j^{(m)}$ favourite" algorithm exists for every $j \in [m]$. For example, an "$\mathscr{H}_1^{(m)}$ favourite" algorithm is as follows: one first sample the arms for $\frac{2\log\frac{1}{0.03}}{\varepsilon^2}$ rounds. If the empirical mean $\widehat{p}_1 < \frac{1}{2} - \frac{\varepsilon}{2}$, terminate and output $\mathscr{H}_1^{(m)}$. Otherwise apply an algorithm which can distinguish $\left\{\mathscr{H}_j^{(m)}\right\}_{j \in [m] \cup \{0\}}$ with probability 0.96. By the Hoeffding's inequality, the error probability in the first stage is at most 0.03. Therefore, this "$\mathscr{H}_1^{(m)}$ favourite" algorithm has success probability 0.925 and with high probability, it only needs to play $\frac{2\log\frac{1}{0.03}}{\varepsilon^2}$ rounds when the input instance is $\mathscr{H}_1^{(m)}$.

Then we are ready to prove Lemma C.1, which is a direct corollary of the following lemma.

Lemma C.6. *Let ε be a number in $\left(0, \frac{1}{8}\right)$ and assume $m \geq 2$. There exists a constant $c_1 > 0$ such that for any algorithm $\mathcal{A}$ which can output an ε-optimal arm on any instance among $\left\{\mathscr{H}_j^{(m)}\right\}_{j \in [m] \cup \{0\}}$ with probability at least 0.95, we have $\mathbf{E}_{\mathscr{H}_0^{(m)}}\left[T\right] \geq \frac{c_1 \log(m+1)}{\varepsilon^2}$.*

Proof. We first consider the case $c_0 \log(m+1) > 4 \log 40$ where c_0 is the universal constant in the definition of t^*. We reduce from the hypothesis testing lower bound in Lemma C.5. Assume $\mathcal{A}$ satisfying $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq \frac{c_0 \log(m+1)}{2\varepsilon^2}\right] < 0.1$. Then we construct an algorithm $\mathcal{A}'$ to distinguish $\left\{\mathscr{H}_j^{(m)}\right\}_{j \in [m] \cup \{0\}}$. Given an instance among $\left\{\mathscr{H}_j^{(m)}\right\}_{j \in [m] \cup \{0\}}$, we first apply $\mathcal{A}$ to get an output arm i. Then we sample $\frac{2 \log \frac{1}{0.025}}{\varepsilon^2}$ rounds and check whether the empirical mean $\widehat{p}_i \leq \frac{1}{2} - \frac{\varepsilon}{2}$. If so, output $\mathscr{H}_i^{(m)}$. Otherwise, output $\mathscr{H}_0^{(m)}$. The success probability of at least 0.925 is guaranteed by Hoeffding's inequality and the union bound.

According to our assumption, with probability larger than 0.9, $\mathcal{A}'$ terminates in $\frac{c_0 \log(m+1)}{2\varepsilon^2} + \frac{2 \log \frac{1}{0.025}}{\varepsilon^2} < \frac{c_0 \log(m+1)}{\varepsilon^2}$ rounds. This violates Lemma C.5.

Then we consider the case $c_0 \log(m+1) \leq 4 \log 40$; that is, when m is bounded by some constant. It then follows from Lemma D.3 that $\mathcal{A}$ satisfies $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq \frac{c_s}{\varepsilon^2}\right] \geq 0.1$ for a universal constant c_s when $m \geq 2$.

Then choosing $c_1 = \min\left\{\frac{c_0}{20}, \frac{c_s}{10 \log(m_0+1)}\right\}$ where $m_0 = \lfloor e^{\frac{4 \log 40}{c_0}} - 1 \rfloor$, we have $\mathbf{E}_{\mathscr{H}_0^{(m)}}[T] \geq \frac{c_1 \log(m+1)}{\varepsilon^2}$ for any algorithms that can output an ε-optimal arm on any instance among $\left\{\mathscr{H}_j^{(m)}\right\}_{j \in [m] \cup \{0\}}$ with probability at least 0.95 when $m \geq 2$. □

C.4. The Lower Bound for $\mathbf{m}$-BAI

Recall that in $\mathbf{m}$-`BAI`, the N arms are partitioned into K groups with size $m_1, m_2, \ldots, m_K$ respectively. Each pull of an arm results in an observation of all the arms in its group. Consider an $\mathbf{m}$-`BAI` instance $\mathscr{H}_0^{\mathbf{m}}$ which consists of all fair coins. Recall that we use $T^{(k)}$ to denote the number of rounds in which the pulled arm belongs to the k-th group.

We then prove the following lemma, which indicates the result of Theorem 1.3 directly.

Lemma C.7 (restate Lemma 4.2)**.** *Let ε be a number in $\left(0, \frac{1}{8}\right)$. For every $(\varepsilon, 0.05)$-PAC algorithm of $\mathbf{m}$-`BAI`, we have $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}\left[T^{(k)}\right] \geq \frac{c_1 \log(m_k+1)}{\varepsilon^2}$ for every $k \in [K]$ with $m_k \geq 2$ and $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}[T] \geq \sum_{k=1}^{K} \frac{c_1 \log(m_k+1)}{2\varepsilon^2}$ if the total number of arms $\sum_{k=1}^{K} m_k \geq 2$, where c_1 is the constant in Lemma C.6.*

Moreover, these lower bounds still hold even the algorithm can identify the ε-optimal arm with probability 0.95 only when the input arms have losses drawn from either `Ber` $\left(\frac{1}{2}\right)$ or `Ber` $\left(\frac{1}{2} - \varepsilon\right)$.

Proof. We only prove the latter case which is stronger. Let $\mathcal{H}$ be the set of all $\mathbf{m}$-`BAI` instances where the input arms have losses drawn from either `Ber` $\left(\frac{1}{2}\right)$ or `Ber` $\left(\frac{1}{2} - \varepsilon\right)$.

Let $\mathcal{A}$ be an algorithm that identifies the ε-optimal arm with probability 0.95 when the input instance is in $\mathcal{H}$. Assume $\mathcal{A}$ satisfies $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}\left[T^{(k)}\right] < \frac{c_1 \log(m_k+1)}{\varepsilon^2}$ for some $k \in [K]$. In the following, we construct an algorithm $\mathcal{A}'$ to find an ε-optimal arm given instances in $\left\{\mathscr{H}_j^{(m_k)}\right\}_{j \in [m] \cup \{0\}}$.

Given any (m_k)-`BAI` instance $\mathscr{H}^{(m_k)} \in \left\{\mathscr{H}_j^{(m_k)}\right\}_{j \in [m] \cup \{0\}}$, we construct an $\mathbf{m}$-`BAI` instance: set $\mathscr{H}^{(m_k)}$ to be the k-th group and all remaining arms are fair ones. Then we apply $\mathcal{A}$ on this instance. The output of $\mathcal{A}'$ is as follows:

$$\text{Output of } \mathcal{A}' = \begin{cases} \text{arm } j, & \text{if the output of } \mathcal{A} \text{ is arm } (k, j); \\ \text{an arbitrary arm}, & \text{otherwise.} \end{cases}$$

Clearly, the correct probability of $\mathcal{A}'$ is at least 0.95. However, $\mathcal{A}'$ satisfies $\mathbf{E}_{\mathscr{H}_0^{(m_k)}}[T] < \frac{c_1 \log(m_k+1)}{\varepsilon^2}$, which violates Lemma C.6.

Therefore, we have $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}\left[T^{(k)}\right] \geq \frac{c_1 \log(m_k+1)}{\varepsilon^2}$ for every $k \in [K]$ with $m_k \geq 2$ and thus have proved $\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}[T] \geq \sum_{k=1}^{K} \frac{c_1 \log(m_k+1)}{\varepsilon^2}$ as long as each $m_k \geq 2$. For those groups of size one, we can pair and merge them so that each group contains at least two arms (in case there are odd number of singleton groups, we merge the remaining one to any other groups). Notice that this operation only makes the problem easier (since one can observe more arms in each round) and only affects the lower bound by a factor of at most 2. Therefore, we still have

$$\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}}}\left[T\right] \geq \sum_{k=1}^{K} \frac{c_1 \log(m_k+1)}{2\varepsilon^2}.$$

□

D. Lower Bound for (m)-BAI with Bounded m

In this section, we will lower bound the number of pulls in $(\varepsilon, 0.05)$-PAC algorithms of (m)-BAI when m is bounded by a constant. To this end, we first prove a likelihood lemma in Appendix D.1.

D.1. Likelihood Lemma

Consider two instances $\mathscr{H}_a$ and $\mathscr{H}_b$ which only differ at one arm (without loss of generality, assume it is the first arm). In $\mathscr{H}_a$, $\ell(1)$ is drawn from $\mathtt{Ber}\left(\frac{1}{2}\right)$ and in $\mathscr{H}_b$, $\ell(1)$ is drawn from $\mathtt{Ber}\left(\frac{1}{2}-\varepsilon\right)$ where $\varepsilon \in \left(0, \frac{1}{2}\right)$ is a fixed number.

Let $\mathcal{A}$ be a PAC algorithm for $\mathtt{BAI}$. Let $K_j^t = \sum_{r=1}^{t} \ell^{(r)}(j)$ be the accumulative loss of arm j before the $(t+1)$-th round and abbreviate $K_j^{N_j}$ as K_j. Let A_j be the event that $N_j < \hat{t}$ for a fixed $\hat{t} \in \mathbb{N}$. Let C_j^a be the event that $\left\{\max_{1\leq t\leq \hat{t}} \left|K_j^t - \frac{1}{2}t\right| < \sqrt{\hat{t}\cdot c\varepsilon^2\hat{t}}\right\}$ and C_j^b be the event $\left\{\max_{1\leq t\leq \hat{t}} \left|K_j^t - \left(\frac{1}{2}-\varepsilon\right)t\right| < \sqrt{\hat{t}\cdot c\varepsilon^2\hat{t}}\right\}$ where c is a positive constant.

Lemma D.1 (Lemma 3 of (Mannor & Tsitsiklis, 2004))**.** *If $0 \leq x \leq \frac{1}{\sqrt{2}}$ and $y > 0$, then $(1-x)^y \geq e^{-dxy}$ where $d = 1.78$.*

Lemma D.2 (Likelihood Lemma)**.** *Let $S^a = A_1 \cap B \cap C_1^a$ and $S^b = A_1 \cap B \cap C_1^b$ where B is an arbitrary event. Then we have*

$$\mathbf{Pr}_{\mathscr{H}_b}\left[S^a\right] \geq e^{-8(1+\sqrt{c})\varepsilon^2\hat{t}}\,\mathbf{Pr}_{\mathscr{H}_a}\left[S^a\right] \tag{16}$$

and

$$\mathbf{Pr}_{\mathscr{H}_a}\left[S^b\right] \geq e^{-8(1+\sqrt{c})\varepsilon^2\hat{t}}\,\mathbf{Pr}_{\mathscr{H}_b}\left[S^b\right] \tag{17}$$

Proof. We first prove Equation (16). For each $\omega \in S^a$ (ω is a history of the algorithm, including the behavior of the algorithm and observed result in each round), we have

$$\begin{aligned}\frac{\mathbf{Pr}_{\mathscr{H}_b}\left[\omega\right]}{\mathbf{Pr}_{\mathscr{H}_a}\left[\omega\right]} &= \frac{\left(\frac{1}{2}-\varepsilon\right)^{K_1}\left(\frac{1}{2}+\varepsilon\right)^{N_1-K_1}}{\left(\frac{1}{2}\right)^{N_1}} = \left(1-2\varepsilon\right)^{K_1}\left(1+2\varepsilon\right)^{N_1-K_1}\\ &= \left(1-4\varepsilon^2\right)^{N_1-K_1}\left(1-2\varepsilon\right)^{2K_1-N_1} \geq \left(1-4\varepsilon^2\right)^{N_1}\left(1-2\varepsilon\right)^{2K_1-N_1}.\end{aligned}$$

From Lemma D.1 and the definition of S^a, we have

$$\left(1-4\varepsilon^2\right)^{N_1} \geq \left(1-4\varepsilon^2\right)^{\hat{t}} \geq e^{-8\varepsilon^2\hat{t}}$$

and

$$\left(1-2\varepsilon\right)^{2K_1-N_1} \geq \left(1-2\varepsilon\right)^{2\sqrt{\hat{t}\cdot c\varepsilon^2\hat{t}}} \geq e^{-8\sqrt{c}\varepsilon^2\hat{t}}.$$

Therefore

$$\frac{\mathbf{Pr}_{\mathscr{H}_b}\left[\omega\right]}{\mathbf{Pr}_{\mathscr{H}_a}\left[\omega\right]} \geq e^{-8(1+\sqrt{c})\varepsilon^2\hat{t}}$$

and thus

$$\mathbf{Pr}_{\mathscr{H}_b}\left[S^a\right] \geq \sum_{\omega\in S^a} \frac{\mathbf{Pr}_{\mathscr{H}_b}\left[\omega\right]}{\mathbf{Pr}_{\mathscr{H}_a}\left[\omega\right]}\cdot\mathbf{Pr}_{\mathscr{H}_a}\left[\omega\right] \geq e^{-8(1+\sqrt{c})\varepsilon^2\hat{t}}\,\mathbf{Pr}_{\mathscr{H}_a}\left[S^a\right].$$

The proof of Equation (17) is similar. □

D.2. Lower Bound for (m)-BAI with Constant m

Lemma D.3. *There exists a constant c_s such that for any algorithm $\mathcal{A}$ which can output an ε-optimal arm on any instance among $\left\{\mathscr{H}_j^{(m)}\right\}_{j\in[m]\cup\{0\}}$ with probability at least 0.95 when $m \geq 2$ and $c_0 \log(m+1) \leq 4\log 40$, we have $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq \frac{c_s}{\varepsilon^2}\right] \geq 0.1$.*

Proof. Note that there must exist $j \in [m]$ such that $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[\mathcal{A} \text{ output arm } j\right] \leq \frac{1}{m}$. Let B be the event that the algorithm output any arm except for arm j. Apply Lemma D.2 with $\hat{t} = \frac{\log 3}{100\varepsilon^2}$, $c = 100$, $\mathscr{H}_b = \mathscr{H}_j^{(m)}$ and $\mathscr{H}_a = \mathscr{H}_0^{(m)}$. Assume that $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq \hat{t}\right] < 0.1$. By the Kolmogorov's inequality, we have $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[\max_{1\leq t\leq \hat{t}} \left|K_j^t - \frac{1}{2}t\right| < \sqrt{\hat{t}\cdot c\varepsilon^2 \hat{t}}\right] \geq 1-0.25$. Therefore, we have $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[S^a\right] \geq 0.9 - \frac{1}{m} - 0.25 \geq 0.15$ by the union bound.

Then from Equation (16), we have

$$\mathbf{Pr}_{\mathscr{H}_j^{(m)}}\left[B\right] \geq e^{-8(1+\sqrt{c})\cdot\frac{\log 3}{100}} \cdot \mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[S^a\right] > 0.15 \cdot \frac{1}{3} = 0.05.$$

However, this is in contradiction with the success probability of $\mathcal{A}$. Therefore, letting $c_s = \frac{\log 3}{100}$, we have $\mathbf{Pr}_{\mathscr{H}_0^{(m)}}\left[T \geq \frac{c_s}{\varepsilon^2}\right] \geq 0.1$. □

E. Regret Lower Bounds

In this section we prove lower bounds for minimax regrets in various settings. All lower bounds for regrets in the section are based on the lower bounds for $\mathbf{m}$-BAI established in Appendix C.

E.1. Regret Lower Bound for $\mathbf{m}$-MAB

Let us fix $\mathbf{m} = (m_1, \ldots, m_K)$. We then derive a regret lower bound for $\mathbf{m}$-MAB and thus prove Theorem 1.4. Let T be the time horizon and c_1 be the constant in Lemma C.6. Consider a set of $\mathbf{m}$-BAI instances where each arm has losses drawn from either $\texttt{Ber}\left(\frac{1}{2}\right)$ or $\texttt{Ber}\left(\frac{1}{2}-\varepsilon\right)$ where $\varepsilon = \sqrt{\frac{c_1\sum_{k=1}^K \log(m_k+1)}{8T}}$. Denote this set by $\mathcal{H}$.

Lemma E.1. *For any algorithm $\mathcal{A}$ of $(m_1, \ldots, m_k)$-MAB, for any sufficiently large $T > 0$, there exists $\mathscr{H} \in \mathcal{H}$ such that the expected regret of $\mathcal{A}$ satisfies*

$$\mathbf{E}_{\mathscr{H}}\left[R(T)\right] \geq c' \cdot \sqrt{T \cdot \sum_{k=1}^K \log(m_k+1)}$$

where $c' > 0$ is a universal constant. Here the expectation is taken over the randomness of losses which are drawn from $\mathscr{H}$ independently in each round.

Proof. Assume $\mathcal{A}$ satisfies

$$\mathbf{E}_{\mathscr{H}}\left[R(T)\right] < \frac{\sqrt{T\cdot\frac{1}{2}\sum_{k=1}^K c_1\log(m_k+1)}}{5000}$$

for every $\mathscr{H} \in \mathcal{H}$ where c_1 is the constant in Lemma C.6. Lemma A.1 shows that $\mathcal{A}$ implies an algorithm to identify the ε-optimal arm for $\mathbf{m}$-BAI instances in $\mathcal{H}$ with probability 0.95 which terminates in $c_1 \cdot \frac{\sum_{k=1}^K \log(m_k+1)}{8\varepsilon^2}$ rounds. We can assume $\varepsilon < \frac{1}{8}$ since T is sufficiently large.

However, according to Lemma C.7, for any such algorithms, there exists some instances in $\mathcal{H}$ that need at least $\frac{c_1\sum_{k=1}^K \log(m_k+1)}{2\varepsilon^2}$ rounds. This violates Lemma A.1 and thus indicates a regret lower bound of $\Omega\left(\sqrt{T\cdot\sum_{k=1}^K \log(m_k+1)}\right)$. □

Theorem 1.4 is a direct corollary of Lemma E.1.

E.2. Regret Lower Bounds for Strongly Observable Graphs

Let $G = (V, E)$ be a strongly observable graph with a self-loop on each vertex. Let $N = |V|$. Assume that there exist K *disjoint* sets $S_1, \dots, S_K \subseteq V$ such that there is no edge between S_i and S_j for any $i \neq j$. For every $k \in [K]$, let $m_k = |S_k|$. Let $S = \bigcup_{k\in[K]} S_k$.

Proof of Theorem 1.6. We present a reduction from $\mathbf{m}$-$\texttt{MAB}$ to bandit with feedback graph G where $\mathbf{m} = (m_1, \dots, m_K)$. Let $\mathcal{A}$ be an algorithm for bandit with feedback graph G. Consider a set of instances where the loss of each arm is drawn from either $\texttt{Ber}\left(\frac{1}{2}\right)$ or $\texttt{Ber}\left(\frac{1}{2} - \varepsilon\right)$ where $\varepsilon = \sqrt{\frac{c_1 \sum_{k=1}^{K} \log(m_k+1)}{8T}}$ (here c_1 is the constant in Lemma C.6). Denote this set by $\mathcal{H}$. When we say the input of $\texttt{MAB}$ is an instance in $\mathcal{H}$, we mean that the loss sequence is drawn from this instance independently in each round.

Then we design an algorithm $\mathcal{A}'$ for $\mathbf{m}$-$\texttt{MAB}$ to deal with instances in $\mathcal{H}$ as follows. For an $\mathbf{m}$-$\texttt{MAB}$ instance $\mathscr{H}^{\mathbf{m}}$ in $\mathcal{H}$, we construct a bandit instance with feedback graph G: the losses of arms in S_k correspond to the losses of arms in the k-th group of $\mathscr{H}^{\mathbf{m}}$ in the $\mathbf{m}$-$\texttt{MAB}$ game and the losses of arms in $V \setminus S$ are always equal to 1.

The algorithm $\mathcal{A}'$ actually makes decisions according to $\mathcal{A}$. If $\mathcal{A}$ pulls an arm in S, $\mathcal{A}'$ pulls the corresponding arm in the $\mathbf{m}$-$\texttt{MAB}$ game. Otherwise, when $\mathcal{A}$ requests to pull an arm $A_t \in V \setminus S$, we replace this action by letting $\mathcal{A}'$ pull the first arm in each group once and then feed the information that A_t should have observed back to $\mathcal{A}$ (Note that all arms outside S have fixed loss 1). We force $\mathcal{A}'$ to terminate after pulling exactly T arms. Note that $\varepsilon \ll \frac{1}{K}$ since T is sufficiently large. If we use $R(T)$ and $R'(T)$ to denote the regret of $\mathcal{A}$ and $\mathcal{A}'$ respectively, then by our choice of ε, we have

$$\mathbf{E}\left[R(T)\right] \geq \mathbf{E}\left[R'(T)\right]$$

where the expectation is taken over the randomness of loss sequences specified above.

Lemma E.1 shows that there exists $\mathscr{H} \in \mathcal{H}$ such that

$$\mathbf{E}_{\mathscr{H}}\left[R'(T)\right] \geq c' \sqrt{T \cdot \sum_{k=1}^{K} \log(m_k + 1)}$$

Therefore, there exist some loss sequences on which $\mathcal{A}$ needs to suffer a regret of $\Omega\left(\sqrt{T \cdot \sum_{k=1}^{K} \log(m_k+1)}\right)$. □

Remark E.2. Although we assume each vertex has a self-loop in Theorem 1.6, it is easy to verify that this result also holds for strongly observable graphs which contain some vertices without self-loops, as long as we can find legal $\{S_k\}_{k\in[K]}$. For example, for the loopless clique, we can also apply Theorem 1.6 with $K = 1$ and $S_1 = V$. It gives a minimax regret lower bound of $\Omega\left(\sqrt{T \log N}\right)$, which matches the previous best upper bound in (Alon et al., 2015).

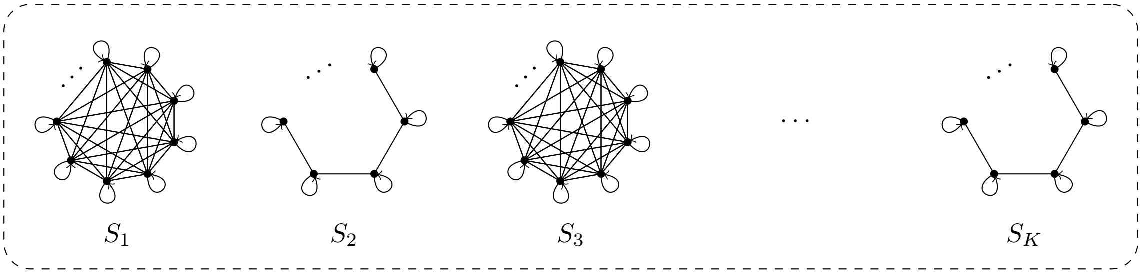

Figure 2. A Feedback Graph Example

Theorem 1.6 gives a general regret lower bound for bandit with arbitrary feedback graphs. Intuitively, it allows us to partition the graph and consider the hardness of each single part respectively.

For example, consider the graph shown in Figure 2: The feedback graph is the disjoint union of K_1 cliques and $K_2 = K - K_1$ cycles where each clique contains m_1 vertices and each cycle contains m_2 vertices. Note that the clique cover of this graph contains K_1 cliques of size m_1 and $\lceil \frac{K_2 m_2}{2} \rceil$ cliques of constant size. According to Theorem 3.1, our Algorithm 1 gives a regret upper bound of $O\left(\sqrt{T\left(K_1\log m_1+K_2m_2\right)}\right)$, which matches the lower bound given in Theorem 1.6. The previous best lower bound ((Alon et al., 2015)) on this feedback graph is $\Omega\left(\sqrt{\left(K_1+K_2m_2\right)T}\right)$. When K_1 and m_1 are large, our result wins by a factor of $\Theta\left(\sqrt{\log m_1}\right)$.

E.3. Regret Lower Bounds for Weakly Observable Graphs

Let $G=(V,E)$ be a weakly observable graph. Assume that V can be partitioned into K disjoint sets $V=V_1\cup V_2\cup\cdots\cup V_K$ and each $G[V_k]$ contains a t_k-packing independent set S_k such that every vertex in S_k does not have a self-loop. Assume there are no edges from V_j to S_i for any $i\neq j$. Let $m_k=|S_k|$ and $S=\bigcup_{k\in[K]}S_k$.

Without loss of generality, we assume in the following proof that each $m_k\geq 2$. When there exists some $m_k=1$, we can pair and merge them into new sets of size at least 2 (in case there are odd number of singleton sets, we merge the remaining one to any other sets). This merging process only affects the result by at most a constant factor. Let $\mathbf{m}=(m_1,\dots,m_K)$. Our proof idea is to embed a certain $\mathbf{m}'$-$\texttt{BAI}$ instance in G so that the lower bound follows from the lower bound of $\mathbf{m}'$-$\texttt{BAI}$.

Proof of Theorem 1.7. Let

$$\xi_k=\max\left\{c_1\log(m_k+1),\frac{c_2m_k}{t_k}\right\}$$

for every $k\in[K]$ where $c_1>0$ is the constant in Lemma C.7 and $c_2=\frac{c_1\log 3}{4}$. Assume there exists an algorithm $\mathcal{A}$ such that

$$R(T)<\frac{1}{2\cdot 1250^{\frac{2}{3}}}\left(\sum_{k=1}^{K}\xi_k\right)^{\frac{1}{3}}\cdot T^{\frac{2}{3}} \tag{18}$$

for every loss sequence. We will construct an $\mathbf{m}'$-$\texttt{BAI}$ game for some $\mathbf{m}'=(m_1',m_2',\dots,m_{K'}')$ and reduce this $\texttt{BAI}$ game to the bandit problem with feedback graph G. The vector $\mathbf{m}'$ is obtained from $\mathbf{m}$ in the following ways. For every $k\in[K]$, we distinguish between two cases:

- Case 1: if $c_1\log(m_k+1)\geq\frac{c_2m_k}{t_k}$, we let the arms in S_k form a group in the $\mathbf{m}'$-$\texttt{BAI}$ instance;
- Case 2: if $c_1\log(m_k+1)<\frac{c_2m_k}{t_k}$, we divide S_k into $\lfloor\frac{m_k}{2}\rfloor$ small sets, each with size at least two. Each small set becomes a group in the $\mathbf{m}'$-$\texttt{BAI}$ instance.

In other words, each group in the $\mathbf{m}'$-$\texttt{BAI}$ instance is either one of S_k (Case 1) or is a subset of a certain S_k (Case 2).

Given an $\mathbf{m}'$-$\texttt{BAI}$ instance and time horizon $T>0$, we now define the loss sequence for bandit with feedback graph G: the losses of arms in S in each round are sampled from the distribution of the corresponding arm in the $\mathbf{m}'$-$\texttt{MAB}$ instance independently, and the losses of arms in $V\setminus S$ are always equal to 1. We then design an algorithm $\mathcal{A}'$ for the $\mathbf{m}'$-$\texttt{BAI}$ game by simulating $\mathcal{A}$ on this graph bandit problem. If $\mathcal{A}$ pulls an arm in $V\setminus S$ and observes arms in S_k, we again consider two cases:

- Case 1: if $c_1\log(m_k+1)\geq\frac{c_2m_k}{t_k}$, we let $\mathcal{A}'$ pull an arbitrary arm in the corresponding group $\mathbf{m}'$-$\texttt{MAB}$ instance;
- Case 2: if $c_1\log(m_k+1)<\frac{c_2m_k}{t_k}$, for each arm in S_k that will be observed, $\mathcal{A}'$ pulls the corresponding arm in the $\mathbf{m}'$-$\texttt{MAB}$ instance once.

Otherwise if $\mathcal{A}$ pulls an arm in S, $\mathcal{A}'$ does nothing and just skips this round. Note that $\mathcal{A}'$ can always observe more information about the feedback of arms in S than $\mathcal{A}$. So $\mathcal{A}'$ can well simulate $\mathcal{A}$ just by feeding the information it observed to $\mathcal{A}$ and making decisions according to the behavior of $\mathcal{A}$ as described above.

Let T_i be the number of times that arm i has been pulled by $\mathcal{A}$. At the end of the game, $\mathcal{A}'$ samples an arm in V according to the distribution $\left(\frac{T_1}{T},\frac{T_2}{T},\dots,\frac{T_N}{T}\right)$. If the sampled arm is in $V\setminus S$, $\mathcal{A}'$ outputs a random arm. Otherwise $\mathcal{A}'$ outputs the sampled arm. Choose $\varepsilon=1250^{\frac{1}{3}}\left(\frac{\sum_{k=1}^{K}\xi_k}{T}\right)^{\frac{1}{3}}$. We can verify that $\mathcal{A}'$ is an $(\varepsilon,0.05)$-PAC algorithm through an argument similar to the one in our proof of Lemma A.1.

Let $T^{(k)}$ be the number of times that the arms in group k have been pulled by $\mathcal{A}'$ in the $\mathbf{m}'$-$\texttt{BAI}$ game. According to Lemma C.7, for each $k \in [K']$,

$$\mathbf{E}_{\mathscr{H}_0^{\mathbf{m}'}}\left[T^{(k)}\right] \geq \frac{c_1 \log(m_k' + 1)}{\varepsilon^2},$$

where $\mathscr{H}_0^{\mathbf{m}'}$ is the $\mathbf{m}'$-$\texttt{BAI}$ instance with all fair coins. Let $\mathscr{I}_0$ denote the graph bandit instance constructed from above rules based on $\mathscr{H}_0^{\mathbf{m}'}$. Recall that one pull of $\mathcal{A}$ corresponds to at most t_k pulls of $\mathcal{A}'$ in Case 2. Therefore, when the input is $\mathscr{I}_0$, $\mathcal{A}$ must pull the arms in $V_k \setminus S_k$ for at least $\frac{c_1 \lfloor \frac{m_k}{2} \rfloor \log 3}{t_k \varepsilon^2} \geq \frac{c_2 m_k}{t_k \varepsilon^2}$ times if k is in Case 2 and at least $\frac{c_1 \log(m_k+1)}{\varepsilon^2}$ times if k is in Case 1. In other words, $\mathcal{A}$ must pull the arms in $V_k \setminus S_k$ for at least $\frac{\xi_k}{\varepsilon^2}$ times for every $k \in [K]$. Plugging in our choice of ε, $\mathcal{A}$ needs to pull the arms in $V \setminus S$ for more than $\frac{1}{1250^{\frac{2}{3}}} \cdot \left(\sum_{k=1}^K \xi_k\right)^{\frac{1}{3}} T^{\frac{2}{3}}$ times in total on $\mathscr{I}_0$. These pulls contribute a regret of at least $\frac{1}{2 \cdot 1250^{\frac{2}{3}}} \left(\sum_{k=1}^K \xi_k\right)^{\frac{1}{3}} \cdot T^{\frac{2}{3}}$, which contradicts the assumption in Equation (18).

Therefore, there exists some loss sequences such that $\mathcal{A}$ satisfies

$$R(T) = \Omega\left(T^{\frac{2}{3}} \cdot \left(\sum_{k=1}^K \max\left\{\log m_k, \frac{m_k}{t_k}\right\}\right)^{\frac{1}{3}}\right).$$

□

Theorem 1.7 confirms a conjecture in (He & Zhang, 2023). It can also generalize the previous lower bound for weakly observable graphs $\Omega\left(T^{\frac{2}{3}}\left(\log|S|, \frac{|S|}{t}\right)^{\frac{1}{3}}\right)$ in (Chen et al., 2021) by applying Theorem 1.7 with $K = 1$ and $V_1 = V$ where $S \subseteq V$ is a t-packing independent set of G. As consequences, Theorem 1.7 provides tight lower bounds for several feedback graphs. For example, when G is the disjoint union of K complete bipartite graphs of size $m_1, m_2, \ldots, m_K$ respectively, it implies a lower bound of $\Omega\left(\left(\sum_{k\in[K]} \log m_k\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$, which matches the upper bound in (He & Zhang, 2023).