Optimal Arm Elimination Algorithms for Combinatorial Bandits

Yuxiao Wen
New York University

Yanjun Han
New York University

Zhengyuan Zhou
New York University

Abstract

Combinatorial bandits extend the classical bandit framework to settings where the learner selects multiple arms in each round, motivated by applications such as online recommendation and assortment optimization. While extensions of upper confidence bound (UCB) algorithms arise naturally in this context, adapting arm elimination methods has proved more challenging. We introduce a novel elimination scheme that partitions arms into three categories (confirmed, active, and eliminated), and incorporates explicit exploration to update these sets. We demonstrate the efficacy of our algorithm in two settings: the combinatorial multi-armed bandit with general graph feedback, and the combinatorial linear contextual bandit. Matching lower bounds are also provided. In both cases, our approach achieves near-optimal regret, whereas UCB-based methods can provably fail due to insufficient explicit exploration.

1 INTRODUCTION

Combinatorial bandits, where the learner picks a subset of actions rather than a single action, have a wide range of applications spanning online recommendations (Wang et al., 2017; Qin et al., 2014), assortment optimization (Han et al., 2021), crowd-sourcing (Lin et al., 2014), webpage optimization (Liu and Li, 2021), online routing (György et al., 2007; Audibert et al., 2014), etc. A common feedback model in this setting is the *semi-bandit feedback*, in which the learner observes the rewards of the chosen actions only. Achieving optimal performance under this limited feedback requires carefully balancing exploration and exploitation: the learner must search for promising new actions while simultaneously committing to actions already known to perform well.

In stochastic bandits, where rewards follow fixed but unknown distributions, two prominent algorithmic frameworks achieve the optimal exploration-exploitation tradeoff: the *upper confidence bound* (UCB) algorithm (Auer et al., 2002) and the *arm elimination* algorithm (Even-Dar et al., 2006). Both approaches maintain confidence intervals for the rewards of each action, but proceed differently: UCB selects the action with the largest upper bound, whereas arm elimination retains only the "active" actions whose confidence intervals overlap with that of the action with the highest lower bound. Consequently, UCB relies on *implicit* exploration, while arm elimination can perform more *explicit* exploration over the active set. For both multi-armed and linear bandits, these algorithms are known to achieve near-optimal regret (Auer and Ortner, 2010; Abbasi-Yadkori et al., 2011).

For more complex bandit problems, the lack of explicit exploration can make UCB less effective than arm elimination methods. One example is the multi-armed bandit with graph feedback, where exploration must be carefully guided by the graph structure. Another is the linear contextual bandit with a finite set of time-varying contexts, where dependencies across rewards undermine the validity of confidence bounds. In both cases, arm elimination provides a remedy. For bandits with graph feedback, arm elimination explicitly leverages the graph structure to explore the active set of actions (Han et al., 2024; Wen et al., 2024). For linear contextual bandits, a master algorithm built on top of an arm elimination subroutine can effectively handle the dependence and achieve the optimal regret (Chu et al., 2011).

On the other hand, while the UCB algorithm admits a relatively direct extension to combinatorial bandits (Kveton et al., 2015; Combes et al., 2015), arm elimination methods in this setting remain largely underexplored. The main difficulty lies in balancing exploration of new arms with exploitation of known good

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Table 1: Regret Bounds in Different Settings

	Graph feedback	Linear context
Our results	$\widetilde{\boldsymbol{\Theta}}(\frac{\boldsymbol{\alpha+S}}{\boldsymbol{\Delta_*}}\log(\boldsymbol{T}))$	$\widetilde{\boldsymbol{\Theta}}(\boldsymbol{\sqrt{dST}})$
Previous best	$\widetilde{O}(\frac{K\log(T)}{\Delta_*}+\frac{S}{\Delta_*^4})^\dagger$	$\widetilde{O}(d\sqrt{ST})^\ddagger$

arms within the same round. In this paper, we study two instances of combinatorial bandits:

1. Combinatorial multi-armed bandits with general graph feedback, where chosen actions can reveal the rewards of other arms according to a given graph structure.

2. Combinatorial linear contextual bandits, where the learner selects a subset of contexts under a linear reward model.

For both problems, existing UCB and arm elimination methods fail to achieve optimal regret, either due to the lack of explicit exploration (UCB) or the challenges posed by the combinatorial structure (arm elimination). To address this, we propose a general arm elimination framework for combinatorial bandits and show that it achieves optimal regret in both settings.

1.1 Our Contributions

The main contributions of this paper are as follows:

1. We propose a general arm elimination framework for combinatorial bandits that partitions actions into three categories (confirmed, active, and eliminated). At each round, the combinatorial budget is allocated between confirmed and active actions, with explicit exploration directed toward the active set.

2. For combinatorial bandits with general graph feedback, we design an arm elimination algorithm based on this framework that achieves near-optimal regret. Specifically, with time horizon T and combinatorial budget S, and for feedback graphs with independence number α, the algorithm simultaneously attains the optimal worst-case regret $\widetilde{\Theta}(\sqrt{\alpha ST}+S\sqrt{T})$ and the optimal gap-dependent regret $\widetilde{\Theta}(\frac{\alpha+S}{\Delta_*}\log(T))$.

3. For combinatorial linear contextual bandits with dimension d and a finite number of contexts, we show that combining our arm elimination method with the master algorithm of Auer (2002) yields the optimal regret $\widetilde{\Theta}(\sqrt{dST})$.

For the sake of clarity, we summarize our regret guarantees in Table 1.[1] This work provides both tight upper and lower bounds for each setting. We also remark that this work proves a tight regret upper bound $\widetilde{O}(S\sqrt{T})$ for the commonly used UCB algorithm under full-information combinatorial bandits in Appendix A.6, which may be of independent interest.

1.2 Related Work

Combinatorial Bandits. The minimax regret for combinatorial bandits under graph feedback is recently shown to be $\widetilde{\Theta}(S\sqrt{T}+\sqrt{\alpha ST})$ even in the adversarial setting (Wen, 2025). In the stochastic regime, Kveton et al. (2015) provides an instance-dependent regret bound $O\Big(\frac{KS\log(T)}{\Delta_*}\Big)$ for a UCB-type algorithm and shows that this rate is tight when the feasible decision set is some constrained subset of the set $\binom{[K]}{S}$. Under the unconstrained decision set $\binom{[K]}{S}$, the bound is later improved to $O\Big(\frac{K\sqrt{S}\log(T)}{\Delta_*}+\frac{KS^3}{\Delta_*^2}\Big)$ by Combes et al. (2015). Their result shaves a factor $\sqrt{S}$ but has an extra term that dominates for small Δ_*. Both of their algorithms are described by the UCB in Algorithm 2 with slightly different input parameter L. Later, Wang and Chen (2018) closes the gap in the logarithmic term. They show that a variant of Thompson Sampling achieves regret $O\Big(\frac{K\log(T)}{\Delta_*}+\frac{S}{\Delta_*^4}\Big)$. However, for small Δ_* this extra term again dominates and leads to a loose upper bound. Under the full-information feedback when the rewards of all actions are revealed, no instance-dependent bound is shown to the best of our knowledge.

Combinatorial Linear Contextual Bandits. Another line of research concerns the setting where contextual information is available to the learner to aid decision-making. A widely adopted (stochastic) reward model assumes that the expected reward is linear in the observed context, whereas no assumption is imposed on how the context is generated. This formulation finds many industrial applications, such as recommender systems (Qin et al., 2014) and assortment management (Han et al., 2021). In terms of regret, Qin et al. (2014) presents a variant of `LinUCB` (Li et al., 2010) that achieves $\widetilde{O}(d\sqrt{ST})$ regret. In comparison, in the classical bandit setting with $S=1$, the near-optimal regret is known to be $\widetilde{O}(\sqrt{dT})$ and is achieved by an elimination-based algorithm (Chu et al., 2011) paired with a master algorithm to handle dependence. The best known result for UCB-type algorithms is $\widetilde{O}(d\sqrt{T})$ by Abbasi-Yadkori et al. (2011).

[1] For previous results in Table 1, see † in Wang and Chen (2018) and ‡ in Qin et al. (2014).

Elimination-based Bandit Algorithms. The idea of arm elimination in bandit algorithms is popularized by Even-Dar et al. (2006). Compared to the arguably more natural UCB-type algorithms, elimination has a demonstrated value in a range of bandit problems, including MAB with graph feedback and contextual MAB. In the former, elimination allows the learner to force exploration as the algorithm runs to achieve the optimal trade-off under general graphs (Han et al., 2024). In the linear contextual case, Auer (2002) develops a hierarchical elimination scheme to address a dependence issue in the reward observations; this scheme has been widely adopted in various settings with contextual information to achieve tight regret (Chu et al., 2011; Han et al., 2024; Wen, 2025). Nonetheless, despite its power in the bandit problems, it remains unclear how to perform elimination, or whether it is possible at all, under the combinatorial setting where the learner selects and compares to $S > 1$ optimal arms.

Top-S Arm Identification. Another relevant line of research is top-S arm identification. In this problem, the learner pulls one arm at a time, and the objective is to identify the S arms with maximal expected rewards, either under a fixed budget or up to a fixed confidence. For best-arm identification where $S = 1$, there is rich literature studying both the complexity under a fixed confidence (Audibert and Bubeck, 2010; Garivier and Kaufmann, 2016) and the probability of error under a fixed budget (Carpentier and Locatelli, 2016; Kato et al., 2022; Komiyama et al., 2022). For general $S \geq 1$, Chen et al. (2014) tackles this problem using UCB under a constrained subset, i.e. the identified S arms must belong to a certain subset. Several works address the unconstrained problem via elimination-based algorithms (Bubeck et al., 2013; Chen et al., 2017; Rejwan and Mansour, 2020; Zhou and Tian, 2022). In particular, Bubeck et al. (2013); Chen et al. (2017); Rejwan and Mansour (2020) propose to sequentially accept "good" arms according to the separated confidence widths, which shares the spirit of the confirmation set in our algorithm. Nonetheless, this arm identification problem does not face the core exploration-exploitation tradeoff *within the same round* in combinatorial bandits.

1.3 Notations

For a positive integer $n \in \mathbb{N}$, let $[n] = \{1, 2, \ldots, n\}$. For a directed graph $G = (V, E)$, let $N_{\text{out}}(a) = \{b \in V : (a, b) \in E\}$ denote the set of out-neighbors of node $a \in V$ and $N_{\text{out}}(U) = \cup_{a \in U} N_{\text{out}}(a)$ denote the set of out-neighbors of a subset U. For a vector $x \in \mathbb{R}^d$ and a positive semi-definite (PSD) matrix $A \in \mathbb{R}^{d \times d}$, the matrix norm is defined by $\|x\|_A = \sqrt{x^\top A x}$. We use $\widetilde{O}$ and $\widetilde{\Theta}$ to denote the usual asymptotic meanings of O and Θ, respectively, but suppress less important poly-logarithmic factors.

2 COMBINATORIAL BANDITS WITH GRAPH FEEDBACK

2.1 Problem Formulation

This section introduces the problem of combinatorial bandits with general graph feedback. At each time t over a horizon of length T, the learner selects a *decision* that is a subset of arms $V_t \subseteq [K]$ such that $|V_t| = S$ for a fixed $S \geq 1$. There is a known directed feedback graph $G = ([K], E)$ over the arms. The learner observes the individual rewards $\{r_{t,a} : a \in N_{\text{out}}(V_t)\}$ and receives the total reward $r_{t,V_t} = \sum_{a \in V_t} r_{t,a}$. For the scope of this work, we assume G contains all self-loops, i.e. $a \in N_{\text{out}}(a)$. We assume $r_{t,a} \in [0, 1]$ and for each arm a, the rewards $\{r_{t,a}\}_{t \in [T]}$ are i.i.d. with a time-invariant mean μ_a.

Without loss of generality (WLOG), we assume the means satisfy $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. For any policy π, the regret measures the expected loss compared to the hindsight optimal decision $[S]$:

$$\mathsf{R}(\pi) = \sum_{t=1}^{T} \bigg(\sum_{i=1}^{S} \mu_i - \sum_{a \in V_t} \mu_a \bigg).$$

2.2 An Elimination-based Algorithm

We start by giving a high-level intuition of the arm elimination algorithm under the classical multi-armed bandit setting $S = 1$ (Even-Dar et al., 2006). In this case, the algorithm maintains an *active set* of "probably good" arms $\mathcal{A}_{\text{act}} \subseteq [K]$ and a minimum count $N = \min_{a \in \mathcal{A}_{\text{act}}} n_{t,a}$ for the active arms, where $n_{t,a}$ denotes the number of observations for arm a up to time t. It then uniformly explores every arm in $\mathcal{A}_{\text{act}}$ with a small $n_{t,a}$ and update N accordingly. By standard concentration results (see Lemma 1 below), the algorithm recognizes a uniform confidence width for each μ_a and eliminates any arm a from the set $\mathcal{A}_{\text{act}}$ that is provably suboptimal based on the confidence widths.

Lemma 1. *Fix any $\delta \in (0, 1)$. With probability at least $1 - \delta$, we have for every arm a at every time t,*

$$|\bar{r}_{t,a} - \mu_a| \leq \sqrt{\log(2KT/\delta)/n_{t,a}} =: w(n_{t,a})$$

where $\bar{r}_{t,a}$ is the empirical mean and $n_{t,a}$ the number of observations at time t.

However, it is challenging to extend this elimination scheme to $S \geq 2$, as it is unclear how to set the elimi-

nation benchmark and how to explore among the uneliminated arms. When $S = 1$, we only need to decide if an arm a is possibly optimal or not, so the algorithm simply eliminates based on the empirically best arm. When $S \geq 2$, an arm a can be worse than μ_1 but still possible to be the S-th optimal, rendering this choice impractical. Additionally, even if we manage to exclude arms that are provably worse than the S-th optimal arm for any decision $V_t \subseteq \mathcal{A}_{\text{act}}$, there is no guarantee that the regret $\sum_{i=1}^{S} \mu_i - \sum_{a \in V_t} \mu_a$ will be small, because each μ_a may be close to μ_S but much worse than μ_{S-1} or μ_1.

To address the aforementioned challenges, we introduce the *confirmed set* $\mathcal{A}_{\text{con}}$ in our Algorithm 1. Suppose the confidence width w_t for every active arm $a \in \mathcal{A}_{\text{act}}$ is uniformly positive at time t. The idea of confirmation is to use this width w_t to identify a subset $\mathcal{A}_{\text{con}} \subseteq [S]$. Each arm $i \in \mathcal{A}_{\text{con}}$ has a mean $\mu_i \gg \mu_S + w_t$ that is much larger than the mean of the S-th optimal arm. Therefore, the learner must have $\mathcal{A}_{\text{con}} \subseteq V_t$ to avoid incurring an instantaneous regret unbounded by w_t. In addition, for every unconfirmed optimal arm $i' \in [S] \backslash \mathcal{A}_{\text{con}}$, it holds that $\mu_{i'} \lesssim \mu_S + w_t$, allowing the learner to include *any* active arm $a \in \mathcal{A}_{\text{act}}$ in the decision V_t and suffer a bounded regret, even if the learner only manages to bound $\mu_S - \mu_a$.

Note that by partitioning the uneliminated arms into $\mathcal{A}_{\text{con}}$ and $\mathcal{A}_{\text{act}} \backslash \mathcal{A}_{\text{con}}$, we are effectively identifying an *exploration-exploitation trade-off* based on the confidence width w_t at every time t. The confirmed arms $\mathcal{A}_{\text{con}}$ are "too good to leave out" and lead to an exploitation of size $|\mathcal{A}_{\text{con}}|$. Meanwhile, we use the remaining $S - |\mathcal{A}_{\text{con}}|$ budget in our decision V_t to explore the remaining active arms and further eliminate the suboptimal ones. For the graph feedback, we adopt the exploration strategy in Han et al. (2024) and successively pull the arm with the largest out-degree (Line 6 of Algorithm 1). This exploration budget $S - |\mathcal{A}_{\text{con}}|$, notably, changes over time.

The overall algorithm is given in Algorithm 1. We remark that Algorithm 1 recovers the standard arm elimination algorithm in the multi-armed bandits ($S = 1$), as $\mathcal{A}_{\text{con}} \equiv \varnothing$ by definition. Recall that $w(n) \coloneqq \sqrt{\log(2KT/\delta)/n}$ is given in Lemma 1.

Algorithm 1: Arm Confirmation and Elimination (ACE)

1 **Input:** failure probability $\delta \in (0, 1)$.
2 **Initialize:** Confirmed set $\mathcal{A}_{\text{con}} \leftarrow \varnothing$, active set $\mathcal{A}_{\text{act}} \leftarrow [K]$, and minimum count $N \leftarrow 0$.
3 **for** $t = 1$ **to** T **do**
4 Let $\mathcal{A}_0 \leftarrow \{a \in \mathcal{A}_{\text{act}} : n_{t-1,a} = N\}$.
5 **for** $j = 1$ **to** $S - |\mathcal{A}_{\text{con}}|$ **do**
6 Let $a_{t,j} \in \mathcal{A}_0$ be any arm with the largest out-degree in $G|_{\mathcal{A}_0}$.
7 Update $\mathcal{A}_0 \leftarrow \mathcal{A}_0 \backslash N_{\text{out}}(a_{t,j})$.
8 Assemble $V_t \leftarrow \mathcal{A}_{\text{con}} \cup \{a_{t,j}\}_{j=1}^{S-|\mathcal{A}_{\text{con}}|}$.
9 Observe feedback $\{r_{t,a} : a \in N_{\text{out}}(V_t)\}$.
10 Update $(\bar{r}_{t,a}, n_{t,a})$ as the average reward and observation count of arm a by the end of time t.
11 **if** $\min_{a \in \mathcal{A}_{\text{act}}} n_{t,a} > N$ **then**
12 Update count $N \leftarrow \min_{a \in \mathcal{A}_{\text{act}}} n_{t,a}$.
13 Let $\bar{r}_{t,(S)}$ be the S-th empirically best reward in the union set $\mathcal{A}_{\text{con}} \cup \mathcal{A}_{\text{act}}$.
14 Update the confirmed set $\mathcal{A}_{\text{con}}$ to be:

$$\mathcal{A}_{\text{con}} \cup \big\{a \in \mathcal{A}_{\text{act}} : \bar{r}_{t,a} > \bar{r}_{t,(S)} + 4w(N)\big\}.$$

Then the active set $\mathcal{A}_{\text{act}}$ to be

$$\big\{a \in \mathcal{A}_{\text{act}} \backslash \mathcal{A}_{\text{con}} : \bar{r}_{t,a} \geq \bar{r}_{t,(S)} - 2w(N)\big\}.$$

2.3 Algorithmic Properties

We now present several key properties of Algorithm 1. For the sake of clarity, we use N^t, $\mathcal{A}_{\text{act}}^t$, and $\mathcal{A}_{\text{con}}^t$ to denote the minimum count of each active arm's observations N, the active set $\mathcal{A}_{\text{act}}$, and the confirmed set $\mathcal{A}_{\text{con}}$ in Algorithm 1 by the end of time t. Let $\bar{r}_{t,(j)}$ denote the j-th empirically best reward in $\mathcal{A}_{\text{con}}^t \cup \mathcal{A}_{\text{act}}^t$, and $a_{t,(j)}$ denote the corresponding arm.[2] For $i \in [S]$ and $a \in [K]$, let $\Delta_{a,i} = \mu_i - \mu_a$ denote the reward gap between arm a and the i-th optimal arm. WLOG, suppose $\Delta_* \triangleq \Delta_{S+1,S} > 0$, which serves as a margin and characterizes the difficulty of distinguishing suboptimal and optimal arms.[3]

Conditioned on the validity of the confidence width in Lemma 1, the following lemma lists three important properties for Algorithm 1.

Lemma 2. *Suppose the event in Lemma 1 holds. Then for each time $t \in [T]$, the following events hold by the end of t:*

(A) The optimal arms remain uneliminated, i.e. $[S] \subseteq \mathcal{A}_{\text{act}}^t \cup \mathcal{A}_{\text{con}}^t$.

(B) $\mathcal{A}_{\text{con}}^t \subseteq [S-1]$.

(C) Let $i_{*,t} = \min \mathcal{A}_{\text{act}}^t$. For every $a \in \mathcal{A}_{\text{act}}^t$, it holds that $\Delta_{a,i_{*,t}} \leq 8w(N^t)$.

[2] For notational simplicity, here we use (j) to denote the j-th largest as opposed to the j-th smallest in the conventional notations for order statistics.

[3] If $\mu_S = \mu_{S+1} = \cdots = \mu_{S+k}$, it is straightforward to extend our analysis to the definition $\Delta_* = \Delta_{S,S+k+1}$.

The first property states that the optimal arms $[S]$ are never eliminated as the algorithm runs. This serves as the basis of the elimination scheme, since otherwise the algorithm would suffer a linear regret.

Second, any confirmed arm in $\mathcal{A}_{\text{con}}^t$ always belongs to the top $S-1$ optimal arms. Recall that $\mathcal{A}_{\text{con}}^{t-1} \subseteq V_t$ for all $t \in [T]$. This optimality ensures that pulling the confirmed arms incurs no instantaneous regret, since now $\mathcal{A}_{\text{con}}^{t-1} \subseteq [S] \cap V_t$ for every time t. Therefore,

$$\begin{aligned}\sum_{i=1}^{S} \mu_i - \sum_{a \in V_t} \mu_a &= \sum_{i \in [S] \setminus \mathcal{A}_{\text{con}}^{t-1}} \mu_i - \sum_{a \in V_t \setminus \mathcal{A}_{\text{con}}^{t-1}} \mu_a \\ &\le \sum_{a \in V_t \setminus \mathcal{A}_{\text{con}}^{t-1}} \left(\mu_{i_{*,t}} - \mu_a\right) \qquad (1)\end{aligned}$$

which follows from that $i_{*,t} = \min \mathcal{A}_{\text{act}}^t$ and so $\mu_i \le \mu_{i_{*,t}}$ for every $i \in [S] \setminus \mathcal{A}_{\text{act}}^t$. The inequality (1) serves as the basis to derive regret bounds.

The last claim concerns the optimal active arm $i_{*,t}$ at each time t in hindsight. Note that by the second claim, $\mathcal{A}_{\text{con}}^t \subseteq [S-1]$ and so $i_{*,t} \in [S]$. The intuition is as follows. In the elimination step (last line) in Algorithm 1, the benchmark is the S-th empirically best reward $\bar{r}_{t,(S)}$. As a result, this choice guarantees that $\Delta_{a,S} = O(w(N^t))$ for any active arm $a \in \mathcal{A}_{\text{act}}^t$. Additionally, the update criterion of the confirmed set implies that $\mu_{i_{*,t}} - \mu_S = O(w(N^t))$ because $i_{*,t} \notin \mathcal{A}_{\text{con}}^t$. Together they give the claimed bound on $\Delta_{a,i_{*,t}}$ and therefore an upper bound of (1).

Intuitively, at time t, the learning process only distinguishes the eliminated and the active arms up to the confidence width $w(N^t)$, where N^t lower bounds the number of observations for all active arms. The instantaneous regret at t is inevitably $O(w(N^t))$. Therefore, at the current time, we may treat any arm $i \in [S-1]$ and S indistinguishably if $\mu_i - \mu_S = O(w(N^t))$. For any arm $i \in [S-1]$ beyond this width, it becomes necessary to identify and include i in the final decision V_t, which motivates the choice of the confirmed set $\mathcal{A}_{\text{con}}^t$.

2.4 Regret Bounds

In this section, we present the regret guarantees for Algorithm 1 and show their tightness. Specifically, our algorithm simultaneously achieves a logarithmic gap-dependent bound $O(\log(KT)(\alpha \log^2 K + S)/\Delta_*)$ and a worst-case regret bound $\widetilde{O}(\sqrt{\alpha ST} + S\sqrt{T})$.

Theorem 2.1 (Instance-dependent regret)**.** *Fix any $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\mathsf{R}(\text{Alg } 1) = O\bigg(\log(2KT/\delta)\frac{\alpha \log^2 K + S}{\Delta_*}\bigg).$$

Specializing to the semi-bandit feedback ($\alpha = K$) and full-information feedback ($\alpha = 1$), we obtain improved or new regret bounds for combinatorial bandits. Note that we do not have remainder terms (such as $O(\Delta_*^{-4})$ in Wang and Chen (2018)) in Theorem 2.1.

Corollary 1. *Under the semi-bandit feedback, we have*

$$\mathsf{R}(\text{Alg } 1) = O\bigg(\log(2KT/\delta)\frac{K \log^2 K}{\Delta_*}\bigg).$$

Under the full-information feedback, we have

$$\mathsf{R}(\text{Alg } 1) = O\bigg(\log(2KT/\delta)\frac{\log^2 K + S}{\Delta_*}\bigg).$$

We also prove matching lower bounds to show that the regret in Theorem 2.1 is near-optimal for all algorithms. Note that in Theorem 2.2, the missing $\log T$ factor for small α is not an artifact of our analysis; in fact, under a full-information feedback ($\alpha = 1$), one can attain a *constant* regret using a simple greedy algorithm (Degenne and Perchet, 2016).

Theorem 2.2. *Let $\mathsf{R}_\nu(\pi)$ denote the regret of policy π under bandit environment ν. Fix any policy π:*

(L1) Suppose the policy satisfies $\max_\nu \mathsf{R}_\nu(\pi) \le CT^p$, $\Delta_* \in (T^{-(1-p)}, \frac{1}{4}]$, and $\alpha \ge 2S$, and for some constants $C > 0$ and $p \in [0,1)$.

Then $\max_\nu \mathsf{R}_\nu(\pi) = \Omega\Big(\log(T\Delta_*) \min\Big\{\frac{\alpha}{\Delta_*}, \frac{1}{\Delta_*^2}\Big\}\Big)$.

(L2) Suppose $K \ge 2S$ and $\Delta_* \le \frac{1}{2}$.

Then $\max_\nu \mathsf{R}_\nu(\pi) = \Omega\Big(\min\Big\{\frac{S}{\Delta_*}, \Delta_* ST\Big\}\Big)$.

We can also derive a minimax regret bound for Algorithm 1, which nearly matches the lower bound $\Omega(\sqrt{\alpha ST} + S\sqrt{T})$ in Theorem 1.3 of Wen (2025).

Theorem 2.3 (Minimax regret)**.** *Fix any $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\mathsf{R}(\text{Alg } 1) = O\Big(\log^2 K \sqrt{\log(2TK/\delta)}\Big(\sqrt{\alpha ST} + S\sqrt{T}\Big)\Big)$$

2.5 Suboptimality of UCB

As discussed in Section 1, a large volume of bandit literature adopts UCB-type algorithms to develop optimal regret guarantees in different settings. In combinatorial bandits, Kveton et al. (2015) proposes a natural UCB algorithm, called `CombUCB1`, that achieves the optimal $\widetilde{O}(\sqrt{KST})$ regret under the semi-bandit feedback. However, this UCB algorithm, described in Algorithm 2, is provably suboptimal under general graph feedback.

Algorithm 2: Combinatorial UCB

Input: width parameter $L > 0$, and failure probability $\delta \in (0, 1)$.
Let $(\bar{r}_{t,a}, n_{t,a})$ be the empirical reward and the observation count of arm a at the end of t.
Let $\mathcal{A} = \{v \subseteq [K] : |v| = S\}$ be the set of feasible decisions.
for $t = 1$ **to** T **do**
  Select $V_t \leftarrow \arg\max_{v \in \mathcal{A}} \sum_{a \in v} \bar{r}_{t-1,a} + \frac{L}{\sqrt{n_{t-1,a}}}$.
  Observe feedback $\{r_{t,a} : a \in N_{\text{out}}(V_t)\}$.
  Update $(\bar{r}_{t,a}, n_{t,a})$ accordingly.

The parameter $L > 0$ is any factor such that with probability at least $1 - \delta$,

$$|\bar{r}_{t,a} - \mu_a| \le \frac{L}{\sqrt{n_{t,a}}}$$

for every arm $a \in [K]$ at every time $t \in [T]$. For instance, Lemma 1 gives one possible option $L = \sqrt{\log(2KT/\delta)}$. It maintains such a UCB for each arm $a \in [K]$ and, at each time t, selects the combination of S arms that maximizes the total UCB.

The high-level reason behind the suboptimality shown in Theorem 2.4 is that, without forced exploration, UCB essentially uses all S arms for either exploitation or exploration *simultaneously* at each time. In contrast, our elimination scheme in Algorithm 1 crucially relies on an exploration-exploitation separation among the S arms at each time.

Theorem 2.4. *Fix any (S, α, K, T) with $S\alpha \le K$ and $\alpha > 1$. There is a problem instance under which*

$$\mathsf{R}(\text{Alg } 2) = \Omega(LS\sqrt{\alpha T}).$$

Theorem 2.4 shows that, in the regime when $S\alpha \le K$, UCB achieves a suboptimal rate compared to $\widetilde{O}(S\sqrt{T} + \sqrt{\alpha ST})$ in Theorem 2.3. When $S\alpha \ge K$, UCB achieves the tight rate $\widetilde{\Theta}(\sqrt{KST})$ (Kveton et al., 2015). Interestingly, this sub-optimality ratio $\min\{S\alpha, K\}/(S + \alpha)$ scales as $\Theta(1)$ at two extremes when $\alpha = K$ and $\alpha = 1$.[4]

[4]To our knowledge, for the full-information feedback ($\alpha = 1$) we do not find an existing regret bound for UCB. For completeness, we show that Algorithm 2 attains $\widetilde{O}(S\sqrt{T})$ regret in this case in Appendix A.6.

3 COMBINATORIAL LINEAR CONTEXTUAL BANDITS

3.1 Problem Formulation

In combinatorial linear contextual bandits, each arm $a \in [K]$ is associated with a context vector $x_{t,a} \in \mathbb{R}^d$ at time t. We consider the linear model where $r_{t,a} = \theta_*^\top x_{t,a} + \varepsilon_{t,a} \in [-1, 1]$ for an unknown parameter $\theta_* \in \mathbb{R}^d$ and mean-zero noise $\varepsilon_{t,a}$. We assume $\|x_{t,a}\|_2 \le 1$ for all $t \in [T]$ and $a \in [K]$, and $\|\theta_*\|_2 \le 1$.

Given the contextual information, we consider a stronger *dynamic* regret where the hindsight optimal oracle knows θ_* and chooses decisions conditioned on the context:

$$\mathsf{R}(\pi) = \sum_{t=1}^{T} \left(\max_{\substack{V_{*,t} \subseteq [K] \\ |V_{*,t}| = S}} \sum_{i \in V_{*,t}} \theta_*^\top x_{t,i} - \sum_{a \in V_t} \theta_*^\top x_{t,a} \right)$$

where the decision V_t is selected by the learner's policy π at time t.

3.2 A Hierarchical Elimination Algorithm

In this section, we introduce an algorithm that builds on a hierarchical elimination idea. This idea stems from Auer (2002) and has been widely applied in the contextual bandit literature thereafter (Chu et al., 2011; Han et al., 2024; Wen et al., 2025). Different from the UCB analysis that bounds $\|\theta_* - \widehat{\theta}_t\|$ uniformly (Abbasi-Yadkori et al., 2011), it bounds the estimation error $|\theta_*^\top x_{t,a} - \widehat{\theta}_t^\top x_{t,a}|$ along the *realized* direction $x_{t,a}$, where $\widehat{\theta}_t$ is the estimated parameter.

To develop bounds only along certain directions, our Algorithm 3 crucially relies on the following property: At time t, it partitions the historical data into $H = \log(\sqrt{ST})$ stages. At stage $h = 1, \ldots, H$, the algorithm builds the final decision V_t by only looking at information that is independent of the reward observations belonging to the current h-th stage (while it is allowed to use rewards belonging to other stages). If V_t is not fully built at the current stage, the algorithm then uses those reward observations to eliminate suboptimal arms and proceed to the next stage $h+1$. The elimination step from the previous stage $h - 1$ guarantees that, even if the algorithm's move at stage h is independent from the rewards at stage h, the suboptimality of the selected arms remains bounded.

The sole purpose of this hierarchical elimination is to guarantee that, when focusing on each stage $h \in [H]$, the reward observations are mutually independent conditioned on the contexts. Then Lemma 3 gives a direction-specific estimation error bound. We refer to

Algorithm 3: Hierarchical Arm Confirmation and Elimination (H-ACE)

Initialize: Set $H \leftarrow \lceil \log_2(\sqrt{ST}) \rceil$, $\Phi_1^{(h)} = \varnothing$ for $h \in [H]$.
for $t = 1$ **to** T **do**
  Observe the contexts $\{x_{t,a}\}_{a\in[K]}$.
  Initialize $A_1 \leftarrow [K]$.
  Initialize the decision $V_t \leftarrow \varnothing$.
  for *stage* $h = 1$ **to** H **do**
    Use observations in $\Phi_t^{(h)}$ and Algorithm 4 to compute reward $\widehat{r}_{t,a}^{(h)}$ and width $w_{t,a}^{(h)}$ for $a \in A_h$.
    (1) Let $U_1 = \{a \in A_h \backslash V_t : w_{t,a}^{(h)} > 2^{-h}\}$.
    if $|U_1| \le S - |V_t|$ **then**
      Set $U_t^{(h,1)} \leftarrow U_1$.
    else
      Set any $U_t^{(h,1)} \subseteq U_1$ with $|U_t^{(h,1)}| = S - |V_t|$.
    Add $V_t \leftarrow V_t \cup U_t^{(h,1)}$.
    Let $a_{1,(S-|V_t|)}$ be the $(S-|V_t|)$-th arm maximizing $\widehat{r}_{t,a}^{(h)} + w_{t,a}^{(h)}$ in $A_h \backslash V_t$.
    (2) Let $U_2 = \left\{a \in A_h \backslash V_t : \widehat{r}_{t,a}^{(h)} > \widehat{r}_{t,a_{1,(S-|V_t|)}}^{(h)} + 4 \cdot 2^{-h}\right\}$
    if $|U_2| \le S - |V_t|$ **then**
      Set $U_t^{(h,2)} \leftarrow U_2$.
    else
      Set any $U_t^{(h,2)} \subseteq U_2$ with $|U_t^{(h,2)}| = S - |V_t|$.
    Add $V_t \leftarrow V_t \cup U_t^{(h,2)}$.
    Let $a_{2,(S-|V_t|)}$ be the $(S-|V_t|)$-th arm maximizing $\widehat{r}_{t,a}^{(h)} + w_{t,a}^{(h)}$ in $A_h \backslash V_t$.
    (3) Find active set A_{h+1} for next stage: $\left\{a \in A_h \backslash V_t : \widehat{r}_{t,a}^{(h)} \ge \widehat{r}_{t,a_{2,(S-|V_t|)}}^{(h)} - 2 \cdot 2^{-h}\right\}$
    Update $\Phi_{t+1}^{(h)} \leftarrow \Phi_t^{(h)} \oplus (U_t^{(h,1)} \cup U_t^{(h,2)})$.
  if $|V_t| < S$ **then**
    Note $w_{t,a}^{(H)} \le \frac{1}{\sqrt{ST}}$ for all $a \in A_H \backslash V_t$.
    Fill in V_t with any arms in $A_H \backslash V_t$.
  Select the decision V_t of S arms.
  Observe feedback $\{r_{t,a} : a \in V_t\}$.

Lemma 1 of Chu et al. (2011) for a proof of this result. Remark 1 explains why UCB-type algorithms (such as `LinUCB` (Li et al., 2010; Abbasi-Yadkori et al., 2011)) fail this assumption.

Lemma 3. *Let $\beta = \sqrt{\log(2KT)}$ and $\lambda = 1$. In Algorithm 4, suppose $\{r_{s,a} : s \in [t-1], a \in \Phi_t(s)\}$ are conditionally independent given contexts $\{x_{s,a} : s \in [t-1], a \in \Phi_t(s)\}$. Then with probability at least $1 - T^{-2}$, it holds that*

$$\left|\widehat{\theta}_t^\top x_{t,a} - \theta_*^\top x_{t,a}\right| \le 2(\beta+\lambda)\|x_{t,a}\|_{A_t^{-1}}.$$

for every $a \in [K]$.

Algorithm 4: Base Algorithm

Input: A sequence of sets of selected arms $\Phi_t = (\Phi_t(s))_{s<t}$ with $\Phi_t(s) \subseteq [K]$ and $|\Phi_t(s)| \le S$, an active set $A \subseteq [K]$.
Set $\beta \leftarrow \sqrt{\log(2KT)}$ and $\lambda \leftarrow 1$.
$A_t \leftarrow \lambda I + \sum_{s=1}^{t-1} \sum_{a\in\Phi_t(s)} x_{s,a} x_{s,a}^\top$;
$z_t \leftarrow \sum_{s=1}^{t-1} \sum_{a\in\Phi_t(s)} r_{s,a} x_{s,a}$;
Compute estimators $\widehat{\theta}_t \leftarrow A_t^{-1} z_t$.
for $a \in A$ **do**
  Compute estimated reward $\widehat{r}_{t,a} \leftarrow \widehat{\theta}_t^\top x_{t,a}$.
  Compute width $w_{t,a} \leftarrow 2(\lambda+\beta)\|x_{t,a}\|_{A_t^{-1}}$.

Remark 1. *We briefly remark on why such conditional independence fails without partitioning: At time $t+1$, we would instead condition on all contexts $\{x_{\tau,a} : \tau \le t, a \in v_\tau\}$. It is true that $\{r_{t,a} : a \in V_t\}$ are independent from others when conditioned on $\{x_{t,a} : a \in V_t\}$. However, the algorithm has used all previous observations $\{r_{\tau,a} : \tau < t, a \in v_\tau\}$ to come up with the decision V_t. When conditioned on $\{x_{t,a} : a \in V_t\}$ which reveals information about V_t, all of the previous reward observations become mutually dependent.*

The hierarchical elimination algorithm is described in Algorithm 3. At each stage $h \in [H]$, given the conditional independence of the reward observations belonging to this stage, we invoke Algorithm 4 to solve the standard ridge regression and derive an estimated parameter $\widehat{\theta}_t$. Thanks to the conditional independence, the estimation error $|\widehat{\theta}_t^\top x_{t,a} - \theta_*^\top x_{t,a}|$ is bounded by Lemma 3. This estimator is then used to compute reward estimators $\widehat{r}_{t,a}$ and their confidence widths $w_{t,a}$.

Construction of V_t. Given the valid reward estimators and their widths, at time t, Algorithm 3 constructs the decision set $V_t \subseteq [K]$ as follows. At the beginning of t, it initializes an empty set V_t, then in step (1) it continuously adds *underexplored* arms U_1 (i.e., with a large width) to V_t as it goes through the hierarchical elimination with H stages. If $|V_t| < S$ and the decision is not yet filled, every remaining arm has a uniformly small width $w_{t,a}^{(h)} \le 2^{-h}$, and in step (2) we identify the *confirmed* (optimal) arms U_2 using this width and add them to the decision V_t. If V_t remains unfilled, we proceed to the elimination step (3) and use the uniform width 2^{-h} to eliminate suboptimal arms. In the end, if there is still space in V_t, we simply add any

remaining arms; note that the remaining arms after H rounds of elimination have a sufficiently small width and hence contribute to a small regret.

3.3 Algorithmic Properties

Similar to the previous section, we establish the key properties of Algorithm 3 that justify the elimination steps and enable our regret analysis. However, there are several important distinctions from the properties in Section 2.3 for noncontextual combinatorial bandits.

First, due to the presence of contextual information and the dynamic nature of the regret, we no longer maintain persistent active and confirmed sets $(\mathcal{A}_{\text{act}}, \mathcal{A}_{\text{con}})$ as in Algorithm 1. Instead, the set of candidate arms $[K]$ is re-eliminated from scratch at every round t after the contexts $\{x_{t,a}\}_{a\in[K]}$ are observed. This substantially complicates the algorithm and the regret analysis, since the confidence width is no longer decreasing over time for any fixed arm a.

Second, to apply the hierarchical elimination framework introduced in Section 3.2 and ensure the conditional independence stated in Lemma 3, we must construct the decision V_t across H stages within each round t. The remaining capacity in V_t, i.e., $S-|V_t|$, is not fixed in advance and evolves indeterministically as the stage $h \in [H]$ proceeds. As a result, every round involves H elimination steps, each with a dynamically varying target size. In particular, to handle this variability, the benchmark arm in each elimination step is defined as the arm with the $(S-|V_t|)$-th largest UCB.

For the sake of clarity, let $V_t^{(h,g)}$ denote the decision set at the end of step (g) at stage $h \in [H]$, for $g = 1, 2, 3$. Namely, the initial set is $V_t^{(0,3)} = \varnothing$, and recursively $V_t^{(h,1)} = V_t^{(h-1,3)} \cup U_t^{(h,1)}$ and $V_t^{(h,2)} = V_t^{(h,3)} = V_t^{(h,1)} \cup U_t^{(h,2)}$. Let $V_{*,t}$ be the top S arms at time t. Note that $V_{*,t}\backslash V_t^{(h,g)}$ are the left-out optimal arms by the time of step (g) at stage h, for $g = 1, 2, 3$. To address the distinctions above, we have the following properties:

Lemma 4. *Suppose the event in Lemma 3 holds for every $t \in [T]$ and stage $h \in [H]$. For each time t, the following events holds for each stage $h \in [H]$:*

(A) The top $S - |V_t^{(h-1,3)}|$ arms of $V_{*,t}\backslash V_t^{(h-1,3)}$ are in A_h, i.e. remain uneliminated.

(B) Confirmed arms are optimal: $U_t^{(h,2)} \subseteq V_{*,t}\backslash V_t^{(h,1)}$.

(C) Let $i_{*,t}^{(h)} = \arg\min\{\theta_*^\top x_{t,a} : a \in A_h\}$ denote the optimal active arm. For every remaining $a \in A_h$, it holds that

$$\theta_*^\top x_{t,i_{*,t}^{(h)}} - \theta_*^\top x_{t,a} \le 16 \cdot 2^{-h}.$$

Lemma 4 summarizes the contextual counterparts of the properties in Lemma 2. Claim (A) states that Algorithm 3 keeps a sufficient number of the optimal arms from $V_{*,t}$ in the active set after each elimination step (3). It guarantees that we never end up with an unfilled $|V_t| < S$ but no arm left in the active set $A_h\backslash V_t$. In addition, it guarantees that the top $S - |V_t^{(h-1,3)}|$ unselected optimal arms are not eliminated, when we proceed from stage $h-1$ to h.

The second claim (B) plays the same role as in Lemma 2, in the sense that the algorithm incurs no instantaneous regret by including the confirmed arms $U_t^{(h,2)}$. Recall that an optimal arm is confirmed if one would suffer an unbounded regret by not including it, so confirmation balances exploration and exploitation.

Finally, (C) bounds the gap between any active arm and the optimal active arm $i_{*,t}^{(h)}$ in A_h at stage h. By claim (A), we always have $i_{*,t}^{(h)} \in V_{*,t}$ at any stage h where the decision V_t is not fully constructed. Crucially, this enables us to select *any* active arm $a \in A_h$ in step (1) of Algorithm 3 solely by looking at its width $w_{t,a}^{(h)}$ and still manage to obtain a bounded regret.

3.4 Regret Bounds

We conclude this section with the near-optimal regret guarantee for Algorithm 3 and a matching lower bound.

Theorem 3.1. *Algorithm 3 (with Algorithm 4 as a subroutine) achieves*

$$\mathsf{R}(\text{Alg } 3) = O\Big(\log(ST)\sqrt{\log(2KT)}(\sqrt{dST} + dS)\Big).$$

Therefore, Algorithm 3 improves on the existing result from $\widetilde{O}(d\sqrt{ST})$ to $\widetilde{O}(\sqrt{dST})$ when K is finite. The following lower bound complements our upper bound.

Theorem 3.2. *Suppose $T \ge \max\{4dS, \frac{d^3}{S}\}$ and $K \ge 2S$. For any policy π, it holds that*

$$\max_{\theta_*, \{x_{t,a}\}} \mathsf{R}(\pi) = \Omega(\sqrt{dST})$$

where the maximum is taken over all problem instances as described in Section 3.1.

4 CONSTRAINED DECISION

The combinatorial bandit literature also considers the constrained setting where the learner can only choose

decision $V_t \in \mathcal{A}_0 \subsetneq \binom{[K]}{S}$. The optimal regret crucially depends on the structure of the specific subset $\mathcal{A}_0$. The sub-optimality gap Δ_* in this case denotes the reward gap between the optimal decision and the second optimal decision. For example, Kveton et al. (2015) shows that $O\Big(\frac{KS\log(T)}{\Delta_*}\Big)$ is optimal for a specific constrained subset $\mathcal{A}_0$. Under general graph feedback, Wen (2025) shows that the minimax regret is $\widetilde{\Theta}(\min\{S\sqrt{\alpha T}, \sqrt{KST}\})$ when $\mathcal{A}_0$ is allowed to be any subset, as opposed to $\widetilde{\Theta}(S\sqrt{T}+\sqrt{\alpha ST})$ when $\mathcal{A}_0 = \binom{[K]}{S}$.

It is straightforward to extend our Algorithm 1 and its analysis to a general $\mathcal{A}_0$:

Theorem 4.1. *There exists a policy π such that, for any constrained decision subset $\mathcal{A}_0 \subseteq \binom{[K]}{S}$, it achieves*

$$\mathsf{R}(\pi) = O\bigg(\log(TK)\frac{S^2\kappa}{\Delta_*}\bigg)$$

where κ is a quantity the depends jointly on the feedback graph G and the subset structure $\mathcal{A}_0$:

$$\kappa \coloneqq \max_{\mathcal{A}'\subseteq\mathcal{A}_0} \min\{n \geq 1 : \exists V_1, \ldots, V_n \in \mathcal{A}' \text{ such that } \cup_{V\in\mathcal{A}'} V \subseteq \cup_{j=1}^n N_{\mathrm{out}}(V_j)\}. \tag{2}$$

In particular, there is a subset $\mathcal{A}_0$ under which

$$\min_{\pi}\max_{\nu} \mathsf{R}_\nu(\pi) = \Omega\bigg(\log(T)\frac{S^2\kappa}{\Delta_*}\bigg)$$

where the maximum is taken over all bandit environments ν.

At a high level, κ denotes the minimum number of decisions needed to observe any subgraph of G. Theorem 4.1 recovers the result of Kveton et al. (2015) where $\kappa = \Theta(K/S)$.

ACKNOWLEDGMENT

This work is supported by NSF (CCF-2312205, ECCS-2419564), ONR-13983263 and 2027 New York University Center for Global Economy and Business grant.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

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

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of machine learning research*, 3(Nov):397–422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.

Auer, P. and Ortner, R. (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

Bubeck, S., Wang, T., and Viswanathan, N. (2013). Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR.

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR.

Chen, L., Li, J., and Qiao, M. (2017). Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Chvatal, V. (1979). A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. (2015). Combinatorial bandits revisited. *Advances in neural information processing systems*, 28.

Degenne, R. and Perchet, V. (2016). Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595. PMLR.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6).

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. (2007). The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10).

Han, Y., Wang, Y., and Chen, X. (2021). Adversarial combinatorial bandits with general non-linear reward functions. In *International Conference on Machine Learning*, pages 4030–4039. PMLR.

Han, Y., Weissman, T., and Zhou, Z. (2024). Optimal no-regret learning in repeated first-price auctions. *Operations Research*.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*.

Kato, M., Ariu, K., Imaizumi, M., Uehara, M., Nomura, M., and Qin, C. (2022). Best arm identification with a fixed budget under a small gap. *stat*, 1050:11.

Komiyama, J., Tsuchiya, T., and Honda, J. (2022). Minimax optimal algorithms for fixed-budget best arm identification. *Advances in Neural Information Processing Systems*, 35:10393–10404.

Koolen, W. M., Warmuth, M. K., Kivinen, J., et al. (2010). Hedging structured concepts. In *COLT*, pages 93–105. Citeseer.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Lin, T., Abrahao, B., Kleinberg, R., Lui, J., and Chen, W. (2014). Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pages 901–909. PMLR.

Liu, Y. and Li, L. (2021). A map of bandits for e-commerce. *KDD Marble Workshop*.

Qin, L., Chen, S., and Zhu, X. (2014). Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM.

Rejwan, I. and Mansour, Y. (2020). Top-k combinatorial bandits with full-bandit feedback. In *Algorithmic Learning Theory*, pages 752–776. PMLR.

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR.

Wang, Y., Ouyang, H., Wang, C., Chen, J., Asamov, T., and Chang, Y. (2017). Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Wen, Y. (2025). Adversarial combinatorial semi-bandits with graph feedback. *International Conference on Machine Learning*.

Wen, Y., Han, Y., and Zhou, Z. (2024). Stochastic contextual bandits with graph feedback: from independence number to mas number. *Advances in Neural Information Processing Systems*.

Wen, Y., Han, Y., and Zhou, Z. (2025). Joint value estimation and bidding in repeated first-price auctions. *arXiv preprint arXiv:2502.17292*.

Zhou, R. and Tian, C. (2022). Approximate top-m arm identification with heterogeneous reward variances. In *International Conference on Artificial Intelligence and Statistics*, pages 7483–7504. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes: Section 2.1 and 3.1.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes: Section 2.3 and 3.3.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including

external libraries. [Not Applicable: there is no numerical experiment.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes: assumptions are included in theorem statements.]
   (b) Complete proofs of all theoretical results. [Yes: included in appendices]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable: no empirical result.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable: no empirical result.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable: no figure.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable: no empirical result.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable: we did not use such asset.]
   (b) The license information of the assets, if applicable. [Not Applicable: we did not use such asset.]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable: we did not use such asset.]
   (d) Information about consent from data providers/curators. [Not Applicable: we did not use such asset.]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable: we did not use such asset.]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable: we did not involve crowdsourcing nor human subjects.]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable: we did not involve crowdsourcing nor human subjects.]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable: we did not involve crowdsourcing nor human subjects.]

Optimal Arm Elimination Algorithms for Combinatorial Bandits: Supplementary Materials

A COMBINATORIAL BANDITS WITH GRAPH FEEDBACK

This section provides proofs for the results in Section 2.

A.1 Proof of Lemma 2

For the readers' convenience, we restate the lemma below:

Lemma 5 (Restatement of Lemma 2)**.** *Suppose the event in Lemma 1 holds. Then for each time $t \in [T]$, the following events hold for Algorithm 1 by the end of time t:*

(A) ***(Optimal are not eliminated)*** *The optimal arms remain uneliminated, i.e. $[S] \subseteq \mathcal{A}_{\text{act}}^t \cup \mathcal{A}_{\text{con}}^t$.*

(B) ***(Confirmed are optimal)*** *$\mathcal{A}_{\text{con}}^t \subseteq [S-1]$.*

(C) ***(Active gaps are bounded)*** *Let $i_{*,t} = \min \mathcal{A}_{\text{act}}^t$. For every $a \in \mathcal{A}_{\text{act}}^t$, it holds that $\Delta_{a,i_{*,t}} \le 8w(N^t)$.*

We prove each claim separately in the remaining section. Recall that the confirmed arms $\mathcal{A}_{\text{con}}^t$ are always pulled, so their observation counts are at least N^t, and by Lemma 1, the same confidence width $w(N^t)$ also holds with high probability for confirmed arm $i \in \mathcal{A}_{\text{con}}^t$.

Lemma 6. *Suppose the event in Lemma 1 holds. Recall that $\bar{r}_{t,(j)}$ denotes the j-th empirically best reward in $\mathcal{A}_{\text{con}} \cup \mathcal{A}_{\text{act}}$ at the end of time t, and $a_{t,(j)}$ denotes the corresponding arm. Then the followings hold by the end each time $t \in [T]$:*

(A) The optimal arms remain uneliminated, i.e. $[S] \subseteq \mathcal{A}_{\text{con}}^t \cup \mathcal{A}_{\text{act}}^t$.

(A') For every index $i \in [S]$, we have $\bar{r}_{t,i} \ge \bar{r}_{t,(i)} - 2w(N^t)$.

Proof. We prove this result via an inductive argument on time t. Note event (A') trivially holds at time $t = 1$, as $w(N^1) \ge 1$, and (A) holds at $t = 0$, namely before the algorithm kicks off.

First, we show that event (A') holds at time t conditioned on event (A) at time $t-1$. Fix any $i \in [S]$ and consider the empirically top i arms $\{a_{t,(j)}\}_{j \le i}$. By the pigeonhole principle, there always exists an index $j \le i$ such that $a_{t,(j)} \ge i$. Then

$$\bar{r}_{t,i} \ge \mu_i - w(N^t) \ge \mu_{a_{t,(j)}} - w(N^t) \ge \bar{r}_{t,(j)} - 2w(N^t) \ge \bar{r}_{t,(i)} - 2w(N^t).$$

It remains to verify event (A) for any time t, conditioned on event (A') at t and event (A) at $t-1$. If the elimination in Line 14 of Algorithm 1 does not occur at the end of time t, event (A) trivially holds for t. Suppose instead the elimination occurs at the end of time t. For every $i \in [S]$, event (A') implies $\bar{r}_{t,i} \ge \bar{r}_{t,(S)} - 2w(N^t)$, so i is not eliminated and event (A) holds for time t. This concludes the induction. □

Lemma 7. *Suppose the event in Lemma 1 holds. At every time $t \in [T]$, it holds that $\mathcal{A}_{\text{con}}^t \subseteq [S-1]$, i.e. every confirmed arm is one of the top $S-1$ optimal arms.*

Proof. It suffices to verify that when any active arm i_0 is added to the confirmed set $\mathcal{A}_{\text{con}}^{\tau_0}$ at time τ_0, it always holds that $i_0 \in [S-1]$. By the elimination step in Algorithm 1, we have $\bar{r}_{\tau_0,i_0} > \bar{r}_{\tau_0,(S)} + 4w(N^{\tau_0})$. Since $\bar{r}_{\tau_0,(S)}$

is the S-th empirically optimal reward in $\mathcal{A}_{\text{con}}^{\tau_0} \cup \mathcal{A}_{\text{act}}^{\tau_0}$ and $[S] \subseteq \mathcal{A}_{\text{con}}^{\tau_0} \cup \mathcal{A}_{\text{act}}^{\tau_0}$ by Lemma 6, there is some $j \in [S]$ with $\bar{r}_{\tau_0,j} \le \bar{r}_{\tau_0,(S)}$. Then

$$\mu_{i_0} \ge \bar{r}_{\tau_0,i_0} - w(N^{\tau_0}) \ge \bar{r}_{\tau_0,(S)} + 3w(N^{\tau_0}) \ge \bar{r}_{\tau_0,j} + 3w(N^{\tau_0}) \ge \mu_j + 2w(N^{\tau_0}) > \mu_j$$

which implies $i_0 \in [S-1]$. □

Lemma 8. *Suppose the event in Lemma 1 holds. Let $i_{*,t} = \min \mathcal{A}_{\text{act}}^t$. For every arm $a \in \mathcal{A}_{\text{act}}^t$, we have $\Delta_{a,i_{*,t}} \le 8w(N^t)$.*

Proof. Lemma 6 and 7 imply that $\mathcal{A}_{\text{act}}^t \cap [S] \neq \varnothing$ for all time t, so $i_{*,t} \in [S]$. Since $i_{*,t} \notin \mathcal{A}_{\text{con}}^t$ by definition, we have

$$\bar{r}_{t,i_{*,t}} \le \bar{r}_{t,(S)} + 4w(N^t).$$

Then for any active arm $a \in \mathcal{A}_{\text{act}}^t$, by the elimination criterion in Algorithm 1 and Lemma 1,

$$\mu_a \ge \bar{r}_{t,a} - w(N^t) \ge \bar{r}_{t,(S)} - 3w(N^t) \ge \bar{r}_{t,i_{*,t}} - 7w(N^t) \ge \mu_{i_{*,t}} - 8w(N^t).$$

□

A.2 Instance-dependent Regret Upper Bound

Theorem A.1 (Restatement of Theorem 2.1)**.** *Fix any $\delta \in (0,1)$. With probability at least $1-\delta$, Algorithm 1 achieves regret*

$$\mathsf{R}(\mathsf{Alg}\ 1) = O\bigg(\log(2KT/\delta)\frac{\alpha \log^2 K + S}{\Delta_*}\bigg).$$

Proof. WLOG, suppose the high-probability event in Lemma 1 holds. Recall the definition $i_{*,t} = \min \mathcal{A}_{\text{act}}^t$ from Lemma 2. An important observation is that, for any suboptimal arm $a \in [K]\backslash[S]$, if the minimum count satisfies $N^t > 64\log(2KT/\delta)/\Delta_{a,i_{*,t}}^2$ at time t, then

$$\begin{aligned} \bar{r}_{t,(S)} - 2w(N^t) &\overset{\text{(a)}}{\ge} \bar{r}_{t,i_{*,t}} - 6w(N^t) \ge \mu_{i_{*,t}} - 7w(N^t) \\ &\overset{\text{(b)}}{>} \mu_a + w(N^t) \ge \bar{r}_{t,a} \end{aligned} \tag{3}$$

where (a) follows from the fact that $i_t \notin \mathcal{A}_{\text{con}}^t$ and the elimination criterion in Line 14 of Algorithm 1, and (b) applies $\mu_{i_{*,t}} - \mu_a = \Delta_{a,i_{*,t}} > 8w(N^t)$. Hence a has already been eliminated from $\mathcal{A}_{\text{con}}^t \cup \mathcal{A}_{\text{act}}^t$.

To bound the cumulative regret, we will partition the horizon into a few sub-horizons, and over each sub-horizon we have a fixed optimal active arm $i_{*,t}$. Specifically, denote two non-repeating sequences of indices as follows: $\pi_1 = 1$ and $t_1 = 0$ since $1 \in \mathcal{A}_{\text{act}}^0$ in the beginning; then $t_{h+1} = \min\{t \in [T] : i_{*,t} \notin \{\pi_1, ..., \pi_h\}\}$ be the next time when the optimal $i_{*,t}$ changes in the active set, and $\pi_{h+1} = i_{*,t_{h+1}}$ be the new optimal active arm (when the previous π_h is sent to the confirmed set). Since $\pi_h \in [S]$ and is increasing by definition, we end up with a sequence $(\pi_h, t_h)_{h\in[H]}$ for some $H \le S$. Namely, during the horizon $t \in [t_h, t_{h+1})$, the optimal active arm is $i_{*,t} = \pi_h$. For the ease of notation, at each of those times t_h, denote the minimum count as $N^{t_h} = n_h$ and define the layers

$$L_n = \{a \in [K]\backslash[S] : a \text{ is pulled as an active arm when } N^t = n\}$$

for $n \ge 1$. Denote $n_1 = 1$ and

$$n_{H+1} = 1 + \max\{n \in \mathbb{N} : L_n \cap [S] \neq \varnothing\}$$

be the last layer that ever pulls a suboptimal arm (and incurs any regret). We set $n_1 = 1$ to keep the following computation compact.

Recall that in Algorithm 1, we construct V_t by repeatedly choosing the arm with the most out-neighbors among the least observed arms $\mathcal{A}_0$ and removing its neighbors. By Lemma 11 and 12, the active arms we have chosen before observing every arm in $\mathcal{A}_0$, while $N^t = n$, form a layer and can be bounded by

$$|L_n| \le 2\log K \max_{A\subseteq[K]} \delta(G|_A) + S \le 100\alpha \log^2 K + S \tag{4}$$

for every layer $n \geq 0$, where $\delta(G|_A)$ is the dominating number of the subgraph G restricted to the subset $A \subseteq [K]$ as defined in (28). Then the cumulative regret is bounded as

$$\begin{aligned}
\mathsf{R}(\mathsf{Alg}\ 1) &\leq |L_0| + \sum_{h=1}^{H} \sum_{n=n_h}^{n_{h+1}-1} \sum_{a \in L_n} \Delta_{a,\pi_h} \\
&\overset{\text{(c)}}{\leq} |L_0| + \sum_{h=1}^{H} \sum_{n=n_h}^{n_{h+1}-1} \sum_{a \in L_n} 8w(n) \\
&\overset{\text{(d)}}{\leq} 100\alpha \log^2 K + S + 8(100\alpha \log^2 K + S) \sum_{h=1}^{H} \sum_{n=n_h}^{n_{h+1}-1} w(n) \\
&= 100\alpha \log^2 K + S + 8(100\alpha \log^2 K + S) \underbrace{\sum_{n=1}^{n_{H+1}-1} w(n)}_{(\spadesuit)}
\end{aligned} \tag{5}$$

where (c) is by Lemma 2, and (d) uses (4).

By (3), we have $n_{H+1} - 1 \leq 64 \log(2KT/\delta)/\Delta_*^2$. Consequently,

$$(\spadesuit) \leq 2\sqrt{\log(2KT/\delta)}\sqrt{n_{H+1}-1} \leq 16 \log(2KT/\delta)\Delta_*^{-1}$$

where the first inequality follows from the elementary inequality that $\sum_{k=1}^{n} 1/\sqrt{k} \leq 2\sqrt{n}$. Putting back to (5), we obtain

$$\mathsf{R}_T(\text{Alg } 1) = O\bigg(\log(2KT/\delta)\frac{\alpha \log^2 K + S}{\Delta_*}\bigg).$$

□

A.3 Instance-dependent Regret Lower Bound

In this section, we prove instance-dependent lower bounds for a given gap Δ_* in Theorem 2.2.

Theorem A.2. *Suppose $\alpha \geq 2S$ and fix a policy π that satisfies $\mathsf{R}(\pi) \leq CT^p$ for some constant $C > 0$ and $p \in [0, 1)$ under any bandit environment. If $\frac{1}{4} \geq \Delta_* > T^{-(1-p)}$, then it holds that*

$$\max_{\nu} \mathsf{R}_\nu(\pi) = \Omega\bigg(\log(T^{1-p}\Delta_*) \min\bigg\{\frac{\alpha}{\Delta_*}, \frac{1}{\Delta_*^2}\bigg\} - \frac{\log(C)}{\Delta_*}\bigg)$$

where the maximum is taken over all bandit environment ν and $\mathsf{R}_\nu(\pi)$ is the expected regret of π under the environment ν.

Proof. Let $I = \{a_1, \ldots, a_\alpha\}$ be a maximum independent subset in G. We will fix the first $a_1, \ldots, a_{S-1}$ arms to be optimal, and construct $\alpha - S + 1$ different environments. Specifically, consider an index $u \in \{S, S+1, \ldots, \alpha\}$. To define an environment based on index u, let the product reward distribution $P^u = \prod_{a \in [K]} \text{Bern}(\mu_a)$ with

$$\mu_a = \begin{cases} 1 & \text{if } a = a_i \text{ for some } i \in [S-1] \\ \frac{1}{4} + \Delta_* & \text{if } a = a_S \\ \frac{1}{4} + 2\Delta_* \mathbb{1}[u > S] & \text{if } u > S \text{ and } a = a_u \\ \frac{1}{4} & \text{else if } a \in I \\ 0 & \text{otherwise} \end{cases}.$$

Fix any policy π. Let $\mathbb{E}_u$ (resp. $\mathbb{P}_u$) be the expectation (resp. probability) taken under environment u. Suppose $\mathsf{R}_u(\pi) \leq CT^p$ for some constants $C > 0$ and $p \in [0, 1)$, where $\mathsf{R}_u(\pi)$ denotes the expected regret of π under

environment u. Denote $N_i(T) = \sum_{t=1}^T \mathbb{1}[i \in V_t]$ where V_t is the decision selected by π at time t. Denote $N_0 = \sum_{t=1}^T \mathbb{1}[V_t \cap ([K]\backslash I) \neq \varnothing]$ the number of times an arm outside I is selected. Then for any $i = S+1, \ldots, \alpha$,

$$\begin{aligned}\mathbb{P}_S\bigg(N_i \geq \frac{T}{2}\bigg) + \mathbb{P}_i\bigg(N_i < \frac{T}{2}\bigg) &\geq \frac{1}{2}\exp\big(-\mathsf{KL}(\mathbb{P}_S^{\otimes T} \| \mathbb{P}_i^{\otimes T})\big) \\ &\overset{\text{(a)}}{\geq} \frac{1}{2}\exp\bigg(-(\mathbb{E}_S[N_i] + \mathbb{E}_S[N_0])\frac{16\Delta_*^2}{3}\bigg)\end{aligned}$$

where (a) uses the inequality $\mathsf{KL}(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) \leq \frac{(p-q)^2}{q(1-q)}$ and $\Delta_* \in (0, \frac{1}{4}]$. By construction, for $i = S+1, \ldots, \alpha$,

$$\mathsf{R}_S(\pi) + \mathsf{R}_i(\pi) \geq \frac{\Delta_* T}{4}\bigg(\mathbb{P}_S\bigg(N_i \geq \frac{T}{2}\bigg) + \mathbb{P}_i\bigg(N_i < \frac{T}{2}\bigg)\bigg) \geq \frac{\Delta_* T}{8}\exp\bigg(-(\mathbb{E}_S[N_i] + \mathbb{E}_S[N_0])\frac{16\Delta_*^2}{3}\bigg)$$

which implies

$$\mathbb{E}_S[N_i] + \mathbb{E}_S[N_0] \geq \frac{3}{16\Delta_*^2}\log\bigg(\frac{\Delta_* T}{8(\mathsf{R}_S(\pi) + \mathsf{R}_i(\pi))}\bigg) \geq \frac{3}{16\Delta_*^2}\log\bigg(\frac{\Delta_* T^{1-p}}{16C}\bigg)$$

Then we have

$$\begin{aligned}\mathsf{R}_S(\pi) &\geq \Delta_* \sum_{i=S+1}^{\alpha} \mathbb{E}_S[N_i] + \frac{1}{4}\mathbb{E}_S[N_0] \\ &\geq \Delta_* \sum_{i=S+1}^{\alpha} (\mathbb{E}_S[N_i] + \mathbb{E}_S[N_0]) + \bigg(\frac{1}{4} - \Delta_*(\alpha - S)\bigg)\mathbb{E}_S[N_0] \\ &\geq \frac{3}{16\Delta_*}\sum_{i=S+1}^{\alpha} \log\bigg(\frac{\Delta_* T^{1-p}}{16C}\bigg) + \bigg(\frac{1}{4} - \Delta_*(\alpha - S)\bigg)\mathbb{E}_S[N_0] \\ &\geq \frac{3\alpha}{32\Delta_*}\log\bigg(\frac{\Delta_* T^{1-p}}{16C}\bigg) + \bigg(\frac{1}{4} - \Delta_*\alpha\bigg)\mathbb{E}_S[N_0] \qquad (6)\end{aligned}$$

where the last inequality uses $\alpha \geq 2S$. On the other hand, we also have

$$\mathsf{R}_S(\pi) \geq \frac{1}{4}\mathbb{E}_S[N_0]. \qquad (7)$$

When $\Delta_* \leq \frac{1}{4\alpha}$, (6) directly gives the desired bound since $1 - \Delta_*\alpha \geq 0$. When $\Delta_* > \frac{1}{4\alpha}$, we have

$$\begin{aligned}\mathsf{R}_S(\pi) &\overset{(6)}{\geq} \frac{3\alpha}{32\Delta_*}\log\bigg(\frac{\Delta_* T^{1-p}}{16C}\bigg) - \Delta_*\alpha\mathbb{E}_S[N_0] \\ &\overset{(7)}{\geq} \frac{3\alpha}{32\Delta_*}\log\bigg(\frac{\Delta_* T^{1-p}}{16C}\bigg) - \frac{\Delta_*\alpha}{4}\mathsf{R}_S(\pi).\end{aligned}$$

Then

$$\frac{17\Delta_*\alpha}{4}\mathsf{R}_S(\pi) \geq \frac{4 + \Delta_*\alpha}{4}\mathsf{R}_S(\pi) \geq \frac{3\alpha}{32\Delta_*}\log\bigg(\frac{\Delta_* T^{1-p}}{16C}\bigg)$$

which concludes the proof. □

Theorem A.3. *Suppose $K \geq 2S$ and $\Delta_* \leq \frac{1}{2}$. For any policy π, it holds that*

$$\max_{\nu} \mathsf{R}_\nu(\pi) = \Omega\big(S\min\{\Delta_* T, \Delta_*^{-1}\}\big)$$

where the maximum is taken over all bandit environment ν and $\mathsf{R}_\nu(\pi)$ is the expected regret of π under the environment ν.

Proof. Note it suffices to prove the lower bound for the full-information feedback, since a policy can recover its performance under other feedback by simply discarding observations obtained under full information. Since $K \geq 2S$, we can fix any two disjoint subsets $V_0, V_1 \subseteq [K]$ with $|V_0| = |V_1| = S$. Write $V_0 = \{a_{0,1}, \ldots, a_{0,S}\}$ and

$V_1 = \{a_{1,1}, \dots, a_{1,S}\}$ respectively. We will construct the bandit environments as follows. For index $u \in \{0,1\}^S$, define the product reward distribution $P^u = \prod_{a\in[K]} \mathrm{Bern}(\mu_a)$ with

$$\mu_a = \begin{cases} \frac{1}{4} + \Delta_* & \text{if } a = a_{u_j,j} \text{ for some } j \in [S] \\ \frac{1}{4} & \text{if } a = a_{1-u_j,j} \text{ for some } j \in [S] \\ 0 & \text{otherwise} \end{cases}.$$

Let $\mathbb{P}_u$ be the probability taken under environment u. Under full information, WLOG, we assume the policy only pulls arms in $V_0 \cup V_1$ that have positive rewards. Consider sampling $u \in \{0,1\}^S$ uniformly. We lower bound the minimax regret by the Bayes regret:

$$\max_u \mathsf{R}_u(\pi) \ge 2^{-S} \sum_{u\in\{0,1\}^S} \mathbb{E}_u\left[\Delta_* \sum_{t=1}^{T}\sum_{j=1}^{S} \mathbb{1}[a_{1-u_{j,j}} \in V_t]\right] \tag{8}$$

Now we consider a uniform random variable $J \sim \mathrm{Unif}([S])$. By (8),

$$\begin{aligned}\max_u \mathsf{R}_u(\pi) &\ge 2^{-S} \sum_{u\in\{0,1\}^S} \mathbb{E}_u\left[S\Delta_* \sum_{t=1}^{T} \mathbb{E}_{J\sim\mathrm{Unif}([S])}[\mathbb{1}[a_{1-u_J,J} \in V_t]]\right] \\ &= S\Delta_* \mathbb{E}_{J\sim\mathrm{Unif}([S])}\left[2^{-S} \sum_{u\in\{0,1\}^S} \mathbb{E}_u\left[\sum_{t=1}^{T} \mathbb{1}[a_{1-u_J,J} \in V_t]\right]\right]. \end{aligned} \tag{9}$$

For each $J \in [S]$ and u, denote u_{-J} as u except the J-th entry, $u^{(0)}_{-J}$ as setting u_J to 0, and $u^{(1)}_{-J}$ as setting u_J to 1. Let $\mathcal{F}_t$ denote the history up to the beginning of time t. Then we have

$$\begin{aligned}(9) &\overset{(a)}{\ge} S\Delta_* \mathbb{E}_{J\sim\mathrm{Unif}([S])}\left[2^{-S} \sum_{u_{-J}\in\{0,1\}^{S-1}} \mathbb{E}_{u^{(0)}_{-J}}\left[\sum_{t=1}^{T} \mathbb{1}[a_{1,J} \in V_t]\right] + \mathbb{E}_{u^{(1)}_{-J}}\left[\sum_{t=1}^{T} \mathbb{1}[a_{1,J} \notin V_t]\right]\right] \\ &= S\Delta_* \mathbb{E}_{J\sim\mathrm{Unif}([S])}\left[2^{-S} \sum_{u_{-J}\in\{0,1\}^{S-1}} \sum_{t=1}^{T} \mathbb{P}_{u^{(0)}_{-J}}(a_{1,J} \in V_t|\mathcal{F}_t) + \sum_{t=1}^{T} \mathbb{P}_{u^{(1)}_{-J}}(a_{1,J} \notin V_t|\mathcal{F}_t)\right] \\ &\overset{(b)}{\ge} S\Delta_* \mathbb{E}_{J\sim\mathrm{Unif}([S])}\left[2^{-S-1} \sum_{u_{-J}\in\{0,1\}^{S-1}} \sum_{t=1}^{T} \exp\left(-\mathsf{KL}\left(\mathbb{P}^{\otimes(t-1)}_{u^{(0)}_{-J}} \middle\| \mathbb{P}^{\otimes(t-1)}_{u^{(1)}_{-J}}\right)\right)\right] \\ &\overset{(c)}{\ge} S\Delta_* \mathbb{E}_{J\sim\mathrm{Unif}([S])}\left[2^{-S-1} \sum_{u_{-J}\in\{0,1\}^{S-1}} \sum_{t=1}^{T} \exp\left(-\frac{16(t-1)\Delta_*^2}{3}\right)\right] \\ &= \frac{S\Delta_*}{4} \sum_{t=1}^{T} \exp\left(-\frac{16(t-1)\Delta_*^2}{3}\right) \\ &\ge \frac{S\Delta_*}{4} \sum_{t=1}^{\min\{T,\Delta_*^{-2}\}} \exp\left(-\frac{16}{3}\right) = \Omega\big(S\Delta_* \min\{T, \Delta_*^{-2}\}\big). \end{aligned}$$

(a) uses a key observation that, for fixed $J \in [S]$, when $a_{J,1}$ is optimal (under environment $u^{(1)}_{-J}$) but $a_{J,1} \notin V_t$, V_t must have included a suboptimal arm (not necessarily $a_{J,0}$) and thereby incur regret Δ_*. (b) applies the Bretagnolle–Huber inequality up to time t. (c) follows from the chain rule of KL, the inequality $\mathsf{KL}(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) \le \frac{(p-q)^2}{q(1-q)}$, and $\Delta_* \in [0, \frac{1}{2}]$. □

A.4 Minimax Regret Upper Bound

Theorem A.4 (Restatement of Theorem 2.3)**.** *Fix any $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\mathsf{R}(\text{Alg 1}) = O\Big(\log^2 K \sqrt{\log(2TK/\delta)}\Big(\sqrt{\alpha ST} + S\sqrt{T}\Big)\Big).$$

Proof. Consider the instantaneous regret $\sum_{i=1}^{S}\mu_i - \sum_{a\in V_t}\mu_a$ at time t. By (1), the instantaneous regret comes from arms in the active set and can be bounded as:

$$\sum_{i=1}^{S}\mu_i - \sum_{a\in V_t}\mu_a \le \sum_{a\in V_t\cap\mathcal{A}_{\rm act}^t}(\mu_{i_t}-\mu_a) \le 8|V_t\cap\mathcal{A}_{\rm act}^t|\cdot w(N^t) = 8\big(S-|\mathcal{A}_{\rm con}^t|\big)w(N^t) \tag{10}$$

where the second inequality follows from Lemma 2. Since $|\mathcal{A}_{\rm con}^t|$ is increasing in t and bounded by $S-1$, we partition the horizon $[T]$ by

$$1 = t_0 \le t_1 \le \cdots \le t_S = T+1.$$

For each $s = 0, 1, \ldots, S-1$, during the sub-horizon $t \in [t_s, t_{s+1})$, it holds that $|\mathcal{A}_{\rm con}^t| = s$. Denote $T(n)$ to be the number of times when the minimum count $N^t = n$, and recall that N^{t_s} is the minimum count at the time $\mathcal{A}_{\rm con}^t$ is updated. Then by (10), the regret is bounded by

$$\mathsf{R}(\text{Alg 1}) \le 8\sum_{s=0}^{S-1}\sum_{n=N^{t_s}}^{N^{t_{s+1}}-1} T(n)(S-s)\min\{1, w(n)\}. \tag{11}$$

Similar to (4), we now bound the number of times when $N^t = n$, i.e. $T(n)$, by bounding the number of active arms pulled during these times. Recall that in Algorithm 1, we construct V_t by repeatedly choosing the arm with the most out-neighbors among the least observed arms $\mathcal{A}_0$ and removing its neighbors. By Lemma 11 and 12, the active arms we have chosen before observing every arm in $\mathcal{A}_0$, while $N^t = n$, can be bounded by

$$T(n)(S-|\mathcal{A}_{\rm con}^t|) \le \max\left\{2\log K \max_{A\subseteq[K]}\delta(G|_A), S-|\mathcal{A}_{\rm con}^t|\right\} \le \max\big\{100\alpha\log^2 K, S-|\mathcal{A}_{\rm con}^t|\big\} \tag{12}$$

where $\delta(G|_A)$ is the dominating number of the subgraph G restricted to A, as defined in (28). The maximum comes from the fact that we experience at least one time step before updating the minimum count N^t (i.e. $T(n) \ge 1$). Then during each sub-horizon $t \in [t_s, t_{s+1})$, as $|\mathcal{A}_{\rm con}^t| = s$, we have

$$T(n) \le \overline{T}_s \coloneqq \left\lceil \frac{100\alpha\log^2 K}{S-s} \right\rceil \tag{13}$$

for $n = N^{t_s}, \ldots, N^{t_{s+1}}-1$. On the other hand, by definition, we also have the following constraints. For each $s = 0, 1, \ldots, S-1$,

$$\sum_{n=N^{t_s}}^{N^{t_{s+1}}-1} T(n) = t_{s+1} - t_s. \tag{14}$$

To bound the regret decomposition in (11), we consider the maximization of the right-hand side over the possible values of $0 = N^{t_0} \le N^{t_1} \le \cdots \le N^{T_S}$ subject to the constraints (13) and (14), for any given partition $\{t_s : s = 0, 1, \ldots, S\}$. The maximum is achieved (or upper bounded, if the sub-horizon $t_{s+1}-t_s$ is not divisible by $\overline{T}_s$) by setting

$$N_*^{t_{s+1}} = N_*^{t_s} + \left\lceil \frac{t_{s+1}-t_s}{\overline{T}_s} \right\rceil, \qquad \text{and } T(n) = \overline{T}_s \tag{15}$$

where $\overline{T}_s$ is the upper bound described in (13). To see this, if $T(n) < \overline{T}_s$ for some "earlier" n at the s-th sub-horizon, the right-hand side of (11) can be made larger by increasing $T(n)$ to $\overline{T}_s$ and decreasing $T(N^{t_{s+1}-1})$ to satisfy (14), because the multiplicative factor $w(n)$ is decreasing in n. Namely, the target quantity in (11) is larger if we fill in $T(n)$ for small n first.

Consequently, the regret in (11) is bounded as

$$\begin{aligned}\mathsf{R}(\text{Alg 1}) &\le 8\sum_{s=0}^{S-1}\sum_{n=N_*^{t_s}}^{N_*^{t_{s+1}}-1}\overline{T}_s(S-s)\min\{1,w(n)\}\\ &\overset{\text{(a)}}{\le} 8S\overline{T}_0 + L\sum_{s=0}^{S-1}\overline{T}_s(S-s)\left(\sqrt{N_*^{t_{s+1}}-1}-\sqrt{N_*^{t_s}-1}\right)\end{aligned} \tag{16}$$

where (a) uses $\min\{1, w(0)\} \le 1$, the elementary inequality $\sum_{t=i}^{j} 1/\sqrt{t} \le 2\big(\sqrt{j} - \sqrt{i-1}\big)$, and for the width factor $L = 16\sqrt{\log(2KT/\delta)}$. Now we distinguish into two cases.

First suppose $\alpha \ge S$, so

$$C_0 \frac{\alpha \log^2 K}{S-s} \le \overline{T}_s \le C_1 \frac{\alpha \log^2 K}{S-s}$$

for some absolute constants $C_0, C_1 > 0$. Then following (16), we have $C_0 \alpha \log^2 K \le \overline{T}_s(S-s) \le C_1 \alpha \log^2 K$ for every $s = 0, 1, ..., S-1$ and

$$\begin{aligned}
\mathsf{R}(\text{Alg } 1) &\le 8C_1 \alpha \log^2 K + C_1 L \alpha \log^2 K \sqrt{N_*^{t_S} - 1} \\
&\overset{\text{(b)}}{\le} 8C_1 \alpha \log^2 K + C_1 L \alpha \log^2 K \sqrt{\sum_{s=0}^{S-1} \frac{S-s}{C_0 \alpha \log^2 K}(t_{s+1} - t_s)} \\
&\le 8C_1 \alpha \log^2 K + \frac{C_1}{\sqrt{C_0}} L \log K \sqrt{\alpha S (T+1)}
\end{aligned}$$

where (b) plugs in the definition of $N_*^{t_S}$ in (15), and the last line simply bounds $S - s \le S$ and $t_S = T + 1$.

When $\alpha < S$, we have $\overline{T}_s(S-s) \le C_2 S \log^2 K$ for an absolute constant $C_2 > 0$ for every $s = 0, 1, ..., S-1$ and also $\overline{T}_s \ge 1$. Then (16) gives

$$\begin{aligned}
\mathsf{R}(\text{Alg } 1) &\overset{\text{(c)}}{\le} 8(100\alpha \log^2 K + S) + C_2 L S \log^2 K \sqrt{N_*^{t_S} - 1} \\
&= 8(100\alpha \log^2 K + S) + C_2 L S \log^2 K \sqrt{\sum_{s=0}^{S-1} \left\lceil \frac{t_{s+1} - t_s}{\overline{T}_s} \right\rceil} \\
&\le 8(100\alpha \log^2 K + S) + C_2 L S \log^2 K \sqrt{\sum_{s=0}^{S-1} (t_{s+1} - t_s)} \\
&= 8(100\alpha \log^2 K + S) + C_2 L S \sqrt{T} \log^2 K
\end{aligned}$$

where (c) substitutes in $\overline{T}_s(S-s) \le C_2 \log^2 KS$ and applies the telescoping sum. □

A.5 Suboptimality Results for UCB

Theorem A.5 (Restatement of Theorem 2.4)**.** *Fix any (S, α, K, T) with $S\alpha \le K$ and $\alpha > 1$. There is a problem instance under which*

$$\mathsf{R}(\text{Alg } 2) = \Omega(LS\sqrt{\alpha T})$$

where L is the width parameter used in Algorithm 2.

Proof. Partition $[K]$ into subsets $V_1, V_2, \ldots, \overline{V}_\alpha$, with $|V_1| = |V_2| = \cdots = |V_{\alpha-1}| = S$ and $|\overline{V}_\alpha| = K - S(\alpha - 1) \ge S$. The feedback graph $G = ([K], E)$ is defined by having each V_k and $\overline{V}_\alpha$ be a clique respectively, for $k \in [\alpha - 1]$. To analyze the trajectory of the UCB algorithm, we consider a *deterministic* reward for each arm. Specifically, fix any subset $V_\alpha \subseteq \overline{V}_\alpha$ with size $|V_\alpha| = S$. For each $a \in V_k$ for $k = 1, 2, \ldots, \alpha$, we define a shared, deterministic reward

$$r_{t,a} \equiv \frac{1}{2} + \mathbb{1}[i = 1]\Delta - (k-1)\varepsilon \tag{17}$$

for some small $\Delta, \varepsilon > 0$ to be specified later, and let $r_{t,b} \equiv 0$ for any $b \in \overline{V}_\alpha \backslash V_\alpha$. These "extra" arms $\overline{V}_\alpha \backslash V_\alpha$ are clearly suboptimal and, as we will argue next, are never pulled for more than once. It is clear that the optimal combinatorial decision is V_1, and

$$\sum_{i \in V_1} \mu_i - \sum_{a \in V_k} \mu_a = S(\Delta + (k-1)\varepsilon) \ge S\Delta \tag{18}$$

for any other $k = 2, 3, \ldots, \alpha$.

We have a few key observations. First, under any tie-breaking rule and this constructed feedback graph G, Algorithm 2 initializes every arm to exactly one observation, i.e. there is a time $t_0 \le \alpha$ such that $n_{t_0,a} = 1$ for every $a \in [K]$. To see why, note that given the graph structure, an arm $a \in V_k$ has $n_{t,a} = 0$ if and only if $n_{t,b} = 0$ for all $b \in V_k$, at any time t. Meanwhile, there are at least S arms in each V_k, so the algorithm never chooses any arm c with $n_{t,c} > 0$ whenever there is an arm $a \in [K]$ with $n_{t,a} = 0$. These imply the existence of the time t_0, at which every arm is observed exactly once. The existence of such t_0, together with the deterministic rewards, allows us to precisely describe the algorithm's trajectory thereafter.

Due to the feedback structure of G, if two arms $a, b \in V_k$ or $\overline{V}_\alpha$ are in the same clique, then $n_{t,a} = n_{t,b}$ for all t. Thus we are guaranteed the following:

- The algorithm never pulls any arm from $\overline{V}_\alpha \backslash V_\alpha$, since $V_\alpha \subseteq \overline{V}_\alpha$ are the S arms with (always) higher UCB (i.e. V_α share the same observation counts but have higher observed empirical rewards than $\overline{V}_\alpha \backslash V_\alpha$).
- The algorithm will only pull one of the cliques $\{V_1, V_2, \ldots, V_\alpha\}$ at each time, due to the deterministic choice of rewards in (17) and if the reward is distinct for each clique, i.e. $r_{t,a} > r_{t,b}$ if $a \in V_i$ and $b \in V_j$ for $i < j$.

Before proceeding to the exact proof, we give an intuition of why UCB fails under the constructed problem instance above. As demonstrated above, the selection rule of UCB guarantees that only one of the α cliques will be selected at each time t, due to the lack of explicit exploration. The number of observations collected is always S at each time. In contrast, our proposed elimination algorithm (Algorithm 1) explicitly leverages the fact that the sub-optimality gap of every clique V_k is of the same order, so it will be forced to distribute the S selection budget over different cliques, paying an instantaneous regret of the same order as selecting only one of the cliques. This is beneficial under graph feedback, as now the number of observations collected is $\min\{S^2, K\}$ i.e. every arm is observed. The speed of information collection is roughly S times faster, shaving the regret by a factor of $\sqrt{S}$. At a high level, this example highlights that the implicit exploration in UCB is insufficient to optimally exploit general graph feedback.

Now we rigorously present the lower bound proof. The exploration-exploitation trade-off is captured by the fact that UCB needs to pull other cliques to realize that they are suboptimal, which incurs sub-optimal regret crucially because UCB always focuses on only one of the cliques. Let t_1 denote the last time at which V_1 is pulled, and $n_{t_1,1} \le t_1$ denote the total number of times V_1 is pulled. Due to the reward gap described in (18), we have

$$\mathsf{R}(\text{Alg } 2) \ge (T - n_{t_1,1}) S\Delta. \tag{19}$$

In addition, since V_1 has the largest UCB at time t_1, for every $k = 2, 3, \ldots, \alpha$, the other cliques have a smaller UCB:

$$\begin{aligned} \frac{1}{2} + \Delta + \frac{L}{\sqrt{n_{t_1,1}}} &\ge \frac{1}{2} - (k-1)\varepsilon + \frac{L}{\sqrt{n_{t_1,k}}} \\ &\ge \frac{1}{2} - \alpha\varepsilon + \frac{L}{\sqrt{n_{t_1,k}}} \end{aligned} \tag{20}$$

where $n_{t_1,k}$ denotes the observation count of any arm of the clique V_k at time t_1, since they are the same. Recall that the observation count of a clique V_k equals the number of times it gets pulled.

Under our graph structure, $T - n_{t_1,1} = \sum_{k=2}^{\alpha} n_{t_1,k}$. So (19) translates to

$$\mathsf{R}(\text{Alg } 2) \ge \sum_{k=2}^{\alpha} S\Delta n_{t_1,k}. \tag{21}$$

For the sake of simplicity, we write $T - n_{t_1,1} = cT$ for some $c \in [0,1]$. Choose $\Delta = L\sqrt{\alpha/(4T)}$ and $\varepsilon = L\sqrt{1/(4\alpha T)}$. Then we have $\Delta + \alpha\varepsilon = L\sqrt{\alpha/T}$. By (20) and some algebra,

$$L\left(\sqrt{\frac{\alpha}{T}} + \sqrt{\frac{1}{(1-c)T}}\right) \ge L\sqrt{\frac{1}{n_{t_1,k}}}$$

for every $k = 2, \dots, \alpha$, which implies

$$n_{t_1,k} \geq \frac{T}{2(\alpha + \frac{1}{1-c})}.$$

Then (21) becomes

$$\mathsf{R}(\text{Alg } 2) \geq LS\sqrt{\alpha T}\frac{\alpha - 1}{4(\alpha + \frac{1}{1-c})} \geq LS\sqrt{\alpha T}\frac{\alpha}{8(\alpha + \frac{1}{1-c})}. \tag{22}$$

Meanwhile, (19) gives a trade-off lower bound

$$\mathsf{R}(\text{Alg } 2) \geq \frac{cL}{2}S\sqrt{\alpha T}. \tag{23}$$

When $c \in [0, \frac{1}{2}]$, (22) gives

$$\mathsf{R}(\text{Alg } 2) \geq LS\sqrt{\alpha T}\frac{\alpha}{8(\alpha+2)} \geq \frac{L}{16}S\sqrt{\alpha T}\frac{\alpha}{\alpha+1} \geq \frac{L}{32}S\sqrt{\alpha T}$$

and when $c \in [\frac{1}{2}, 1]$, (23) gives

$$\mathsf{R}(\text{Alg } 2) \geq \frac{L}{4}S\sqrt{\alpha T}.$$

Since they hold simultaneously, the regret $\mathsf{R}(\text{Alg } 2)$ is lower bounded by the maximum of the two, which concludes the proof. □

A.6 UCB Regret Upper Bound under Full Information

Theorem A.6. *Suppose we have full-information feedback. Let $L = \sqrt{\log(2KT/\delta)}$. With probability at least $1 - \delta$,*

$$\mathsf{R}(\text{Alg } 2) = O(S\sqrt{\log(2KT/\delta)T}).$$

Proof. WLOG, suppose the event in the concentration result Lemma 1 holds. The number of observations for every arm $a \in [K]$ at time t is $t - 1$ under full-information. Fix any $i \in [S]$. Consider the i-th arm $a_{t,(i)}$ with the largest UCB (if there is a tie, consider the selected ones with arbitrary ordering) and compare with the arm i. Suppose $a_{t,(i)} > i$. By the pigeonhole principle, there exists an arm $j \leq i$ such that $\bar{r}_{t,(i)} \geq \bar{r}_{t,j}$. Then by Lemma 1,

$$\mu_{a_{t,(i)}} \geq \bar{r}_{t,(i)} - \frac{L}{\sqrt{t-1}} \geq \bar{r}_{t,j} - \frac{L}{\sqrt{t-1}} \geq \mu_j - 2\frac{L}{\sqrt{t-1}} \geq \mu_i - 2\frac{L}{\sqrt{t-1}}.$$

The instantaneous regret at time t is thus bounded by

$$\sum_{i=1}^{S}\mu_i - \sum_{a\in V_t}\mu_a \leq \sum_{i=1}^{S}\mathbb{1}[a_{t,(i)} > i]\big(\mu_i - \mu_{a_{t,(i)}}\big) \leq 2\sum_{i=1}^{S}\mathbb{1}[a_{t,(i)} > i]\frac{L}{\sqrt{t-1}} \leq 2S\frac{L}{\sqrt{t-1}}.$$

Then the cumulative regret is easily bounded as

$$\mathsf{R}(\text{Alg } 2) \leq 2S\sum_{t=1}^{T}\min\left\{1, \frac{L}{\sqrt{t-1}}\right\} \leq 4S\sqrt{\log(2KT/\delta)T} + 2S.$$

□

B COMBINATORIAL BANDITS WITH LINEAR CONTEXTS

This section provides proofs for the results in Section 3.

B.1 Proof of Lemma 4

Lemma 9 (Restatement of Lemma 4)**.** *Suppose the event in Lemma 3 holds for every $t \in [T]$ and stage $h \in [H]$. For each time t, the following events holds in Algorithm 3 for each stage $h \in [H]$:*

(A) The top $S - |V_t^{(h-1,3)}|$ arms of $V_{*,t} \backslash V_t^{(h-1,3)}$ are in A_h, i.e. remain uneliminated.

(B) The confirmed arms are optimal: $U_t^{(h,2)} \subseteq V_{*,t} \backslash V_t^{(h,1)}$.

(C) Let $i_{*,t}^{(h)} = \arg\min\{\theta_*^\top x_{t,a} : a \in A_h\}$ denote the optimal active arm. For every $a \in A_h$, it holds that

$$\theta_*^\top x_{t,i_{*,t}^{(h)}} - \theta_*^\top x_{t,a} \le 16 \cdot 2^{-h}.$$

Proof. We will prove the result via an inductive argument. At $h = 1$, the claims (A) and (C) trivially hold.

We first show that claim (B) holds for any stage $h \in [H]$ conditioned on (A). By the time of step (2), the width of every remaining active arm satisfies $w_{t,a}^{(h)} \le 2^{-h}$ for $a \in A_h \backslash V_t^{(h,1)}$. By claim (A), the top $S - |V_t^{(h,1)}|$ arms of $V_{*,t} \backslash V_t^{(h,1)}$ remain in the active set $A_h \backslash V_t^{(h,1)} = A_h \backslash U_t^{(h,1)}$, since the number of selected optimal arms in step (1) of Algorithm 3 is no more than the number of selected active arms $U_t^{(h,1)}$; denote the set of these arms as $S_t^{(h)}$ for simplicity. Intuitively, this subset of optimal arms serves as the benchmark that the arms added to V_t later will be compared against. Under the concentration event in Lemma 3, if any remaining active arm $i_0 \in A_h \backslash V_t^{(h,1)}$ is confirmed, by definition

$$\widehat{r}_{t,i_0}^{(h)} > \widehat{r}_{t,a^*_{(|S_t^{(h)}|)}}^{(h)} + 4 \cdot 2^{-h}.$$

By the pigeonhole principle, there exists an optimal arm $j \in S_t^{(h)}$ with $\widehat{r}_{t,j}^{(h)} + w_{t,j}^{(h)} \le \widehat{r}_{t,a^*_{(|S_t^{(h)}|)}}^{(h)} + w_{t,a^*_{(|S_t^{(h)}|)}}^{(h)}$ where $a^*_{(k)}$ is the arm with the k-th largest $\widehat{r}_{t,a}^{(h)} + w_{t,a}^{(h)}$ by definition in Algorithm 3. Then we have

$$\begin{aligned}
\theta_*^\top x_{t,i_0} &\ge \widehat{r}_{t,i_0}^{(h)} - w_{t,i_0}^{(h)} \\
&\overset{\text{(a)}}{\ge} \widehat{r}_{t,i_0}^{(h)} - 2^{-h} \\
&> \widehat{r}_{t,a^*_{(|S_t^{(h)}|)}}^{(h)} + 3 \cdot 2^{-h} \\
&\ge \widehat{r}_{t,a^*_{(|S_t^{(h)}|)}}^{(h)} + w_{t,a^*_{(|S_t^{(h)}|)}}^{(h)} + 2 \cdot 2^{-h} \\
&\ge \widehat{r}_{t,j}^{(h)} + w_{t,j}^{(h)} + 2 \cdot 2^{-h} \\
&> \theta_*^\top x_{t,j}
\end{aligned}$$

where (a) follows since $i_0 \in A_h \backslash V_t^{(h,1)}$ implies $w_{t,i_0}^{(h)} \le 2^{-h}$. This leads to $i_0 \in S_t^{(h)} \subseteq V_{*,t}$ and proves claim (B). It also shows that the confirmed set $U_t^{(h,2)} \subsetneq S_t^{(h)}$ never covers all remaining optimal arms, much like Lemma 7 in the graph-feedback case.

Next, we show that the claims (A) and (C) for stage h hold conditioned on (A,B,C) at stage $h-1$. Since $i_{*,t}^{(h)} \in A_h \backslash V_t^{(h,2)}$, it is not confirmed in step (2) at stage $h-1$, so

$$\widehat{r}_{t,i_{*,t}^{(h)}}^{(h-1)} \le \widehat{r}_{t,a_{1,(|S_t^{(h-1)}|)}}^{(h-1)} + 4 \cdot 2^{-(h-1)}.$$

Recall that $a_{1,(|S_t^{(h-1)}|)}$ is the arm with the $|S_t^{(h-1)}|$-th largest UCB $\widehat{r}_{t,a}^{(h-1)} + w_{t,a}^{(h-1)}$ at the beginning of the confirmation step (2), and $a_{1,(|S_t^{(h-1)}|)}$ is not confirmed by definition of step (2). In addition, observe that $a_{2,(|S_t^{(h-1)}|)} = a_{1,(|S_t^{(h-1)}|)}$, because every arm confirmed and selected in step (2) has a larger UCB than $a_{1,(|S_t^{(h-1)}|)}$: If i is confirmed in step (2), then

$$\begin{aligned}
\widehat{r}_{t,i}^{(h-1)} + w_{t,i}^{(h-1)} &> \widehat{r}_{t,a_{1,(|S_t^{(h-1)}|)}}^{(h-1)} + 4 \cdot 2^{-(h-1)} + w_{t,i}^{(h-1)} \\
&> \widehat{r}_{t,a_{1,(|S_t^{(h-1)}|)}}^{(h-1)} + w_{t,a_{1,(|S_t^{(h-1)}|)}}^{(h-1)}.
\end{aligned}$$

After the elimination step in (3) at stage $h-1$, we get the active set A_h. Recall that $w_{t,a}^{(h-1)} \le 2^{-(h-1)}$ for every arm $a \in A_h \subseteq A_{h-1} \backslash V_t^{(h-1,1)}$. Thus

$$\begin{aligned}\theta_*^\top x_{t,a} &\ge \widehat{r}_{t,a}^{(h-1)} - 2^{-(h-1)} \ge \widehat{r}_{t,a_{2,(|S_t^{(h-1)}|)}}^{(h-1)} - 3 \cdot 2^{-(h-1)} \\ &= \widehat{r}_{t,a_{1,(|S_t^{(h-1)}|)}}^{(h-1)} - 3 \cdot 2^{-(h-1)} \ge \widehat{r}_{t,i_{*,t}^{(h)}}^{(h-1)} - 7 \cdot 2^{-(h-1)} \\ &\ge \theta_*^\top x_{t,i_{*,t}^{(h)}} - 8 \cdot 2^{-(h-1)}\end{aligned}$$

which proves claim (C) for stage h.

To show (A) for stage h, conditioned on (A) at stage $h-1$, we have

$$\begin{aligned}|(A_{h-1} \backslash V_t^{(h-1,3)}) \cap V_{*,t}| &= |(A_{h-1} \backslash (U_t^{(h-1,1)} \cup U_t^{(h-1,2)})) \cap V_{*,t}| \\ &\ge |A_{h-1} \cap V_{*,t}| - |U_t^{(h-1,1)} \cup U_t^{(h-1,2)}| \\ &\ge S - |V_t^{(h-2,3)}| - |U_t^{(h-1,1)} \cup U_t^{(h-1,2)})| \\ &= S - |V_t^{(h-2,3)} \cup U_t^{(h-1,1)} \cup U_t^{(h-1,2)})| = S - |V_t^{(h-1,3)}| \end{aligned} \tag{24}$$

where the first equality holds because $V_t^{(h-2,3)}$ has been removed from the active set A_{h-1}, and the last inequality uses (A) at stage $h-1$ (recall that $V_t^{(h-1,2)} = V_t^{(h-1,3)}$). Let $E_t^{(h-1)}$ denote the top $S - |V_t^{(h-1,3)}|$ arms in $A_{h-1} \backslash V_t^{(h-1,3)}$ (which is defined at the end of step (2) at stage $h-1$, as opposed to $S_t^{(h-1)}$ defined at the start of step (2)). By (24), $E_t^{(h-1)} \subseteq V_{*,t}$ is optimal. For any arm $i \in E_t^{(h-1)}$, by the pigeonhole principle again, there is another arm $j \in A_{h-1} \backslash V_t^{(h-1,3)}$ that satisfies $\theta_*^\top x_{t,j} \le \theta_*^\top x_{t,i}$ and $\widehat{r}_{t,j}^{(h-1)} \ge \widehat{r}_{t,a_{2,(|E_t^{(h-1)}|)}}^{(h-1)}$. Since $w_{t,a}^{(h-1)} \le 2^{-(h-1)}$ for every remaining arm $a \in A_{h-1} \backslash V_t^{(h-1,3)}$ at the time of elimination, we have

$$\begin{aligned}\widehat{r}_{t,i}^{(h-1)} &\ge \theta_*^\top x_{t,i} - 2^{-(h-1)} \ge \theta_*^\top x_{t,j} - 2^{-(h-1)} \ge \widehat{r}_{t,j}^{(h-1)} - 2 \cdot 2^{-(h-1)} \\ &\ge \widehat{r}_{t,a^*_{(|E_t^{(h-1)}|)}}^{(h-1)} - 2 \cdot 2^{-(h-1)}.\end{aligned}$$

As a consequence, this arm $i \in E_t^{(h-1)}$ is uneliminated by the end of stage $h-1$, i.e. $E_t^{(h-1)} \subseteq A_h$, which proves (A) and hence completes the induction. □

B.2 Regret Upper Bound

First, we state an elliptical potential lemma for the batched contextual bandits. It allows us to upper bound the sum of the self-normalized confidence widths defined in Algorithm 4.

Lemma 10 (Lemma 3 in Han et al. (2020)). *Fix any collection of d-dimensional vectors $\{\{x_{t,\ell}\}_{\ell \in [L_t]}\}_{t \in [T]}$ with $\|x_{t,\ell}\|_2 \le 1$. Suppose $L_t \le L$ for every $t \in [T]$ for some $L > 0$. Define*

$$A_t = I + \sum_{s<t} \sum_{\ell \in [L_t]} x_{t,\ell} x_{t,\ell}^\top.$$

It holds that

$$\sum_{t=1}^{T} \sqrt{\sum_{\ell \in [L_t]} \|x_{t,\ell}\|^2_{A_t^{-1}}} \le \sqrt{10} \log(T+1) \Big(\sqrt{dT} + d\sqrt{L} \Big).$$

Now we are in the position of proving the near-optimal regret upper bound for Algorithm 3.

Theorem B.1 (Restatement of Theorem 3.1). *Algorithm 3 combined with Algorithm 4 achieves*

$$\mathsf{R}(\text{Alg } 3) = O\Big(\log(ST) \sqrt{\log(2KT)} (\sqrt{dST} + dS) \Big).$$

Proof. WLOG, we assume the event in Lemma 3 holds for every $t \in [T]$. First consider the instantaneous regret at time t. Since $i_{*,t}^{(h)} = \arg\min\{\theta_*^\top x_{t,a} : a \in A_h\}$ is the optimal active arm at stage h, by the first claim in Lemma 4, $i_{*,t}^{(h)} \in V_{*,t}$. To bound the instantaneous regret, we will define a partition of the optimal $V_{*,t}$: Let $G_t^{(h)} \subseteq A_h$ denote the top $S - |V_t^{(h-1,3)}|$ arms in A_h (as in the first claim of Lemma 4) and define $V_{*,t}^{(h)} = G_t^{(h)} \backslash G_t^{(h+1)}$ the optimal arms that are in the top optimal arms at stage h but are not the top ones at next stage $h+1$. Observe that $G_t^{(h+1)} \subseteq G_t^{(h)}$ as the size $S - |V_t^{(h-1,3)}|$ is decreasing. By Lemma 4, $G_t^{(h)} \subseteq V_{*,t}$, and by definition we have $|G_t^{(h)}| = S - |V_t^{(h-1,3)}|$. Since $G_t^{(1)} = V_{*,t}$, we have

$$|V_{*,t}^{(1)}| = |V_{*,t}| - |G_t^{(2)}| = S - (S - |V_t^{(1,3)}|) = |V_t^{(1,3)}|$$

and recursively

$$|V_{*,t}^{(h)}| = |G_t^{(h)}| - |G_t^{(h-1)}| = |V_t^{(h,3)}| - |V_t^{(h-1,3)}| = |U_t^{(h,1)} \cup U_t^{(h,2)} \cup (\mathbb{1}[h = H] U_t^{(H,3)})|,$$

which matches the number of newly added arms to the decision V_t during stage h. By definition, $\theta_*^\top x_{t,i_{*,t}^{(h)}} \geq \theta_*^\top x_{t,i}$ for any $i \in V_{*,t}^{(h)} \subseteq A_h$ since $i_{*,t}^{(h)}$ is the optimal active arm at stage h. Then we can bound the instantaneous regret using this partition:

$$\sum_{i \in V_{*,t}} \theta_*^\top x_{t,i} - \sum_{a \in V_t} \theta_*^\top x_{t,a} = \sum_{h=1}^{H} \left(\sum_{i \in V_{*,t}^{(h)}} \theta_*^\top x_{t,i} - \sum_{a \in V_t^{(h,3)} \backslash V_t^{(h-1,3)}} \theta_*^\top x_{t,a} \right). \tag{25}$$

Recall that the confirmed arms are optimal i.e. $U_t^{(h,2)} \subseteq V_{*,t}$ for every stage $h \in [H]$, by Lemma 4.

We now show $U_t^{(h,2)} \subseteq V_{*,t}^{(h)}$: By step (2) at stage h, every remaining arm $a \in A_h \backslash V_t^{(h,1)}$ has a width $w_{t,a}^{(h)} \leq 2^{-h}$. By pigeonhole principle, there is an arm $j \in A_h$ with $\theta_*^\top x_{t,j} \geq \theta_*^\top x_{t,i}$ and Then if arm i is confirmed in step (2), we have

$$\begin{aligned}
\theta_*^\top x_{t,i} &\geq \widehat{r}_{t,i}^{(h)} - w_{t,i}^{(h)} > \widehat{r}^{(h)}_{t,a_{1,(S-|V_t^{(h,1)}|)}} + 4 \cdot 2^{-h} - w_{t,i}^{(h)} \\
&\geq \widehat{r}^{(h)}_{t,a_{1,(S-|V_t^{(h,1)}|)}} + w^{(h)}_{t,a_{1,(S-|V_t^{(h,1)}|)}} + 2 \cdot 2^{-h} \\
&\geq \widehat{r}^{(h)}_{t,a_{1,(S-|V_t^{(h-1,3)}|)}} + w^{(h)}_{t,a_{1,(S-|V_t^{(h-1,3)}|)}} + 2 \cdot 2^{-h} \\
&\overset{\text{(a)}}{\geq} \widehat{r}_{t,j}^{(h)} + w_{t,j}^{(h-1)} + 2 \cdot 2^{-h} \\
&\geq \theta_*^\top x_{t,j} + 2 \cdot 2^{-h} > \theta_*^\top x_{t,j}.
\end{aligned}$$

In (a), by pigeonhole principle, there is an arm $j \in G_t^{(h)}$ that satisfies $\theta_*^\top x_{t,j} \geq \theta_*^\top x_{t,a_{1,(S-|V_t^{(h-1,3)}|)}}$ but $\widehat{r}_{t,j}^{(h)} + w_{t,j}^{(h)} \leq \widehat{r}^{(h)}_{t,a_{1,(S-|V_t^{(h-1,3)}|)}} + w^{(h)}_{t,a_{1,(S-|V_t^{(h-1,3)}|)}}$, i.e. the expected reward is better than the benchmark, but the UCB is smaller, because $G_t^{(h)}$ is the top $S - |V_t^{(h-1,3)}|$ arms in A_h in terms of expected rewards. Then $\theta_*^\top x_{t,i} > \theta_*^\top x_{t,j}$ implies $i \in G_t^{(h)}$ and thereby $U_t^{h,2} \subseteq G_t^{(h)} \backslash G_t^{(h+1)} = V_{*,t}^{(h)}$.

We can proceed to remove $U_t^{(h,2)}$ from both the optimal arm partition $V_{*,t}^{(h)}$ and the arms selected by the learners

at each stage h in (25). This leads to

$$\begin{aligned}
\sum_{i\in V_{*,t}} \theta_*^\top x_{t,i} - \sum_{a\in V_t} \theta_*^\top x_{t,a} &= \sum_{h=1}^{H} \left(\sum_{i\in V_{*,t}^{(h)}\setminus U_t^{(h,2)}} \theta_*^\top x_{t,i} - \sum_{a\in V_t^{(h,3)}\setminus V_t^{(h-1,3)}\setminus U_t^{(h,2)}} \theta_*^\top x_{t,a} \right) \\
&\overset{\text{(b)}}{\le} \sum_{h=1}^{H} \left(\sum_{i\in V_{*,t}^{(h)}\setminus U_t^{(h,2)}} \theta_*^\top x_{t,i_{*,t}^{(h)}} - \sum_{a\in V_t^{(h,3)}\setminus V_t^{(h-1,3)}\setminus U_t^{(h,2)}} \theta_*^\top x_{t,a} \right) \\
&\le \sum_{h=1}^{H} \sum_{a\in U_t^{(h,1)}} \left(\theta_*^\top x_{t,i_{*,t}^{(h)}} - \theta_*^\top x_{t,a} \right) + \sum_{a\in U_t^{(H,0)}} \left(\theta_*^\top x_{t,i_{*,t}^{(H)}} - \theta_*^\top x_{t,a} \right) \\
&\overset{\text{(c)}}{\le} 16 \sum_{h=1}^{H} \sum_{a\in U_t^{(h,1)}} 2^{-h} + |U_t^{(H,0)}| \frac{16}{\sqrt{ST}} \\
&\overset{\text{(d)}}{\le} 16 \sum_{h=1}^{H} \sum_{a\in U_t^{(h,1)}} w_{t,a}^{(h)} + 16\sqrt{\frac{S}{T}}
\end{aligned}$$

where (b) follows from the definition that $i_{*,t}^{(h)}$ is optimal active arm in $A_h \supseteq V_{*,t}^{(h)}$, (c) applies property (C) in Lemma 4 and the fact that every arm in $U_t^{(H,0)}$ has width $\le 2^{-H}$, and (d) is by the selection criterion of $U_t^{(h,1)}$ in Algorithm 3 such that $w_{t,a}^{(h)} > 2^{-h}$ for every $a \in U_t^{(h,1)}$. In particular, by the choice of $H = \log(\sqrt{ST})$, Lemma 4 implies

$$\theta_*^\top x_{t,i_{*,t}^{(H)}} - \theta_*^\top x_{t,a} \le 16 \cdot 2^{-H} \le \frac{16}{\sqrt{ST}}$$

for every $a \in U_t^{(H,0)}$.

Then the cumulative regret is bounded as follows.

$$\begin{aligned}
\mathsf{R}(\text{Alg 3}) &\le 16\sqrt{ST} + 16 \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{a\in U_t^{(h,1)}} w_{t,a}^{(h)} \\
&\le 16\sqrt{ST} + 32(1+\beta) \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{a\in U_t^{(h,1)}} \|x_{t,a}\|_{\left(A_t^{(h)}\right)^{-1}}
\end{aligned}$$

where $A_t^{(h)} = I + \sum_{s<t} \sum_{a\in\Phi_t^{(h)}(s)} x_{s,a} x_{s,a}^\top$ is the Gram matrix for stage h, as defined in Algorithm 4. To handle the sum, note that

$$\begin{aligned}
\sum_{a\in U_t^{(h,1)}} \|x_{t,a}\|_{\left(A_t^{(h)}\right)^{-1}} &\le \sqrt{S} \sqrt{\sum_{a\in U_t^{(h,1)}} \|x_{t,a}\|^2_{\left(A_t^{(h)}\right)^{-1}}} \\
&\le \sqrt{S} \sqrt{\sum_{a\in \Phi_{T+1}^{(h)}(t)} \|x_{t,a}\|^2_{\left(A_t^{(h)}\right)^{-1}}}
\end{aligned}$$

where the first inequality is by Cauchy-Schwartz inequality and that $|U_t^{(h,1)}| \le S$, and the second follows from $U_t^{(h,1)} \subseteq \Phi_{T+1}^{(h)}(t)$ as $U_t^{(h,1)}$ is added to the index set at step (1) in Algorithm 3. Finally, we have

$$\begin{aligned}
\mathsf{R}(\text{Alg 3}) &\le 16\sqrt{ST} + 32(1+\beta)\sqrt{S} \sum_{h=1}^{H} \sum_{t=1}^{T} \sqrt{\sum_{a\in \Phi_{T+1}^{(h)}(t)} \|x_{t,a}\|^2_{\left(A_t^{(h)}\right)^{-1}}} \\
&\overset{\text{(c)}}{\le} 16\sqrt{ST} + c_0 \log(ST) \sqrt{\log(2KT)} \sqrt{S} \left(\sqrt{dT} + d\sqrt{S} \right)
\end{aligned}$$

where (c) follows from Lemma 10, with some absolute constant $c_0 > 0$. □

B.3 Minimax Lower Bound

Theorem B.2 (Restatement of theorem 3.2)**.** *Suppose $K \geq 2S$ and $T \geq \max\{4dS, \frac{d^3}{S}\}$. For any policy π, it holds that*

$$\max_{\theta_*, \{x_{t,a}\}} \mathsf{R}(\pi) = \Omega(\sqrt{dST})$$

where the maximum is taken over all problem instances as described in Section 3.1.

Proof. Fix any policy π. We will prove the lower bound by a reduction argument to the combinatorial bandits under graph feedback.

($d > 2S$). First suppose $d > 2S$. WLOG, let $d = 2nS + 1$ for some $n \geq 1$. We partition the horizon $[T]$ into n sub-horizons of equal length $\frac{T}{n}$ (WLOG, suppose $\frac{T}{n}$ is an integer). Specifically, define

$$t_1 = 1, \quad t_{j+i} = t_j + \frac{T}{n}$$

and the j-th sub-horizon be $T_j = [t_j, t_{j+1})$.

Let $e(k) \in \mathbb{R}^d$ denote the one-hot canonical basis $[0, \ldots, 0, \underbrace{1}_{k\text{-th}}, 0, \ldots, 0]^\top$. The context sequence over sub-horizon T_j is defined as follows: for arm $a \in [K]$, let $x_{t,a} = \frac{1}{\sqrt{2}}(e(1) + e(1 + (j-1)2S + a))$ for all $t \in T_j$. Then during T_j, the expected reward of every arm a is $\theta_*^\top x_{t,a} = \frac{1}{\sqrt{2}}\theta_{*,1} + \frac{1}{\sqrt{2}}\theta_{*,1+(j-1)2S+a} =: \frac{1}{\sqrt{2}}\theta_{*,1} + \Delta_a^{(j)}$ which is a constant over this sub-horizon, i.e.

$$\theta_* = \left[\theta_{*,1}, \underbrace{\Delta_1^{(1)}, \ldots, \Delta_{2S}^{(1)}}_{\text{relevant during } T_1}, \underbrace{\Delta_1^{(2)}, \ldots, \Delta_{2S}^{(2)}}_{\text{relevant during } T_2}, \Delta_1^{(3)}, \ldots, \Delta_{2S}^{(n)}\right].$$

The reward $r_{t,a} \sim \mathrm{Bern}(\theta_*^\top x_{t,a})$ is drawn from a Bernoulli distribution.

Consequently, the regret of policy π can be decomposed into the sum of regrets of n independent sub-problems over the n sub-horizons. During each T_j, the learning problem now becomes a combinatorial semi-bandit with Bernoulli rewards. Following the lower bound construction in Theorem 1.3 of Wen (2025), for any $j \in [n]$, there is an instance with $\theta_{*,1} = \frac{1}{2\sqrt{2}}$ and $0 \leq \Delta_a^{(j)} \leq \frac{1}{64}\sqrt{\frac{2S}{ST}} = \frac{1}{64}\sqrt{\frac{2}{T}}$ such that the regret of policy π over T_j is $\Omega(\sqrt{(2S)S|T_j|})$, subject to $|T_j| \geq \frac{(2S)^3}{S}$. The constraint $|T_j| = \frac{(2S)T}{d-1} \geq \frac{(2S)^3}{S}$ is satisfied since $T \geq 4Sd$. Then the total regret over the n sub-horizons is bounded by

$$\mathsf{R}(\pi) \geq \sum_{j=1}^{n} \Omega\left(\sqrt{S^2|T_j|}\right) = \Omega\left(\frac{d-1}{S}\sqrt{S^2\frac{ST}{d-1}}\right) = \Omega\left(\sqrt{(d-1)ST}\right) = \Omega\left(\sqrt{dST}\right)$$

when $d > 2S$. We can verify that the parameter norm is

$$\|\theta_*\|_2^2 = \frac{1}{8} + \frac{d-1}{2048}\frac{1}{T} \leq 1$$

as $T \geq 4Sd \geq \frac{d}{1024}$.

($1 < d \leq 2S$). Now suppose $1 < d \leq 2S$. We instead consider the entire horizon $[T]$ with the following construction. For arm $a \in [d-1]$, let $x_{t,a} = \frac{1}{\sqrt{2}}(e(1) + e(1+a))$ so its expected reward is $\theta_*^\top x_{t,a} = \frac{1}{\sqrt{2}}\theta_{*,1} + \frac{1}{\sqrt{2}}\theta_{*,1+a}$. The reward $r_{t,a} \sim \mathrm{Bern}(\theta_*^\top x_{t,a})$ still follows a Bernoulli distribution. For $a \geq d$, let $r_{t,a} \equiv 0$ by setting $x_{t,a} \equiv 0$. Note that this problem reduces to a combinatorial bandit with graph feedback, and the feedback graph G includes edge $a \to a'$ only when both $a, a' > d-1$ (which share the same mean reward 0). Following the construction in Theorem 1.3 of Wen (2025) again gives a regret lower bound $\Omega(\sqrt{(d-1)ST}) = \Omega(dST)$ subject to $T \geq \frac{(d-1)^3}{S}$.[5]

[5]Note that the lower bound $\Omega(S\sqrt{T})$ in Wen (2025) does not appear in our case. For this part of the lower bound, Wen

($d = 1$). Finally, consider the special case $d = 1$. Since $K \geq 2S$, we can partition the arms into two subsets V_+ and V_0, with $|V_+| \geq S$ and $|V_0| \geq S$. For $a \in V_+$, let the context be $x_{t,a} \equiv \Delta$ for some $\Delta \in (0, \frac{1}{2}]$ to be specified later, and let $x_{t,a} \equiv 0$ for $a \in V_0$. For each arm a, the reward is drawn from a scaled Bernoulli such that

$$r_{t,a} = \begin{cases} 1 & \text{with probability } \frac{1+\theta_* x_{t,a}}{2} \\ -1 & \text{otherwise} \end{cases}.$$

We will consider two environments indexed by $\theta_* \in \{-1, 1\}$. Since the contexts are time-invariant, the expected reward for arm a is $\mu_a = \theta_* x_{t,a} = \mathbb{1}[a \in V_+]\theta_* \Delta$. Let $\mathbb{P}_+$ (resp. $\mathbb{P}_-$) denote the law of the policy π under environment $\theta_* = 1$ (resp. $\theta_* = -1$), and the expectations $\mathbb{E}_+$ and $\mathbb{E}_-$ respectively under each law.

The regret under environment $\theta_* = 1$ is denoted by $\mathsf{R}_+(\pi) \geq \Delta \mathbb{E}_+[N_0]$, and the regret under environment $\theta_* = -1$ is $\mathsf{R}_-(\pi) \geq \Delta \mathbb{E}_-[N_+]$, where $N_0 = \sum_{t=1}^T |V_t \cap V_0|$ and $N_+ = \sum_{t=1}^T |V_t \cap V_+| = ST - N_0$ count the number of pulls in the set V_0 and V_+ respectively. Let θ_* be drawn uniformly from $\{-1, 1\}$. The minimax regret is lower bounded by the Bayes regret

$$\begin{aligned} \max_{\theta_*, \{x_{t,a}\}} \mathsf{R}(\pi) &\geq \frac{1}{2}(\mathsf{R}_+(\pi) + \mathsf{R}_-(\pi)) \\ &\geq \frac{\Delta}{2}(ST - \mathbb{E}_+[N_+] + \mathbb{E}_-[N_+]). \end{aligned} \tag{26}$$

Let $R^T = \{r_{t,a} : t \in [T], a \in [K]\}$ denote the generated reward sequence. By Pinsker's inequality and the chain rule of KL divergence, we have

$$\begin{aligned} \mathbb{E}_+[N_+] - \mathbb{E}_-[N_+] &\leq ST\sqrt{\frac{1}{2}\mathsf{KL}(\mathbb{P}_-(R^T) \| \mathbb{P}_+(R^T))} \\ &= ST\sqrt{\frac{1}{2}\sum_{t=1}^T \sum_{R^{t-1}} \mathbb{P}_-(R^{t-1})\mathsf{KL}(\mathbb{P}_-(r^t|R^{t-1}) \| \mathbb{P}_+(r^t|R^{t-1}))} \end{aligned} \tag{27}$$

where $r^t = \{r_{t,a} : a \in [K]\}$ is the generated rewards at time t. Since the feedback is semi-bandit, the observed distribution only makes a difference when $V_t \cap V_+ \neq \varnothing$. Denote $V_t = \{v_{t,1}, \ldots, v_{t,S}\}$. Thus we have

$$\begin{aligned} \sum_{t=1}^T \sum_{R^{t-1}} \mathbb{P}_-(R^{t-1})\mathsf{KL}\big(\mathbb{P}_-(r^t|R^{t-1}) \| \mathbb{P}_+(r^t|R^{t-1})\big) &= \sum_{t=1}^T \sum_{s=1}^S \mathbb{E}_-[\mathbb{1}[v_{t,s} \in V_+]]\mathsf{KL}\left(\mathrm{Bern}\left(\frac{1}{2}\right) \| \mathrm{Bern}\left(\frac{1+\Delta}{2}\right)\right) \\ &\overset{\text{(a)}}{\leq} \sum_{t=1}^T \sum_{s=1}^S \mathbb{E}_-[\mathbb{1}[v_{t,s} \in V_+]]\frac{4\Delta^2}{3} \\ &\leq \frac{4}{3}\Delta^2 ST \end{aligned}$$

where (a) uses the inequality $\mathsf{KL}(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) \leq \frac{(p-q)^2}{q(1-q)}$ and $\Delta \in (0, \frac{1}{2}]$. Plugging back into (27) gives

$$\mathbb{E}_+[N_+] - \mathbb{E}_-[N_+] \leq ST\sqrt{\frac{4}{3}\Delta^2 ST}.$$

Then we have the lower bound in (26) to be

$$\max_{\theta_*, \{x_{t,a}\}} \mathsf{R}(\pi) \geq \frac{\Delta}{2}\left(ST - ST\sqrt{\frac{4}{3}\Delta^2 ST}\right) = \frac{\Delta ST}{2}\left(1 - \sqrt{\frac{4}{3}\Delta^2 ST}\right).$$

Take $\Delta = 1/\sqrt{12ST} \in (0, \frac{1}{2}]$, we have $\max_{\theta_*, \{x_{t,a}\}} \mathsf{R}(\pi) \geq \frac{\sqrt{ST}}{3}$. This concludes the proof. □

(2025) directly adopts the full information lower bound $\Omega(S\sqrt{T})$ from Koolen et al. (2010), whose construction crucially relies on the fact that no two arms share the same reward mean (at least not to the learner's knowledge). This is however not the case here, as the learner can cluster these one-hot contexts and identify which arms share the same mean. If n arms are known to share the same mean or share the mean up to some known scalars, there are roughly n times more observations for each of those arms, leading to an n times faster concentration.

C CONSTRAINED DECISION SUBSETS IN SECTION 4

As promised in Section 4, we consider a potentially constrained decision subset $\mathcal{A}_0 \subseteq \binom{[K]}{S}$ and another elimination-based algorithm.

Algorithm 5: Combinatorial Arm Elimination (Wen, 2025)

Input: time horizon T, decision subset $\mathcal{A}_0 \subseteq \binom{[K]}{S}$, arm set $[K]$, combinatorial budget S, feedback graph G, and failure probability $\delta \in (0,1)$.
Initialize: Active set $\mathcal{A}_{\text{act}} \leftarrow \mathcal{A}_0$, minimum count $N \leftarrow 0$, time $t \leftarrow 1$.
Let $(\bar{r}_{t,a}, n_{t,a})$ be the empirical reward and the observation count of arm $a \in [K]$ at the end of time t.
For each combinatorial decision $V \in \mathcal{A}_{\text{act}}$, let $\bar{r}_{t,V} = \sum_{a\in V} \bar{r}_{t,a}$ be the empirical reward and $n_{t,V} = \min_{a\in v} n_{t,a}$ be the its observation count.
while $t \le T$ **do**
  Let $\mathcal{A}_N \leftarrow \{V \in \mathcal{A}_{\text{act}} : n_{t,V} = N\}$ be the decisions that have been observed least.
  Let $U_t = \{a \in [K] : \exists V \in \mathcal{A}_N \text{ with } a \in V\} = \bigcup_{V\in\mathcal{A}_N} V$.
  Find a minimal set of decisions $\{V_1, \dots, V_n\} \subseteq \mathcal{A}_0$ such that $U_t \subseteq \cup_{m=1}^{n} N_{\text{out}}(V_m)$.
  for $m = 1$ **to** n **do**
    Select decision V_m.
    Collect feedback $\{r_{t,a} : a \in N_{\text{out}}(V_m)\}$ and update $(\bar{r}_{t,a}, n_{t,a})$ accordingly.
    Update time $t \leftarrow t + 1$.
  Note by the choice of $\{V_1, \dots, V_n\}$, now $\min_{V\in\mathcal{A}_{\text{act}}} n_{t,V} > N$.
  Update the minimum count $N \leftarrow \min_{V\in\mathcal{A}_{\text{act}}} n_V^t$.
  Let $\bar{r}_{t,\max} \leftarrow \max_{V\in\mathcal{A}_{\text{act}}} \bar{r}_{t,V}$ be the maximum empirical reward in the active set.
  Update the active set as follows:

$$\mathcal{A}_{\text{act}} \leftarrow \left\{ V \in \mathcal{A}_{\text{act}} : \bar{r}_{t,V} \ge \bar{r}_{t,\max} - 2S\sqrt{\frac{\log(2KT/\delta)}{N}} \right\}.$$

Before presenting the regret guarantee, we remark that finding a minimal set of decisions $\{V_1, \dots, V_n\}$ that covers the set U_t in Algorithm 5 can be computationally hard in general. Therefore, the final algorithm can be inefficient.

Theorem C.1 (Restatement of Theorem 4.1)**.** *Fix any $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\mathsf{R}(\mathsf{Alg}\ 5) = O\left(\log(2KT/\delta)\frac{S^2\kappa}{\Delta_*}\right)$$

where κ is defined as in (2) *and Δ_* is the reward gap between the optimal and the second optimal decisions in $\mathcal{A}_0$.*

In particular, there is a subset $\mathcal{A}_0$ under which

$$\min_{\pi}\max_{\nu} \mathsf{R}_\nu(\pi) = \Omega\left(\log(T)\frac{S^2\kappa}{\Delta_*}\right)$$

where the maximum is taken over all bandit environments ν.

Proof. WLOG, suppose the confidence bound in Lemma 1 holds. Let N^t denote the minimum count N at the end of time t, and $w(n) \coloneqq \sqrt{\log(2KT/\delta)/n}$ denote the confidence width. For a decision $V \in \mathcal{A}_0$, define its expected reward as $\mu_V = \sum_{a\in V} \mu_a$. Then we have the following two observations:

(1) The optimal decision $V_* \in \mathcal{A}_{\text{act}}$ always remains uneliminated. This follows from

$$\bar{r}_{t,V_*} \ge \mu_{V_*} - Sw(N^t) \ge \mu_V - Sw(N^t) \ge \bar{r}_{t,V} - 2Sw(N^t)$$

for any other decision $V \in \mathcal{A}_{\text{act}}$.

(2) At any time t, any active decision $V \in \mathcal{A}_{\text{act}}$ satisfies $\mu_{V_*} - \mu_V \le 4Sw(N^t)$. This follows from

$$\mu_{V_*} \le \bar{r}_{t,V_*} + Sw(N^t) \le \bar{r}_{t,V} + 3Sw(N^t) \le \mu_V + 4Sw(N^t)$$

for any $V \in \mathcal{A}_{\text{act}}$ by the elimination criterion.

Define a layer $L_n \coloneqq \{V \in \mathcal{A}_0 : \text{ V is selected when the minimum count is } N^t = n\} \subseteq \mathcal{A}_0 \subseteq \binom{[K]}{S}$. Note that when $4Sw(N^t) < \Delta_*$, the active set contains only the optimal decision, so

$$\max\{n \ge 1 : |L_n| > 1\} \le \frac{16S^2 \log(2KT/\delta)}{\Delta_*^2} =: M.$$

We have $|L_n| \le \kappa$ by the choice of V_t in Algorithm 5. Then the cumulative regret is bounded as

$$\begin{aligned}
\mathsf{R}(\mathsf{Alg}\ 5) &\le S|L_0| + \sum_{n=1}^{M} \sum_{V \in L_n} (\mu_{V_*} - \mu_V) \\
&\le S\kappa + 4S \sum_{n=1}^{M} |L_n| w(n) \\
&\le S\kappa + 8S\kappa\sqrt{\log(2KT/\delta)}\sqrt{M} \\
&= S\kappa + 32 \log(2KT/\delta) \frac{S^2\kappa}{\Delta_*}.
\end{aligned}$$

For the second claim, Kveton et al. (2015) shows the following in their Proposition 1. Under a subset $\mathcal{A}_0$ that partitions the K arms, i.e. $\forall V_1, V_2 \in \mathcal{A}_0, V_1 \cap V_2 = \varnothing$ and $\sum_{V \in \mathcal{A}_0} |V| = \Omega(K)$, the minimax regret is $\Omega(\log(T)KS/\Delta_*)$. Under this subset, we have $\kappa = \Omega(K/S)$ and hence $S^2\kappa = \Omega(SK)$ recovers the minimax lower bound. □

D AUXILIARY LEMMATA

We first define the *dominating number* as a relevant graph-theoretic quantity. For any directed graph $G = (V, E)$, the dominating number is defined by the cardinality of the smallest subset whose out-neighbor covers the entire graph:

$$\delta(G) \coloneqq \min\{|D| : D \subseteq V \text{ and } V \subseteq N_{\text{out}}(D)\}. \tag{28}$$

For any directed graph $G = (V, E)$, consider the following greedy approximation algorithm that finds a dominating subset: one starts with an empty set $D \leftarrow \varnothing$, recursively finds an "unobserved" node $v \notin N_{\text{out}}(D)$ that has the largest out-degree in the remaining subgraph on $V \backslash N_{\text{out}}(D)$, and add it to the set $D \leftarrow D \cup \{v\}$. The following result characterizes the performance of this well-known approximation.

Lemma 11 (Chvatal (1979))**.** *For any directed graph $G = (V, E)$, the above greedy algorithm gives a dominating subset $D \subseteq V$ that satisfies*

$$|D| \le (1 + \log|V|)\delta(G).$$

Lemma 12 (Lemma 8 in Alon et al. (2015))**.** *For any directed graph $G = (V, E)$, it holds that $\delta(G) \le 50 \log|V| \alpha(G)$.*