Online Learning under Partial Feedback

by

Yifan Wu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

In an online learning problem a player makes decisions in a sequential manner. In each round, the player receives some reward that depends on his action and an outcome generated by the environment while some feedback information about the outcome is revealed. The goal of the player can be various.

In this thesis we investigate several variants of online learning problems with different feedback models and objectives. First we consider the pure exploration problem with multi-action probes. We design algorithms that can find the best one or several actions with high probability while using as few probes as possible. Then we study the side observation model in the regret minimization scenario. We derive a novel finite time distribution dependent lower bound and design asymptotically optimal and minimax optimal algorithms. Last we investigate the conservative bandit problem where the objective is to minimize the regret while maintaining the cumulative reward above a baseline. We design algorithms for several variants of the problem and derive a lower bound.

In each of the three variants of the online learning problem we consider, our problem setting generalizes some previous work. The theoretical results successfully recover existing results in special cases as well as propose novel perspectives in the more general settings.

Preface

Chapter 2 and Chapter 3 are joint works with András György and Csaba Szepesvári, and were published in Wu et al. (2015a) and Wu et al. (2015b). Chapter 4 is a joint work with Roshan Shariff, Tor Lattimore and Csaba Szepesvári, and will be published in Wu et al. (2016).

Acknowledgements

I sincerely thank my supervisors, Csaba Szepesvári and András György, for their encouragement and patient support on my research. I am also grateful to my collaborators, Roshan Shariff and Tor Lattimore, for working together on this work. Furthermore I would like to thank my thesis committee chair Micheal Bowling for his comments on how to improve the thesis. Finally I would like to thank my parents and the University of Alberta for supporting my study during the master program.

Table of Contents

List of Figures

Chapter 1

Introduction

In this chapter we first introduce the online learning framework. Next we present different learning objectives along with more general feedback models that generalize the full-information and the bandit setting. Then we summarize the contributions that will be presented in the following chapters of this thesis.

1.1 Online learning

In an online learning problem, a player (or learner) needs to interact with the environment in a round-by-round case. In each round the player makes a decision and the environment generates an outcome. After the player takes the action some feedback information about the outcome is revealed then the player can make a decision for the next round. There are two types of environments: stochastic and non-stochastic. In stochastic environments the outcome is generated from some probability distribution while in non-stochastic environments there is no probabilistic assumption about the outcome, which may be generated in an adversarial way.

The formulation of an online learning problem typically includes an environment, an action space, a feedback model and a learning objective. In this thesis we mainly focus on environments that have a finite set of K options, which are referred to *learning with expert advice* in the literature. In this framework, there are two basic feedback models: the *full-information* setting

and the *multi-armed bandit* [1] setting. In both settings the action space is the same as the set of options[2]: in each round, the player picks an option and observes some feedback about the outcome of the environment. In the full-information setting, the player can observe the outcome associated with each of the options while in the bandit setting the player can only observe the outcome associated with the option that is picked in that round. In the next section we will talk about different types of learning objectives and introduce some more general feedback models.

1.2 Different objectives

In this section we will first introduce two different learning objectives — pure exploration and regret minimization as well as feedback models that generalize the full-information and the bandit setting. Furthermore we will introduce *conservative bandits* which aim at minimizing the regret under some additional constraint.

1.2.1 Pure exploration

In pure exploration problems the player aims at extracting information about the environment regardless of the reward/loss incurred during the process. A most basic pure exploration problem is the *best arm identification* problem in the stochastic multi-armed bandit setting, where the goal is to find the option with highest reward mean with high probability. The history of the best arm identification problem goes back more than half a century (Bechhofer, 1958; Paulson, 1964), and with much activity in the last decade (e.g. Even-Dar et al., 2002; Mannor and Tsitsiklis, 2004; Audibert et al., 2010; Kalyanakrishnan and Stone, 2010; Bubeck et al., 2011; Kalyanakrishnan et al., 2012; Gabillon et al., 2012; Karnin et al., 2013; Kaufmann and Kalyanakrishnan, 2013; Bubeck et al., 2013; Jamieson et al., 2014; Kaufmann et al., 2015a; Zhou

[1] In the rest of this thesis we will mostly use *bandit* instead of *multi-armed bandit* for simplicity.

[2] When the action space is the same as the set of options we will simply use the term "action" for an option.

et al., 2014).

Multi-option probes

In addition to finding the best option in the standard bandit setting, some of the recent work also studies other variants of objectives such as finding the best multiple options and different settings such as the combinatorial setting (Chen et al., 2014; Gabillon et al., 2016). In this thesis we will present our work on the *multi-option probe* setting where the goal is to find the best one or multiple option(s) by using as few multi-option probes as possible.

The motivation of the multi-option probe setting is as follows: Consider the problem of identifying the most rewarding option(s) out of finitely many. At your disposal are a number of probing devices, or just *probes*, that give you noisy measurements of the quality of a select set of options. More precisely, each *probe* is associated with a *known subset* of options whose quality the probe will measure. In a sequential process, the goal is to select the probes so that one can stop early to return, with high probability, a sufficiently rewarding option (or a set of options). As a specific example, consider the problem of identifying the segment on a road network that is in the worst shape after a long winter. Measurements can be obtained by sending trucks checking the road for potholes along the paths they travel on. The trucks must return to their garage every day. Here, the options correspond to road segments, the probes correspond to a closed walk in the road network that starts from the garage. Somewhat ironically, a road segment is "rewarding" (from the point of view of how beneficial it is to sending there the repair team) if it has many potholes. [3] Measurements are noisy, as potholes are easy to miss.

Problems like the above one abound. Numerous quality assurance and surveying tasks are such that measurements give simultaneous information about multiple entities due to physical constraints on the measurement process. Application areas include technical computing (e.g., networking), biology (ecology, microbiology, etc.), physics, etc. Of course, even though individual

[3] In practice, one may want a whole "plan" at the end for the repair team. As often, we took the liberty of simplifying the problem to be able to focus on how the structure of probes should be used.

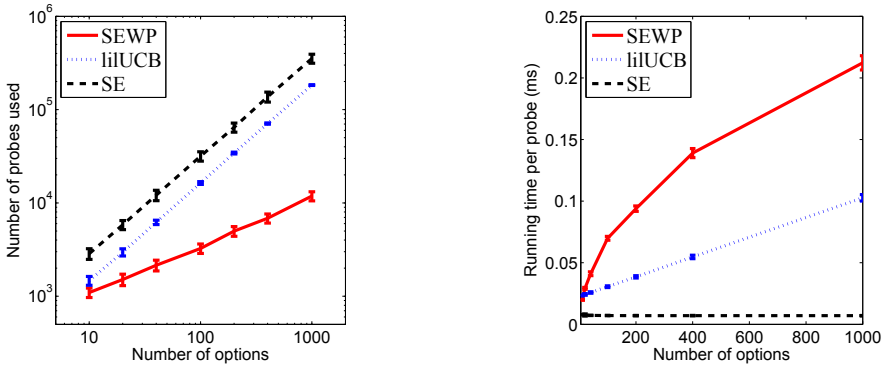

Figure 1.1: A specialized algorithm (SEWP) proposed in this thesis can take nontrivial advantage of the probe structure as compared with simple adaptations of earlier algorithms, while being only marginally more expensive. All algorithms maintain the same error-rate. The plot on the left-hand-side uses a log-log-scale. Due to the special structure of the problem, the expected stopping time of the specialized algorithm scale linearly with $\sqrt{K}$, while the others scale linearly with K, the number of options.

measurements might be impossible, it is always possible to treat each probe as one that gives individual measurements for the options associated with it, though this could be wasteful (cf. Fig. 1.1). The main topic of this part of work is how to exploit, with efficient algorithms, when probes give information about multiple options.

Compared to our multi-option probe setting, Chen et al. (2014) and Gabillon et al. (2016) study a "reversed" setting where the goal is to find the best "probe" by pulling "options". After our work is published, Jun et al. (2016) studies a similar scenario where options can be experimented in "batches". The difference between their work and ours is that they allow repeated options in a "batch" while their "batch space" is more restricted than ours (e.g. of the same size, contains any combination).

(In Figure 1.1, the *lilUCB* algorithm comes from Jamieson et al. (2014). The parameters we used in experiments is the *lilUCB Heuristic* setting, which performs the best in the experiments of Jamieson et al. (2014). The *SE* algorithm is short for the *successive elimination* algorithm of Even-Dar et al. (2002). As these algorithms select options for measurements, we adapt them

to the probe setting by choosing the first probe in some arbitrary ordering of probes that gives a measurement for the selected option . In experiments, all distributions we used are Gaussian with variance 1/4. Each point reported in the figure is based on 100 repeated experiments under the same reward distributions, where we set one of the means to be 0.5 and the others to be 0.)

1.2.2 Regret minimization

In regret minimization problems the player receives some reward, or, interchangeably, payoff, in each round after taking an action. The goal is to maximize the cumulative reward during the learning process. The performance of algorithms is defined in terms of cumulative *regret*: the difference between the total reward received by the player and that when the player constantly takes some "best" action. Lower regret means better performance.

In both full-information and bandit setting the reward of taking an action is just the outcome associated with that action: In the full information setting the player observes the reward of all possible actions at the end of every round. In the bandit setting the player only observes its own reward and receives no information about the reward of other actions (Bubeck and Cesa-Bianchi, 2012).

Graph-structured feedback

Recently, several papers considered a more refined setup, called graph-structured feedback, that interpolates between the full-information and the bandit case: here the feedback structure is described by a (possibly directed) graph, and choosing an action reveals the payoff of all actions that are connected to the selected one, including the chosen action itself. This problem, motivated for example by social networks, has been studied extensively in both the adversarial (Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014; Alon et al., 2015) and the stochastic cases (Caron et al., 2012; Buccapatnam et al., 2014). However, most algorithms presented heavily depend on the self-observability assumption, that is, that the payoff of the selected action can be observed.

Removing this self-loop assumption leads to the so-called partial monitoring case (Alon et al., 2015). In the absolutely general partial monitoring setup the learner receives some general feedback that depends on its choice (and the environment), with some arbitrary (but known) dependence (Cesa-Bianchi and Lugosi, 2006; Bartók et al., 2014). While the partial monitoring setup covers all other problems, its analysis has concentrated on the finite case where both the set of actions and the set of feedback signals are finite (Cesa-Bianchi and Lugosi, 2006; Bartók et al., 2014), which is in contrast to the standard full information and bandit settings where the feedback is typically assumed to be real-valued. To our knowledge there are only a few exceptions to this case: in Alon et al. (2015), graph-structured feedback is considered without the self-loop assumption, while continuous action spaces are considered in Lin et al. (2014) and Lattimore et al. (2014) with special feedback structure (linear and censored observations, resp.).

Gaussian side observations

In this thesis we consider a generalization of the graph-structured feedback model, called the *Gaussian side observation model*, which can also be viewed as a general partial monitoring model with real-valued feedback. In the Gaussian side observation model, after selecting an action i, the learner receives information about the payoff of every action j in the form of Gaussian observations whose mean is the same as the mean payoff, but the variance depends on the pair (i, j) (and may be infinite). The setup allows a more refined information transfer from one action to another than previous partial monitoring setups, including the recently introduced graph-structured feedback case.

After our work is published, Kocák et al. (2016) generalized the graph-structured feedback setting in a similar fashion by allowing different level of noise in observations. Compared with our work, they are studying a non-stochastic payoff setting and presenting different type of regret bounds.

1.2.3 Conservative bandits

In this thesis we also study a variant of the regret minimization problem in the multi-armed bandit setting called *conservative bandits*, where the goal is to minimize the cumulative regret under some additional constraint (lower bound) on the cumulative reward over time. This problem is motivated by the challenge faced by a company wishing to explore new strategies to maximize revenue whilst simultaneously maintaining their revenue above a fixed baseline, uniformly over time. For example, the manager of Zonlex, a fictional company, has just learned about bandit algorithms and is very excited about the opportunity to use this advanced technology to maximize Zonlex's revenue by optimizing the content on the landing page of the company's website. Every click on the content of their website pays a small reward; thanks to the high traffic that Zonlex's website enjoys, this translates into a decent revenue stream. Currently, Zonlex chooses the website's contents using a strategy designed over the years by its best engineers, but the manager suspects that some alternative strategies could potentially extract significantly more revenue. The manager is willing to explore bandit algorithms to identify the winning strategy. The manager's problem is that Zonlex cannot afford to lose more than 10% of its current revenue during its day-to-day operations and *at any given point in time*, as Zonlex needs a lot of cash to support its operations. The manager is aware that standard bandit algorithms experiment "wildly", at least initially, and as such may initially lose too much revenue and jeopardize the company's stable operations. As a result, the manager is afraid of deploying cutting-edge bandit methods, but notes that this just seems to be a chicken-and-egg problem: a learning algorithm cannot explore due to the potential high loss, whereas it must explore to be good in the long run.

The problem described in the previous paragraph is ubiquitous. It is present, for example, when attempting to learn better human-computer interaction strategies, say in dialogue systems or educational games. In these cases a designer may feel that experimenting with sub-par interaction strategies could cause more harm than good (Rieser and Lemon, 2008; Liu et al.,

2014). Similarly, optimizing a production process in a factory via learning (and experimentation) has much potential (Gabel and Riedmiller, 2011), but deviating too much from established "best practices" will often be considered too dangerous. For examples from other domains see the survey paper of García and Fernández (2015).

Our constraint here is equivalent to a constraint on the regret to a default strategy, or in the language of prediction-with-expert-advice, or bandit literature, regret to a default action. In the full information setting, mostly studied in the adversarial setting, much work has been devoted to understanding the price of such constraints (Hutter and Poland, 2005; Even-Dar et al., 2008; Koolen, 2013; Sani et al., 2014). In particular, Koolen (2013) studies the Pareto frontier of regret vectors (which contains the non-dominated worst-case regret vectors of all algorithms). The main lesson of these works is that in the full information setting even a constant regret to a fixed default action can be maintained with essentially no increase in the regret to the best action. The situation quickly deteriorates in the bandit setting as shown by Lattimore (2015a). This is perhaps unsurprising given that, as opposed to the full information setting, in the bandit setting one needs to actively explore to get improved estimates of the actions' payoffs. Lattimore (2015a) describes two learning algorithms relevant to our setting: In the stochastic setting we consider, Unbalanced MOSS (and its relative, Unbalanced UCB) are able to achieve a constant regret penalty while maintaining the return constraint while Exp3-γ achieves a much better regret as compared to our strategy for the adversarial setting. However, *neither of these algorithms maintain the return constraint uniformly in time*. Neither will the constraint hold with high probability. While Unbalanced UCB achieves problem-dependent bounds, it has the same issues as Unbalanced MOSS with maintaining the return constraint. Also, all these strategies rely heavily on knowing the payoff of the default action.

More broadly, the issue of staying safe while exploring has long been recognized in reinforcement learning (RL). García and Fernández (2015) provides a comprehensive survey of the relevant literature. Lack of space prevents us

from including much of this review. However, the short summary is that while the issue has been considered to be important, no previous approach addresses the problem from a theoretical angle. Also, while it has been recognized that adding constraints on the return is one way to ensure safety, as far as we know, maintaining the constraints during learning (as opposed to imposing them as a way of restricting the set of feasible policies) has not been considered in this literature. Our work, while it considers a much simpler setting, suggest a novel approach to address the safe exploration problem in RL. Another line of work considers safe exploration in the related context of optimization (Sui et al., 2015). However, the techniques and the problem setting (e.g., objective) in this work is substantially different from ours.

1.3 Summary of contributions

Chapter 2 is based on the work of Wu et al. (2015a). We investigate the pure exploration problem in the multi-option probe setting: In each round, a subset of the options, from an available set of subsets, can be selected to receive noisy information about the quality of the options in the chosen subset. The goal is to identify the highest quality option, or a group of options of the highest quality, with a small error probability, while using the smallest number of measurements. The problem generalizes best-arm identification problems. By extending previous work, we design new algorithms that are shown to be able to exploit the combinatorial structure of the problem in a nontrivial fashion, while being unimprovable in special cases. The algorithms call a set multi-covering oracle, hence their performance and efficiency is strongly tied to whether the associated set multi-covering problem can be efficiently solved.

Chapter 3 is based on the work of Wu et al. (2015b): We investigate the regret minimization problem with Gaussian side observations in stochastic environments: For the first time in the literature, we provide non-asymptotic problem-dependent lower bounds on the regret of any algorithm, which recover existing asymptotic problem-dependent lower bounds and finite-time minimax lower bounds available in the literature. We also provide algorithms that

achieve the problem-dependent lower bound (up to some universal constant factor) or the minimax lower bounds (up to logarithmic factors).

Chapter 4 is based on the work of Wu et al. (2016): We study the conservative bandit problem. We consider both the stochastic and the adversarial settings, where we propose natural yet novel strategies and analyze the price for maintaining the constraints. Amongst other things, we prove both high probability and expectation bounds on the regret, while we also consider both the problem of maintaining the constraints with high probability or expectation. For the adversarial setting the price of maintaining the constraint appears to be higher, at least for the algorithm considered. A lower bound is given showing that the algorithm for the stochastic setting is almost optimal. Empirical results obtained in synthetic environments complement our theoretical findings.

Chapter 2

Pure Exploration with Multi-option Probes

In this chapter we present our work on the pure exploration problem in the multi-option probe setting (Wu et al., 2015a). We consider two basic settings: identifying the best option with a prespecified error probability while using the smallest possible number of probes, and identifying a group of options of a fixed size, again with a prespecified error probability with the smallest possible number of probes. For the first setting, we propose two algorithms, SEWP and EGEWP described in Section 2.2, extending the works of Even-Dar et al. (2002) and Karnin et al. (2013). They work by constructing coverings with the probes of the sets of options not eliminated. The second algorithm removes a logarithmic term from the upper bound and it required a non-trivial extension of the median elimination method of Even-Dar et al. (2002). For the second setting, in Section 2.3, the quality of a group returned is assessed either by the quality of the worst option in the group (following Kalyanakrishnan and Stone (2010)), or by the average quality of options in the group (Zhou et al., 2014). We propose a single algorithm (SARWP) that essentially covers both cases. For the average quality, our distribution dependent upper bound is novel even in the bandit case and also near optimal in the worst case compared with the lower bound proposed by Zhou et al. (2014). For simple probe structures (singletons, or when a probe that covers all options is available), our algorithms are shown to be essentially unimprovable. We also give lower bounds for general probe structures. While both our lower and upper bounds express

how the structure of the probes interferes with the structure of payoffs, they differ in subtle ways and it remains for future work to see whether there is a gap between them.

2.1 Preliminaries

In this section, we formulate the problem studied, as well as introducing the set covering problem, which will play an important role in our algorithms and analysis. We start by defining some notation.

2.1.1 Notation

The set of natural numbers will be denoted by $\mathbb{N}$, which includes zero. For a positive natural number n, $[n]$ denotes the set of integers between 1 and n: $[n] = \{1, \ldots, n\}$. The power set, i.e., the set of all subsets of a set S, will be denoted by 2^S. As usual, functions, mapping set X to set Y will be viewed as elements of Y^X. For $v \in Y^X$, we will often write v_x instead of $v(x)$ to minimize clutter. This also helps with the next convention: When $U \subset X$, we will use v_U to denote the restriction of $v \in Y^X$ to U: $v_U(u) = v(u)$, $u \in U$. We identify $Y^{[n]}$ with Y^n (the set of n-tuples) in the natural way, which allows us to use notation v_U for $v \in Y^n \equiv Y^{[n]}$. The cardinality of a set S is denoted by $|S|$. Certain symbols will be reserved to denote elements of certain sets (i.e., p will always be an element of set $\mathcal{P}$). When using such reserved symbols, we will abbreviate (e.g.) $\sum_{p \in \mathcal{P}} f(p)$ to $\sum_p f(p)$. We will use $\log(\cdot)$ to denote the natural logarithm function.

2.1.2 Problem Formulation

A decision maker is given a pair $([K], \mathcal{P})$, where elements of $[K]$ are called options, and $\mathcal{P} \subset 2^{[K]}$ such that the sets in $\mathcal{P}$ cover $[K]$: $\cup\mathcal{P} = [K]$. Elements of $\mathcal{P}$ are called *probes*. A problem instance D, or *environment*, is specified by K distributions over the reals, $D = (D_1, \ldots, D_K)$. The decision maker does not have direct access to these distributions. For $1 \leq i \leq K$, we think of distribution D_i as the distribution of "rewards" associated with option i. We

assume that the mean reward $\mu_i = \int x D_i(dx)$ of each option is well defined. Further assumptions on D_i will be given later.

The goal of the decision maker is to find options with the largest mean reward. For this, the decision maker can query the rewards of the options by using the probes in a sequential manner. In particular, for each round $t = 1, 2, \ldots$, first a random reward $X_{t,i} \sim D_i$ is generated for each option i from its associated distribution. It is assumed that $X_{t,i}$ is independent of the other rewards $(X_{s,j})_{s \neq t \text{ or } j \neq i}$. We set $X_t = (X_{t,1}, \ldots, X_{t,K}) \in \mathbb{R}^K$. In round $t = 1, 2, \ldots$, the decision maker chooses a probe $p_t \in \mathcal{P}$ based on her past observations, to observe the values $X_{t,i}$ for each option i in p_t; with our earlier introduced notation we can write that the decision maker observes $X_{t,p_t} \doteq (X_t)_{p_t} \in \mathbb{R}^{p_t}$. At the end of each round, the decision maker can decide between continuing or stopping to return a list of guesses (or a single guess) on the indices of the good options. The goal is to stop as soon as possible, while avoiding poor guesses.

The following specific problem settings will be considered:

(i) *Fixed confidence, best-option identification.* The *optimal option* is unique: If $\mu^\star = \max_{i \in [K]} \mu_i$, $\max_{i:\mu_i \neq \mu^\star} \mu_i < \mu^\star$. The goal of the decision maker is to identify the index $i^\star = \operatorname{argmax}_{i \in [K]} \mu_i$ of the optimal option. The decision maker is given a *confidence* parameter $0 \leq \delta < 1$ and it is required that the guess returned after τ probes must be correct on an event $\mathcal{E}$ with probability at least $1 - \delta$. Decision makers are compared based on their *probe complexity*, i.e., the number of probes they use when the "good event" $\mathcal{E}$ happens.

(ii) *PAC subset selection.* There are two subproblems that we consider. In both cases the decision maker is given a confidence, $0 \leq \delta < 1$, a suboptimality threshold $\varepsilon > 0$ and a subset cardinality $1 \leq m \leq K$. The problems differ in how a quality $q(S, \mu)$ measure is assigned to a subset $S \subset [K]$ of options. In both problems, the goal is to find a subset of options of cardinality m such that $q(S, \mu) \geq \max_{P \subset [K]:|P|=m} q(P, \mu) - \varepsilon$ and with probability $1 - \delta$, the decision maker must return a subset satisfying

the above quality constraint. As before, decision makers are compared based on how many probes they use before stopping. The two quality measures considered are the reward of the worst option in the set and the average reward: $q_{\min}(S,\mu) = \min_{i\in S}\mu_i$ and $q_{\text{avg}}(S,\mu) = \frac{1}{|S|}\sum_{i\in S}\mu_i$, $S \subset [K]$, $|S| = m$. We call the corresponding problems the *strong* and the *average* PAC subset selection problems.

An algorithm used by a decision maker to select probes, stop and return a guess will be said to be *admissible* with respect to a class of environments, if, for *any* environment within the class and any $0 \le \delta < 1$, the guess computed is correct (according to the previous requirements) with probability $1-\delta$.

The above problems have been considered in the past in the special case when $\mathcal{P}$ contains singletons only, by a number of authors (see Section 1.2.1 for some references). We shall call these the "bandit" problems. While one can readily apply the algorithms developed for the bandit case to our problem, the expectation is that the probe complexity of reasonable algorithms should improve considerably as $\mathcal{P}$ becomes "richer" (this was illustrated in Fig. 1.1). The question is how the structure of $\mathcal{P}$ together with the problem instance influences the problem complexity. For example, in the extreme case when $\mathcal{P}$ contains $[K]$, we expect the probe complexity of reasonable algorithms to scale sublinearly with K, whereas in the bandit case a linear scaling is unavoidable. The case when $\mathcal{P} = \{[K]\}$ will be called the *full information case*.

Note that since all probes "cost" the same amount (one unit of time), a reasonable algorithm will avoid any probe p that is entirely included in some other probe $p' \in \mathcal{P}$. Hence, we may as well assume that the set of probes does not have nontrivial chains in it.

We will present results for the class of environments $\mathcal{D}_{\text{sg}}$ with the following restrictions: For each $1 \le i \le K$, D_i is sub-Gaussian with common parameter $\sigma^2 = 1/4$:

$$\log \int_{\mathbb{R}} e^{-\lambda(x-\mu_i)} D_i(dx) \le \lambda^2\sigma^2/2 = \lambda^2/8$$

for all $\lambda \in \mathbb{R}$. To simplify the presentation of our results, without loss of generality, we *assume that* $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_K$. (note that, obviously, the

algorithms do not use this assumption). For further simplicity, we assume that $\Delta_i \in [0,1]$ for all $i \in [K]$ where $\Delta_i = \mu_1 - \mu_i$, $2 \le i \le K$. Our assumptions on the reward distributions D_i are satisfied if, for example, D_i has bounded support.

We will present algorithms, which will be shown to be admissible for $\mathcal{D}_{\mathrm{sg}}$ and we will bound their probe complexities. The bounds on the probe complexities will be given in terms of the *(suboptimality) gaps* Δ_i, $2 \le i \le K$, i.e., they will be dependent on the distributions $D = (D_1, \ldots, D_K)$. Hence, we call them distribution dependent bounds. We will accompany our constructive results with lower bounds, putting a lower limit on the probe complexity of all admissible algorithms. Again, these will be given in terms of the gaps Δ_i.

2.1.3 Set Multi-Cover Problems

Probes allow one to "explore" multiple options simultaneously. Clever algorithms should use the probes in a smart way to guarantee the necessary number of samples for each of the options while using the smallest number of probes. If, for example, $n \in \mathbb{N}$ observations are enough from each of the options to distinguish their mean payoff from that of the optimal option, then an intelligent algorithm would try to create the smallest *covering* of $[K]$ using the subsets in $\mathcal{P}$ to meet this requirement. More generally, for $J \subset [K]$, we define

$$\min\left\{ \textstyle\sum_p s_p : s \in \mathbb{N}^{\mathcal{P}}, \sum_{p:i\in p} s_p \ge n, i \in J \right\}$$

to be the *cost* of the smallest n-fold multi-covering of elements of J. Any $s \in \mathbb{N}^{\mathcal{P}}$ achieving the minimum is called an *optimal (integral) n-cover* of J, while a feasible vector s is called an n-cover. Given an n-cover $s \in \mathbb{N}^{\mathcal{P}}$, we will say that probe p belongs to s (writing $p \in s$) if $s_p > 0$. The optimization problem defining $\mathcal{C}_{\mathrm{IP}}$ is a linear integer program (hence the IP in $\mathcal{C}_{\mathrm{IP}}$). Relaxing the *integrality* constraint $s \in \mathbb{N}^{\mathcal{P}}$ to the nonnegativity constraint $s \in [0,\infty)^{\mathcal{P}}$, we get a so-called *fractional optimal n-cover* of J by solving the otherwise identical optimization problem. The resulting optimal value will be denoted by $\mathcal{C}_{\mathrm{LP}}(J,n)$. Note that the relaxed problem is a linear program, explaining

"LP" in $\mathcal{C}_{\mathrm{LP}}$. While this linear program has potentially exponentially many variables in K, it can still be efficiently solved provided an efficiently computable membership oracle is available for its dual (Grötschel et al., 1993). Both $\mathcal{C}_{\mathrm{IP}}(J,n)$ and $\mathcal{C}_{\mathrm{LP}}(J,n)$ can be extended to non-integer values of n.

It follows immediately from the definitions that $\mathcal{C}_{\mathrm{LP}}(J,n) \leq \mathcal{C}_{\mathrm{IP}}(J,n)$. Further, for any $a > 0$, $\mathcal{C}_{\mathrm{LP}}(J, a\,n) = a\,\mathcal{C}_{\mathrm{LP}}(J,n) = an\,\mathcal{C}_{\mathrm{LP}}(J,1)$. The *integrality gap* for a set multi-covering problem instance is given by $(\mathcal{P}, J, n)$ is $\mathcal{C}_{\mathrm{IP}}(J,n)/\mathcal{C}_{\mathrm{LP}}(J,n)$ (Vazirani, 2001).

Our algorithms will need "small" n-covers for various subsets $J \subset [K]$. Depending on the structure of $\mathcal{P}$, calculating an optimal multi-cover of J may be easy or hard [1] (Slavik, 1998; Schrijver, 2003; Korte and Vygen, 2006). Thus, to keep the presentation general, our algorithms will rely on a set multi-covering *oracle* COrcl, which given $J, n, \mathcal{P}$, returns an n-fold multi-cover of J using the sets in $\mathcal{P}$. Denote by $\mathcal{C}_{\mathrm{O}}(J,n)$ the cost of the multi-cover returned by the oracle on J, n (as with $\mathcal{C}_{\mathrm{IP}}$ and $\mathcal{C}_{\mathrm{LP}}$ the dependence on $\mathcal{P}$ is suppressed). The oracle's integral (fractional) approximation gap, $\mathcal{G}_{IP}(O,\mathcal{P})$ ($\mathcal{G}_{LP}(O,\mathcal{P})$), is the worst-case multiplicative loss due to using COrcl in place of an optimal integral (fractional) cover. In particular, with $\star \in \{IP, LP\}$,

$$\mathcal{G}_{\star}(O,\mathcal{P}) = \sup_{n\in\mathbb{N}^{+}, J\subset[K]} \frac{\mathcal{C}_{\mathrm{O}}(J,n)}{\mathcal{C}_{\star}(J,n)}\,.$$

Let $d = \max_{p\in\mathcal{P}} |p|$ be the maximum number of actions that can be covered by a single probe. If the set-system $\mathcal{P}$ has no special structure, one possibility is to use the greedy algorithm G as the oracle. This algorithm works by sequentially setting $s_p = n$ for the probe $p \in \mathcal{P}$ that covers the maximum number of active options in J and then deactivates the options that are covered by p, until all options are deactivated. Further, $\mathcal{G}_{LP}(O,\mathcal{P}) \leq 1 + \log(d) \leq 1 + \log(K)$. Lovász (1975) showed that $\mathcal{C}_{\mathrm{G}}(J,1) \leq (1+\log d)\mathcal{C}_{\mathrm{LP}}(J,1)$. Then, $\mathcal{C}_{\mathrm{G}}(J,n) = n\,\mathcal{C}_{\mathrm{G}}(J,1) \leq (1+\log d) n\,\mathcal{C}_{\mathrm{LP}}(J,1) = (1+\log d)\mathcal{C}_{\mathrm{LP}}(J,n)$, showing that the required inequality indeed holds. Raz and Safra (1997) proved that the exists some constant $c > 0$ such that, unless $P = NP$, no approximation ratio of

[1] Computing the exact solution for the decision version of set covering (i.e., when $n = 1$), when $\mathcal{P}$ can be any covering system, is known to be NP-hard (Vazirani, 2001).

$c \log(K)$ can be achieved, so in a worst-case the greedy algorithm is a near-optimal approximation algorithm.

2.2 Finding the Best Option

In this section we present two algorithms and their analysis for the fixed confidence, best-option identification problem. Recall that in this problem, given a set of probes $\mathcal{P}$ and a confidence $\delta \in (0, 1]$, we need to design a sequential procedure that identifies the best option $i^\star$ with probability at least $1-\delta$ using as few probes as possible.

2.2.1 Successive Elimination with Probes

The first algorithm modifies the successive elimination algorithm of Even-Dar et al. (2002) to take into account the richer observation structure of our problem. Recall that the algorithm of Even-Dar et al. (2002) works in phases, in each phase observing a certain number of rewards for each remaining candidate actions. At the end of the phase the provably suboptimal actions are eliminated. The number of observations in each phase depends only on the phase index. The process stops when the candidate set contains a single element. The main difference to the algorithm of Even-Dar et al. (2002) is that in each phase our algorithm, which we call Successive Elimination with Probes (SEWP), computes a set multi-covering for the remaining candidate actions given the probes, with a requirement adjusted to the phase index. The returned multi-cover is then used to get the observations for the remaining actions.

Our first result shows that Algorithm 1 is admissible and gives an upper bound on its probe complexity. To state it, define the scheduling and confidence functions

$$f(t) = 2^t, \qquad g(t, \delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}}\,. \tag{2.1}$$

For simplicity, assume that the options are ordered in decreasing order of their mean rewards and $\Delta_2 > 0$, i.e., the optimal option is unique. For $2 \le i \le K$

Algorithm 1 SuccessiveEliminationWithProbes (SEWP)

Inputs: K, δ, $\mathcal{P}$, observation scheduling function $f : \mathbb{N} \to \mathbb{N}$ and confidence function $g : \mathbb{N} \times (0, 1] \to [0, \infty)$.
Initialize candidate set: $A_1 = [K]$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, f(t), \mathcal{P})$.
  Use each p in $C(t)$ for $C_p(t)$-times to get new observations.
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations so far for option i.
  $A_{t+1} \leftarrow \{i \in A_t : \hat{\mu}_i(t) + 2g(t, \delta) > \max_{j \in A_t} \hat{\mu}_j(t)\}$.
  if $|A_{t+1}| = 1$ **then**
    Return the option in A_{t+1}.
  end if
end for

define

$$\widehat{T}_i(\delta) = 1 + \max\left\{ s \,:\, g(s, \delta) \geq \frac{\Delta_i}{4} \right\}, \tag{2.2}$$
$$\hat{N}_i(\delta) = \frac{128}{\Delta_i^2} \log\left(\frac{54K}{\delta} \log \frac{4}{\Delta_i} \right)$$

and let $\widehat{T}_{K+1}(\delta) = 0$ and $\hat{N}_{K+1}(\delta) = 0$. Note that $2^{\widehat{T}_i(\delta)+1} \leq \hat{N}_i(\delta)$, and both are decreasing with $i \geq 2$ increasing.

Theorem 1. *Pick any $0 \leq \delta < 1$ and let SEWP run with inputs $(K, \delta, \mathcal{P}, f, g)$ with f, g given by (2.1). Then, with probability at least $1 - \delta$, SEWP returns the optimal option $i^\star = 1$ within N probes, where N satisfies*

$$N \leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}(\delta)+1}^{\widehat{T}_i(\delta)} \mathcal{C}_{\mathrm{IP}}([i], 2^t). \tag{2.3}$$

Furthermore, with $\hat{M}_i(\delta) \doteq \hat{N}_i(\delta) - \hat{N}_{i+1}(\delta)$,

$$N \leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_i(\delta)\, \mathcal{C}_{\mathrm{LP}}([i], 1). \tag{2.4}$$

The proof borrows ideas from Even-Dar et al. (2002). To prove that SEWP is admissible, one only needs to show that when none of the confidence intervals based on g used in the elimination step fail, the optimal option will not be eliminated. This essentially relied on Hoeffding's inequality, union bounds

and calculations. To calculate the bound on the probe complexity bound, one shows that option i will be eliminated after phase $\hat{T}_i(\delta)$. This happens because in each phase the confidence sets of all options decrease at a uniform rate.

We start with a technical lemma:

Lemma 2. *Let* $0 < a < 1/e$, $b \geq 2$. *Then, for any* $n \geq n^*(a,b) \doteq \frac{2}{a}\log(2b\log\frac{1}{a})$, $an \geq \log(b\log n)$.

Proof. Let $q_1(x) = ax$, $q_2(x) = \log(b\log x) = \log(\log x) + \log b$, $x > e$. The claim to be proven is that for any $n \geq n^* \doteq n^*(a,b)$, $q_1(n) \geq q_2(n)$. By differentiation, it is easy to verify that the function $f(x) = q_1(x) - q_2(x)$ is non-decreasing if and only if $x\log x \geq 1/a$. Hence, it suffices to show that $n^*\log n^* \geq 1/a$, $n^* > e$ so that $q_2(n^*)$ is well-defined and $q_1(n^*) \geq q_2(n^*)$.

From the assumptions and the definition of n^*, we get that $n^* \geq 2\log(4)/a \geq \frac{1}{a} > e$. Hence $q_2(n^*)$ is well-defined. Now, from $n^* > e$, we also get $n^*\log n^* \geq n^*$, which together with $n^* \geq 1/a$ proves that $n^*\log n^* \geq 1/a$.

To verify $q_1(n^*) \geq q_2(n^*)$ note first that from our assumptions on a and b, $2b\log\frac{1}{a} \geq 4 \geq \sqrt{e}$. Hence,

$$q_1(n^*) = 2\log(2b\log\tfrac{1}{a}) = \log(4b^2\log^2\tfrac{1}{a}) \geq \log(2b^2\log\tfrac{1}{a}) = \log b + \log(2b\log\tfrac{1}{a})$$

which holds, as by our condition on a, $\log(1/a) \geq \frac{1}{2}$. On the other hand,

$$\begin{aligned} q_2(n^*) &= \log b + \log(\log n^*) = \log b + \log\log\left(\tfrac{2}{a}\log(2b\log\tfrac{1}{a})\right) \\ &< \log b + \log\log(\tfrac{2b}{a}\log\tfrac{1}{a}) && (\log 2x < x) \\ &< \log b + \log\log(\tfrac{2b}{a^2})\,. && (\log\tfrac{1}{a} < \tfrac{1}{a}) \end{aligned}$$

Now, using again that $\log(2x) < x$,

$$\log(\tfrac{2b}{a^2}) = \log(2b) + \log\tfrac{1}{a^2} < b + \log\tfrac{1}{a^2} \leq b\log\tfrac{1}{a^2}\,,$$

where in the last inequality we also used $b \geq 2$ and $\log\frac{1}{a^2} \geq 2$ and that for $x, y \geq 2$, $x + y \leq x\frac{y}{2} + y\frac{x}{2} = xy$. Putting together all the inequalities, we obtain $q_2(n^*) < q_1(n^*)$.

. . . . . □

With this, we are ready to prove Theorem 1:

Proof of Theorem 1. Let T denote the number of phases before the algorithm exits, i.e., $|A_T| > 1$ and $|A_{T+1}| = 1$. Let U denote the event that for any phase $1 \le t \le T$, and for any option $i \in A_t$ that is not yet eliminated, the mean reward μ_i of option i is within the $g(t,\delta)$ vicinity of its estimate $\hat{\mu}_i(t)$:

$$U = \{|\hat{\mu}_i(t) - \mu_i| \le g(t,\delta) \text{ for all } (i,t) \text{ s.t. } 1 \le t \le T \text{ and } i \in A_t\}\,.$$

First, we will argue about the correctness and cost of the algorithm assuming that U happens and then we will show that U indeed happens with large probability.

Assume therefore that U happens. We claim that on this event, the optimal option $i^\star = 1$ cannot be eliminated, i.e., $1 \in A_1, \ldots, A_{T+1}$. That $1 \in A_1$ holds since $A_1 = [K]$. Now, given that $1 \in A_t$ for some $1 \le t \le T$, we have that $\hat{\mu}_1(t) + 2g(t,\delta) \ge \mu_1 + g(t,\delta) > \max_{j\in A_t} \mu_j + g(t,\delta) \ge \max_{j\in A_t} \hat{\mu}_j(t)$, showing that $1 \in A_{t+1}$ and hence option 1 indeed will not be eliminated.

Now, still assuming that U happens, consider bounding N. We start by asking how big t can be for a suboptimal option $i \neq 1$ to be still included in A_{t+1}. Intuitively, if an option is still considered as a candidate, its suboptimality "gap" Δ_i cannot be large. Indeed, defining $\hat{\mu}^*(t) = \max_{j\in A_t} \hat{\mu}_j(t)$, from $i \in A_{t+1}$ we derive

$$\begin{aligned}\Delta_i = \mu_1 - \mu_i &\le \hat{\mu}_1(t) + g(t,\delta) - (\hat{\mu}_i(t) - g(t,\delta)) \le \hat{\mu}^*(t) - \hat{\mu}_i(t) + 2g(t,\delta) \\ &\le 4g(t,\delta)\,,\end{aligned}$$

where the second inequality used that $1 \in A_t$ and hence $\hat{\mu}^*(t) \ge \hat{\mu}_1(t)$, while the last inequality used that $i \in A_{t+1}$. Hence, by the definition of $\widehat{T}_i \doteq \widehat{T}_i(\delta)$, from $i \in A_{t+1}$ it follows that $t < \widehat{T}_i$. In particular, for any $t \ge \widehat{T}_i + 1$, $i \notin A_t$. As a matter of fact, for any $i > 2$, $t \ge \widehat{T}_i + 1$, and $j \ge i$, j cannot be in A_t. Hence, $A_t \subset \{1, \ldots, i-1\}$. By reindexing, for $1 \le i \le K$ and using $\widehat{T}_{K+1} = 0$, we conclude that

$$t \ge \widehat{T}_{i+1} + 1 \text{ implies that } A_t \subset [i]\,, \quad 1 \le i \le K\,. \tag{2.5}$$

Since (2.5) implies that for $t \geq \widehat{T}_2 + 1$, A_t is a singleton, $T \geq \widehat{T}_2 + 1$ cannot hold. Hence, $T \leq \widehat{T}_2$. Now, we can bound N, the total number of probes used before termination:

$$\begin{aligned} N &= \sum_{t=1}^{T} \sum_{p=1}^{P} C_p(t) = \sum_{t=1}^{T} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \leq \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \\ &\leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{IP}}(A_t, f(t)) \,, \end{aligned} \tag{2.6}$$

where we set $A_t = \{1\}$ for $t > T$. Now, we divide the set $\{1, \ldots, \widehat{T}_2\}$ into the disjoint intervals $S_i = \{\widehat{T}_{i+1} + 1, \ldots, \widehat{T}_i\}$, $i = 2, \ldots, K$. Using that, by (2.5), for any $t \in S_i$ it holds that $A_t \subset [i]$ and thus $\mathcal{C}_{\mathrm{IP}}(A_t, f(t)) \leq \mathcal{C}_{\mathrm{IP}}([i], f(t))$ (where we used that for any $A \subset B$, $n \in \mathbb{N}$, $\mathcal{C}_{\mathrm{IP}}(A, n) \leq \mathcal{C}_{\mathrm{IP}}(B, n)$), we get

$$N \leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}+1}^{\widehat{T}_i} \mathcal{C}_{\mathrm{IP}}([i], 2^t) \,,$$

proving (2.3).

It remains to lower bound the probability that U happens by $1 - \delta$. As usual, we do this by upper bounding the probability of the complementer event $U^c = \{\exists s \in [T], \exists i \in A_s \text{ s.t. } |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\}$. For the sake of simplicity, let us now assume that in each phase t, for each option in A_t, we use only the first $f(t)$ observed rewards and drop the potential "overflow". In fact, by dropping additional observations, the probability of failure can only increase, hence we may make this assumption without loss of generality.

We have

$$\begin{aligned} \Pr(U^c) &= \Pr\left(\exists s \in [T], \exists i \in A_s, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \\ &= \sum_{t=1}^{\infty} \Pr\left(T = t, \exists s \in [t], \exists i \in A_s, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \,. \end{aligned}$$

Note that $\hat{\mu}_i(s)$ is defined only when $i \in A_s$. Without loss of generality we can assume that $\hat{\mu}_i(s)$ when $i \in A_s$ is calculated based on taking the average of the first $n(s) = \sum_{q=1}^{s} f(q)$ elements of an infinite i.i.d. sequence of random variables drawn from the distribution of option i. Hence, defining $\hat{\mu}_i(s)$ as the

average of the first $n(s)$ random variables in this infinite sequence, we get a consistent extension of the definition of $\hat{\mu}_i(s)$ for arbitrary $s \geq 1$.

We have

$$\begin{aligned}
\Pr(U^c) &= \sum_{t=1}^{\infty} \Pr\left(T = t, \exists s \in [t], \exists i \in A_s, |\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) \\
&\leq \sum_{t=1}^{\infty} \Pr\left(T = t, \exists s \in [t], \exists i \in [K], |\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) \\
&\leq \sum_{t=1}^{\infty} \sum_{i=1}^{K} \Pr\left(T = t, \exists s \in [t], |\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) \\
&\leq \sum_{i=1}^{K} \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} \Pr\left(T = t, |\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) \\
&= \sum_{i=1}^{K} \sum_{s=1}^{\infty} \Pr\left(|\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) .
\end{aligned}$$

According to Hoeffding's inequality,

$$\begin{aligned}
\Pr\left(|\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) &\leq 2\exp\left(-2\sum_{t=1}^{s} f(t) g(s,\delta)^2\right) \\
&\leq 2\exp\left(-2^{s+1} \cdot \frac{\log(4Ks^2/\delta)}{2^{s+1}}\right) \\
&= \frac{\delta}{2Ks^2}
\end{aligned}$$

and hence

$$\Pr(U^c) \leq \sum_{i=1}^{K} \sum_{s=1}^{\infty} \frac{\delta}{2Ks^2} < \delta\,.$$

Thus, it remains to upper bound $\widehat{T}_i = 1 + \max\left\{t \,:\, g(t,\delta) \geq \frac{\Delta_i}{4}\right\}$:

$$\begin{aligned}
\widehat{T}_i &= 1 + \max\left\{t \,:\, g(t,\delta) \geq \frac{\Delta_i}{4}\right\} \\
&= 1 + \max\left\{t \,:\, \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}} \geq \frac{\Delta_i}{4}\right\} \\
&\leq 1 + \max\left\{\log_2 n \,:\, \sqrt{\frac{\log(4K(\log_2 n)^2/\delta)}{2n}} \geq \frac{\Delta_i}{4}\right\} \\
&\leq 1 + \log_2 \max\left\{n \,:\, \frac{\Delta_i^2}{16} n \leq \log\left(\frac{4K}{\delta \cdot \log 2} \log n\right)\right\} \,.
\end{aligned}$$

To bound the maximum above, we use Lemma 2. In our problem both $b = \frac{4K}{\delta \cdot \log 2} > 2$ and $\frac{1}{a} = \frac{16}{\Delta_i^2} > e$ satisfy the conditions in this lemma. Plugging in these values of a and b in $n^* = n^*(a, b)$, we get an upper bound of $\widehat{T}_i$ in the form of $1 + \log_2 \left(\frac{32}{\Delta_i^2} \log \left(\frac{16K}{\delta \cdot \log 2} \log \frac{4}{\Delta_i} \right) \right) \leq \log_2 \left(\frac{64}{\Delta_i^2} \log \left(\frac{54K}{\delta} \log \frac{4}{\Delta_i} \right) \right)$, which concludes the proof of the upper bound on $\widehat{T}_i$.

Let us now turn to proving (2.4). According to (2.6), we also have

$$N = \sum_{t=1}^{T} \sum_{p=1}^{P} C_p(t) = \sum_{t=1}^{T} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \leq \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{O}}(A_t, f(t))$$

$$\leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{LP}}(A_t, f(t)) \,, \tag{2.7}$$

Since for $A_t \in [i]$, $\mathcal{C}_{\mathrm{LP}}(A_t, f(t)) \leq \mathcal{C}_{\mathrm{LP}}([i] \,, f(t))$ also holds,

$$\begin{aligned}
N &\leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}+1}^{\widehat{T}_i} \mathcal{C}_{\mathrm{LP}}([i] \,, 2^t) \\
&= \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \left(\sum_{t=\widehat{T}_{i+1}+1}^{\widehat{T}_i} 2^t \right) \mathcal{C}_{\mathrm{LP}}([i] \,, 1) \\
&\leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \left(2^{\widehat{T}_i+1} - 2^{\widehat{T}_{i+1}+1} \right) \mathcal{C}_{\mathrm{LP}}([i] \,, 1) \\
&= \mathcal{G}_{LP}(O, \mathcal{P}) \left(\sum_{i=2}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i] \,, 1) - \sum_{i=2}^{K} 2^{\widehat{T}_{i+1}+1} \mathcal{C}_{\mathrm{LP}}([i] \,, 1) \right) \\
&\leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \left(2^{\widehat{T}_i+1} - 2^{\widehat{T}_{i+1}+1} \right) \mathcal{C}_{\mathrm{LP}}([i] \,, 1) \\
&\leq \mathcal{G}_{LP}(O, \mathcal{P}) \left(\sum_{i=2}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i] \,, 1) - \sum_{i=2}^{K-1} 2^{\widehat{T}_{i+1}+1} \mathcal{C}_{\mathrm{LP}}([i] \,, 1) \right) \\
&= \mathcal{G}_{LP}(O, \mathcal{P}) \left(\sum_{i=2}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i] \,, 1) - \sum_{i=3}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i-1] \,, 1) \right) \\
&= \mathcal{G}_{LP}(O, \mathcal{P}) \left(2^{\widehat{T}_2+1} \mathcal{C}_{\mathrm{LP}}([2] \,, 1) + \sum_{i=3}^{K} 2^{\widehat{T}_i+1} \left(\mathcal{C}_{\mathrm{LP}}([i] \,, 1) - \mathcal{C}_{\mathrm{LP}}([i-1] \,, 1) \right) \right) \\
&\leq \mathcal{G}_{LP}(O, \mathcal{P}) \left(\hat{N}_2(\delta) \mathcal{C}_{\mathrm{LP}}([2] \,, 1) + \sum_{i=3}^{K} \hat{N}_i(\delta) \left(\mathcal{C}_{\mathrm{LP}}([i] \,, 1) - \mathcal{C}_{\mathrm{LP}}([i-1] \,, 1) \right) \right) \\
&= \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \left(\hat{N}_i(\delta) - \hat{N}_{i+1}(\delta) \right) \mathcal{C}_{\mathrm{LP}}([i] \,, 1)
\end{aligned}$$

$$= \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_i(\delta) \mathcal{C}_{\mathrm{LP}}([i], 1)$$

where $2^{\widehat{T}_i+1} \le \frac{128}{\Delta_i^2} \log\left(\frac{54K}{\delta} \log \frac{4}{\Delta_i}\right) = \hat{N}_i(\delta)$ and $\hat{N}_{K+1}(\delta) = 0$.

. . . . . □

The bound (2.3) may be tighter than that shown in (2.4), but perhaps the second is a bit easier to understand. [2] For simplicity, let us explain (2.4). Once (2.4) is explained, the meaning of (2.3) follows. The term $\mathcal{G}_{LP}(O, \mathcal{P})$ is the price of using an oracle combined with some upper bounding that allowed us to arrive at this simpler result by resorting to the linearity properties of $\mathcal{C}_{\mathrm{LP}}$. The rest is what we call a sequential fractional multi-cover with the requirements that option i be covered $\hat{N}_i(\delta)$ times: In a sequential multi-cover, the covering is not done in a single-shot, but is done in phases. In the first phase, all the options must be covered $\hat{M}_K(\delta)$ times. In the next phase, all the options but the last must be covered $\hat{M}_{K-1}(\delta)$ times, etc., up to the last phase when options one and two must be covered $\hat{M}_2(\delta)$ times. Note that the total requirements for an option i are $\hat{M}_K(\delta) + \hat{M}_{K-1}(\delta) + \cdots + \hat{M}_i(\delta) = \hat{N}_K(\delta) - \hat{N}_{K+1}(\delta) + \hat{N}_{K-1}(\delta) - \hat{N}_K(\delta) + \cdots + \hat{N}_i(\delta) - \hat{N}_{i+1}(\delta) = \hat{N}_i(\delta)$. Roughly $\hat{N}_i(\delta) \approx O(1/\Delta_i^2)$ is the number of observations needed from option i (and one) in order to be able to tell which of the two options has a bigger mean reward. Now, compared to (2.4), (2.3) uses a more precise expression for the number of probes, by relying on the the phase structure of the algorithm.

An alternative choice of $f(t)$ and $g(t, \delta)$ is that $f(t) = 1$ and $g(t, \delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{t}}$, which leads to $\hat{N}_i(\delta) = O\left(\frac{1}{\Delta_i^2} \log \frac{K}{\delta \Delta_i}\right)$ instead.

Now, we argue that this bound is tight up to a $\log K$ factor, at least in some cases. In particular, in the bandit case, the covering problem is trivial and we can use an optimal covering oracle. Then, $\mathcal{C}_{\mathrm{O}}([i], 2^t) = i2^t$, and hence the bound becomes $O\left(\sum_{i=1}^{K} \frac{1}{\Delta_i^2} \log\left(\frac{K}{\delta} \log \frac{1}{\Delta_i}\right)\right)$. Up to a log factor, this matches the lower bound of Kaufmann et al. (2015a) which takes the form $\Omega\left(\sum_{i=1}^{K} \Delta_i^{-2} \log(1/\delta)\right)$. Furthermore, as noted by Jamieson et al. (2014)

[2] In fact, if $\mathcal{C}_{\mathrm{O}}(\cdot, n)$ is monotone increasing, (2.3) will hold with $\mathcal{C}_{\mathrm{O}}$ replacing $\mathcal{G}_{IP} \cdot \mathcal{C}_{\mathrm{IP}}$, further tightening the bound.

(based on a result of Farrell (1964)) the $\log\log\Delta^{-1}$ term is necessary.

To examine the tightness of the upper bound, we derive a distribution dependent lower bound on the probe complexity of algorithms admissible for $\mathcal{D}_{\text{sg}}$. Call an environment D a Gaussian environment with common variance σ^2 if for any $1 \le i \le K$, D_i is a Gaussian with variance σ^2.

Theorem 3 (Distribution-dependent lower bound)**.** *For any algorithm admissible for $\mathcal{D}_{sg}$, any confidence $0 < \delta < 1/2$, any probe set $\mathcal{P}$, any sequence $0 = \Delta_1 < \Delta_2 \le \ldots \Delta_K$, if D is a Gaussian environment with common variance $\sigma^2 = 1/4$ and means $\mu_1 = \mu_2 + \Delta_2 = \cdots = \mu_K + \Delta_K$, if N is the number of probes used by the algorithm on D then*

$$\mathbb{E}[N] \ge \min_{s\in[0,\infty)^{\mathcal{P}}} \sum_{p\in\mathcal{P}} s_p \quad s.t. \sum_{p:1\in p} s_p \ge \frac{1}{4\Delta_2^2}\log\frac{1}{6\delta}\,,$$
$$\text{and} \sum_{p:i\in p} s_p \ge \frac{1}{4\Delta_i^2}\log\frac{1}{6\delta}\,, \quad 2 \le i \le K\,.$$

The proofs of our lower bounds are based on the following lemma, a specialized version of Lemma 1 of Kaufmann et al. (2015a). In the lemma we need the Kullback-Leibler divergence (or relative entropy) $KL(P_1, P_2)$ of two distributions: $KL(P_1, P_2) = \int P_1(dx)\log\frac{dP_1}{dP_2}(x)$ if the Radon-Nikodym derivative $\frac{dP_1}{dP_2}$ exists and is $+\infty$ otherwise. Specializing this to two Bernoulli distributions, we get the binary relative entropy function, $d(x, y) = x\log(x/y) + (1 - x)\log((1 - x)/(1 - y))$ defined for $x, y \in [0, 1]$. (Define $d(0, 0) = d(1, 1) = 0$, $d(0, 1) = d(1, 0) = +\infty$.)

Lemma 4. *Let $\hat{i}^* \in [K]$ be the option returned by some algorithm after observing reward from option $i \in [K]$ M_i times and let $\hat{i}^* = 0$ if the algorithm never stops. For any $a \in [K]$, let U_a denote the event that $\hat{i}^* = a$. Then, for any two environments D^1 and D^2, and for any $a \in [K]$,*

$$\sum_{i=1}^{K} \mathbb{E}_{D^1}[M_i] KL(D_i^1, D_i^2) \ge d(\Pr_{D^1}(U_a), \Pr_{D^2}(U_a))\,,$$

where $\mathbb{E}_{D^j}$ and $\Pr_{D^j}$ denote expectation and probability, respectively, under the assumptions that the environment is D^j.

Proof of Theorem 3. The relative entropy of two one-dimensional Gaussian distributions with common variance $\sigma^2 = 1/4$ and mean difference m is $m^2/(2\sigma^2)$. Let G_μ denote the Gaussian distribution with mean μ. Hence, $KL(G_\mu, G_{\mu+a}) = 2a^2$ for any $\mu, a \in \mathbb{R}$. Further, for any $\delta \in (0, 1/2)$,

$$\begin{aligned} d(1-\delta,\delta) &= (1-\delta)\log\frac{1-\delta}{\delta} + \delta\log\frac{\delta}{1-\delta} > \frac{1}{2}\log\frac{1}{2\delta} + \delta\log\delta \\ &\geq \frac{1}{2}\log\frac{1}{2\delta} - \frac{1}{e} > \frac{1}{2}\log\frac{1}{6\delta}\,. \end{aligned} \tag{2.8}$$

Pick $\mu_1 = 1/2$, $\mu_i = \mu_1 - \Delta_i$ and let $D^0 = (D^1, \ldots, D^K) \doteq (G_{\mu_1}, \ldots, G_{\mu_K})$. Define D^1 to be the modification of D^0 when D_2 is replaced by $G_{\mu_1+\varepsilon}$ and let D^i with $2 \leq i \leq K$ be the modification of D^0 when D_i is replaced by $G_{\mu_1+\varepsilon}$ with some $\varepsilon > 0$. As in the proof of Theorem 2 in Kaufmann et al. (2015a), we apply Lemma 4 to the K pairs of environments $(D^0, D^1), \ldots, (D^0, D^K)$ and option $a = 1$.

We have $KL(D^0_j, D^1_j) = 0$ unless $j = 2$ in which case $KL(D^0_2, D^1_2) = KL(G_{\mu_2}, G_{\mu_1+\varepsilon}) = (\Delta_2 + \varepsilon)^2$. Also, for any $2 \leq i \leq K$, $1 \leq j \leq K$, $KL(D^0_j, D^i_j) = 0$ unless $i = j$ in which case $KL(D^0_i, D^i_i) = KL(G_{\mu_i}, G_{\mu_1+\varepsilon}) = 2(\Delta_i + \varepsilon)^2$. Further, the optimal option in D^0 is option one, while the optimal option in D^i is option i because $\varepsilon > 0$. Hence, if U is the event that the algorithm picks option one, then, since the algorithm is admissible, $\Pr_{D^0}(U) \geq 1 - \delta$, and $\Pr_{D^i}(U) \leq \delta$. Combined with (2.8), letting M_i denote the number of observations from option i, we get

$$\mathbb{E}_{D^0}[M_1] \geq \frac{1}{4(\Delta_2+\varepsilon)^2}\log\frac{1}{6\delta}\,, \quad \mathbb{E}_{D^0}[M_i] \geq \frac{1}{4(\Delta_i+\varepsilon)^2}\log\frac{1}{6\delta}\,, \quad 2 \leq i \leq K\,.$$

Define N_p the number of times probe p is used. Then, $N = \sum_{p\in\mathcal{P}} N_p$ and $M_i = \sum_{p:i\in p} N_p$. Combining this with the previous inequalities leads to the linear program as shown in Theorem 3.

. . . . . □

Note that the lower bound clearly reflects the structure of $\mathcal{P}$. However, even disregarding the constants and logarithmic factors, there is still a gap between our upper and lower bounds: In the upper bound, as explained before, the

size of a *sequential* cover appears, while in the lower bound, the size of a "one-shot" cover is seen. Note that in either the bandit or the full information case, there is no gap between these quantities. We were able to establish a gap of $\log(K)$ when considering sequential and one-shot *integral* covers. However, it remains a very interesting open question whether the gap can be closed in the fractional case.

2.2.2 An Alternative Algorithm to Find the Best Option

The second algorithm is a generalization of the exponential gap elimination algorithm of Karnin et al. (2013), which improves the logarithmic term in the sample complexity from $\log(\frac{K}{\delta}\log\frac{1}{\Delta})$ to $\log(\frac{1}{\delta}\log\frac{1}{\Delta})$ for the bandit problem. So we expect that generalizing that algorithm to our setting will have a similar improvement regarding the $\log K$ term.

The exponential gap elimination algorithm of Karnin et al. (2013) calls the median elimination algorithm of Even-Dar et al. (2002) as a subroutine, which finds an ε-optimal option using $O(K\varepsilon^{-2}\log(1/\delta))$ samples with probability at least $1-\delta$ (an option is ε-optimal iff its expected reward is at least $\mu_1-\varepsilon$). So before generalizing the exponential gap elimination algorithm, we need to first design a counterpart for the median elimination algorithm.

Median Elimination With Probes

The median elimination algorithm (ME) of Even-Dar et al. (2002) works as follows: The algorithm runs in phases. In every phase t each potentially good arm is sampled $4\varepsilon_t^{-2}\log(3/\delta_t)$ times, where $\delta_t = \delta 2^{-t-1}$ and $\varepsilon_t = \varepsilon(3/4)^t/3$; then the lower half of the arms with inferior performance is eliminated, and the next phase is run with the remaining arms only. The algorithm terminates when a single arm remains.

A tempting approach to address our problem would be, instead of sampling each remaining arm n times in one phase, we sample a set of probes that is a minimum n-cover of those arms. We will call this naive modification of the median elimination algorithm the naive-ME algorithm. While "naive-

ME" preserves the same $O(K\varepsilon^{-2}\log(1/\delta))$ performance in the bandit case, the following proposition shows that in the full information case this algorithm requires $K^{1/2}$-times more probes than expected.

Proposition 5. *In the full information case where $\mathcal{P} = \{[K]\}$, the probe complexity of the naive-ME algorithm is at least*

$$\Omega\left(\frac{K^{1/2}}{\varepsilon^2}\log\frac{K}{\delta}\right).$$

Proof. The median elimination algorithm deterministically runs $\lceil \log_2 K \rceil$ phases since it eliminates half of the arms in each phase. In phase t, the algorithm collects $\frac{4}{\varepsilon_t^2}\log\frac{3}{\delta_t}$ samples for each arm in the set of arms A_t considered, where $\varepsilon_t = \frac{\varepsilon}{3}\left(\frac{3}{4}\right)^t$ and $\delta_t = \frac{\delta}{2^{t+1}}$, and then selects A_{t+1} to contain half of the arms with better estimated mean rewards. Under the full information setting, there is only one probe that covers all arms, so the algorithm uses that probe the probe $\frac{4}{\varepsilon_t^2}\log\frac{3}{\delta_t}$ times in each phase. Then the total probe complexity N is

$$\begin{aligned}
N &= \sum_{t=1}^{\lceil \log_2 K \rceil} \frac{4}{\varepsilon_t^2}\log\frac{3}{\delta_t} = \sum_{t=1}^{\lceil \log_2 K \rceil} \frac{36}{\varepsilon^2}\left(\frac{16}{9}\right)^t \log\frac{6\cdot 2^t}{\delta} \\
&\geq \frac{36}{\varepsilon^2}\left(\frac{16}{9}\right)^{\log_2 K}\log\frac{6K}{\delta} \qquad \text{(only take the last term)} \\
&= \frac{36}{\varepsilon^2}K^{\log_2\frac{16}{9}}\log\frac{6K}{\delta} > \frac{36}{\varepsilon^2}K^{1/2}\log\frac{6K}{\delta} \\
&= \Omega\left(\frac{K^{1/2}}{\varepsilon^2}\log\frac{K}{\delta}\right).
\end{aligned}$$

. . . . . □

Intuitively, the presence of the $K^{1/2}$ term is not expected since the full information case gives K times more information than the bandit case.

We have shown that simply replacing the uniform sampling in each phase in the median elimination algorithm of Even-Dar et al. (2002) with a set multi-cover does not work, so a more careful design is needed. Our proposed algorithm, called Median Elimination With Probes (MEWP) is shown in Algorithm 2. It essentially runs the original median elimination algorithm for bandits over a one-cover of all options (that is, each probe in the cover is

treated as an option in the bandit setting), and in each phase we eliminate half of the *probes* that do not seem to cover a good option. We stop running median elimination when a single probe covers all the remaining options. Then the algorithm enters its second stage where we use this probe until we identify an almost optimal option from the remaining ones. In the next theorem we prove that the algorithm is admissible, and give an upper bound on the number of probes required to find an ε-optimal option.

Algorithm 2 MedianEliminationWithProbes

Inputs: K, $\delta \in (0,1]$, $\varepsilon > 0$, $\mathcal{P}$.
Set $\varepsilon_t = \frac{\varepsilon}{6}(\frac{3}{4})^t$, $\delta_t = \frac{\delta}{2^{t+1}}$.
$C \leftarrow \mathrm{COrcl}([K], 1, \mathcal{P})$, and define a partition of the options as $A_1 = \{\pi_p \subset p : p \in C, \cup_{p\in C}\pi_p = [K]\}$.
for $t = 1, 2, \ldots$ **do**
  for all $\pi \in A_t$ **do**
    Use $\frac{4}{\varepsilon_t^2}\log\frac{3|\pi|}{\delta_t}$-times $p \in C$ that covers π to get observations for each option in p.
    Let $\hat{\mu}_\pi(t) = \max_{i\in\pi}\hat{\mu}_i(t)$, where $\hat{\mu}_i(t)$ is the empirical mean reward of option i based on the observations in the actual phase t.
  end for
  Find the median $m(t)$ of $\{\hat{\mu}_\pi(t) : \pi \in A_t\}$.
  Let $A_{t+1} = \{\pi \in A_t : \hat{\mu}_\pi(t) \geq m(t)\}$.
  if $|A_{t+1}| = 1$ **then**
    terminate the loop and let $\widehat{\pi}^*$ be the single element of A_{t+1}
  end if
end for
If $|\widehat{\pi}^*| > 1$, use the probe that covers $\widehat{\pi}^*$ for $\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}$-times.
Return the option $\hat{i}^* \in \widehat{\pi}^*$ with the highest empirical mean based on these observations.

Theorem 6. *With probability at least $1-\delta$, MEWP returns an ε-optimal option $\hat{i}^*$, and N, the total number of probes used by the algorithm is*

$$N = O\left(\frac{\mathcal{C}_{\mathrm{O}}([K], 1)}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right). \tag{2.9}$$

where $|\pi_{\max}| = \max_{\pi\in A_1}|\pi|$.

Proof of Theorem 6. The algorithm contains two stages: First, in the for loop, we aim to find a probe that contains an $\varepsilon/2$-optimal option with probability at least $1-\delta/2$, in $O(|A_1|\varepsilon^{-2}\log(|\pi_{\max}|/\delta))$; then we find an option that is

$\varepsilon/2$-optimal option within this probe with probability at least $1-\delta/2$ after $O(\varepsilon^{-2}\log(|\pi_{\max}|/\delta))$ probes.

First we will analyze the algorithm on the first stage. We need to show that

$$\Pr\left(\hat{\mu}_{\hat{\pi}^*} > \mu_{\pi^*} - \frac{\varepsilon}{2}\right) \geq 1 - \frac{\delta}{2} \tag{2.10}$$

where $\mu_\pi = \max_{i\in\pi}\mu_i$ for all $\pi \in A_1$ and $\pi^* = \operatorname{argmax}_\pi \mu_\pi$. Clearly, $\mu_{\pi^*} = \mu_1$, the expectation of the best option.

Let $\pi_t = \operatorname{argmax}_{\pi\in A_t}\mu_\pi$. Let $\Pr_t$ and $\mathbb{E}_t$ denote the conditional probability and conditional expectation given all randomness before phase t. To prove (2.10), we will first show that

$$\Pr_t\left(\mu_{\pi_{t+1}} > \mu_{\pi_t} - \varepsilon_t\right) \geq 1 - \delta_t\ .$$

Define $A_t^\varepsilon = \{\pi \in A_t : \mu_\pi \leq \mu_{\pi_t} - \varepsilon_t\}$ and $A_t^* = \{\pi \in A_t : \hat{\mu}_\pi(t) > \hat{\mu}_{\pi_t}(t)\}$. Then, for any $\pi \in A_t$, the event $\{\pi \in A_t^* \cap A_t^\varepsilon\} \wedge \{\hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \varepsilon_t/2\}$ implies $\{\hat{\mu}_\pi(t) > \mu_\pi + \varepsilon_t/2\}$. Thus, for any $\pi \in A_t, \pi \neq \pi_t$,

$$\begin{aligned}
&\Pr_t\left(\pi \in A_t^* \cap A_t^\varepsilon \middle| \hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) \\
&\leq \Pr_t\left(\hat{\mu}_\pi(t) > \mu_\pi + \frac{\varepsilon_t}{2} \middle| A_t, \hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) \\
&= \Pr_t\left(\hat{\mu}_\pi(t) > \mu_\pi + \frac{\varepsilon_t}{2}\right) \leq \frac{\delta_t}{3},
\end{aligned}$$

where (i) the equality holds since the samples from the options in π and π_t are independent, and (ii) the last inequality holds, since by Hoeffding's inequality (Cesa-Bianchi and Lugosi, 2006),

$$\Pr_t\left(\hat{\mu}_i(t) > \mu_\pi + \frac{\varepsilon_t}{2}\right) \leq \Pr_t\left(\hat{\mu}_i(t) > \mu_i + \frac{\varepsilon_t}{2}\right) < \frac{\delta_t}{3|\pi|}$$

for all $i \in \pi$, since $\hat{\mu}_i(t)$ is estimated from $(2/\varepsilon_t)^2\log(3|\pi|/\delta_t)$ samples, and the union bound implies that this inequality simultaneously holds for all $i \in \pi$ with probability $\delta_t/3$. Furthermore, by definition $\pi_t \notin A_t^\varepsilon$, hence

$$\Pr_t\left(\pi_t \in A_t^* \cap A_t^\varepsilon \middle| \hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) = 0\,.$$

Therefore,

$$\mathbb{E}_t\left[\frac{|A_t^* \cap A_t^\varepsilon|}{|A_t|}\middle|\hat{\mu}_{\pi_t}(t) > \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right] \le \frac{\delta_t}{3}$$

Applying Markov's inequality, we have

$$\Pr_t\left(\frac{|A_t^* \cap A_t^\varepsilon|}{|A_t|} \ge \frac{1}{2}\middle|\hat{\mu}_{\pi_t}(t) > \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) \le \frac{2\delta_t}{3}.$$

Note that again by Hoeffding's inequality and the union bound, $\Pr_t(\hat{\mu}_{\pi_t}(t) > \mu_{\pi_t} - \frac{\varepsilon_t}{2}) \ge 1 - \frac{\delta_t}{3}$, and $\{\frac{|A_t^* \cap A_t^\varepsilon|}{|A_t|} < \frac{1}{2}\}$ implies $\{\mu_{\pi_{t+1}} > \mu_{\pi_t} - \varepsilon_t\}$. Then, by union bound, we get

$$\Pr_t\left(\mu_{\pi_{t+1}} > \mu_{\pi_t} - \varepsilon_t\right) \ge 1 - \delta_t$$

Since the bound is constant, the unconditional probability also satisfies this inequality, and so, by the union bound,

$$\Pr\left(\mu_{\hat{\pi}^*} \le \mu_{\pi^*} - \frac{\varepsilon}{2}\right) \le \sum_{t=1}^{\log_2|A_1|} \Pr\left(\mu_{\pi_{t+1}} \le \mu_{\pi_t} - \varepsilon_t\right) \le \sum_{t=1}^{\log_2|A_1|} \delta_t < \frac{\delta}{2},$$

proving (2.10).

Next we will calculate the probe complexity until $\widehat{\pi}^*$ is found:

$$\sum_{t=1}^{\log_2|A_1|} \sum_{\pi\in A_t} \frac{4}{\varepsilon_t^2}\log\frac{3|\pi|}{\delta_t} \le \sum_{t=1}^{\log_2|A_1|} \frac{4|A_t|}{\varepsilon_t^2}\log\frac{3|\pi_{\max}|}{\delta_t} = O\left(\frac{|A_1|}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right)$$

Now we will analyze the second stage, by showing that it finds an $\varepsilon/2$-optimal option from $\widehat{\pi}^*$ with probability at least $1 - \delta/2$. Assume that the first stage ran for T phases, so we will consider conditional probabilities $\Pr_T$ conditioned on the first T phases of the first stage.

Let $i^*_{\widehat{\pi}^*}$ denote the optimal option in $\widehat{\pi}^*$, $\hat{\mu}_i$ be the empirical mean reward for option $i \in \widehat{\pi}^*$ in the second stage, computed from $\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}$ samples, $\hat{i}^* = \operatorname{argmax}_{i\in\widehat{\pi}^*}\hat{\mu}_i$, $A^\varepsilon = \{i \in \widehat{\pi}^* : \mu_i \le \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{2}\}$ and $A^* = \{\hat{\mu}_i > \hat{\mu}_{i^*_{\widehat{\pi}^*}}\}$. Clearly, $\{\hat{\mu}_{i^*_{\widehat{\pi}^*}} \ge \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{4}\}$ and $\{\forall i \in A^\varepsilon, \hat{\mu}_i \le \mu_i + \frac{\varepsilon}{4}\}$ imply $\{|A^\varepsilon \cap A^*| = \varnothing\}$, which in turn implies $\{\hat{i}^* \notin A^\varepsilon\}$. Therefore,

$$\Pr_T\left(\mu_{\hat{i}^*} > \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{2}\right) \ge \Pr_T\left(\hat{\mu}_{i^*_{\widehat{\pi}^*}} \ge \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{4} \wedge \forall i \in A^\varepsilon, \hat{\mu}_i \le \mu_i + \frac{\varepsilon}{4}\right)$$

$$\geq 1 - \Pr_T\left(\hat{\mu}_{i^*_{\widehat{\pi}^*}} < \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{4}\right) - \sum_{i \in A^\varepsilon} \Pr_T\left(\hat{\mu}_i > \mu_i + \frac{\varepsilon}{4}\right).$$

Applying Hoeffding's inequality, we have

$$\Pr_T\left(\hat{\mu}_i - \mu_i > \frac{\varepsilon}{4}\right) \leq e^{-\frac{n\varepsilon^2}{8}} = \frac{\delta}{2|\widehat{\pi}^*|}$$

where $n = \frac{8}{\varepsilon^2} \log \frac{2|\widehat{\pi}^*|}{\delta}$. Note that the same probability bound holds for $\mu_i - \hat{\mu}_i > \frac{\varepsilon}{4}$. Therefore,

$$\Pr_T\left(\mu_{\hat{i}^*} \geq \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{2}\right) \geq 1 - \frac{(|A^\varepsilon| + 1)\delta}{2|\widehat{\pi}^*|} \geq 1 - \frac{\delta}{2}.$$

Since the bound is independent of the condition, we also have

$$\Pr\left(\mu_{\hat{i}^*} \geq \mu_{i^*_{\widehat{\pi}^*}} - \frac{\varepsilon}{2}\right) \geq 1 - \frac{\delta}{2}.$$

Combining with (2.10), we obtain

$$\Pr(\mu_{\hat{i}^*} \geq \mu_{i^\star} - \varepsilon) \geq 1 - \delta\ .$$

Finally, the total number of probes can be bounded as

$$N = O\left(\frac{|A_1|}{\varepsilon^2} \log \frac{|\pi_{\max}|}{\delta}\right) + \frac{8}{\varepsilon^2} \log \frac{2|\widehat{\pi}^*|}{\delta} = O\left(\frac{\mathcal{C}_{\mathrm{O}}([K], 1)}{\varepsilon^2} \log \frac{|\pi_{\max}|}{\delta}\right)$$
$$(|A_1| = \mathcal{C}_{\mathrm{O}}([K], 1))$$

. . . . . □

Note that we have $|\pi_{\max}|$ inside the log term instead of the expected 1. It can be shown that this worst case upper bound is unimprovable in both the bandit and full information setting by the following theorem.

Theorem 7. *In the full information case, for every $K \geq 2$, $\varepsilon > 0$ and $\delta \in (0, 1/2)$, and for any algorithm that returns an ε-optimal option with probability at least $1 - \delta$, there exist reward distributions $(D_1, \ldots, D_K)$ such that*

$$\mathbb{E}[N] \geq \frac{1}{16\varepsilon^2} \log \frac{K}{12\delta} \tag{2.11}$$

where N is the total number of probes used by the algorithm.

Moreover, for any general observation structure $\mathcal{P}$, a lower bound is

$$\mathbb{E}[N] \geq \frac{\mathcal{C}_{\mathrm{LP}}([K], 1)}{16\varepsilon^2} \log \frac{1}{6\delta}\,. \tag{2.12}$$

Proof of Theorem 7. We prove (2.11) and (2.12) separately.

First we prove a modification of Lemma 4 for the full information case.

Lemma 8. *Consider the full information case. Let $\hat{i}^* \in [K]$ be the option returned by some algorithm after N trials if the algorithm stops and let $\hat{i}^* = 0$ if the algorithm never stops. Furthermore, for any $a \in [K]$, let U_a denote the event that $\hat{i}^* = a$. Then, for any two environments D^1 and D^2, and for any $a \in [K]$,*

$$\mathbb{E}_{D^1}[N] \sum_{i=1}^{K} KL(D_i^1, D_i^2) \geq d(\Pr_{D^1}(U_a), \Pr_{D^2}(U_a)),$$

where $\mathbb{E}_{D^j}$ and $\Pr_{D^j}$ denote expectation and probability under the assumptions that the environment is D^j, $KL(D_i^1, D_i^2)$ denotes the relative entropy (or Kullback-Leibler divergence) between D_i^1 and D_i^2 for all $i \in [K]$, and $d(x, y) = x \log(x/y) + (1-x) \log((1-x)/(1-y))$ is the binary relative entropy.

Proof. Assume that the full information algorithm is applied in the bandit case in a naive way: trying each option once in the bandit case when it would choose to try the only probe in the full information case. Then $N_1 = \ldots = N_K = N$, and the statement of the lemma follows immediately from Lemma 4. □

Proof of (2.11). We prove the theorem by applying Lemma 8. In order to do so, we need to construct the environments D^1 and D^2. We assume that for any $i \in [K], k \in \{1, 2\}$, D_i^k is Gaussian with mean μ_i^k and variance $\sigma^2 = 1/4$. In D_2 we set $\mu_1^2 = \varepsilon, \mu_2^2 = ... = \mu_K^2 = 0$. Now consider any algorithm that returns an ε-optimal option with probability $1 - \delta$. Then we have $\Pr_{D^2}(U_1) \geq 1 - \delta$. Furthermore, $\sum_{i=2}^{K} \Pr_{D^2}(U_i) < \delta$, and so there exists some $j \in \{2, \ldots, K\}$ such that $\Pr_{D^2}(U_j) < \delta/(K-1)$. We use this j to select the expected values of the distributions D_i^1: in particular, we let $\mu_1 = \varepsilon, \mu_j = 2\varepsilon$, and $\mu_i = 0$ for all other i. Then we have $\Pr_{D^1}(U_j) \geq 1 - \delta$.

Since the relative entropy of two 1-dimensional Gaussian distributions with common variance σ^2 and mean difference m is $m^2/(2\sigma^2)$, we have

$$\sum_{i=1}^{K} KL(D_i^1, D_i^2) = KL(D_j^1, D_j^2) = (2\varepsilon)^2/(2\sigma^2) = 8\varepsilon^2 .$$

Furthermore, by the monotonicity properties of the binary entropy function d, and since $\Pr_{D^2}(U_j) < \delta/(K-1) < 1-\delta \leq \Pr_{D^1}(U_j)$, we have

$$d(\Pr_{D^1}(U_j), \Pr_{D^2}(U_j)) \geq d(1-\delta, \delta/(K-1))\,.$$

Thus, applying Lemma 8, we get

$$\mathbb{E}_{D^1}[N] \geq \frac{d\left(1-\delta, \frac{\delta}{K-1}\right)}{8\varepsilon^2}\ . \tag{2.13}$$

The last step is to bound $d\left(1-\delta, \frac{\delta}{K-1}\right)$ from below:

$$\begin{aligned} d\left(1-\delta, \frac{\delta}{K-1}\right) &= (1-\delta)\log\frac{1-\delta}{\frac{\delta}{K-1}} + \delta\log\frac{\delta}{1-\frac{\delta}{K-1}} \\ &> \frac{1}{2}\log\frac{K-1}{2\delta} + \delta\log\delta \geq \frac{1}{2}\log\frac{K-1}{2\delta} - \frac{1}{e} \\ &> \frac{1}{2}\log\frac{K-1}{6\delta} \geq \frac{1}{2}\log\frac{K}{12\delta}\ . \end{aligned}$$

Combined with (2.13) we have $\mathbb{E}_{D^1}[N] \geq \frac{1}{16\varepsilon^2}\log\frac{K}{12\delta}$, which concludes the proof. □

Proof of (2.12). Let D be an environment such that $D_i, i \in [K]$ is Gaussian with mean μ_i and variance $\sigma^2 = 1/4$, where $\mu_1 = \varepsilon$ and $\mu_i = 0$ for all $i \neq 1$.

We create K environments, $D^1, \ldots, D^K$, such that D_i^k is Gaussian with mean μ_i^k and variance $\sigma^2 = 1/4$, and use Lemma 4 to lower bound the number of trials needed in environment D. For D^1, let $\mu_1^1 = -\varepsilon$ and $\mu_i^1 = \mu_i$ for all $i \neq 1$. For D^k, $k \neq 1$, let $\mu_k^k = 2\varepsilon$ and $\mu_j^k = \mu_j$ for all $j \neq k$.

Consider an algorithm A that, with probability at least $1-\delta$, returns an ε-optimal solution (in any environment satisfying the assumptions of our setting). Then, using the notation of Lemma 4, we have $\Pr_{D^k}(U_1) < \delta$ for all $k \in [K]$ and $\Pr_D(U_1) \geq 1-\delta$.

Let N_i be the number of samples observed by algorithm A for option i. Similarly to the proof of Lemma 8, we construct a bandit algorithm from A using probes in such a way that whenever A decides to try a probe p in the original problem, the bandit version tries each option $i \in p$ once in the bandit problem. Then the number of samples for each option i will be the same in the

original and in the bandit problem, and so, similarly to the proof of Theorem 7, Lemma 4 implies that

$$\mathbb{E}_D[N_i] \geq \frac{d(1-\delta,\delta)}{8\varepsilon^2}$$

for all $i \in [K]$. Using the derivation in Theorem 3, we get $d(1-\delta,\delta) > \frac{1}{2}\log\frac{1}{6\delta}$. Therefore, we have

$$\mathbb{E}_D[N_i] = \sum_{p \ni i} \mathbb{E}_D[N_p] \geq \frac{1}{16\varepsilon^2}\log\frac{1}{6\delta}$$

where N_p is the number of times that probe p is played. Since $\mathbb{E}_D[N] = \sum_{p\in\mathcal{P}} \mathbb{E}_D[N_p]$, lower bounding $\mathbb{E}_D[N]$ leads to

$$\mathbb{E}_D[N] \geq \mathcal{C}_{\mathrm{LP}}\left([K], \frac{1}{16\varepsilon^2}\log\frac{1}{6\delta}\right) = \frac{\mathcal{C}_{\mathrm{LP}}([K],1)}{16\varepsilon^2}\log\frac{1}{6\delta}\ . \tag{2.14}$$

□

. . . . . □

Compared to the upper bound of Theorem 6 in general cases, lower bound (2.12) has a $|\pi_{\max}|$ gap inside the log term. However, (2.12) is not tight since in the full information case we have a tighter lower bound $\Omega(\varepsilon^{-2}\log(K/\delta))$ in (2.11). Therefore, although whether the $|\pi_{\max}|$ term is tight or not is still an open question there has to be some quantity between 1 and K in the log term. Note that MEWP may not be the best choice for only finding an ε-optimal option in practice since it does not provide distribution dependent performance. However, the worst case upper bound is theoretically good enough (has a better log term) for being a subroutine of our later algorithm EGEWP.

Exponential Gap Elimination Algorithm

Given the MEWP algorithm, we continue with generalizing the exponential gap elimination algorithm. The new algorithm, called Exponential Gap Elimination with Probes (EGEWP), is shown in Algorithm 3. The new idea here is to use the partition-based exploration technique (as in the MEWP algorithm) and replace the bandit-case median elimination subroutine with

Algorithm 3 ExpGapEliminationWithProbes

Inputs: K, δ, $\mathcal{P}$.
Initialize candidate set: $A_1 = [K]$. Set $\varepsilon_t = \frac{1}{4 \cdot 2^t}$, $\delta_t = \frac{\delta}{50t^3}$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, 1, \mathcal{P})$.
  Create a partition Π_t of A_t such that $\Pi_t = \left\{\pi_p \subset p \,:\, p \in C(t), \cup_{p \in C(t)} \pi_p = A_t\right\}$.
  for $\pi_p \in \Pi_t$ **do**
    Use probe p for $\frac{2}{\varepsilon_t^2} \log \frac{2|\pi_p|}{\delta_t}$-times to get observations for each option in p.
  end for
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations in phase t for option i.
  $i_t \leftarrow \mathrm{MedianEliminationWithProbes}(A_t, \frac{\varepsilon_t}{2}, \delta_t)$.
  Let $A_{t+1} = \{i \in A_t \,:\, \hat{\mu}_i(t) \geq \hat{\mu}_{i_t}(t) - \varepsilon_t\}$.
  if $|A_{t+1}| = 1$ **then**
    Return the option in A_{t+1}.
  end if
end for

MEWP. The analysis follows a combination of the techniques of Karnin et al. (2013) and the proof of Theorem 6. However, due to the more complicated observation structure, we are only able to prove a Δ_2 dependent upper bound on the number of probes:

Theorem 9. *If the oracle COrcl always returns the optimal solution for integer programming, EGEWP finds the optimal option with probability at least* $1 - \delta$ *after using*

$$O\left(\frac{\mathcal{C}_{\mathrm{O}}([K], 1)}{\Delta_2^2} \log\left(\frac{|p_{\max}|}{\delta} \log \frac{1}{\Delta_2}\right)\right) \tag{2.15}$$

probes where $|p_{\max}| = \max_{p \in \mathcal{P}} |p|$.

Proof of Theorem 9. As in earlier proofs, we are going to use $\Pr_t$ and $\mathbb{E}_t$ to denote the conditional probability and conditional expectation, respectively, given all randomness before phase t, and we denote the σ-algebra corresponding to the latter by $\mathcal{F}_{t-1}$.

First we are going to bound the number of phases in running EGEWP. We start with the following simple observation: For any $i \neq 1$ and t such that

$T \geq t \geq \log_2 \frac{1}{\Delta_i}$, $i \in A_t$ and $1 \in A_t$, the event $C_{t,i} = \{\mu_{i_t} \geq \mu_1 - \frac{\varepsilon_t}{2}\} \wedge \{\mu_{i_t} \leq \hat{\mu}_{i_t}(t) + \frac{\varepsilon_t}{2}\} \wedge \{\mu_i \geq \hat{\mu}_i(t) - \frac{\varepsilon_t}{2}\}$ implies $i \notin A_{t+1}$. This holds since given $C_{t,i}$,

$$\hat{\mu}_{i_t}(t) \geq \mu_{i_t} - \frac{\varepsilon_t}{2} \geq \mu_1 - \varepsilon_t \geq \mu_i + 3\varepsilon_t \geq \hat{\mu}_i(t) + \frac{5}{2}\varepsilon_t > \hat{\mu}_i(t) + \varepsilon_t$$

where in the third step we used that $\mu_1 - \mu_i = \Delta_i \geq 2^{-t} = 4\varepsilon_t$ for $t \geq \log_2 \frac{1}{\Delta_i}$. Now assume that $\mathcal{F}_{t-1}$ is such that $1 \in A_t$ and $\pi \in \Pi_t$. Then, for any $t \geq \log_2 \frac{1}{\Delta_2}$,

$$\begin{aligned}
&\Pr_t\left(\exists i \in \pi, i \neq 1, i \in A_{t+1}\right) \\
&\leq \Pr_t\left(\mu_{i_t} < \mu_1 - \frac{\varepsilon_t}{2}\right) + \Pr_t\left(\mu_{i_t} > \hat{\mu}_{i_t}(t) + \frac{\varepsilon_t}{2}\right) + \sum_{i \in \pi, i \neq 1} \Pr_t\left(\mu_i < \hat{\mu}_i(t) - \frac{\varepsilon_t}{2}\right).
\end{aligned} \tag{2.16}$$

Now, for any $t \geq 1$ and $\mathcal{F}_{t-1}$ as above,

$$\Pr_t\left(\mu_{i_t} < \mu_1 - \frac{\varepsilon_t}{2}\right) \leq \delta_t$$

by the high probability guarantee for the success of MEWP and the fact that new samples are used in each phase. Furthermore, for any $i \in \pi$ and $t \geq 1$,

$$\Pr_t\left(\mu_i < \hat{\mu}_i(t) - \frac{\varepsilon_t}{2}\right) \leq \frac{\delta_t}{2|\pi|} \quad \text{and} \quad \Pr_t\left(\mu_i > \hat{\mu}_i(t) + \frac{\varepsilon_t}{2}\right) \leq \frac{\delta_t}{2|\pi|} \tag{2.17}$$

by Hoeffding's inequality since $\hat{\mu}_i$ is computed from $2\varepsilon_t^{-2}\log(2|\pi|/\delta_t)$ new samples. Finally, since i_t is selected based on different samples than the ones used in estimating $\hat{\mu}_{i_t}$, denoting by $\pi_t(j)$ the partition cell of A_t containing j, we have

$$\begin{aligned}
\Pr_t\left(\mu_{i_t} > \hat{\mu}_{i_t}(t) + \frac{\varepsilon_t}{2}\right) &= \sum_{j \in A_t} \Pr_t\left(\mu_j > \hat{\mu}_j(t) + \frac{\varepsilon_t}{2} \middle| i_t = j\right) \Pr_t\left(i_t = j\right) \\
&= \sum_{j \in A_t} \Pr_t\left(\mu_j > \hat{\mu}_j(t) + \frac{\varepsilon_t}{2}\right) \Pr_t\left(i_t = j\right) \\
&\leq \sum_{j \in A_t} \frac{\delta_t}{2|\pi_t(j)|} \Pr_t\left(i_t = j\right) \leq \frac{\delta_t}{2}.
\end{aligned} \tag{2.18}$$

Continuing (2.16) with the above inequalities, we obtain that for any $t \geq \log_2 \frac{1}{\Delta_2}$ and $\mathcal{F}_{t-1}$ such that $1 \in A_t$ and $\pi \in \Pi_t$,

$$\Pr_t\left(\exists i \in \pi, i \neq 1, i \in A_{t+1}\right) \leq \delta_t + \frac{\delta_t}{2} + \frac{\delta_t}{2} = 2\delta_t. \tag{2.19}$$

Since the same $2\delta_t$ bound holds for ny $\mathcal{F}_{t-1}$ with $1 \in A_t$ and $\pi \in \Pi_t$, we also have

$$\Pr\left(\exists i \in \pi, i \neq 1, i \in A_{t+1} | \pi \in \Pi_t, 1 \in A_t\right) \leq 2\delta_t. \tag{2.20}$$

Furthermore, for any $t \geq 1$, the events $\{1 \in A_t\}$, $\{\hat{\mu}_1(t) \geq \mu_1 - \frac{\varepsilon_t}{2}\}$, and $\{\hat{\mu}_{i_t}(t) \leq \mu_{i_t} + \frac{\varepsilon_t}{2}\}$ imply that $1 \in A_{t+1}$, since

$$\hat{\mu}_1(t) \geq \mu_1 - \frac{\varepsilon_t}{2} \geq \mu_{i_t} - \frac{\varepsilon_t}{2} \geq \hat{\mu}_{i_t}(t) - \varepsilon_t \,.$$

Therefore, from (2.17) (for $i = 1$) and (2.18) we get

$$\Pr(1 \in A_{t+1} | 1 \in A_t) \geq 1 - \frac{\delta_t}{2} - \frac{\delta_t}{2} = 1 - \delta_t \,. \tag{2.21}$$

The above inequality shows that the optimal option 1 is not eliminated with high probability, while (2.20) shows that for large enough t, the suboptimal options are eliminated with high probability. Thus, it remains to quantify how fast the suboptimal options are eliminated. To this end, we show that the number of probes used in every phase decays exponentially fast for $t \geq \log_2 \frac{1}{\Delta_2}$. Let $\Pi_t^+ = \{\pi \in \Pi_t : \exists i \in \pi, i \in A_{t+1}\}$. Then, for any $t \geq \log_2 \frac{1}{\Delta_2}$ and $\mathcal{F}_{t-1}$ with $1 \in A_t$, we have $\mathbb{E}_t\left[|\Pi_t^+ - \pi_t(1)|\right] \leq 2\delta_t |\Pi_t - \pi_t(1)|$ by (2.19), and so

$$\mathbb{E}_t\left[\frac{|\Pi_t^+ - \pi(1)|}{|\Pi_t - \pi_t(1)|}\right] \leq 2\delta_t \,.$$

Again, since the right hand side is independent of the conditioning in $\mathbb{E}_t$, we can replace the conditioning on $\mathcal{F}_{t-1}$ with conditioning on $1 \in A_t$; then, by Markov's inequality, for any $z > 0$ and $t \geq \log_2 \frac{1}{\Delta_2}$,

$$\Pr\left(\frac{|\Pi_t^+ - \pi(1)|}{|\Pi_t - \pi_t(1)|} > \frac{1}{z} \middle| 1 \in A_t\right) \leq \frac{\mathbb{E}\left[\frac{|\Pi_t^+ - \pi(1)|}{|\Pi_t - \pi_t(1)|}\right]}{z} \leq 2z\delta_t \,. \tag{2.22}$$

Now define the event

$$B(t) = \left\{\frac{|\Pi_t^+ - \pi_t(1)|}{|\Pi_t - \pi_t(1)|} \leq \frac{1}{z}\right\} \wedge \{\forall i \in \pi_t(1), i \neq 1, i \notin A_{t+1}\};$$

note that $\pi(1)$, and hence $B(t)$, is defined when $1 \in A_t$. Then, by (2.20),(2.22), and the union bound,

$$\Pr(B(t) | 1 \in A_t) \geq 1 - 2z\delta_t - 2\delta_t = 1 - 2(z+1)\delta_t. \tag{2.23}$$

Next we consider when the algorithm stops if $1 \in A_t$ and $B(t)$ happen in each phase $t \geq \log_2 \frac{1}{\Delta_2}$. Note that, denoting the last phase of the algorithm by T, the probability of this event can be bounded from below as

$$\begin{aligned}
&\Pr\left(\{\forall t \in [T], 1 \in A_t\} \wedge \{\forall \log_2 \frac{1}{\Delta_2} \leq t \leq T, B(t)\}\right) \\
&\geq 1 - \sum_{t=1}^{\infty}(2z+3)\delta_t = 1 - \sum_{t=1}^{\infty}\frac{(2z+3)\delta}{50t^3} \geq 1 - \frac{3\delta(2z+3)}{100}\,, \qquad (2.24)
\end{aligned}$$

by (2.21), (2.23), the union bound, and since $\sum_{t=1}^{\infty} 1/t^3 < 1 + \int_1^{\infty} 1/t^3 dt = 3/2$.

If $z > 1$ and $|\Pi_t| \leq z$, then

$$|\Pi_t^+ - \pi(1)| \leq \frac{|\Pi_t - \pi(1)|}{z} \leq \frac{z-1}{z} < 1\,,$$

which means that $\Pi_t^+ \subset \{\pi(1)\}$. Also, $B(t)$ implies that all suboptimal options in $\pi(1)$ are eliminated, which leads to the fact that only the optimal option 1 can survive after phase t. According to the algorithm, there must be at least one option left after the elimination of each phase, so we can conclude that if for some $z > 1$ and phase $t > \log_2 \frac{1}{\Delta_2}$, $\{1 \in A_t\}$, $|\Pi_t| \leq z$, and $B(t)$ holds, the algorithm must stop after this phase and return the optimal option $i^\star = 1$.

If $|\Pi_t| > z$, and $\{1 \in A_t\}$ and $B(t)$ holds, then

$$\begin{aligned}
\frac{|\Pi_t^+|}{|\Pi_t|} &\leq \frac{|\Pi_t^+ - \pi(1)| + 1}{|\Pi_t|} \leq \frac{|\Pi_t - \pi(1)| + z}{z|\Pi_t|} = \frac{|\Pi_t| + z - 1}{z|\Pi_t|} \\
&\leq \frac{(z-1)+(z+1)}{z(z+1)} = \frac{2}{z+1}\,,
\end{aligned}$$

Since repartitioning in the next phase will not increase the number of probes needed to cover A_{t+1} compared to Π_t^+, we have $|\Pi_{t+1}| \leq |\Pi_t^+|$. Therefore, for any $z > 1$ and $t \geq \log_2 \frac{1}{\Delta_2}$ such that $\{1 \in A_t\} \wedge B(t)$ holds, $|\Pi_t| > z$, implies

$$\frac{|\Pi_{t+1}|}{|\Pi_t|} \leq \frac{2}{z+1}\,. \qquad (2.25)$$

For simplicity, we choose $z = 15$. Then, by (2.24), the probability of the event $\{\forall t \in [T], 1 \in A_t\} \wedge \{\forall \log_2 \frac{1}{\Delta_2} \leq t \leq T, B(t)\}$ is at least $1 - \delta$; thus, it is enough to bound the probe complexity of the algorithm under the latter event. Assuming the event holds, by the choice of z we have that after $t \geq \log_2 \frac{1}{\Delta_2}$ phases, $|\Pi_t| \geq 16$ implies $\frac{|\Pi_{t+1}|}{|\Pi_t|} \leq \frac{1}{8}$, and the algorithm stops after phase t

if $|\Pi_t| \leq 15$. Let $s = \log_2 \frac{1}{\Delta_2}$. Then the algorithm must run into one of the following three cases: (a) $T < s$, (b) $T \geq s$ and $|\Pi_t| \geq 16$ for $s \leq t \leq T$, (c) $T \geq s$ and $|\Pi_t| \geq 16$ for $s \leq t \leq T-1$, $|\Pi_T| \leq 15$.

Here we only consider the last two cases where $T \geq s$; the upper bound obtained this way trivially hold for case (a), as well. For $T \geq s$, we divide the T phases into two parts: $1 \leq t < s$ and $s \leq t \leq T$. In the second part, by definition, $|\Pi_t| \geq 16$ for $s \leq t \leq T-1$, and so $|\Pi_t| \leq \mathcal{C}_{\mathrm{O}}([K], 1) \left(\frac{1}{8}\right)^{t-s}$ for $s \leq t \leq T$ by (2.25). Therefore, the probe complexity of the algorithm, without the samples used by the MEWP subroutine, is

$$\begin{aligned}
&\sum_{t=1}^{s-1} \sum_{\pi \in \Pi_t} \frac{2}{\varepsilon_t^2} \log \frac{2|\pi|}{\delta_t} + \sum_{t=s}^{T} \sum_{\pi \in \Pi_t} \frac{2}{\varepsilon_t^2} \log \frac{2|\pi|}{\delta_t} \\
&\leq 32 \mathcal{C}_{\mathrm{O}}([K], 1) \sum_{t=1}^{s-1} 4^t \log \frac{100 |p_{\max}| t^3}{\delta} \\
&+ 32 \mathcal{C}_{\mathrm{O}}([K], 1) \sum_{r=0}^{T-s} \left(\frac{1}{8}\right)^r 4^{r+s} \log \frac{100 |p_{\max}| (r+s)^3}{\delta} \, . \qquad (2.26)
\end{aligned}$$

Here the first term on the right hand side is clearly bounded from above by

$$C_1 \frac{\mathcal{C}_{\mathrm{O}}([K], 1)}{\Delta_2^2} \log \left(\frac{|p_{\max}|}{\delta} \log \frac{1}{\Delta_2} \right)$$

for some universal positive constant C_1, while the second term can be bounded as

$$\begin{aligned}
&C_2 \cdot \mathcal{C}_{\mathrm{O}}([K], 1) 4^s \left(\sum_{r=0}^{T-s} \frac{1}{2^r} \log \frac{s \cdot |p_{\max}|}{\delta} + \sum_{r=0}^{T-s} \frac{\log r}{2^r} \right) \\
&\leq C_3 \left(\frac{\mathcal{C}_{\mathrm{O}}([K], 1)}{\Delta_2^2} \log \left(\frac{|p_{\max}|}{\delta} \log \frac{1}{\Delta_2} \right) \right) \qquad (r + s \leq rs)
\end{aligned}$$

for universal contants $C_2, C_3 > 0$. In conclusion, the total probe complexity without the samples used by median elimination is

$$O \left(\frac{\mathcal{C}_{\mathrm{O}}([K], 1)}{\Delta_2^2} \log \left(\frac{|p_{\max}|}{\delta} \log \frac{1}{\Delta_2} \right) \right) .$$

The last thing is to show that the probe complexity of the MEWP subroutine is dominated by the above quantity. To show this, consider each phase t, the

number of probes used outside median elimination is $\sum_{\pi\in\Pi_t}\frac{2}{\varepsilon_t^2}\log\frac{2|\pi|}{\delta_t}$ which is relaxed to $\frac{2\mathcal{C}_{\mathrm{O}}(A_t,1)}{\varepsilon_t^2}\log\frac{2|p_{\max}|}{\delta_t}$ in our analysis. According to Theorem 6, MEWP in phase t uses $O\left(\frac{\mathcal{C}_{\mathrm{O}}(A_t,1)}{\varepsilon_t^2}\log\frac{|\pi_{\max}|}{\delta_t}\right)$ probes, where $|\pi_{\max}| = \max_{\pi\in\Pi_t}|\pi| \leq |p_{\max}|$. So taking the probe complexity of median elimination processes into account we still have the total probe complexity as

$$N = O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)\right).$$

. . . . . □

If COrcl is not guaranteed to return the optimal integer cover, the above theorem still holds by making the following modification to the algorithm to ensure that Π_{t+1} is not worse than Π_t for every t: if $|\{\pi\in\Pi_t : \pi\cap A_{t+1}\neq\varnothing\}| < \mathcal{C}_{\mathrm{O}}(A_{t+1},1)$, then use the same partition pattern from Π_t for Π_{t+1}.

Compared to the bound for SEWP, the $\log K$ term is replaced with $\log|p_{\max}|$. More specifically, in the full information case, the upper bound becomes $O\left(\frac{1}{\Delta_2^2}\log\left(\frac{K}{\delta}\log\frac{1}{\Delta_2}\right)\right)$, which is the same as the upper bound for SEWP. In the bandit case, the algorithm is exactly the same as the exponential gap elimination algorithm of Karnin et al. (2013), which enjoys an optimal $O\left(\sum_{i=1}^{K}\frac{1}{\Delta_i^2}\log\left(\frac{1}{\delta}\log\frac{1}{\Delta_i}\right)\right)$ upper bound on the number of probes, and is better than the upper bound for SEWP in bandit case. Therefore, although not formally proved, we expect that EGEWP enjoys an improved probe complexity compared with SEWP.

Comparing SEWP and EGEWP empirically

To empirically compare the performance between SEWP and EGEWP algorithms, we investigate the performance under three different probe settings: (a) the bandit case; (b) the full information case; and (c) an intermediate case where every subset of size $|p| = \sqrt{K}$ is a probe. For each scenario we consider two environments: (a) an *easy* problem where $\mu_1 = 0.3$ and $\mu_2 = ... = \mu_K = 0$ and (b) a *hard* problem where $\mu_1 = 1$ and $\mu_i = 1 - (i/n)^{0.5}$ for $i \neq 1$. Each reward distribution is Gaussian with variance $\sigma^2 = 1/4$. Under each combination of probe and distribution settings, we test the sample complexity for

different values of K with $\delta = 0.1$. In the experiments we report average probe usage over 100 runs. The results are shown in Figure 2.4.

The results show that EGEWP performs worse than the SEWP in all settings considered, despite its favorable asymptotic performance guarantees. This phenomenon is supported by the experimental studies by Jamieson et al. (2014) in the bandit case, in which the exponential gap elimination algorithm of Karnin et al. (2013) is shown to be worse than the successive elimination algorithm of Even-Dar et al. (2002).

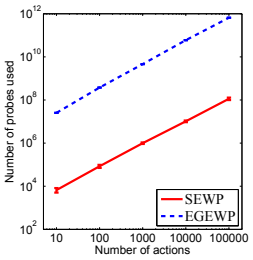

Figure 2.1: $|p| = 1$, easy case

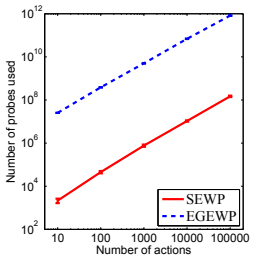

Figure 2.2: $|p| = 1$, hard case

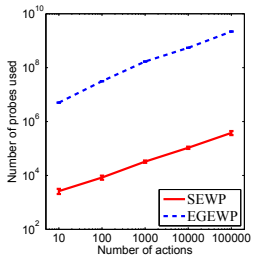

Figure 2.3: $|p| = \sqrt{K}$, easy case

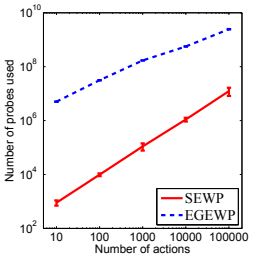

Figure 2.4: $|p| = \sqrt{K}$, hard case

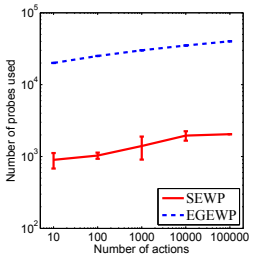

Figure 2.5: $|p| = K$, easy case

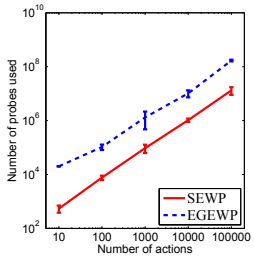

Figure 2.6: $|p| = K$, hard case

2.3 PAC Subset Selection

In this section, we consider the two PAC subset selection problems introduced in Section 2.1. The first, named *strong* PAC subset selection, is the same as the Explore-m problem introduced by Kalyanakrishnan and Stone (2010) where the goal is to find m (ε, m)-optimal options. The second problem, named *average* PAC subset selection, is to select a subset of m options with ε-optimal average reward, introduced by Zhou et al. (2014).

The basic idea of our approach is to generalize our SEWP algorithm with two modifications: (i) First, besides rejecting the options that cannot be in the best m options after each phase, we also accept options that have enough confidence to be one of the best m options, which shares a similar idea with the Racing algorithm in Kaufmann and Kalyanakrishnan (2013). (ii) Specific stopping conditions are designed to meet the ε-relaxation in the problem definition.

To make it easier to express the probe complexity, we introduce a new symbol $\Delta_i^{(\varepsilon,m)}$ defined by $\Delta_i^{(\varepsilon,m)} = \max\{\mu_i - \mu_{m+1}, \varepsilon\}$ if $i \leq m$ and $\Delta_i^{(\varepsilon,m)} = \max\{\mu_m - \mu_i, \varepsilon\}$ if $i > m$. We then sort $\Delta_i^{(\varepsilon,m)}$ for all $i \in [K]$ in ascending order and let $S_{(i)}$ be the first i options in the list, while $\Delta_{(i)}^{(\varepsilon,m)}$ denotes the i-th smallest entry.

Analogously to Theorem 1, let $f(t) = 2^t$, $g(t,\delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}}$, and define

$$\hat{N}_{(i)}(\varepsilon,\delta) = \frac{128}{\left(\Delta_{(i)}^{(\varepsilon,m)}\right)^2} \log\left(\frac{54K}{\delta}\log\frac{4}{\Delta_{(i)}^{(\varepsilon,m)}}\right) \tag{2.27}$$

and let $\hat{N}_{(K+1)}(\varepsilon,\delta) = 0$.

Note that $\hat{N}_{(1)}(\varepsilon,\delta) = \hat{N}_{(2)}(\varepsilon,\delta)$ since $\Delta_{(1)}^{(\varepsilon,m)} = \Delta_{(2)}^{(\varepsilon,m)} = \max\{\mu_m - \mu_{m+1}, \varepsilon\}$. Also let $\hat{M}_{(i)}(\varepsilon,\delta) \doteq \hat{N}_{(i)}(\varepsilon,\delta) - \hat{N}_{(i+1)}(\varepsilon,\delta)$.

2.3.1 *Strong* PAC Subset Selection

First we propose an algorithm that returns a subset $\hat{S}^*$ containing m (ε,m)-optimal options with high probability. An option i is defined to be (ε,m)-optimal iff $\mu_i \geq \mu_m - \varepsilon$. This requirement is the same as $q_{\min}(\hat{S}^*,\mu) \geq q_{\min}([m],\mu) - \varepsilon$ where $q_{\min}(S,\mu) = \min_{i\in S}\mu_i$.

The algorithm, called Successive Accept Reject with Probes (SARWP) is shown in Algorithm 4. The following theorem shows that Algorithm 4 is admissible and the probe complexity is bounded.

Theorem 10. *With probability at least $1-\delta$, SARWP returns a subset $\hat{S}^*$ of size m within N probes, where $q_{\min}(\hat{S}^*,\mu) \geq q_{\min}([m],\mu) - \varepsilon$ and N satisfies $N \leq \mathcal{G}_{LP}(O,\mathcal{P})\sum_{i=2}^{K}\hat{M}_{(i)}(\varepsilon,\delta)\mathcal{C}_{\mathrm{LP}}\left(S_{(i)},1\right)$.*

Algorithm 4 SuccessiveAcceptRejectWithProbes

Inputs: K, m, ε, δ, $\mathcal{P}$, observation scheduling function $f : \mathbb{N} \to \mathbb{N}$ and confidence function $g : \mathbb{N} \times (0, 1] \to [0, \infty)$.
Initialize candidate set $A_1 = [K]$, accepted options $A_1^a = \varnothing$, rejected options $A_1^r = \varnothing$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, f(t), \mathcal{P})$.
  Use each $p \in C(t)$ for $C_p(t)$-times to get new observations.
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations so far for option i. Sort the options in A_t in descending order of $\hat{\mu}_i(t)$. Let H_t be the first $m - |A_t^a|$ options and $L_t = A_t \setminus H_t$.
  if $\min_{i \in H_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) + 2g(t, \delta) - \varepsilon$ **then**
    Return $\hat{S}^* = A_t^a \cup H_t$ as selected subset.
  end if
  Let
  $A_{t+1}^a = A_t^a \cup \{i \in H_t : \hat{\mu}_i(t) > \max_{j \in L_t} \hat{\mu}_j(t) + 2g(t, \delta)\}$,
  $A_{t+1}^r = A_t^r \cup \{i \in L_t : \hat{\mu}_i(t) < \min_{j \in H_t} \hat{\mu}_j(t) - 2g(t, \delta)\}$,
  and $A_{t+1} = [K] - A_{t+1}^a - A_{t+1}^r$
end for

Proof of Theorem 10. Let T denote the number of phases that the algorithm runs until the stopping condition is satisfied and U denote the event that all confidence bounds hold throughout the process:

$$U = \{|\hat{\mu}_i(t) - \mu_i| \leq g(t, \delta) \text{ for all } (i, t) \text{ s.t. } 1 \leq t \leq T \text{ and } i \in A_t\}.$$

In the proof of Theorem 1, we have already shown that $\Pr(U) \geq 1 - \delta$. So the remaining of the proof contains two parts given the fact that U holds: (i) if T is finite thus $\hat{S}^*$ is returned, each option in $\hat{S}^*$ must be (ε, m)-optimal, and (ii) the probe complexity is upper bounded.

First we will show that if $T < \infty$ then each option $i \in \hat{S}^*$ must be (ε, m)-optimal. Since $\hat{S}^* = A_T^a \cup H_T$, if $i \in \hat{S}$, i belongs to either A_T^a or H_T. If $i \in A_T^a$, we use the following proposition to show that $1 \leq i \leq m$.

Proposition 11. *If U holds, then for any $2 \leq t \leq T$, if $i \in A_t^a$ then $1 \leq i \leq m$, if $i \in A_t^r$ then $m + 1 \leq i \leq K$.*

Proof. For $t = 2$, if $i \in A_2^a$, then $\hat{\mu}_i(1) > \max_{j \in L_1} \hat{\mu}_j(1) + 2g(1, \delta)$. Since $|L_1| = K - m$, we can find at least $K - m$ options such that for each j of them $\hat{\mu}_i(1) > \hat{\mu}_j(1) + 2g(1, \delta)$. From U we know that $\mu_i > \mu_j$, which means there

are at least $K-m$ options worse than i, hence $1 \leq i \leq m$ holds. On the other hand, if $i \in A_T^r$, for similar reason, we can find at least m options better than i and thus $m+1 \leq i \leq K$. Next we will show that if it holds for t then it also holds for $t+1$.

If it holds for t, then we have $\#\{i : 1 \leq i \leq m, i \in A_t\} = m - |A_t^a|$ and $\#\{i : m+1 \leq i \leq K, i \in A_t\} = K-m-|A_t^r|$. For $i \in A_t$ and $i \in A_{t+1}^a$, we can find at least $|L_t| = K-m-|A_t^r|$ options in A_t worse than i so $1 \leq i \leq m$ must hold. Similarly, for $i \in A_t$ and $i \in A_{t+1}^r$, we can find at least $|H_t| = m - |A_t^a|$ options in A_t better than i so $m+1 \leq i \leq K$ must hold. Then by induction, Proposition 11 holds. □

We now continue the proof of Theorem 10. Proposition 11 shows that if $i \in A_T^a$ then $1 \leq i \leq m$. For the other case, if $i \in H_T$, then $\hat{\mu}_i(T) \geq \max_{j \in L_T} \hat{\mu}_j(T) + 2g(T,\delta) - \varepsilon$. Next we will show that $\mu_i \geq \mu_m - \varepsilon$ by discussing the following two cases:

If $1 \leq i \leq m$ then $\mu_i \geq \mu_m - \varepsilon$ must hold. If $m+1 \leq i \leq K$, since $i \in H_T$ and all options in A_T^r must be $K-m$ worst, then there exists $1 \leq j \leq m$ such that $j \in L_T$ and thus $\hat{\mu}_i(T) \geq \hat{\mu}_j(T) + 2g(T,\delta) - \varepsilon$. Therefore $\mu_i \geq \mu_j - \varepsilon \geq \mu_m - \varepsilon$.

Now we have shown that if $T < \infty$, every option in $\hat{S}^* = A_T^a \cup H_T$ must be (ε, m)-optimal. Next we will prove that if U holds then the probe complexity is upper bounded by the following propositions.

Proposition 12. *For $1 \leq t < T$, $g(t,\delta) > \varepsilon/2$.*

Proof. If $g(t,\delta) \leq \varepsilon/2$, then $\min_{i \in H_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) + 2g(t,\delta) - \varepsilon$. The stopping condition is satisfied, thus $T = t$. □

Proposition 13. *For $1 \leq t < T$, if $i \in A_{t+1}$, then $g(t,\delta) \geq (\mu_i - \mu_{m+1})/4$ if $1 \leq i \leq m$, and $g(t,\delta) \geq (\mu_m - \mu_i)/4$ if $m+1 \leq i \leq K$.*

Proof. For $i \in A_t$, $1 \leq i \leq m$, if $g(t,\delta) < (\mu_i - \mu_{m+1})/4$, since $\#\{i : m+1 \leq i \leq K, i \in A_t\} = K - m - |A_t^r|$, there exist at least $K - m - |A_t^r|$ options in A_t such that for each j of them $\mu_i - \mu_j > 4g(t,\delta)$. Then

$$\hat{\mu}_i(t) - \hat{\mu}_j(t) \geq (\mu_i - g(t,\delta)) - (\mu_j + g(t,\delta)) = \mu_i - \mu_j - 2g(t,\delta) > 2g(t,\delta)\,.$$

Given the fact that L_t contains $K - m - |A_t^r|$ options with the lowest $\hat{\mu}_j(t)$s for $j \in A_t$, we have $\hat{\mu}_i(t) > \max_{j \in L_t} \hat{\mu}_j(t) + 2g(t,\delta)$, which means i must be accepted to A_{t+1}^a thus $i \notin A_{t+1}$.

Similarly, we can prove that for $i \in A_t$, $m+1 \leq i \leq K$, if $g(t,\delta) < (\mu_m - \mu_i)/4$, then i must be rejected to A_{t+1}^r. Now Proposition 13 has been proved. □

Combining Propositions 12 and 13 and the definition of $\Delta_i^{(\varepsilon,m)}$ we get that for $1 \leq t < T$, if $i \in A_{t+1}$, $g(t,\delta) \geq \Delta_i^{(\varepsilon,m)}/4$. Then following the proof of Theorem 1 gives the result of Theorem 10.

. . . . . □

The upper bound on the probe complexity is in a similar form to the one for SEWP in Theorem 1, while here the number of samples required for option i is determined by $\Delta_i^{(\varepsilon,m)}$ instead of Δ_i. This complexity measure matches existing work for the bandit case (Kalyanakrishnan et al., 2012; Kaufmann and Kalyanakrishnan, 2013). In the bandit case, the upper bound matches the worst case lower bound in Kalyanakrishnan et al. (2012): $\Omega(K\varepsilon^{-2}\log(m/\delta))$, up to logarithmic factors. We do not have a distribution dependent lower bound like Theorem 3 and even in the bandit case a distribution dependent lower bound for $\varepsilon > 0$ is still an open question (Kaufmann and Kalyanakrishnan, 2013).

2.3.2 *Average* PAC Subset Selection

Next we consider the problem that aims to find a subset whose aggregate regret is ε-optimal. Given a subset $S \subset [K]$ and $|S| = m$, the *aggregate regret* of S is defined as $R_S = \frac{1}{m}\left(\sum_{i\in[m]} \mu_i - \sum_{i \in S} \mu_i\right) = q_{\text{avg}}([m], \mu) - q_{\text{avg}}(S, \mu)$ where $q_{\text{avg}}(S,\mu) = \frac{1}{|S|}\sum_{i \in S} \mu_i$. The aggregate regret of S is said to be ε-optimal iff $R_S \leq \varepsilon$.

To address the problem of finding an average ε-optimal subset, Algorithm 4 can still be employed by only modifying the stopping condition according to the different objective. The new stopping condition is described as follows:

Stopping condition for average PAC subset selection: First for each $i \in A_t$, we construct an adversarial estimation $\hat{\mu}'_i(t)$ by setting $\hat{\mu}'_i(t) = \hat{\mu}_i(t) - g(t,\delta)$ if $i \in H_t$ and $\hat{\mu}'_i(t) = \hat{\mu}_i(t) + g(t,\delta)$ if $i \in L_t$. Then we sort the options in descending order according to $\hat{\mu}'_i(t)$ and let H'_t be the first $m - |A^a_t|$ options while L'_t be the remaining. The algorithm stops and returns $\hat{S}^* = A^a_t \cup H_t$ if

$$\sum_{i \in H_t \setminus H'_t} (\hat{\mu}_i(t) - g(t,\delta)) \geq \sum_{i \in H'_t \setminus H_t} (\hat{\mu}_i(t) + g(t,\delta)) - m\varepsilon\,.$$

This way of constructing "adversarial estimation" is similar to the one in the CLUCB algorithm of Chen et al. (2014), where the goal is to identify a subset with the highest reward sum without ε relaxation.

The next theorem shows that with the modified stopping condition, Algorithm 4 is admissible and bounds its probe complexity. Define

$$b(m,\varepsilon) = \max\left\{a \in \mathbb{N}^+ : \mu_{m-a+1} - \mu_{m+a} \leq \frac{m\varepsilon}{a}\right\}, \tag{2.28}$$

or $b(m,\varepsilon) = 1$ if $\mu_m - \mu_{m+1} > m\varepsilon$. Then we have the following result:

Theorem 14. *With probability at least $1-\delta$, SARWP with modified stopping condition returns a subset $\hat{S}^*$ of size m within N probes, where $q_{avg}(\hat{S}^*, \mu) \geq q_{avg}([m], \mu) - \varepsilon$ and N satisfies $N \leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_{(i)}(m\varepsilon/b, \delta) \mathcal{C}_{\mathrm{LP}}\left(S_{(i)}, 1\right)$, where $b = b(m,\varepsilon)$.*

Proof of Theorem 14. Let T denote the number of phases that the algorithm runs until the stopping condition is satisfied and U denote the event that all confidence bounds hold throughout the process:

$$U = \{|\hat{\mu}_i(t) - \mu_i| \leq g(t,\delta) \text{ for all } (i,t) \text{ s.t. } 1 \leq t \leq T \text{ and } i \in A_t\}\,.$$

We have $\Pr(U) \geq 1 - \delta$. Similar with the proof of Theorem 10, the remaining of the proof contains two parts given the fact that U holds: (i) if T is finite thus $\hat{S}^*$ is returned, the aggregate regret of $\hat{S}^*$ must be ε-optimal, and (ii) the probe complexity is upper bounded.

First we will show that if $T < \infty$ then $\frac{1}{m}\left(\sum_{i\in[m]} \mu_i - \sum_{i \in \hat{S}^*} \mu_i\right) \leq \varepsilon$. Recall that $\hat{S}^* = A^a_T \cup H_T$. The options in A^a_T incur no regret since Proposition 11

still holds and says that $A_T^a \subset [m]$. So we only need to show that

$$\sum_{i\in[m]\setminus A_T^a} \mu_i - \sum_{i\in H_T} \mu_i \le m\varepsilon\,.$$

Furthermore, it is equivalent to show

$$\sum_{i\in[m]\setminus A_T^a\setminus H_T} \mu_i - \sum_{i\in H_T\setminus[m]} \mu_i \le m\varepsilon\,.$$

Recall the stopping condition

$$\sum_{i\in H_T\setminus H_T'} (\hat{\mu}_i(T) - g(T,\delta)) \ge \sum_{i\in H_T'\setminus H_T} (\hat{\mu}_i(T) + g(T,\delta)) - m\varepsilon\,.$$

To show that the stopping condition is sufficient, we introduce some new notations:

Consider the sequence of options in A_t sorted by their $\hat{\mu}_i(t)$, let $a_t(i)$ be the option at the i-th position. Let

$$b_t = \max\left\{a \in \mathbb{N} \,:\, \hat{\mu}_{a_t(m_t-a+1)} - \hat{\mu}_{a_t(m_t+a)} < 2g(t,\delta)\right\}\,,$$

where $m_t = m - |A_t^a|$.

According to the construction of H_t' we know that $H_t = \{a_t(1), ..., a_t(m_t)\}$, $H_t' = \{a_t(1), ..., a_t(m_t - b_t), a_t(m_t+1), ..., a_t(m_t+b_t)\}$ and $|H_t\setminus H_t'| = |H_t'\setminus H_t| = b_t$.

Next we construct a set of pairs $Pair_T = \{(i,j)\}$ for $i \in [m]\setminus A_T^a \setminus H_T$ and $j \in H_T\setminus[m]$ as follows: sort $H_T\setminus[m]$ and $[m]\setminus A_T^a\setminus H_T$ both in descending order according to their $\hat{\mu}_i(T)$s (this is valid since $[m]\setminus A_T^a \subset A_T$ by Proposition 11) , then take last of $i \in [m]\setminus A_T^a \setminus H_T$ and the first $j \in H_T \setminus [m]$ as a pair into $Pair_T$, then repeat this procedure until no option remains (Note that $|\,[m]\setminus A_T^a \setminus H_T| = |H_T \setminus [m]\,|$). Since for each pair (i,j), $i \notin H_T$ and $j \in H_T$, we have $\hat{\mu}_j(T) \ge \hat{\mu}_i(T)$.

Then

$$\begin{aligned}&\sum_{i\in[m]\setminus A_T^a\setminus H_T} \mu_i - \sum_{i\in H_T\setminus[m]} \mu_i\\ &\le \sum_{(i,j)\in Pair_T} (\hat{\mu}_i(T) - \hat{\mu}_j(T) + 2g(T,\delta))\end{aligned}$$

$$\leq \sum_{(i,j)\in Pair_T^+} (2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T)))$$

where $Pair_T^+ = \{(i,j) \in Pair_T : \hat{\mu}_j(T) - \hat{\mu}_i(T) < 2g(T,\delta)\}$. Then we will show $|Pair_T^+| \leq b_T$. This is because, if $|Pair_T^+| > b_T$, then there must be a pair $(i,j) \in Pair_T^+$ such that $j \in H_T \cap H'_T$ and $i \notin H_T \cup H'_T$. Thus $\hat{\mu}_{a_T(m_T - b_T)}(T) - \hat{\mu}_{a_T(m_T + b_T + 1)}(T) \leq \hat{\mu}_j(T) - \hat{\mu}_i(T) < 2g(T,\delta)$ which contradicts the definition of b_t.

Next we construct another set of $|Pair_T^+|$ pairs (i,j) between $i \in H'_T \setminus H_T$ and $j \in H_T \setminus H'_T$ in the similar fashion: Let

$$Pair'_T = \left\{(a_T(m_T + |Pair_T^+|), a_T(m_T - |Pair_T^+| + 1)), ..., (a_T(m_T + 1), a_T(m_T))\right\} .$$

If we consider the pairs in $Pair_T^+$ and $Pair'_T$ in the order that they are constructed, then for each corresponding $(i,j) \in Pair_T^+$ and $(i',j') \in Pair'_T$, we have $\hat{\mu}_j(T) - \hat{\mu}_i(T) \geq \hat{\mu}_{j'}(T) - \hat{\mu}_{i'}(T)$. Therefore,

$$\begin{aligned}\sum_{i\in[m]\setminus A_T^a\setminus H_T} \mu_i - \sum_{i\in H_T\setminus[m]} \mu_i &\leq \sum_{(i,j)\in Pair_T^+} (2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T))) \\ &\leq \sum_{(i,j)\in Pair'_T} (2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T)))\end{aligned}$$

Consider the remaining pairs (i,j) between $i \in H'_T \setminus H_T$ and $j \in H_T \setminus H'_T$ which are not in $Pair'_T$, $2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T)) > 0$ still holds. Then we have

$$\begin{aligned}&\sum_{i\in[m]\setminus A_T^a\setminus H_T} \mu_i - \sum_{i\in H_T\setminus[m]} \mu_i \\ &\leq \sum_{i\in H'_T\setminus H_T} (\hat{\mu}_i(T) + g(T,\delta)) - \sum_{i\in H_T\setminus H'_T} (\hat{\mu}_i(T) - g(T,\delta)) \\ &\leq m\varepsilon\,.\end{aligned}$$

Now we have proved that the aggregate regret of $\hat{S}^*$ is ε-optimal. The remaining task is to upper bound the probe complexity.

Proposition 15. *For $1 \leq t < T$, $g(t,\delta) > m\varepsilon/4b$, where*

$$b = \max\left\{a \in \mathbb{N}^+ : \mu_{m-a+1} - \mu_{m+a} \leq \frac{m\varepsilon}{a}\right\} ,$$

or $b = 1$ if $\mu_m - \mu_{m+1} > m\varepsilon$.

Proof. The proposition is proved by showing that if $g(t,\delta) \leq m\varepsilon/4b$, the stopping condition must be satisfied after this phase. Recall the definition of $b_t = |H_t \setminus H'_t| = |H'_t \setminus H_t|$, we will first show that $b_t \leq b$.

If $b = \min\{m, K-m\}$, $b_t \leq b$ must hold. Next we discuss the case when $b < \min\{m, K-m\}$: Since $\mu_{m-b} - \mu_{m+b+1} > m\varepsilon/b \geq 4g(t,\delta)$, for any $1 \leq i \leq m-b$ and $m+b+1 \leq j \leq K$, if $i, j \in A_t$, then $\hat{\mu}_i(t) - \hat{\mu}_j(t) \geq \mu_i - \mu_j - 2g(t,\delta) > 2g(t,\delta)$. So there are at least $m - |A_t^a| - b = m_t - b$ options in A_t such that for each i of them $1 \leq i \leq m-b$, as well as at least $|A_t| - m_t - b$ options such that for each j of them $m+b+1 \leq j \leq K$. Since for each pair of such i, j, $\hat{\mu}_i(t) > \hat{\mu}_j(t)$, if $b_t > b$ then there must exist $1 \leq i \leq m-b$ and $m+b+1 \leq j \leq K$ such that $i, j \in (H_t \setminus H'_t) \cup (H'_t \setminus H_t)$. This is impossible because

$$\hat{\mu}_{a_t(m_t - b_t + 1)}(t) \geq \hat{\mu}_i(t) > \hat{\mu}_j(t) + 2g(t,\delta) \geq \hat{\mu}_{a_t(m_t+b_t)}(t) + 2g(t,\delta)\,,$$

which contradicts the definition of b_t. Hence $b_t \leq b$ holds.

Then

$$\begin{aligned}
&\sum_{i \in H'_t \setminus H_t} (\hat{\mu}_i(t) + g(t,\delta)) - \sum_{i \in H_t \setminus H'_t} (\hat{\mu}_i(t) - g(t,\delta)) \\
&= 2b_t g(t,\delta) + \sum_{i \in H'_t \setminus H_t} \hat{\mu}_i(t) - \sum_{i \in H_t \setminus H'_t} \hat{\mu}_i(t) \\
&\leq 2b_t g(t,\delta) \leq 2b_t \leq 2b \cdot \frac{m\varepsilon}{4b} \\
&\leq m\varepsilon\,,
\end{aligned}$$

which shows that the stopping condition is satisfied and thus the proposition holds. □

Note that Proposition 13 still holds here, together with Proposition 15 we get that for $1 \leq t < T$, if $i \in A_{t+1}$, $g(t,\delta) \geq \Delta_i^{(m\varepsilon/b, m)}/4$. Then following the proof of Theorem 1 gives the result of Theorem 14.

. . . . . □

Compared with Theorem 10, the complexity here is measured by $\Delta_i^{(m\varepsilon/b, m)}$ instead. This distribution dependent complexity measure is novel even in the

bandit case since the algorithm in Zhou et al. (2014) comes with distribution independent guarantee only. Regarding the worst case performance, since $b(m, \varepsilon) \leq \min\{m, K - m\}$, in bandit case our upper bound can be further relaxed to $O\left(\frac{K}{\varepsilon^2} \log\left(\frac{K}{\delta} \log \frac{1}{\varepsilon}\right)\right)$ if $m \leq K/2$ and $O\left(\frac{K(K-m)^2}{m^2\varepsilon^2} \log\left(\frac{K}{\delta} \log \frac{K-m}{m\varepsilon}\right)\right)$ if $m > K/2$. Compared with the worst case lower bound in Zhou et al. (2014): $\Omega\left(\frac{K}{\varepsilon^2}\left(1 + \frac{\log(1/\delta)}{m}\right)\right)$ for $m \leq K/2$ and $\Omega\left(\frac{K-m}{m} \cdot \frac{K}{\varepsilon^2}\left(\frac{K-m}{m} + \frac{\log(1/\delta)}{m}\right)\right)$ for $m > K/2$, although our upper bound does not exactly match this worse case lower bound, our distribution dependent quantity $b(m, \varepsilon)$ shows how the different $\frac{K}{\varepsilon^2}$ and $\frac{K(K-m)^2}{m^2\varepsilon^2}$ terms appear for small m and large m compared with $K/2$.

2.4 Summary

We introduced a generalized version of the best arm identification problem, where a decision maker can query multiple arms at a time. This generalization describes several real world problems that are not adequately modeled by the standard best-arm identification problem. We generalized several existing algorithms and provided distribution dependent upper and lower bounds on the probe complexity, and showed that our algorithms achieve essentially the best possible performance in special cases. In the PAC subset selection problems our complexity measure either matches existing works for the bandit case or provides some new insights. One very interesting question that remains for future work is whether there is a real gap between our lower and upper bounds. However, much work remains to be done: We view our paper as opening a whole new practical and exciting research area of investigating richer feedback structures in "winner selection" problems. Interesting questions include how to change the algorithms when the subsets to be returned are restricted, or when probes are associated with costs.

Chapter 3

Regret Minimization with Gaussian Side Observations

In this chapter we present our work on the regret minimization problem in the Gaussian side observation setting (Wu et al., 2015b). We assume that selecting an action i the learner can observe a random variable X_{ij} for each action j whose mean is the same as the payoff of j, but its variance σ_{ij}^2 depends on the pair (i, j). For simplicity, throughout the chapter we assume that all the payoffs and the X_{ij} are Gaussian. While in the graph-structured feedback case one either has observation on an action or not, but the observation always gives the same amount of information, our model is more refined: Depending on the value of σ_{ij}^2, the information can be of different quality. For example, if $\sigma_{ij}^2 = \infty$, trying action i gives no information about action j. In general, for any $\sigma_{ij}^2 < \infty$, the value of the information depends on the time horizon T of the problem: when σ_{ij}^2 is large relative to T (and the payoff differences of the actions) essentially no information is received, while a small variance results in useful observations.

After defining the problem formally in Section 3.1, we provide non-asymptotic problem-dependent lower bounds in Section 3.2, which depend on the distribution of the observations through their mean payoffs and variances. To our knowledge, these are the first such bounds presented for any stochastic partial monitoring problem beyond the full-information setting: previous work either presented asymptotic problem-dependent lower bounds (e.g., Graves and Lai, 1997; Buccapatnam et al., 2014), or finite-time minimax bounds (e.g., Bartók

et al., 2014; Alon et al., 2013, 2015). Our bounds can recover all previous bounds up to some universal constant factors not depending on the problem. In Section 3.3, we present two algorithms with finite-time performance guarantees for the case of graph-structured feedback without the self-observability assumption. While due to their complicated forms it is hard to compare our finite-time upper and lower bounds, we show that our first algorithm achieves the asymptotic problem-dependent lower bound up to problem-independent multiplicative factors. Regarding the minimax regret, the hardness ($\widetilde{\Theta}(T^{1/2})$ or $\widetilde{\Theta}(T^{2/3})$ regret[1]) of partial monitoring problems is characterized by their global/local observability property (Bartók et al., 2014) or, in case of the graph-structured feedback model, by their strong/weak observability property (Alon et al., 2015). In the same section we present another algorithm that achieves the minimax regret (up to logarithmic factors) under both strong and weak observability, and achieves an $O(\log^{3/2} T)$ problem-dependent regret. Earlier results for the stochastic graph-structured feedback problems (Caron et al., 2012; Buccapatnam et al., 2014) provided only asymptotic problem-dependent lower bounds and performance bounds that did not match the asymptotic lower bounds or the minimax rate up to constant factors. A related combinatorial partial monitoring problem with linear feedback was considered in Lin et al. (2014), where the presented algorithm was shown to satisfy both an $\widetilde{O}(T^{2/3})$ minimax bound and a logarithmic problem dependent bound. However, the dependence on the problem structure in that paper is not optimal, and, in particular, the paper does not achieve the $O(\sqrt{T})$ minimax bound for easy problems.

3.1 Problem Formulation

Formally, we consider an online learning problem with *Gaussian payoffs and side observations*: Suppose a learner has to choose from K actions in every round. When choosing an action, the learner receives a random payoff and also some side observations corresponding to other actions. More precisely, each

[1]Tilde denotes order up to logarithmic factors.

action $i \in [K] = \{1, \ldots, K\}$ is associated with some parameter θ_i, and the payoff $Y_{t,i}$ to action i in round t is normally distributed random variable with mean θ_i and variance σ_{ii}^2, while the learner observes a K-dimensional Gaussian random vector $X_{t,i}$ whose jth coordinate is a normal random variable with mean θ_j and variance σ_{ij}^2 (we assume $0 \leq \sigma_{ij} \leq \infty$) and the coordinates of $X_{t,i}$ are independent of each other. We assume the following: (i) the random variables $(X_t, Y_t)_t$ are independent for all t; (ii) the parameter vector θ is unknown to the learner but the variance matrix $\Sigma = (\sigma_{ij}^2)_{i,j \in [K]}$ is known in advance; (iii) $\theta \in [0, D]^K$ for some $D > 0$; (iv) $\min_{i \in [K]} \sigma_{ij} \leq \sigma < \infty$ for all $j \in [K]$, that is, the expected payoff of each action can be observed.

The goal of the learner is to maximize its payoff or, in other words, minimize the expected regret

$$R_T = T \max_{i \in [K]} \theta_i - \sum_{t=1}^{T} \mathbb{E}\left[Y_{t,i_t}\right]$$

where i_t is the action selected by the learner in round t. Note that the problem encompasses several common feedback models considered in online learning (modulo the Gaussian assumption), and makes it possible to examine more delicate observation structures:

Full information: $\sigma_{ij} = \sigma_j < \infty$ for all $i, j \in [K]$.

Bandit: $\sigma_{ii} < \infty$ and $\sigma_{ij} = \infty$ for all $i \neq j \in [K]$.

Partial monitoring with feedback graphs (Alon et al., 2015): Each action $i \in [K]$ is associated with an observation set $S_i \subset [K]$ such that $\sigma_{ij} = \sigma_j < \infty$ if $j \in S_i$ and $\sigma_{ij} = \infty$ otherwise.

We will call the *uniform variance* version of these problems when all the finite σ_{ij} are equal to some $\sigma \geq 0$. Some interesting features of the problem can be seen when considering the *generalized full information* case , when all entries of Σ are finite. In this case, the greedy algorithm, which estimates the payoff of each action by the average of the corresponding observed samples and selects the one with the highest average, achieves at most a constant regret

for any time horizon T.[2] On the other hand, the constant can be quite large: in particular, when the variance of some observations are large relative to the gaps $d_j = \max_i \theta_i - \theta_j$, the situation is rather similar to a partial monitoring setup for a smaller, finite time horizon. In this chapter we are going to analyze this problem and present algorithms and lower bounds that are able to "interpolate" between these cases and capture the characteristics of the different regimes.

3.1.1 Notation

Define $C_T^{\mathbb{N}} = \{c \in \mathbb{N}^K : c_i \geq 0\,, \sum_{i\in[K]} c_i = T\}$ and let $N(T) \in C_T^{\mathbb{N}}$ denote the number of plays over all actions taken by some algorithm in T rounds. Also let $C_T^{\mathbb{R}} = \{c \in \mathbb{R}^K : c_i \geq 0\,, \sum_{i\in[K]} c_i = T\}$. We will consider environments with different expected payoff vectors $\theta \in \Theta$, but the variance matrix Σ will be fixed. Therefore, an environment can be specified by θ; oftentimes, we will explicitly denote the dependence of different quantities on θ: The probability and expectation functionals under environment θ will be denoted by $\Pr(\cdot;\theta)$ and $\mathbb{E}[\cdot;\theta]$, respectively. Furthermore, let $i_j(\theta)$ be the jth best action (ties are broken arbitrarily, i.e., $\theta_{i_1} \geq \theta_{i_2} \geq \cdots \geq \theta_{i_K}$) and define $d_i(\theta) = \theta_{i_1(\theta)} - \theta_i$ for any $i \in [K]$. Then the expected regret under environment θ is $R_T(\theta) = \sum_{i\in[K]} \mathbb{E}[N_i(T);\theta]\, d_i(\theta)$. For any action $i \in [K]$, let $S_i = \{j \in [K] : \sigma_{ij} < \infty\}$ denote the set of actions whose parameter θ_j is observable by choosing action i. Throughout the chapter, log denotes the natural logarithm and Δ^n denotes the n-dimensional simplex for any positive integer n.

3.2 Lower Bounds

The aim of this section is to derive generic, problem-dependent lower bounds to the regret, which are also able to provide minimax lower bounds. The hardness in deriving such bounds is that for any fixed θ and Σ, the dumb algorithm that always selects $i_1(\theta)$ achieves zero regret (obviously, the regret of this algorithm

[2] To see this, notice that the error of identifying the optimal action decays exponentially with the number of rounds.

is linear for any θ' with $i_1(\theta) \neq i_1(\theta')$), so in general it is not possible to give a lower bound for a single instance. When deriving asymptotic lower bounds, this is circumvented by only considering *consistent* algorithms whose regret is sub-polynomial for any problem (Graves and Lai, 1997). However, this asymptotic notion of consistency is not applicable to finite-horizon problems. Therefore, following ideas of Li et al. (2015), for any problem we create a family of *related* problems (by perturbing the mean payoffs) such that if the regret of an algorithm is "too small" in one of the problems than it will be "large" in another one, while it still depends on the original problem parameters (note that deriving minimax bounds usually only involves perturbing certain special "worst-case" problems).

As a warm-up, and to show the reader what form of a lower bound can be expected, first we present an asymptotic lower bound for the uniform-variance version of the problem of *partial monitoring with feedback graphs*. The result presented below is an easy consequence of Graves and Lai (1997), hence its proof is omitted. An algorithm is said to be *consistent* if $\sup_{\theta\in\Theta} R_T(\theta) = o(T^\gamma)$ for every $\gamma > 0$. Now assume for simplicity that there is a unique optimal action in environment θ, that is, $\theta_{i_1(\theta)} > \theta_i$ for all $i \neq i_1$ and let

$$C_\theta = \left\{ c \in [0,\infty)^K \,:\, \sum_{i:j\in S_i} c_i \geq \frac{2\sigma^2}{d_j^2(\theta)} \;\text{ for all } j \neq i_1(\theta)\,,\; \sum_{i:i_1(\theta)\in S_i} c_i \geq \frac{2\sigma^2}{d_{i_2(\theta)}^2(\theta)} \right\}.$$

Then, for any consistent algorithm and for any θ with $\theta_{i_1(\theta)} > \theta_{i_2(\theta)}$,

$$\liminf_{T\to\infty} \frac{R_T(\theta)}{\log T} \geq \inf_{c\in C_\theta} \langle c, d(\theta)\rangle \;. \tag{3.1}$$

Note that the right hand side of (3.1) is 0 for any *generalized full information* problem (recall that the expected regret is bounded by a constant for such problems), but it is a finite positive number for other problems. Similar bounds have been provided in Caron et al. (2012); Buccapatnam et al. (2014) for graph-structured feedback with self-observability (under non-Gaussian assumptions on the payoffs). In the following we derive finite time lower bounds that are also able to replicate this result.

3.2.1 A General Finite Time Lower Bound

First we derive a general lower bound. For any $\theta, \theta' \in \Theta$ and $q \in \Delta^{|C_T^{\mathbb{N}}|}$, define $f(\theta, q, \theta')$ as

$$f(\theta, q, \theta') = \inf_{q' \in \Delta^{|C_T^{\mathbb{N}}|}} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \langle a, d(\theta') \rangle$$

$$\text{such that } \sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \leq \sum_{i \in [K]} \left(I_i(\theta, \theta') \sum_{a \in C_T^{\mathbb{N}}} q(a) a_i \right),$$

where $I_i(\theta, \theta')$ is the KL-divergence between $X_{t,i}(\theta)$ and $X_{t,i}(\theta')$, given by $I_i(\theta, \theta') = \mathrm{KL}(X_{t,i}(\theta); X_{t,i}(\theta')) = \sum_{j=1}^{K} (\theta_j - \theta'_j)^2 / 2\sigma_{ij}^2$. Clearly, $f(\theta, q, \theta')$ is a lower bound on $R_T(\theta')$ for any algorithm for which the distribution of $N(T)$ is q. The intuition behind the allowed values of q' is that we want q' to be as similar to q as the environments θ and θ' look like for the algorithm (through the feedback $(X_{t,i_t})_t$). Now define

$$g(\theta, c) = \inf_{q \in \Delta^{|C_T^{\mathbb{N}}|}} \sup_{\theta' \in \Theta} f(\theta, q, \theta'), \qquad \text{such that } \sum_{a \in C_T^{\mathbb{N}}} q(a) a = c \in C_T^{\mathbb{R}}.$$

$g(\theta, c)$ is a lower bound of the worst-case regret of any algorithm with $\mathbb{E}[N(T); \theta] = c$. Finally, for any $x > 0$, define

$$b(\theta, x) = \inf_{c \in C_{\theta,x}} \langle c, d(\theta) \rangle \qquad \text{where} \quad C_{\theta,x} = \{c \in C_T^{\mathbb{R}} \,;\, g(\theta, c) \leq x\}.$$

Here $C_{\theta,x}$ contains all the possible values of $\mathbb{E}[N(T); \theta]$ that can be achieved by some algorithm whose lower bound g on the worst-case regret is smaller than x. These definitions give rise to the following theorem:

Theorem 16. *Given any $B > 0$, for any algorithm such that $\sup_{\theta' \in \Theta} R_T(\theta') \leq B$, we have, for any environment $\theta \in \Theta$, $R_T(\theta) \geq b(\theta, B)$.*

Proof of Theorem 16 . Let $\phi_{\theta,\sigma}$ denote the density function of a K-dimensional Gaussian random variable with mean vector θ and independent components wehere the variance of the ith coordinate is σ_i^2, and define

$$L_T = \sum_{t=1}^{T} \log \frac{\phi_{\theta,\sigma_{i_t}}(X_{t,i_t})}{\phi_{\theta',\sigma_{i_t}}(X_{t,i_t})}$$

where i_t is the choice of the algorithm in round t. Let $q, q' \in \Delta^{|C_T^{\mathbb{N}}|}$ denote the joint distribution over the number of plays for each action under environment θ and $\theta' \in \Theta$, respectively, that is, $q(a) = \Pr(N(T) = a; \theta)$ and $q'(a) = \Pr(N(T) = a; \theta')$ for each $a \in C_T^{\mathbb{N}}$.

For any $a \in C_T^{\mathbb{N}}$, applying a standard change of measure equality (see, e.g., Kaufmann et al., 2015a, Lemma 15), we obtain

$$\begin{aligned}
q'(a) &= \Pr(N(T) = a; \theta') = \mathbb{E}\left[\mathbb{I}\left\{N(T) = a\right\} \exp(-L_T); \theta\right] \\
&= \mathbb{E}\left[\mathbb{I}\left\{N(T) = a\right\} \mathbb{E}\left[\exp(-L_T) | N(T) = a; \theta\right]; \theta\right] \\
&\geq \mathbb{E}\left[\mathbb{I}\left\{N(T) = a\right\} \exp\left(\mathbb{E}\left[-L_T | N(T) = a; \theta\right]\right); \theta\right] \\
&= \Pr(N(T) = a; \theta) \exp\left(\mathbb{E}\left[-L_T | N(T) = a; \theta\right]\right) \\
&= q(a) \exp\left(\mathbb{E}\left[-L_T | N(T) = a; \theta\right]\right) .
\end{aligned}$$

Thus $\mathbb{E}\left[L_T | N(T) = a; \theta\right] \geq \log \frac{q(a)}{q'(a)}$ and so

$$\begin{aligned}
\sum_{i \in [K]} \mathbb{E}\left[N_i(T); \theta\right] I_i(\theta, \theta') &= \mathbb{E}\left[L_T; \theta\right] \\
&= \sum_{a \in C_T^{\mathbb{N}}} \Pr(N(T) = a; \theta) \, \mathbb{E}\left[L_T | N(T) = a; \theta\right] \\
&\geq \sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} ,
\end{aligned}$$

where $\mathbb{E}\left[N_i(T); \theta\right] = \sum_{a \in C_T^{\mathbb{N}}} q(a) a_i$. Therefore, according to the definition of $f(\theta, q, \theta')$, we have $f(\theta, q, \theta') \leq \sum_{a \in C_T^{\mathbb{N}}} q'(a) \langle a, d(\theta') \rangle = R_T(\theta')$ for any $\theta' \in \Theta$. Then $\sup_{\theta' \in \Theta} f(\theta, q, \theta') \leq \sup_{\theta' \in \Theta} R_T(\theta') \leq B$ must hold. Since $\mathbb{E}\left[N(T); \theta\right] = \sum_{a \in C_T^{\mathbb{N}}} q(a) a$ we have $g(\theta, \mathbb{E}\left[N(T); \theta\right]) \leq \sup_{\theta' \in \Theta} f(\theta, q, \theta') \leq B$. Thus $\mathbb{E}\left[N(T); \theta\right] \in C_{\theta, B}$ and so $R_T(\theta) \geq b(\theta, B)$, which concludes the proof of Theorem 16.

. . . . . □

Remark 17. If B is picked as the minimax value of the problem given the observation structure Σ, the theorem states that for any minimax optimal algorithm the expected regret for a certain θ is lower bounded by $b(\theta, B)$.

3.2.2 A Relaxed Lower Bound

Now we introduce a relaxed but more interpretable version of the finite-time lower bound of Theorem 16, which can be shown to match the asymptotic lower bound (3.1). The idea of deriving the lower bound is the following: instead of ensuring that the algorithm performs well in the most adversarial environment θ', we consider a set of "bad" environments and make sure that the algorithm performs well on them, where each "bad" environment θ' is the most adversarial one by only perturbing one coordinate θ_i of θ.

However, in order to get meaningful finite-time lower bounds, we need to perturb θ more carefully than in the case of asymptotic lower bounds. The reason for this is that for any sub-optimal action i, if θ_i is very close to $\theta_{i_1(\theta)}$, then $\mathbb{E}[N_i(T);\theta]$ is not necessarily small for a good algorithm for θ. If it is small, one can increase θ_i to obtain an environment θ' where i is the best action and the algorithm performs bad; otherwise, when $\mathbb{E}[N_i(T);\theta]$ is large, we need to decrease θ_i to make the algorithm perform badly in θ'. Moreover, when perturbing θ_i to be better than $\theta_{i_1(\theta)}$, we cannot make $\theta'_i - \theta_{i_1(\theta)}$ arbitrarily small as in asymptotic lower-bound arguments, because when $\theta'_i - \theta_{i_1(\theta)}$ is small, large $\mathbb{E}\left[N_{i_1(\theta)};\theta'\right]$, and not necessarily large $\mathbb{E}[N_i(T);\theta']$, may also lead to low finite-time regret in θ'. In the following we make this argument precise to obtain an interpretable lower bound.

Formulation

We start with defining a subset of $C_T^{\mathbb{R}}$ that contains the set of "reasonable" values for $\mathbb{E}[N(T);\theta]$. For any $\theta \in \Theta$ and $B > 0$, let

$$C'_{\theta,B} = \left\{ c \in C_T^{\mathbb{R}} \,:\, \sum_{j=1}^{K} \frac{c_j}{\sigma_{ji}^2} \geq m_i(\theta, B) \;\text{ for all } i \in [K] \right\}$$

where m_i, the minimum sample size required to distinguish between θ_i and its worst-case perturbation, is defined as follows: For $i \neq i_1$, if $\theta_{i_1} = D$,[3] then $m_i(\theta, B) = 0$. Otherwise let

$$m_{i,+}(\theta, B) = \max_{\varepsilon \in (d_i(\theta), D-\theta_i]} \tfrac{1}{\varepsilon^2} \log \tfrac{T(\varepsilon - d_i(\theta))}{8B},$$

[3]Recall that $\theta_i \in [0, D]$.

$$m_{i,-}(\theta, B) = \max_{\varepsilon \in (0, \theta_i]} \tfrac{1}{\varepsilon^2} \log \tfrac{T(\varepsilon + d_i(\theta))}{8B},$$

and let $\varepsilon_{i,+}$ and $\varepsilon_{i,-}$ denote the value of ε achieving the maximum in $m_{i,+}$ and $m_{i,-}$, respectively. Then, define

$$m_i(\theta, B) = \begin{cases} m_{i,+}(\theta, B) & \text{if } d_i(\theta) \geq 4B/T; \\ \min\left\{m_{i,+}(\theta, B), m_{i,-}(\theta, B)\right\} & \text{if } d_i(\theta) < 4B/T \ . \end{cases}$$

For $i = i_1$, then $m_{i_1}(\theta, B) = 0$ if $\theta_{i_2(\theta)} = 0$, else the definitions for $i \neq i_1$ change by replacing $d_i(\theta)$ with $d_{i_2(\theta)}(\theta)$ (and switching the $+$ and $-$ indices):

$$m_{i_1(\theta),-}(\theta, B) = \max_{\varepsilon \in (d_{i_2(\theta)}(\theta), \theta_{i_1(\theta)}]} \tfrac{1}{\varepsilon^2} \log \tfrac{T(\varepsilon - d_{i_2(\theta)}(\theta))}{8B},$$
$$m_{i_1(\theta),+}(\theta, B) = \max_{\varepsilon \in (0, D - \theta_{i_1(\theta)}]} \tfrac{1}{\varepsilon^2} \log \tfrac{T(\varepsilon + d_{i_2(\theta)}(\theta))}{8B}$$

where $\varepsilon_{i_1(\theta),-}$ and $\varepsilon_{i_1(\theta),+}$ are the maximizers for ε in the above expressions. Then, define

$$m_{i_1(\theta)}(\theta, B) = \begin{cases} m_{i_1(\theta),-}(\theta, B) & \text{if } d_{i_2(\theta)}(\theta) \geq 4B/T; \\ \min\left\{m_{i_1(\theta),+}(\theta, B), m_{i_1(\theta),-}(\theta, B)\right\} & \text{if } d_{i_2(\theta)}(\theta) < 4B/T \ . \end{cases}$$

Note that $\varepsilon_{i,+}$ and $\varepsilon_{i,-}$ can be expressed in closed form using the Lambert $W : \mathbb{R} \to \mathbb{R}$ function satisfying $W(x)e^{W(x)} = x$: for any $i \neq i_1(\theta)$,

$$\begin{aligned} \varepsilon_{i,+} &= \min\left\{ D - \theta_i \, , \, 8\sqrt{e}Be^{W\left(\frac{d_i(\theta)T}{16\sqrt{e}B}\right)}/T + d_i(\theta) \right\} \ , \\ \varepsilon_{i,-} &= \min\left\{ \theta_i \, , \, 8\sqrt{e}Be^{W\left(-\frac{d_i(\theta)T}{16\sqrt{e}B}\right)}/T - d_i(\theta) \right\}, \end{aligned} \tag{3.2}$$

and similar results hold for $i = i_1$, as well.

Now we can give the main result of this section, a simplified version of Theorem 16:

Corollary 18. *Given $B > 0$, for any algorithm such that $\sup_{\lambda \in \Theta} R_T(\lambda) \leq B$, we have, for any environment $\theta \in \Theta$, $R_T(\theta) \geq b'(\theta, B) = \min_{c \in C'_{\theta,B}} \langle c, d(\theta) \rangle$.*

Proof of Corollary 18. We start the proof with two technical lemmas on the Lambert W function.

Lemma 19. *Let $a, b > 0$ with $ab < 1$ and $f(x) = \frac{1}{x^2}\log((x+a)b)$ for $x > 0$. Then $f(x) \le f(x_*)$ for all $x > 0$ where*

$$x_* = \frac{\sqrt{e}}{b} e^{W\left(-\frac{ab}{2\sqrt{e}}\right)} - a \ .$$

Proof.

$$f'(x) = \frac{x^{-3}}{x+a}\left(x - 2(x+a)\log((x+a)b)\right) \ .$$

Let $g(y) = y - a - 2y\log by$ defined on $y > a$.

$$g'(y) = -2\log yb - 1 \, .$$

If $ab \le \frac{1}{\sqrt{e}}$ then $g(y)$ is increasing when $a < y < \frac{1}{b\sqrt{e}}$ and decreasing when $y > \frac{1}{b\sqrt{e}}$. If $\frac{1}{\sqrt{e}} < ab < 1$ then $g(y)$ is decreasing on $y > a$.

Since $\lim_{y\to a} g(y) = -2a\log ab > 0$ and $\lim_{y\to+\infty} g(y) = -\infty$ we know that there exists a unique $y_* > a$ such that $g(y_*) = 0$, $g(y) > 0$ for $a < y < y_*$ and $g(y) < 0$ for $y > y_*$. It can be verified that $y_* = x_* + a = \frac{\sqrt{e}}{b} e^{W\left(-\frac{ab}{2\sqrt{e}}\right)}$ satisfies $g(y_*) = 0$ and $y_* > a$ (which comes from $W(-\frac{ab}{2\sqrt{e}}) > W(-\frac{1}{2\sqrt{e}}) = -\frac{1}{2}$ and thus $y_* > \frac{1}{b} > a$). Therefore $f'(x) > 0$ when $0 < x < x_*$ and $f'(x) < 0$ when $x > x_*$. Since $f(x)$ is continuous when $x > 0$ we have proved that $f(x) \le f(x_*)$ for all $x > 0$.

□

Lemma 20. *Let $a, b > 0$ and $f(x) = \frac{1}{x^2}\log((x-a)b)$ for $x > a$. Then $f(x) \le f(x_*)$ for all $x > a$ where*

$$x_* = \frac{\sqrt{e}}{b} e^{W\left(\frac{ab}{2\sqrt{e}}\right)} + a \ .$$

Proof.

$$f'(x) = \frac{x^{-3}}{x-a}\left(x - 2(x-a)\log((x-a)b)\right) \ .$$

Let $g(y) = y + a - 2y\log by$ defined on $y > 0$.

$$g'(y) = -2\log yb - 1 \, .$$

So $g(y)$ is increasing when $0 < y < \frac{1}{b\sqrt{e}}$ and decreasing when $y > \frac{1}{b\sqrt{e}}$.

Since $\lim_{y\to 0} g(y) = a > 0$ and $\lim_{y\to+\infty} g(y) = -\infty$ we know that there exists a unique $y_* > 0$ such that $g(y_*) = 0$, $g(y) > 0$ for $0 < y < y_*$ and $g(y) < 0$ for $y > y_*$. It can be verified that $y_* = x_* - a = \frac{\sqrt{e}}{b} e^{W\left(\frac{ab}{2\sqrt{e}}\right)}$ satisfies $g(y_*) = 0$. Therefore $f'(x) > 0$ when $a < x < x_*$ and $f'(x) < 0$ when $x > x_*$. Since $f(x)$ is continuous when $x > a$ we have proved that $f(x) \le f(x_*)$ for all $x > a$.

□

To prove Corollary 18, it suffices to show $b'(\theta, B) \le b(\theta, B)$.

Define $C'_{\theta,B} = \left\{ c \in C_T^{\mathbb{R}} : \sum_{j=1}^K \frac{c_j}{\sigma_{ji}^2} \ge m_i(\theta, B) \text{ for all } i \in [K] \right\}$. We will prove $C_{\theta,B} \subset C'_{\theta,B}$ by showing that if $c \in C_T^{\mathbb{R}}$ satisfies $g(\theta, c) \le B$ then $c \in C'_{\theta,B}$.

For $c \in C_T^{\mathbb{R}}$, if $g(\theta, c) \le B$, then there exists $q \in \Delta^{|C_T^{\mathbb{N}}|}$ such that

$$\sup_{\theta' \in \Theta} f(\theta, q, \theta') \le B \text{ and } \sum_{a \in C_T^{\mathbb{N}}} q(a) a = c\,.$$

We will next derive K constraints on c to show that $c \in C'_{\theta,B}$ by picking different θ's. Before proceeding with the proof, we introduce the following technical lemma:

Lemma 21. *For any $A \subset C_T^{\mathbb{N}}$ and $q, q' \in \Delta^{|C_T^{\mathbb{N}}|}$, if $q(A) \ge 1/2$ then*

$$\sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \ge \frac{1}{2} \log \frac{1}{4q'(A)},$$

where $q'(A) = \sum_{a \in A} q'(a)$.

Proof. Let $A^c = C_T^{\mathbb{N}} - A$. By the log-sum inequality we have

$$\sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \ge \mathrm{KL}(q(A), q'(A))\,, \tag{3.3}$$

where for $x, y \in [0,1]$, $\mathrm{KL}(x,y) = x \log(x/y) + (1-x)\log((1-x)/(1-y))$ denotes the binary KL-divergence. Now for such x, y, since $x \log x + (1-x)\log(1-x)$ is minimized for $x = 1/2$, we have

$$\mathrm{KL}(x,y) \ge \log\frac{1}{2} + x \log\frac{1}{y} + (1-x)\log(\frac{1}{1-y}) \ge \log\frac{1}{2} + \frac{1}{2}\log\frac{1}{y} = \frac{1}{2}\log\frac{1}{4y}\,.$$

Combining with (3.3) proves the lemma. □

Now we continue the proof of Corollary 18. First consider $i \neq i_1(\theta)$.

If $\sum_{a:a_i \leq T/2} q(a) \geq 1/2$, construct $\theta^{(i,+)}$ by replacing θ_i with $\theta_i + \varepsilon_{i,+}$. Then $f(\theta, q, \theta^{(i,+)}) \leq B$ holds, so there exists $q' \in \Delta^{|C_T^{\mathbb{N}}|}$ such that

$$\sum_{a \in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,+)}) \right\rangle \leq B$$

and

$$\sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \leq \sum_{j \in [K]} c_j I_j(\theta, \theta^{(i,+)}) \,.$$

Applying Lemma 21 with $A = \{a : a_i \leq T/2\}$ gives

$$\sum_{j \in [K]} c_j I_j(\theta, \theta^{(i,+)}) \geq \frac{1}{2} \log \frac{1}{4q'(A)} \,,$$

where

$$\begin{aligned}
q'(A) &= \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ \sum_{j \neq i} a_j \geq T/2 \right\} q'(a) \leq \frac{2}{T} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \sum_{j \neq i} a_j \\
&= \frac{2}{T(\varepsilon_{i,+} - d_i(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \sum_{j \neq i} a_j (\varepsilon_{i,+} - d_i(\theta)) \\
&\leq \frac{2}{T(\varepsilon_{i,+} - d_i(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,+)}) \right\rangle \\
&\leq \frac{2B}{T(\varepsilon_{i,+} - d_i(\theta))} \,.
\end{aligned}$$

Since $I_j(\theta, \theta^{(i,+)}) = \varepsilon_{i,+}^2 / 2\sigma_{ji}^2$, we get

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\varepsilon_{i,+}^2} \log \frac{T(\varepsilon_{i,+} - d_i(\theta))}{8B} \,. \tag{3.4}$$

If $\sum_{a:a_i \leq T/2} q(a) < 1/2$ and $d_i(\theta) \geq 4B/T$, then

$$\begin{aligned}
f(\theta, q, \theta) &= \sum_{a \in C_T^{\mathbb{N}}} q(a) \left\langle a, d(\theta) \right\rangle \geq \sum_{a \in C_T^{\mathbb{N}}} q(a) a_i d_i(\theta) \\
&\geq d_i(\theta) \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ a_i \geq T/2 \right\} q(a) a_i \\
&\geq \frac{4B}{T} \frac{T}{2} \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ a_i \geq T/2 \right\} q(a) > B \,,
\end{aligned}$$

which contradicts the fact that $\sup_{\theta' \in \Theta} f(\theta, q, \theta') \leq B$.

If $\sum_{a:a_i \leq T/2} q(a) < 1/2$ and $d_i(\theta) < 4B/T$, construct $\theta^{(i,-)}$ by replacing θ_i with $\theta_i - \varepsilon_{i,-}$. Then there exists $q' \in \Delta^{|C_T^{\mathbb{N}}|}$ such that $\sum_{a \in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,-)}) \right\rangle \leq B$ and $\sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \leq \sum_{j \in [K]} c_j I_j(\theta, \theta^{(i,-)})$. Applying Lemma 21 with $A = \{a : a_i > T/2\}$ gives

$$\sum_{j \in [K]} c_j I_j(\theta, \theta^{(i,-)}) \geq \frac{1}{2} \log \frac{1}{4q'(A)} \,,$$

where

$$\begin{aligned} q'(A) &= \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{a_i > T/2\right\} q'(a) \leq \frac{2}{T} \sum_{a \in C_T^{\mathbb{N}}} a_i q'(a) \\ &\leq \frac{2}{T(\varepsilon_{i,-} + d_i(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) a_i (\varepsilon_{i,-} + d_i(\theta)) \\ &\leq \frac{2}{T(\varepsilon_{i,-} + d_i(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,-)}) \right\rangle \leq \frac{2B}{T(\varepsilon_{i,-} + d_i(\theta))} \,. \end{aligned}$$

Using $I_j(\theta, \theta^{(i,-)}) = \varepsilon_{i,-}^2 / 2\sigma_{ji}^2$ gives

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\varepsilon_{i,-}^2} \log \frac{T(\varepsilon_{i,-} + d_i(\theta))}{8B} \,. \tag{3.5}$$

Now consider $i = i_1(\theta)$.

If $\sum_{a:a_i \geq T/2} q(a) \geq 1/2$, construct $\theta^{(i_1,-)}$ by replacing θ_i with $\theta_i - \varepsilon_{i,-}$. Then there exists $q' \in \Delta^{|C_T^{\mathbb{N}}|}$ such that $\sum_{a \in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,-)}) \right\rangle \leq B$ and $\sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \leq \sum_{j \in [K]} c_j I_j(\theta, \theta^{(i,-)})$. Applying Lemma 21 with $A = \{a : a_i \geq T/2\}$ and

$$\begin{aligned} q'(A) = \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{a_i \geq T/2\right\} q'(a) &\leq \frac{2}{T(\varepsilon_{i,-} - d_{i_2(\theta)}(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) a_i (\varepsilon_{i,-} - d_{i_2(\theta)}(\theta)) \\ &\leq \frac{2B}{T(\varepsilon_{i,-} - d_{i_2(\theta)}(\theta))} \end{aligned}$$

gives

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\varepsilon_{i,-}^2} \log \frac{T(\varepsilon_{i,-} - d_{i_2(\theta)}(\theta))}{8B} \,. \tag{3.6}$$

If $\sum_{a:a_i \geq T/2} q(a) < 1/2$ and $d_{i_2(\theta)}(\theta) \geq 4B/T$, then

$$\begin{aligned} f(\theta, q, \theta) &= \sum_{a \in C_T^{\mathbb{N}}} q(a) \langle a, d(\theta) \rangle \geq \sum_{a \in C_T^{\mathbb{N}}} q(a) d_{i_2(\theta)} \sum_{j \neq i} a_j \\ &\geq d_{i_2(\theta)} \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ \sum_{j \neq i} a_j > T/2 \right\} q(a) \sum_{j \neq i} a_j \\ &> \frac{4B}{T} \frac{T}{2} \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ \sum_{j \neq i} a_j > T/2 \right\} q(a) \\ &\geq B \,, \end{aligned}$$

which contradicts the fact that $\sup_{\theta' \in \Theta} f(\theta, q, \theta') \leq B$.

If $\sum_{a:a_i \geq T/2} q(a) < 1/2$ and $d_{i_2(\theta)}(\theta) < 4B/T$, construct $\theta^{(i,+)}$ by replacing θ_i with $\theta_i + \varepsilon_{i,+}$. Then there exists $q' \in \Delta^{|C_T^{\mathbb{N}}|}$ such that $\sum_{a \in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,+)}) \right\rangle \leq B$ and $\sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \leq \sum_{j \in [K]} c_j I_j(\theta, \theta^{(i,+)})$. Applying Lemma 21 with $A = \{a : a_i < T/2\}$ and

$$\begin{aligned} q'(A) &= \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ \sum_{j \neq i} a_j > T/2 \right\} q'(a) \leq \frac{2}{T} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \sum_{j \neq i} a_j \\ &= \frac{2}{T(\varepsilon_{i,+} + d_{i_2(\theta)}(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \sum_{j \neq i} a_j (\varepsilon_{i,+} + d_{i_2(\theta)}) \leq \frac{2B}{T(\varepsilon_{i,+} + d_{i_2(\theta)})} \end{aligned}$$

gives

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\varepsilon_{i,+}^2} \log \frac{T(\varepsilon_{i,+} + d_{i_2(\theta)})}{8B} \,. \tag{3.7}$$

Combining (3.4) (3.5) (3.6) (3.7) gives $c \in C'_{\theta,B}$, which concludes the proof. . . . . □

Next we compare this bound to existing lower bounds.

Comparison to the Asymptotic Lower Bound (3.1)

Now we will show that our finite time lower bound in Corollary 18 matches the asymptotic lower bound in (3.1) up to some constants. Pick $B = \alpha T^\beta$ for some $\alpha > 0$ and $0 < \beta < 1$. For simplicity, we only consider θ which is "away from" the boundary of Θ (so that the minima in (3.2) are achieved by

the second terms) and has a unique optimal action. Then, for $i \neq i_1(\theta)$, it is easy to show that $\varepsilon_{i,+} = \frac{d_i(\theta)}{2W(d_i(\theta)T^{1-\beta}/(16\alpha\sqrt{e}))} + d_i(\theta)$ by (3.2) and $m_i(\theta, B) = \frac{1}{\varepsilon_{i,+}^2} \log \frac{T(\varepsilon_{i,+} - d_i(\theta))}{8B}$ for large enough T. Then, using the fact that $\log x - \log\log x \leq W(x) \leq \log x$ for $x \geq e$, it follows that $\lim_{T\to\infty} m_i(\theta, B)/\log T = (1-\beta)/d_i^2(\theta)$, and similarly we can show that $\lim_{T\to\infty} m_{i_1(\theta)}(\theta, B)/\log T = (1-\beta)/d_{i_2(\theta)}^2(\theta)$. Thus, $C'_{\theta,B} \to \frac{(1-\beta)\log T}{2} C_\theta$, under the assumptions of (3.1), as $T \to \infty$. This implies that Corollary 18 matches the asymptotic lower bound of (3.1) up to a factor of $(1-\beta)/2$.

Comparison to Minimax Bounds

Now we will show that our θ-dependent finite-time lower bound reproduces the minimax regret bounds of Mannor and Shamir (2011) and Alon et al. (2015), except for the generalized full information case.

The minimax bounds depend on the following notion of observability: An action i is *strongly observable* if either $i \in S_i$ or $[K] \setminus \{i\} \subset \{j : i \in S_j\}$. i is *weakly observable* if it is not strongly observable but there exists j such that $i \in S_j$ (note that we already assumed the latter condition for all i). Let $\mathcal{W}(\Sigma)$ be the set of all weakly observable actions. Σ is said to be strongly observable if $\mathcal{W}(\Sigma) = \varnothing$. Σ is weakly observable if $\mathcal{W}(\Sigma) \neq \varnothing$.

Next we will define two key qualities introduced by Mannor and Shamir (2011) and Alon et al. (2015) that characterize the hardness of a problem instance with feedback structure Σ: A set $A \subset [K]$ is called an independent set if for any $i \in A$, $S_i \cap A \subset \{i\}$. The *independence number* $\kappa(\Sigma)$ is defined as the cardinality of the largest independent set. For any pair of subsets $A, A' \subset [K]$, A is said to be *dominating* A' if for any $j \in A'$ there exists $i \in A$ such that $j \in S_i$. The *weak domination number* $\rho(\Sigma)$ is defined as the cardinality of the smallest set that dominates $\mathcal{W}(\Sigma)$.

Corollary 22. *Assume that $\sigma_{ij} = \infty$ for some $i, j \in [K]$, that is, we are not in the generalized full information case. Then,*

(i) if Σ is strongly observable, with $B = \alpha\sigma\sqrt{\kappa(\Sigma)T}$ for some $\alpha > 0$, we have $\sup_{\theta\in\Theta} b'(\theta, B) \geq \frac{\sigma\sqrt{\kappa(\Sigma)T}}{64e\alpha}$ for $T \geq 64e^2\alpha^2\sigma^2\kappa(\Sigma)^3/D^2$.

(ii) If Σ is weakly observable, with $B = \alpha(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}\log^{-2/3} K$ for some $\alpha > 0$, we have $\sup_{\theta\in\Theta} b'(\theta, B) \geq \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}\log^{-2/3} K}{51200e^2\alpha^2}$.

Proof of Corollary 22. Define $\varepsilon = \frac{8eB}{T}$. First consider the case that Σ is strongly observable.

If the maximum independence number $\kappa(\Sigma) \geq 2$, there exists an independent set $A_\kappa \subset [K]$ such that $|A_\kappa| = \kappa(\Sigma)$. We construct θ as follows: Let $\theta_{i_1} = D/2$ for some $i_1 \in A_\kappa$ and $\theta_i = D/2 - \varepsilon$ for $i \in A_\kappa \setminus \{i_1\}$. For the remaining $i \notin A_\kappa$, let $\theta_i = 0$. Note that each i in A_κ must be self observable since otherwise it is a weakly observable action. Also in A_κ i can be observed only by itself according to the definition of independent sets.

Then we will lower bound $b'(\theta, B)$. According to our choice of ε, we have

$$\frac{8\sqrt{e}B}{T} e^{W\left(\frac{\varepsilon T}{16\sqrt{e}B}\right)} + \varepsilon = 2\varepsilon\,.$$

Therefore, for $i = i_1$ we have $\varepsilon_{i,-} = 2\varepsilon$ and $\varepsilon_{i,+} = 2\varepsilon$ for $i \in A_\kappa \setminus \{i_1\}$. Thus for any $i \in A_\kappa$,

$$m_i(\theta, B) = \frac{1}{4\varepsilon^2}\log\frac{T\varepsilon}{8B} = \frac{1}{4\varepsilon^2}\,.$$

Recall that we defined

$$C'_{\theta,B} = \left\{ c \in C_T^{\mathbb{R}} \,:\, \sum_{j:i\in S_j} c_j \geq \sigma^2 m_i(\theta, B) \;\text{ for all } i \in [K] \right\}$$

and $b'(\theta, B) = \inf_{c\in C'_{\theta,B}} \langle c, d(\theta)\rangle$. For any $c \in C'_{\theta,B}$, let $a = \sum_{i\notin A_\kappa} c_i$. Then we have for any $i \in A_\kappa$, $\sum_{j:i\in S_j} c_j \leq a + c_i$ and thus $c_i \geq \sigma^2 m_i(\theta, B) - a = \frac{\sigma^2}{4\varepsilon^2} - a$. Since $d_i(\theta) = \varepsilon$ for all $i \in A_\kappa \setminus \{i_1\}$ and $d_i(\theta) = D/2$ for all $i \notin A_\kappa$, we get

$$\begin{aligned}
\langle c, d(\theta)\rangle &= \sum_{i\in A_\kappa\setminus\{i_1\}} c_i\varepsilon + \frac{aD}{2} \geq (\kappa(\Sigma) - 1)\left(\frac{\sigma^2}{4\varepsilon^2} - a\right)\varepsilon + \frac{aD}{2} \\
&\geq \frac{\kappa(\Sigma)}{2}\left(\frac{\sigma^2}{4\varepsilon^2} - a\right)\varepsilon + \frac{aD}{2} = \frac{\kappa(\Sigma)\sigma^2}{8\varepsilon} + \frac{D - \kappa(\Sigma)\varepsilon}{2}a \\
&\geq \frac{\kappa(\Sigma)\sigma^2}{8\varepsilon} \qquad (3.8)
\end{aligned}$$

if $\kappa(\Sigma)\varepsilon < D$. Applying our particular choice of ε and B, we get the conclusion that for $T \geq \frac{64e^2\alpha^2\sigma^2\kappa(\Sigma)^3}{D^2}$, $b'(\theta, B) \geq \frac{\sigma\sqrt{\kappa(\Sigma)T}}{64e\alpha}$.

If $\kappa(\Sigma) = 1$, since we exclude the full information case, there always exists a pair of actions i_1 and i_2 such that $i_2 \notin S_{i_1}$ (here $i_1 \neq i_2$ is not necessary). We construct θ by setting $\theta_{i_1} = D/2$ and $\theta_i = D/2 - \varepsilon$ for all $i \neq i_1$. Then $m_i(\theta, B) = \frac{1}{4\varepsilon^2}$ holds for all $i \in [K]$. For any $c \in C'_{\theta,B}$, let $a = \sum_{i \neq i_1} c_i$, then $\sum_{j: i_2 \in S_j} c_j \leq a$. Hence $a \geq \sigma^2 m_{i_2}(\theta, B) = \frac{\sigma^2}{4\varepsilon^2}$ and

$$\langle c, d(\theta) \rangle = a\varepsilon \geq \frac{\sigma^2}{4\varepsilon} > \frac{\kappa(\Sigma)\sigma^2}{8\varepsilon}\,. \tag{3.9}$$

Combining (3.8) and (3.9) gives the first part of Corollary 22.

Now we turn to the case that Σ is weakly observable. The idea of constructing the worst θ comes from the proof of Theorem 7 in Alon et al. (2015) which based on the following graph-theoretic lemma:

Lemma 23 (Restated from Lemma 8 in Alon et al. (2015))**.** *Let $G = (V, E)$ be a directed graph with K vertices and let $W \subset V$ be a subset of vertices with domination number ρ. Then there exists an independent set $U \subset W$ with the property that $|U| \geq \frac{\rho}{50 \log K}$ and any vertex of G dominates at most $\log K$ vertices of U.*

Let $\mathcal{W}(\Sigma) \subset [K]$ be the set of all weakly observable actions. By Lemma 23 we know that there exists an independent set $A_\rho \subset \mathcal{W}(\Sigma)$ such that $|A_\rho| \geq \frac{\rho(\Sigma)}{50 \log K}$ and for any $i \in [K]$, $|S_i \cap U| \leq \log K$.

If $\rho(\Sigma) \geq 100 \log K$ such that $|A_\rho| \geq 2$, we can construct θ as follows: Let $\theta_{i_1} = D/2$ for some $i_1 \in A_\rho$ and $\theta_i = D/2 - \varepsilon$ for $i \in A_\rho \setminus \{i_1\}$. For the remaining $i \notin A_\rho$, let $\theta_i = 0$. Note that actions in A_ρ cannot be observed by any action inside A_ρ. For any $c \in C'_{\theta,B}$, let $a = \sum_{i \notin A_\rho} c_i$. Since for any i, $|S_i \cap U| \leq \log K$, we have $\sum_{i \in A_\rho} \sum_{j: i \in S_j} c_j \leq a \log K$ and

$$a \log K \geq |A_\rho| \min_{i \in A_\rho} \sum_{j: i \in S_j} c_j \geq |A_\rho| \min_{i \in A_\rho} \sigma^2 m_i(\theta, B) \geq \frac{\rho(\Sigma)\sigma^2}{200 \log K \varepsilon^2}\,.$$

Therefore,

$$\langle c, d(\theta) \rangle \geq \frac{aD}{2} \geq \frac{\rho(\Sigma)\sigma^2 D}{200\varepsilon^2 \log^2 K} = \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3} \log^{-2/3} K}{12800 e^2 \alpha^2}\,. \tag{3.10}$$

If $\rho(\Sigma) < 100 \log K$, then we pick a weakly observable action as i_2. There must be another action i_1 such that $i_2 \notin S_{i_1}$ due to the definition of weakly

observable actions. Then we set θ as $\theta_{i_1} = D/2$, $\theta_{i_2} = D/2 - \varepsilon$ and $\theta_i = 0$ for the remaining actions. So for any $c \in C'_{\theta,B}$, let $a = \sum_{i \neq i_1, i_2} c_i \geq \sigma^2 m_{i_2}(\theta, B)$. Then

$$\begin{aligned}\langle c, d(\theta) \rangle &\geq \frac{aD}{2} \geq \frac{\sigma^2 m_{i_2}(\theta, B) D}{2} = \frac{D\sigma^2}{8\varepsilon^2} = \frac{D^{1/3}(\sigma T)^{2/3}}{512 e^2 \alpha^2} \cdot \frac{\log^{4/3} K}{\rho(\Sigma)^{2/3}} \\ &\geq \frac{(\rho(\Sigma) D)^{1/3} (\sigma T)^{2/3} \log^{-2/3} K}{51200 e^2 \alpha^2} \,. \end{aligned} \tag{3.11}$$

In the last step we used the fact that $K \geq 3$ for any weakly observable Σ.

Combining (3.10) and (3.11) gives the second part of Corollary 22.

. . . . . □

Remark 24. In Corollary 22, picking $\alpha = \frac{1}{8\sqrt{e}}$ for strongly observable Σ and $\alpha = \frac{1}{73}$ for weakly observable Σ gives formal minimax lower bounds: (i) If Σ is strongly observable, for any algorithm we have $\sup_{\theta \in \Theta} R_T(\theta) \geq \frac{\sigma \sqrt{\kappa(\Sigma) T}}{8\sqrt{e}}$ for $T \geq e\sigma^2 \kappa(\Sigma)^3 / D^2$. (ii) If Σ is weakly observable, for any algorithm we have $\sup_{\theta \in \Theta} R_T(\theta) \geq \frac{(\rho(\Sigma) D)^{1/3} (\sigma T)^{2/3}}{73 \log^{2/3} K}$.

3.3 Algorithms

In this section we present two algorithms and their finite-time analysis for the uniform variance version of our problem (where σ_{ij} is either σ or ∞). The upper bound for the first algorithm matches the asymptotic lower bound in (3.1) up to constants. The second algorithm achieves the minimax lower bounds of Corollary 22 up to logarithmic factors, as well as $O(\log^{3/2} T)$ problem-dependent regret. In the problem-dependent upper bounds of both algorithms, we assume that the optimal action is unique, that is, $d_{i_2(\theta)}(\theta) > 0$.

3.3.1 An Asymptotically Optimal Algorithm

Let $c(\theta) = \operatorname{argmin}_{c \in C_\theta} \langle c, d(\theta) \rangle$; note that increasing $c_{i_1(\theta)}(\theta)$ does not change the value of $\langle c, d(\theta) \rangle$ (since $d_{i_1(\theta)}(\theta) = 0$), so we take the minimum value of $c_{i_1(\theta)}(\theta)$ in this definition. Let $n_i(t) = \sum_{s=1}^{t-1} \mathbb{I}\{i \in S_{i_s}\}$ be the number of observations for action i before round t and $\hat{\theta}_{t,i}$ be the empirical estimate of θ_i based on the first $n_i(t)$ observations. Let $N_i(t) = \sum_{s=1}^{t-1} \mathbb{I}\{i_s = i\}$ be the

number of plays for action i before round t. Note that this definition of $N_i(t)$ is different from that in the previous sections since it excludes round t.

Algorithm 5

Inputs: Σ, α, $\beta : \mathbb{N} \to [0, \infty)$.
For $t = 1, ..., K$, observe each action i at least once by playing i_t such that $t \in S_{i_t}$.
Set exploration count $n_e(K + 1) = 0$.
for $t = K + 1, K + 2, ...$ **do**
  if $\frac{N(t)}{4\alpha \log t} \in C_{\hat{\theta}_t}$ **then**
    Play $i_t = i_1(\hat{\theta}_t)$.
    Set $n_e(t + 1) = n_e(t)$.
  else
    if $\min_{i \in [K]} n_i(t) < \beta(n_e(t))/K$ **then**
      Play i_t such that $\operatorname{argmin}_{i \in [K]} n_i(t) \in S_{i_t}$.
    else
      Play i_t such that $N_i(t) < c_i(\hat{\theta}_t) 4\alpha \log t$.
    end if
    Set $n_e(t + 1) = n_e(t) + 1$.
  end if
end for

Our first algorithm is presented in Algorithm 5. The main idea, coming from Magureanu et al. (2014), is that by forcing exploration over all actions, the solution $c(\theta)$ of the linear program can be well approximated while paying a constant price. This solves the main difficulty that, without getting enough observations on each action, we may not have good enough estimates for $d(\theta)$ and $c(\theta)$. One advantage of our algorithm compared to that of Magureanu et al. (2014) is that we use a nondecreasing, sublinear exploration schedule $\beta(n)$ ($\beta : \mathbb{N} \to [0, \infty)$) instead of a constant rate $\beta(n) = \beta n$. This resolves the problem that, to achieve asymptotically optimal performance, some parameter of the algorithm needs to be chosen according to $d_{\min}(\theta)$ as in Magureanu et al. (2014). The expected regret of Algorithm 5 is upper bounded as follows:

Theorem 25. *For any $\theta \in \Theta$, $\varepsilon > 0$, $\alpha > 2$ and any non-decreasing $\beta(n)$ that satisfies $0 \le \beta(n) \le n/2$ and $\beta(m + n) \le \beta(m) + \beta(n)$ for $m, n \in \mathbb{N}$,*

$$R_T(\theta) \le \big(2K + 2 + 4K/(\alpha - 2)\big) d_{\max}(\theta) + 4K d_{\max}(\theta) \sum_{s=0}^{T} \exp\Big(- \frac{\beta(s)\varepsilon^2}{2K\sigma^2} \Big)$$

$$+ 2d_{\max}(\theta)\beta\Big(4\alpha \log T \sum_{i\in[K]} c_i(\theta,\varepsilon) + K\Big) + 4\alpha \log T \sum_{i\in[K]} c_i(\theta,\varepsilon) d_i(\theta)\,.$$

where $c_i(\theta,\varepsilon) = \sup\{c_i(\theta') \,:\, |\theta'_j - \theta_j| \le \varepsilon \;\; \textit{for all } j \in [K]\}$.

Proof of Theorem 25. Define the events

$$U_t = \left\{ |\hat{\theta}_{t,i} - \theta_i| \le \sqrt{\frac{2\alpha\sigma^2 \log t}{n_i(t)}} \;\text{ for all } i \in [K] \right\},$$

$$V_t = \left\{ |\hat{\theta}_{t,i} - \theta_i| \le \varepsilon \;\text{ for all } i \in [K] \right\},$$

$$W_t = \left\{ \frac{N(t)}{4\alpha \log t} \in C(\hat{\theta}_t) \right\},$$

$$Y_t = \left\{ \min_{i\in[K]} n_i(t) < \beta(n_e(t))/K \right\},$$

and let U_t^c, V_t^c, W_t^c, Y_t^c denote their complements, respectively. Using these events, the regret of the algorithm can be decomposed as

$$\begin{aligned} R_T(\theta) &= \sum_{t=1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\right] \le K d_{\max}(\theta) + \sum_{t=K+1}^{n} \mathbb{E}\left[d_{i_t}(\theta)\right] \\ &= K d_{\max}(\theta) + \sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\left(\mathbb{I}\left\{U_t^c\right\} + \mathbb{I}\left\{U_t, W_t\right\} + \mathbb{I}\left\{U_t, W_t^c, Y_t\right\}\right.\right. \\ &\left.\left. + \mathbb{I}\left\{U_t, W_t^c, Y_t^c, V_t^c\right\} + \mathbb{I}\left\{U_t, W_t^c, Y_t^c, V_t\right\}\right)\right]. \end{aligned} \tag{3.12}$$

We will upper bound each quantity in (3.12) separately.

By Hoeffding's inequality and the union bound, we have for $t \ge K+1$,

$$\begin{aligned} &\Pr\left(|\hat{\theta}_{t,i} - \theta_i| > \sqrt{\frac{2\alpha\sigma^2 \log t}{n_i(t)}} \right) \\ &\le \sum_{s=1}^{t} \Pr\left(|\hat{\theta}_{t,i} - \theta_i| > \sqrt{\frac{2\alpha\sigma^2 \log t}{n_i(t)}} \;\middle|\; n_i(t) = s \right) \\ &\le 2t^{1-\alpha}. \end{aligned}$$

Then, using $\alpha > 2$, $\sum_{t=K+1}^{n} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t^c\right\}\right]$ can be bounded as

$$\begin{aligned} \sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t^c\right\}\right] &\le d_{\max}(\theta) \sum_{t=K+1}^{T} \Pr(U_t^c) \\ &\le d_{\max}(\theta) \sum_{t=K+1}^{T} 2Kt^{1-\alpha} \le \frac{2K d_{\max}(\theta)}{\alpha - 2}. \end{aligned} \tag{3.13}$$

Next consider $\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t, W_t\right\}\right]$. If U_t and W_t hold, first we have

$$n_{i_1(\hat{\theta}_t)} \geq \frac{8\alpha\sigma^2 \log t}{d^2_{i_2(\hat{\theta}_t)}(\hat{\theta}_t)}\,,$$

and

$$\hat{\theta}_{t,i_1(\hat{\theta}_t)} - \theta_{i_1(\hat{\theta}_t)} \leq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_{i_1(\hat{\theta}_t)}(t)}} \leq \frac{d_{i_1(\hat{\theta}_t)}(\hat{\theta}_t)}{2} \leq \frac{d_i(\hat{\theta}_t)}{2} \tag{3.14}$$

for any $i \neq i_1(\hat{\theta}_t)$. Similarly, for $i \neq i_1(\hat{\theta}_t)$ we have

$$\theta_i - \hat{\theta}_{t,i} \leq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_i(t)}} \leq \frac{d_i(\hat{\theta}_t)}{2}\,. \tag{3.15}$$

Combining (3.14) and (3.15) gives $\theta_i \leq \theta_{i_1(\hat{\theta}_t)}$ for any $i \neq i_1(\hat{\theta}_t)$, which means $i_1(\hat{\theta}_t) = i_1(\theta)$, hence

$$\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t, W_t\right\}\right] = 0\,. \tag{3.16}$$

Consider the next term in (3.12),

$$\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t, W_t^c, Y_t\right\}\right] \leq d_{\max}(\theta)\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{U_t, W_t^c, Y_t\right\}\right]. \tag{3.17}$$

To bound (3.17), first we prove two auxiliary results:

Proposition 26. *Let $K < t_1 < t_2$. If $\sum_{s=t_1}^{t_2-1} \mathbb{I}\left\{W_s^c, Y_s\right\} \geq K$, then $\min_{i\in[K]} n_i(t_2) \geq \min_{i\in[K]} n_i(t_1) + 1$.*

Proof. We prove the proposition by contradiction. Assume $\min_{i\in[K]} n_i(t_2) = \min_{i\in[K]} n_i(t_1)$. Then there exists a j such that $n_j(t_1) = n_j(t_2)$ and $n_j(s) = \min_{i\in[K]} n_i(s)$ for all $t_1 \leq s \leq t_2$. Since $\sum_{s=t_1}^{t_2-1} \mathbb{I}\left\{W_s^c, Y_s\right\} \geq K$, there exist K time instants $t_1 \leq s_1 < s_2 < ... < s_K \leq t_2 - 1$ such that $\left\{W_{s_k}^c, Y_{s_k}\right\}$ happens for $1 \leq k \leq K$. According to the algorithm, for each s_k, there exists $j_k \neq j$ such that $j_k \in S_{i_{s_k}}$ and $n_{j_k}(s_k) = n_j(s_k) = \min_{i\in[K]} n_i(s_k)$. Note that each action appears at most once as such j_k for $1 \leq k \leq K$ since $n_{j_k}(s_k + 1) = n_{j_k}(s_k) + 1$, but there are only $K-1$ actions other than j, which means that such j cannot exist. This proves the proposition. □

Proposition 27.

$$\sum_{t=K+1}^{T} \mathbb{I}\{W_t^c, Y_t\} \leq 1 + \beta\left(\sum_{t=K+1}^{T} \mathbb{I}\{W_t^c\}\right). \tag{3.18}$$

Proof. According to the algorithm we have $n_e(t) = \sum_{s=K+1}^{t-1} \mathbb{I}\{W_s^c\}$ for $t > K$. Now define

$$t' = \max\{K+1 \leq t \leq T \,:\, W_t^c, Y_t\}.$$

If such t' does not exist, then the left hand side of (3.18) becomes 0, and the proposition holds. If such t' exists, by Proposition 26,

$$\min_{i\in[K]} n_i(t') \geq \min_{i\in[K]} n_i(K+1) + \left\lfloor \frac{1}{K} \sum_{t=K+1}^{t'-1} \mathbb{I}\{W_t^c, Y_t\} \right\rfloor \geq \frac{1}{K} \sum_{t=K+1}^{t'-1} \mathbb{I}\{W_t^c, Y_t\}.$$

Therefore,

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{I}\{W_t^c, Y_t\} &= 1 + \sum_{t=K+1}^{t'-1} \mathbb{I}\{W_t^c, Y_t\} \leq 1 + K \min_{i\in[K]} n_i(t') < 1 + \beta(n_e(t')) \\ &\leq 1 + \beta(n_e(T)) \leq 1 + \beta\left(\sum_{t=K+1}^{T} \mathbb{I}\{W_t^c\}\right).\end{aligned}$$

□

Now we return to bounding the right hand side of (3.17). Using the above proposition and the properties of β, we get

$$\begin{aligned}&\sum_{t=K+1}^{T} \mathbb{I}\{U_t, W_t^c, Y_t\} \\ &\leq \sum_{t=K+1}^{T} \mathbb{I}\{W_t^c, Y_t\} \leq 1 + \beta\left(\sum_{t=K+1}^{T} \mathbb{I}\{W_t^c\}\right) \\ &\leq 1 + \beta\left(\sum_{t=K+1}^{T} \mathbb{I}\{U_t^c\} + \mathbb{I}\{U_t, W_t^c, Y_t\} + \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t^c\} + \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\}\right) \\ &\leq 1 + \frac{1}{2}\sum_{t=K+1}^{T} (\mathbb{I}\{U_t^c\} + \mathbb{I}\{U_t, W_t^c, Y_t\} + \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t^c\}) \\ &+ \beta\left(\sum_{t=K+1}^{n} \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\}\right).\end{aligned}$$

Reordering gives

$$\sum_{t=K+1}^{T} \mathbb{I}\{U_t, W_t^c, Y_t\}$$
$$\leq 2 + \sum_{t=K+1}^{T} \mathbb{I}\{U_t^c\} + \sum_{t=K+1}^{T} \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t^c\} + 2\beta\left(\sum_{t=K+1}^{n} \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\}\right),$$

and, by (3.13), we get

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\{U_t, W_t^c, Y_t\}\right] \leq d_{\max}(\theta)\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{U_t, W_t^c, Y_t\}\right] \\
&\leq 2d_{\max}(\theta) + \frac{2Kd_{\max}(\theta)}{\alpha - 2} + d_{\max}(\theta) \sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\{U_t, W_t^c, Y_t^c, V_t^c\}\right] \\
&+ 2d_{\max}(\theta)\mathbb{E}\left[\beta\left(\sum_{t=K+1}^{n} \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\}\right)\right]. \qquad (3.19)
\end{aligned}$$

To bound $\sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\{U_t, W_t^c, Y_t^c, V_t^c\}\right]$, we first introduce two lemmas based on Combes and Proutiere (2014):

Lemma 28 (Lemma 4.3 of Combes and Proutiere (2014))**.** *Let $\{Z_t\}_{t\in\mathbb{N}^+}$ be a sequence of independent zero-mean normal random variables with variance σ^2. Let $\mathcal{F}_t$ denote the σ-algebra generated by $\{Z_s\}_{s\leq t}$ and define the filtration $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{N}^+}$. Consider $r, n_0 \in \mathbb{N}^+$ and $T \geq n_0$. Define $Y_t = \sum_{s=n_0}^{t-1} B_s Z_s$ where $B_t \in \{0,1\}$ is an $\mathcal{F}_{t-1}$-measurable random variable. Furthermore, let $n(t) = \sum_{s=n_0}^{t-1} B_s$ and let ϕ be a stopping time with respect to the filtration $\mathcal{F}$, which satisfies either $n(\phi) \geq r$ or $\phi = T+1$. Then, for any $\varepsilon > 0$ we have*

$$\Pr\left(|Y_\phi| > n(\phi)\varepsilon, \phi \leq T\right) \leq 2\exp\left(-\frac{r\varepsilon^2}{2\sigma^2}\right).$$

Lemma 29. *Define $\mathcal{F}_t$ the σ-algebra generated by $\{X_{i,s}\}_{s\in[t],i\in[K]}$. Let $\Lambda \subset [1,T] \cap \mathbb{N}$ be a set of (random) time instants. Assume there exists a sequence of (random) sets $\{\Lambda_s\}_{0\leq s\leq T}$ such that (i) $\Lambda \subset \cup_{0\leq s\leq T}\Lambda_s$, (ii) for all $0 \leq s \leq T$, $|\Lambda_s| \leq 1$, (iii) for all $0 \leq s \leq T$, if $t \in \Lambda_s$ then $n_i(t) \geq \beta(s)/K$, and (iv) the event $\{t \in \Lambda_s\}$ is $\mathcal{F}_t$ measurable. Then for any $\varepsilon > 0$ and $i \in [K]$:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{t \in \Lambda, |\hat{\theta}_{t,i} - \theta_i| > \varepsilon\right\}\right] \leq \sum_{s=0}^{T} 2\exp\left(-\frac{\beta(s)\varepsilon^2}{2K\sigma^2}\right).$$

Proof of Lemma 29. We adapt the proof of Lemma 2.2 from Combes and Proutiere (2014). For $0 \le s \le T$, define $\phi_s = t$ if $\Lambda_s = \{t\}$ or $\phi_s = T+1$ if $\Lambda_s = \varnothing$. Then

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{t \in \Lambda, |\hat{\theta}_{t,i} - \theta_i| > \varepsilon\right\}\right] &\le \mathbb{E}\left[\sum_{s=0}^{T} \mathbb{I}\left\{\phi_s \le T, |\hat{\theta}_{\phi_s,i} - \theta_i| > \varepsilon\right\}\right] \\ &= \sum_{s=0}^{T} \Pr\left(\phi_s \le T, |\hat{\theta}_{\phi_s,i} - \theta_i| > \varepsilon\right). \end{aligned} \tag{3.20}$$

Since ϕ_s can be viewed as an $\mathcal{F}$-stopping time and satisfies either $n_i(\phi_s) \ge \lceil \beta(s)/K \rceil$ or $\phi_s = T+1$, if $\lceil \beta(s)/K \rceil \ge 1$ then applying Lemma 28 gives

$$\Pr\left(\phi_s \le T, |\hat{\theta}_{\phi_s,i} - \theta_i| > \varepsilon\right) \le 2\exp\left(-\frac{\lceil \beta(s)/K \rceil \varepsilon^2}{2\sigma^2}\right) \le 2\exp\left(-\frac{\beta(s)\varepsilon^2}{2K\sigma^2}\right).$$

If $\lceil \beta(s)/K \rceil = 0$ then $\Pr\left(\phi_s \le T, |\hat{\theta}_{\phi_s,i} - \theta_i| > \varepsilon\right) < 2 = 2\exp\left(-\frac{\beta(s)\varepsilon^2}{2K\sigma^2}\right)$ still holds. Now proceeding from (3.20) we can get the result of Lemma 29.

□

Now we define $\Lambda = \{t : K+1 \le t \le T, U_t, W_t^c, Y_t^c\}$, and $\Lambda_s = \{t : K+1 \le t \le T, U_t, W_t^c, n_e(t) = s, \min_{i\in[K]} n_i(t) \ge \beta(s)/K\}$. It can be verified that Λ_s satisfies the conditions in Lemma 29: (i) If $t \in \Lambda$ then there must be some $0 \le s \le T$ such that $n_e(t) = s$ and thus $t \in \Lambda_s$. (ii) If $t \in \Lambda_s$ then for $t' > t$, $n_e(t') \ge n_e(t+1) = n_e(t) + 1 = s+1$, so $t' \notin \Lambda_s$. Condition (iii) and (iv) are also satisfied from the definition of Λ_s.

Then

$$\begin{aligned}&\sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\left\{U_t, W_t^c, Y_t^c, V_t^c\right\}\right] = \sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\left\{t \in \Lambda, V_t^c\right\}\right] \\ &\le \sum_{i=1}^{K}\sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\left\{t \in \Lambda, |\hat{\theta}_{t,i} - \theta_i| > \varepsilon\right\}\right] \le 2K\sum_{s=0}^{T} \exp\left(-\frac{\beta(s)\varepsilon^2}{2K\sigma^2}\right). \end{aligned} \tag{3.21}$$

Finally we will upper bound $\sum_{t=K+1}^{n} d_{i_t}(\theta)\mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\}$.

Recall that in the algorithm, if W_t^c and Y_t^c happens, some i_t satisfying $N_i(t) < c_i(\hat{\theta}_t)4\alpha\log t$ is played. Such i_t must exist because otherwise $\frac{N_i(t)}{4\alpha\log t} \ge c_i(\hat{\theta}_t)4\alpha\log t$ holds for any $i \in [K]$ and thus $W_t = \left\{\frac{N(t)}{4\alpha\log t} \in C(\hat{\theta}_t)\right\}$ happens, which causes contradiction.

Define

$$\Theta(\theta,\varepsilon)=\{\lambda\in\Theta\,:\,|\lambda_i-\theta_i|\le\varepsilon\ \text{ for all } i\in[K]\}\,,$$

and

$$c_i(\theta,\varepsilon)=\sup_{\lambda\in\Theta(\theta,\varepsilon)}c_i(\lambda)\,.$$

Let T_i be the maximum $t\le T$ such that $i_t=i$ and $\mathbb{I}\{U_t,W_t^c,Y_t^c,V_t\}=1$. Then

$$N_i(T_i)=\sum_{s=1}^{T_i-1}\mathbb{I}\{i_s=i\}\le c_i(\hat{\theta}_{T_i})4\alpha\log T_i\le c_i(\theta,\varepsilon)4\alpha\log T\,.$$

Thus

$$\sum_{t=K+1}^{T}\mathbb{I}\{i_t=i,U_t,W_t^c,Y_t^c,V_t\}\le c_i(\theta,\varepsilon)4\alpha\log T+1\,.$$

So we have

$$\sum_{t=K+1}^{T}d_{i_t}(\theta)\mathbb{I}\{U_t,W_t^c,Y_t^c,V_t\}\le 4\alpha\log T\sum_{i\in[K]}c_i(\theta,\varepsilon)d_i(\theta)+\sum_{i\in[K]}d_i(\theta)\,,\quad(3.22)$$

and

$$\sum_{t=K+1}^{T}\mathbb{I}\{U_t,W_t^c,Y_t^c,V_t\}\le 4\alpha\log T\sum_{i\in[K]}c_i(\theta,\varepsilon)+K\,.\quad(3.23)$$

Now plugging (3.23) (3.21) into (3.19) and plugging (3.13) (3.16) (3.19) (3.21) (3.22) into (3.12) we get

$$\begin{aligned}R_T(\theta)\le&\left(2K+2+\frac{4K}{\alpha-2}\right)d_{\max}(\theta)+4Kd_{\max}(\theta)\sum_{s=0}^{T}\exp\left(-\frac{\beta(s)\varepsilon^2}{2K\sigma^2}\right)\\&+2d_{\max}(\theta)\beta\left(4\alpha\log T\sum_{i\in[K]}c_i(\theta,\varepsilon)+K\right)+4\alpha\log T\sum_{i\in[K]}c_i(\theta,\varepsilon)d_i(\theta)\,.\end{aligned}$$

. . . . . □

Further specifying $\beta(n)$ and using the continuity of $c(\theta)$ around θ, it immediately follows that Algorithm 5 achieves asymptotically optimal performance:

Corollary 30. *Suppose the conditions of Theorem 25 hold. Assume, furthermore, that* $\beta(n)$ *satisfies* $\beta(n) = o(n)$ *and* $\sum_{s=0}^{\infty} \exp\left(-\frac{\beta(s)\varepsilon^2}{2K\sigma^2}\right) < \infty$ *for any* $\varepsilon > 0$*, then for any* θ *such that* $c(\theta)$ *is unique,*

$$\limsup_{T\to\infty} R_T(\theta)/\log T \leq 4\alpha \inf_{c\in C_\theta} \langle c, d(\theta)\rangle \ .$$

Note that any $\beta(n) = an^b$ with $a \in (0, \frac{1}{2}]$, $b \in (0, 1)$ satisfies the requirements in Theorem 25 and Corollary 30. Also note that the algorithms presented in Caron et al. (2012); Buccapatnam et al. (2014) do not achieve this asymptotic bound.

3.3.2 A Minimax Optimal Algorithm

Next we present an algorithm achieving the minimax bounds. For any $A, A' \subset [K]$, let $c(A, A') = \text{argmax}_{c\in\Delta^{|A|}} \min_{i\in A'} \sum_{j:i\in S_j} c_j$ (ties are broken arbitrarily) and $m(A, A') = \min_{i\in A'} \sum_{j:i\in S_j} c_j(A, A')$. For any $A \subset [K]$ and $|A| \geq 2$, let $A^{\mathcal{S}} = \{i \in A : \exists j \in A, i \in S_j\}$ and $A^{\mathcal{W}} = A - A^{\mathcal{S}}$. Furthermore, let $g_{r,i}(\delta) = \sigma\sqrt{\frac{2\log(8K^2r^3/\delta)}{n_i(r)}}$ where $n_i(r) = \sum_{s=1}^{r-1} i_{s,i}$ and $\hat{\theta}_{r,i}$ be the empirical estimate of θ_i based on first $n_i(r)$ observations (i.e., the average of the samples).

The algorithm is presented in Algorithm 6. It follows a successive elimination process: it explores all possibly optimal actions (called "good actions" later) based on some confidence intervals until only one action remains. While doing exploration, the algorithm first tries to explore the good actions by only using good ones. However, due to weak observability, some good actions might have to be explored by actions that have already been eliminated. To control this exploration-exploitation trade off, we use a sublinear function γ to control the exploration of weakly observable actions.

In the following we present high-probability bounds on the performance of the algorithm, so, with a slight abuse of notation, $R_T(\theta)$ will denote the regret without expectation in the rest of this section.

Algorithm 6

Inputs: Σ, δ.
Set $t_1 = 0$, $A_1 = [K]$.
for $r = 1, 2, ...$ **do**
  Let $\alpha_r = \min_{1 \le s \le r, A_s^{\mathcal{W}} \neq \varnothing} m([K], A_s^{\mathcal{W}})$ and $\gamma(r) = (\sigma \alpha_r t_r / D)^{2/3}$. (Define $\alpha_r = 1$ if $A_s^{\mathcal{W}} = \varnothing$ for all $1 \le s \le r$.)
  if $A_r^{\mathcal{W}} \neq \varnothing$ and $\min_{i \in A_r^{\mathcal{W}}} n_i(r) < \min_{i \in A_r^{\mathcal{S}}} n_i(r)$ and $\min_{i \in A_r^{\mathcal{W}}} n_i(r) < \gamma(r)$ **then**
    Set $c_r = c([K], A_r^{\mathcal{W}})$.
  else
    Set $c_r = c(A_r, A_r^{\mathcal{S}})$.
  end if
  Play $i_r = \lceil c_r \cdot \|c_r\|_0 \rceil$ and set $t_{r+1} \leftarrow t_r + \|i_r\|_1$.
  $A_{r+1} \leftarrow \{i \in A_r : \hat{\theta}_{r+1,i} + g_{r+1,i}(\delta) \ge \max_{j \in A_r} \hat{\theta}_{r+1,j} - g_{r+1,j}(\delta)\}$.
  if $|A_{r+1}| = 1$ **then**
    Play the only action in the remaining rounds.
  end if
end for

Theorem 31. *For any $\delta \in (0,1)$ and any $\theta \in \Theta$,*

$$R_T(\theta) \le (\rho(\Sigma) D)^{1/3} (\sigma T)^{2/3} \cdot 7\sqrt{6 \log(2KT/\delta)} + 125\sigma^2 K^3 / D + 13 K^3 D$$

with probability at least $1 - \delta$ if Σ is weakly observable, while

$$R_T(\theta) \le 2KD + 80\sigma \sqrt{\kappa(\Sigma) T \cdot 6 \log K \log \frac{2KT}{\delta}}$$

with probability at least $1 - \delta$ if Σ is strongly observable.

Proof of Theorem 31. For every $r > 0$, define the events

$$U_r = \left\{ |\hat{\theta}_{r,i} - \theta_i| \le g_{r,i}(\delta) \ \text{ for all } i \in [K] \right\} .$$

Then, by Hoeffding's inequality and union bound, we have

$$\Pr\left(\cap_{r \ge 2} U_r\right) \ge 1 - \delta \ .$$

Next we will upper bound the regret based on the fact that U_r holds for all $r \ge 2$. Define $r_T = \max\{r : t_t < T, |A_r| \ge 2\}$, the event

$$V_r = \left\{ A_r^{\mathcal{W}} \neq \varnothing, \min_{i \in A_r^{\mathcal{W}}} n_i(r) < \min\{\min_{i \in A_r^{\mathcal{S}}} n_i(r), \gamma(r)\} \right\}$$

and its complement V_r^c. Then consider the regret:

$$\begin{aligned} R_T(\theta) &\leq \sum_{r=1}^{r_T} \mathbb{I}\{V_r\} \langle i_r, d(\theta) \rangle + \sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \langle i_r, d(\theta) \rangle \\ &\leq \sum_{r=1}^{r_T} \mathbb{I}\{V_r\} \|i_r\|_1 D + \sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \|i_r\|_1 \max_{i \in A_r} d_i(\theta) . \end{aligned} \tag{3.24}$$

We upper bound the two terms in (3.24) separately. Before proceeding, we introduce the following proposition which lower bounds $n_i(r)$ for $i \in A_r^{\mathcal{W}}$.

Proposition 32. *For any i, r such that $i \in A_r^{\mathcal{W}}$,*

$$n_i(r) \geq \frac{\alpha_{r-1}}{2} \sum_{s=1}^{r-1} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_r - 1)K , \tag{3.25}$$

where $\beta_r = \left| \bigcup_{1 \leq s \leq r} A_s^{\mathcal{W}} \right|$.

Proof of Proposition 32. The proof is done by induction. Let W_r denote the event that for any $1 \leq s \leq r$ and any $i \in A_s^{\mathcal{W}}$, (3.25) holds. W_1 holds because $A_1^{\mathcal{W}} = \varnothing$. Now we assume W_r holds and consider W_{r+1}.

If $A_{r+1}^{\mathcal{W}} = \varnothing$, then W_{r+1} holds. If $A_{r+1}^{\mathcal{W}} \neq \varnothing$, for $i \in A_{r+1}^{\mathcal{W}}$, consider $n_i(r+1)$ in different cases:

If $i \in A_r^{\mathcal{W}}$, then $n_i(r) \geq \frac{\alpha_{r-1}}{2} \sum_{s=1}^{r-1} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_r - 1)K$. Recall that $\alpha_r = \min_{1 \leq s \leq r, A_s^{\mathcal{W}} \neq \varnothing} m([K], A_s^{\mathcal{W}})$. So we have

$$n_i(r+1) \geq n_i(r) + \mathbb{I}\{V_r\} \|c_r\|_0 \alpha_r \geq \frac{\alpha_r}{2} \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_{r+1} - 1)K ,$$

where we use the fact that α_r is non-increasing, β_r is non-decreasing as well as

$$\|i_r\|_1 = \|\lceil c_r \cdot \|c_r\|_0 \rceil\|_1 \leq \|c_r\|_0 + \|c_r\|_0 \cdot \|c_r\|_1 = 2 \|c_r\|_0 . \tag{3.26}$$

If $i \notin A_r^{\mathcal{W}}$, then $i \in A_s^{\mathcal{S}}$ for all $1 \leq s \leq r$ and thus $\beta_{r+1} \geq \beta_r + 1$. Let $r' = \max\{s \leq r : V_s\}$. If such r' does not exist, then

$$n_i(r+1) \geq 0 \geq \frac{\alpha_r}{2} \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_{r+1} - 1)K .$$

If such r' exists

$$\begin{aligned} n_i(r+1) \geq n_i(r') > \min_{j \in A_{r'}^{\mathcal{W}}} n_j(r') &\geq \frac{\alpha_{r'-1}}{2} \sum_{s=1}^{r'-1} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_{r'}-1)K \\ &\geq \frac{\alpha_r}{2} \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 - \frac{\alpha_r}{2} \|i_{r'}\|_1 - (\beta_{r'}-1)K \\ &\geq \frac{\alpha_r}{2} \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 - \beta_{r'} K \\ &\geq \frac{\alpha_r}{2} \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_{r+1}-1)K\,, \end{aligned}$$

where the facts $\alpha_r \leq 1$, $\|i_{r'}\|_1 \leq 2K$ and $\beta_{r'} \leq \beta_{r+1} - 1$ are used.

Now we have proved that W_{r+1} holds based on the assumption of W_r, hence W_r holds for any r, which gives the result of Proposition 32.

□

Based on Proposition 32, $\sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1$ can be upper bounded by the following fact:

Proposition 33. *For any $r \geq 1$, $\sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 \leq \frac{2\gamma(r)+2K\beta_r}{\alpha_r}$.*

Proof of Proposition 33. Let $r' = \max\{s \leq r : V_s\}$. Then

$$\gamma(r') > \min_{i \in A_{r'}^{\mathcal{W}}} n_i(r') \geq \frac{\alpha_{r'-1}}{2} \sum_{s=1}^{r'-1} \mathbb{I}\{V_s\} \|i_s\|_1 - (\beta_{r'}-1)K\,.$$

Hence

$$\begin{aligned} \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 \leq \sum_{s=1}^{r'-1} \mathbb{I}\{V_s\} \|i_s\|_1 + \|i_{r'}\|_1 &\leq \frac{2\gamma(r') + 2K(\beta_{r'}-1)}{\alpha_{r'}} + 2K \\ &\leq \frac{2\gamma(r') + 2K\beta_{r'}}{\alpha_{r'}}\,. \end{aligned}$$

Since α_r is non-increasing, β_r is non-decreasing and $\gamma(r)/\alpha_r = \alpha_r^{-1/3}(\sigma t_r/D)^{2/3}$ is non-decreasing, we have $\sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 \leq \frac{2\gamma(r)+2K\beta_r}{\alpha_r}$.

□

Now we are ready to upper bound the first term in (3.24):

$$\sum_{r=1}^{r_T} \mathbb{I}\{V_r\} \|i_r\|_1 D \leq \frac{2\gamma(r_T) + 2K\beta_{r_T}}{\alpha_{r_T}} D = 2\alpha_{r_T}^{-1/3} D^{1/3} (\sigma T)^{2/3} + 2KD\frac{\beta_{r_T}}{\alpha_{r_T}}\,. \tag{3.27}$$

Next consider the second term in (3.24): $\sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \|i_r\|_1 \max_{i\in A_r} d_i(\theta)$. Given U_r holds for all r we know that $i_1(\theta)$ is never eliminated. Then for any $i \in A_r$, we have $|\hat{\theta}_{r,i} - \theta_i| \le g_{r,i}(\delta)$ and $\hat{\theta}_{r,i} + g_{r,i}(\delta) \ge \hat{\theta}_{i_1(\theta)} - g_{r,i_1(\theta)}(\delta)$. Therefore,

$$\begin{aligned} d_i(\theta) &\le \min\left\{D, 2g_{r,i}(\delta) + 2g_{r,i_1(\theta)}(\delta)\right\} \\ &\le \min\left\{D, 4\sigma\sqrt{6\log\frac{2KT}{\delta}}\left(\min_{i\in A_r} n_i(r)\right)^{-1/2}\right\}. \end{aligned}$$

So

$$\sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \|i_r\|_1 \max_{i\in A_r} d_i(\theta) \le \sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \|i_r\|_1 \min\left\{D, C(\min_{i\in A_r} n_i(r))^{-1/2}\right\}, \tag{3.28}$$

where $C = 4\sigma\sqrt{6\log\frac{2KT}{\delta}}$.

The next step is to lower bound $\min_{i\in A_r} n_i(r)$ when V_r^c happens. Define $\eta_{\min} = \min_{A\in[K],|A|\ge 2} m(A, A^{\mathcal{S}})$. For $i \in A_r^{\mathcal{S}}$,

$$n_i(r) \ge \sum_{s=1}^{r-1} \mathbb{I}\{V_s^c\} \|c_s\|_0 m(A_s, A_s^{\mathcal{S}}) \ge \frac{\eta_{\min}}{2} \sum_{s=1}^{r-1} \mathbb{I}\{V_s^c\} \|i_s\|_1. \tag{3.29}$$

For $i \in A_r^{\mathcal{W}}$, since V_r^c happens and $A_r^{\mathcal{W}} \ne \varnothing$, we have

$$n_i(r) \ge \min\{\min_{i\in A_r^{\mathcal{S}}} n_i(r), \gamma(r)\} \ge \min\left\{\frac{\eta_{\min}}{2}\sum_{s=1}^{r-1} \mathbb{I}\{V_s^c\} \|i_s\|_1, \gamma(r)\right\}.$$

By Proposition 33,

$$\begin{aligned} \frac{\eta_{\min}}{2}\sum_{s=1}^{r-1} \mathbb{I}\{V_s^c\} \|i_s\|_1 &\ge \frac{1}{2K}\left(t_r - \sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1\right) \\ &\ge \frac{1}{2K}\left(t_r - \frac{2\gamma(r) + 2K\beta_r}{\alpha_r}\right) \\ &= \frac{1}{2K}\left(t_r - 2\alpha_r^{-1/3}\left(\frac{\sigma t_r}{D}\right)^{2/3} - 2K\beta_r/\alpha_r\right) \\ &\ge \frac{1}{2K}t_r - \left(\frac{\sigma t_r}{D}\right)^{2/3} - K^2, \end{aligned}$$

where we used $\alpha_r, \eta_{\min} \ge 1/K$ and $\beta_r \le K$.

For $t_r \geq \frac{125\sigma^2}{D^2}K^3 + 10K^3$, we have $\frac{4}{5}t_r \geq 4K\left(\frac{\sigma t_r}{D}\right)^{2/3}$ and $\frac{1}{5}t_r \geq 2K^3$, so

$$\begin{aligned}\frac{\eta_{\min}}{2}\sum_{s=1}^{r-1}\mathbb{I}\left\{V_s^c\right\}\|i_s\|_1 &\geq \frac{1}{2K}t_r - \left(\frac{\sigma t_r}{D}\right)^{2/3} - K^2 \\ &\geq 2\left(\frac{\sigma t_r}{D}\right)^{2/3} + K^2 - \left(\frac{\sigma t_r}{D}\right)^{2/3} - K^2 \\ &= \left(\frac{\sigma t_r}{D}\right)^{2/3} \geq \left(\frac{\sigma\alpha_r t_r}{D}\right)^{2/3} = \gamma(r)\,.\end{aligned}$$

So we have proved that for any $r \leq r_T$ such that $t_r \geq T_0 = \frac{125\sigma^2}{D^2}K^3 + 10K^3$ and V_r^c happens, $\min_{i\in A_r} n_i(r) \geq \gamma(r) \geq (\sigma\alpha_{r_T}t_r/D)^{2/3}$. Therefore, following (3.28) gives

$$\begin{aligned}&\sum_{r=1}^{r_T}\mathbb{I}\left\{V_r^c\right\}\|i_r\|_1 \max_{i\in A_r} d_i(\theta) \\ &\leq \sum_{r=1}^{r_T}\mathbb{I}\left\{V_r^c\right\}\|i_r\|_1 \min\left\{D, C(\min_{i\in A_r} n_i(r))^{-1/2}\right\} \\ &\leq \sum_{r\geq 1: t_r<T_0}\|i_r\|_1 D + \sum_{r\leq r_T: t_r\geq T_0}\|i_r\|_1 C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3} t_r^{-1/3} \\ &\leq (T_0+2K)D + C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3}\sum_{r\leq r_T: t_r\geq T_0}(t_{r+1}-t_r)(t_{r+1}-2K)^{-1/3} \\ &\leq (T_0+2K)D + C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3}\int_{T_0}^{t_{r_T+1}}(x-2K)^{-1/3}dx \\ &\leq (T_0+2K)D + C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3}\int_{T_0-2K}^{t_{r_T}}x^{-1/3}dx \\ &\leq (T_0+2K)D + \frac{3}{2}C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3}T^{2/3} \\ &= \frac{125\sigma^2K^3}{D} + (10K^3+2K)D + \alpha_{r_T}^{-1/3}D^{1/3}(\sigma T)^{2/3}\cdot 6\sqrt{6\log\frac{2KT}{\delta}}\,. \quad (3.30)\end{aligned}$$

Now plugging (3.27) and (3.30) into (3.24) gives

$$R_T(\theta) \leq \alpha_{r_T}^{-1/3}D^{1/3}(\sigma T)^{2/3}\cdot 7\sqrt{6\log\frac{2KT}{\delta}} + \frac{125\sigma^2K^3}{D} + 13K^3D\,.$$

If Σ is strongly observable, then $A_r^{\mathcal{W}}$ is always empty and V_r^c always happens. According to (3.24) (3.28) and (3.29) we have

$$R_T(\theta) \leq \sum_{r=1}^{r_T}\|i_r\|_1 \max_{i\in A_r} d_i(\theta)$$

$$
\begin{aligned}
&\leq \sum_{r=1}^{r_T} (t_{r+1} - t_r) \min\left\{ D, C\left(\frac{\eta_{\min}}{2}\right)^{-1/2} t_r^{-1/2} \right\} \\
&\leq 2KD + C\left(\frac{\eta_{\min}}{2}\right)^{-1/2} \int_0^{t_{r_T}} x^{-1/2} dx \\
&\leq 2KD + 8\sigma\sqrt{\frac{T}{\eta_{\min}} \cdot 12 \log\frac{2KT}{\delta}}\,.
\end{aligned}
$$

To finish the proof, it suffices to show that $\frac{1}{\alpha_{r_T}} \leq \rho(\Sigma)$ and $\frac{1}{\eta_{\min}} \leq \kappa(\Sigma) 50 \log K$, which is based on the following fact:

Proposition 34. *For any $A, A' \subset [K]$ Let $\rho_{LP}(A, A')$ denote the minimum fractional cover number from A to A', that is*

$$
\begin{aligned}
\rho_{LP}(A, A') = &\min_{b \in [0,\infty)^A} \sum_{i \in A} b_i \\
&\text{s.t.} \sum_{i: j \in S_i} b_i \geq 1 \quad \text{for all } j \in A'\,.
\end{aligned}
$$

Then $m(A, A') = \frac{1}{\rho_{LP}(A, A')}$.

Proof of Proposition 34. Recall that

$$
\begin{aligned}
m(A, A') &= \max_{c \in \Delta^A} \min_{i \in A'} \sum_{j: i \in S_j} c_j \\
&= \max_{c \in \Delta^A, a} a \text{ s.t. } \sum_{i: j \in S_i} c_i \geq a \quad \text{for all } j \in A'\,.
\end{aligned}
$$

Let $b = c/a$, then

$$
\begin{aligned}
m(A, A') &= \max_{b \in [0,\infty)^A, a} a \text{ s.t. } \sum_{i: j \in S_i} b_i \geq 1 \quad \text{for all } j \in A' \text{ and } \sum_{i \in A} b_i = \frac{1}{a} \\
&= \max_{b \in [0,\infty)^A} \frac{1}{\sum_{i \in A} b_i} \text{ s.t. } \sum_{i: j \in S_i} b_i \geq 1 \quad \text{for all} j \in A' \\
&= \frac{1}{\rho_{\mathrm{LP}}(A, A')}\,.
\end{aligned}
$$

□

To lower bound α_{r_T}, let $\rho(A, A')$ be the integer version of $\rho_{\mathrm{LP}}(A, A')$ by restricting $b \in \mathbb{N}^A$. Then we have $\rho(\Sigma) = \rho([K], \mathcal{W}(\Sigma))$ and

$$
\alpha_{r_T} \geq m([K], \mathcal{W}(\Sigma)) = \frac{1}{\rho_{\mathrm{LP}}([K], \mathcal{W}(\Sigma))} \geq \frac{1}{\rho(\Sigma)}\,,
$$

where we used the fact that $A_r^{\mathcal{W}} \subset \mathcal{W}(\Sigma)$ for any $r \leq r_T$.

To lower bound $\eta_{\min}$, we use

$$\eta_{\min} = \min_{A\in[K],|A|\geq 2} m(A, A^{\mathcal{S}}) = \min_{A\in[K],|A|\geq 2} m(A, A) = \frac{1}{\max_{A\in[K],|A|\geq 2} \rho_{\mathrm{LP}}(A, A)}$$

($A^{\mathcal{S}} = A$ for strongly observable Σ), thus

$$\max_{A\in[K],|A|\geq 2} \rho(A, A) \geq \frac{1}{\eta_{\min}} .$$

For any $A \in [K], |A| \geq 2$, let Σ_A be the subgraph of Σ on A. We apply Lemma 23 on Σ_A with the subset $W = A$. Then the lemma states that A contains an independent set U of size at least $\frac{\rho(A,A)}{50 \log |A|}$. Since an independent set of Σ_A is also an independent set of Σ, for each subset A there exists an independent set of Σ with size at least $\frac{\rho(A,A)}{50 \log |A|}$. So the independence number

$$\kappa(\Sigma) \geq \max_{A\in[K],|A|\geq 2} \frac{\rho(A, A)}{50 \log |A|} \geq \frac{1}{50 \log K} \max_{A\in[K],|A|\geq 2} \rho(A, A) \geq \frac{1}{\eta_{\min} 50 \log K} ,$$

which indicates $\frac{1}{\eta_{\min}} \leq \kappa(\Sigma) 50 \log K$.

. . . . . □

Theorem 35 (Problem-dependent upper bound)**.** *For any $\delta \in (0, 1)$ and any $\theta \in \Theta$ such that the optimal action is unique, with probability at least $1 - \delta$,*

$$\begin{aligned} R_T(\theta) \leq & \frac{1603 K D \sigma^2}{d_{\min}^2(\theta)} \left(\log(2KT/\delta)\right)^{3/2} + 14 K^3 D + 125 \sigma^2 K^3 / D \\ & + 15 \left(K D \sigma^2\right)^{1/3} \left(125 \sigma^2 / D^2 + 10\right) K^2 \left(\log(2KT/\delta)\right)^{1/2} . \end{aligned}$$

Proof of Theorem 35. Similarly to the proof of Theorem 35, we define high probability events

$$U_r = \left\{ |\hat{\theta}_{r,i} - \theta_i| \leq g_{r,i}(\delta) \ \text{ for all } i \in [K] \right\} .$$

and upper bound the regret based on the fact that for all $r \geq 2$, U_r holds. The rest of the proof will be based on upper bounding the number of round before all sub-optimal actions are eliminated.

Define $r_T = \max\{r \,:\, t_t < T, |A_r| \geq 2\}$, event

$$V_r = \left\{ A_r^{\mathcal{W}} \neq \varnothing, \min_{i \in A_r^{\mathcal{W}}} n_i(r) < \min\{\min_{i \in A_r^{\mathcal{S}}} n_i(r), \gamma(r)\} \right\}$$

and V_r^c be its complement.

For any $r \leq r_T$ and any $i \in A_r$, $i \neq i_1(\theta)$, we have $2g_{r,i}(\delta) + 2g_{r,i_1(\theta)}(\delta) \geq d_i(\theta) \geq d_{\min}(\theta)$, where $d_{\min}(\theta)$ denotes $d_{i_2(\theta)}(\theta)$. From $g_{r,i}(\delta) = \sigma\sqrt{\frac{2\log(8K^2r^3/\delta)}{n_i(r)}}$ we get

$$d_{\min}(\theta) \leq 2\sigma\sqrt{2\log(8K^2r^3/\delta)}\left(\frac{1}{\sqrt{n_i(r)}} + \frac{1}{\sqrt{n_{i_1(\theta)}(r)}}\right) \leq C_r \left(\min_{i \in A_r} n_i(r)\right)^{-1/2},$$

where $C_r = 4\sigma\sqrt{6\log\frac{2Kr}{\delta}}$, and thus

$$\min_{i \in A_r} n_i(r) \leq \frac{C_r^2}{d_{\min}^2(\theta)}\,. \tag{3.31}$$

Then consider the regret:

$$\begin{aligned} R_T(\theta) &\leq \sum_{r=1}^{r_T} \mathbb{I}\{V_r\} \langle i_r, d(\theta)\rangle + \sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \langle i_r, d(\theta)\rangle \\ &\leq \sum_{r=1}^{r_V} \mathbb{I}\{V_r\} \|i_r\|_1 d_{\max}(\theta) + \sum_{r=1}^{r_W} \mathbb{I}\{V_r^c\} \|i_r\|_1 \max_{i \in A_r} d_i(\theta)\,. \end{aligned} \tag{3.32}$$

where $r_V = \max\{r \leq r_T \,:\, V_r\}$ and $r_W = \max\{r \leq r_T \,:\, V_r^c\}$.

Since $\min_{i \in A_{r_V}^{\mathcal{W}}} n_i(r_V) < \min_{i \in A_{r_V}^{\mathcal{S}}} n_i(r_V)$ we have

$$\min_{i \in A_{r_V}} n_i(r_V) = \min_{i \in A_{r_V}^{\mathcal{W}}} n_i(r_V) \geq \frac{1}{2\rho^+(\Sigma)} \sum_{s=1}^{r_V - 1} \mathbb{I}\{V_s\} \|i_s\|_1 - K^2$$

by applying Proposition 32, where $\rho^+(\Sigma) = \max\{\rho(\Sigma), 1\}$. Then we can upper bound the first term in (3.32) by

$$\sum_{r=1}^{r_V} \mathbb{I}\{V_r\} \|i_r\|_1 \leq \frac{2\rho^+(\Sigma) C_{r_V}^2}{d_{\min}^2(\theta)} + 2\rho^+(\Sigma)K^2 + 2K\,. \tag{3.33}$$

Regarding the second term in (3.32), recall that for any $r \leq r_T$ such that $t_r \geq T_0 = \frac{125\sigma^2}{D^2}K^3 + 10K^3$ and V_r^c happens, $\min_{i \in A_r} n_i(r) \geq \gamma(r) \geq (\sigma\alpha_{r_T} t_r / D)^{2/3} \geq \left(\frac{\sigma t_r}{\rho^+(\Sigma) D}\right)^{2/3}$.

Using the fact that $\max_{i\in A_r} d_i(\theta) \le \min\left\{d_{\max}(\theta), C_r\left(\min_{i\in A_r} n_i(r)\right)^{-1/2}\right\}$ gives

$$
\begin{aligned}
&\sum_{r=1}^{r_W} \mathbb{I}\left\{V_r^c\right\} \|i_r\|_1 \max_{i\in A_r} d_i(\theta) \\
&\le \sum_{r=1}^{r_W} \mathbb{I}\left\{V_r^c\right\} \|i_r\|_1 \min\left\{d_{\max}(\theta), C_r(\min_{i\in A_r} n_i(r))^{-1/2}\right\} \\
&\le \sum_{r\ge 1: t_r < T_0} \|i_r\|_1 d_{\max}(\theta) + \sum_{r\le r_W: t_r \ge T_0} \|i_r\|_1 C_{r_W}\left(\frac{\sigma}{\rho^+(\Sigma)D}\right)^{-1/3} t_r^{-1/3} \\
&\le (T_0+2K) d_{\max}(\theta) + C_{r_W}\left(\frac{\sigma}{\rho^+(\Sigma)D}\right)^{-1/3} \sum_{r\le r_W: t_r\ge T_0} (t_{r+1}-t_r)(t_{r+1}-2K)^{-1/3} \\
&\le (T_0+2K) d_{\max}(\theta) + C_{r_W}\left(\frac{\sigma}{\rho^+(\Sigma)D}\right)^{-1/3} \int_{T_0}^{t_{r_W+1}} (x-2K)^{-1/3} dx \\
&\le (T_0+2K) d_{\max}(\theta) + C_{r_W}\left(\frac{\sigma}{\rho^+(\Sigma)D}\right)^{-1/3} \int_{T_0-2K}^{t_{r_W}} x^{-1/3} dx \\
&\le (T_0+2K) d_{\max}(\theta) + \frac{3}{2} C_{r_W}\left(\frac{\sigma}{\rho^+(\Sigma)D}\right)^{-1/3} t_{r_W}^{2/3}. \qquad (3.34)
\end{aligned}
$$

Now we upper bound t_{r_W}. If $t_{r_W} \ge T_0$ then $\frac{C_{r_W}^2}{d_{\min}^2(\theta)} \ge \min_{i\in A_{r_W}} n_i(r_W) \ge \left(\frac{\sigma t_{r_W}}{\rho^+(\Sigma)D}\right)^{2/3}$. Hence

$$
t_{r_W}^{2/3} \le \left(\frac{\sigma}{\rho^+(\Sigma)D}\right)^{-2/3} \frac{C_{r_W}^2}{d_{\min}^2(\theta)} + T_0^{2/3}. \qquad (3.35)
$$

Combining (3.32) (3.33) (3.34) and (3.35) with $C_{r_W} \le C_{r_T}$ gives

$$
\begin{aligned}
R_T(\theta) \le{}& \frac{1603\rho^+(\Sigma)D\sigma^2}{d_{\min}^2(\theta)} \left(\log\frac{2Kr_T}{\delta}\right)^{3/2} + 14K^3D + \frac{125\sigma^2K^3}{D} \\
&+ 15\left(\rho^+(\Sigma)D\sigma^2\right)^{1/3}\left(\frac{125\sigma^2}{D^2}+10\right)K^2\left(\log\frac{2Kr_T}{\delta}\right)^{1/2}. \qquad (3.36)
\end{aligned}
$$

Applying $r_T \le T$ and $\rho^+(\Sigma) \le K$ gives the result of Theorem 35.

Note that using $r_T \le T$ here is only for simplicity, actually r_T can be upper bounded by some constant by more careful analysis. This is because, according to Proposition 33, $\sum_{s=1}^{r_T} \mathbb{I}\left\{V_s\right\} \|i_s\|_1 = O\left(t_{r_T}^{2/3}\right)$, and $t_{r_W} = O\left((\log t_{r_T})^{3/2}\right)$, we have

$$
t_{r_T} \le t_{r_W} + \sum_{s=1}^{r_T} \mathbb{I}\left\{V_s\right\} \|i_s\|_1 = O\left(t_{r_T}^{2/3}\right) + O\left((\log t_{r_T})^{3/2}\right),
$$

which mean t_{r_T} must be upper bounded by some constant independent with T.

. . . . . □

Remark 36. Picking $\delta = 1/T$ gives an $O(\log^{3/2} T)$ upper bound on the expected regret.

3.4 Summary

We considered a novel partial-monitoring setup with Gaussian side observations, which generalizes the recently introduced setting of graph-structured feedback, allowing finer quantification of the observed information from one action to another. We provided non-asymptotic problem-dependent lower bounds that imply existing asymptotic problem-dependent and non-asymptotic minimax lower bounds (up to some constant factors) beyond the full information case. We also provided an algorithm that achieves the asymptotic problem-dependent lower bound (up to some universal constants) and another algorithm that achieves the minimax bounds under both weak and strong observability.

However, we think this is just the beginning. For example, we currently have no algorithm that achieves both the problem dependent and the minimax lower bounds at the same time. Also, our upper bounds only correspond to the graph-structured feedback case. It is of great interest to go beyond the weak/strong observability in characterizing the hardness of the problem, and provide algorithms that can adapt to any correspondence between the mean payoffs and the variances (the hardness is that one needs to identify suboptimal actions with good information/cost trade-off).

Chapter 4

Conservative Bandits

In this chapter we present our work on the conservative bandit problem (Wu et al., 2016). Our contributions are as follows: **(i)** Starting from multi-armed bandits, we first formulate what we call the family of "conservative bandit problems". As expected in these problems, the goal is to design learning algorithms that minimize regret under the additional constraint that at any given point in time, the total reward (return) must stay above a fixed percentage of the return of a fixed default arm, i.e., the return constraint must hold *uniformly in time*. The variants differ in terms of how stringent the constraint is (i.e., should the constraint hold in expectation, or with high probability?), whether the bandit problem is stochastic or adversarial, and whether the default arm's payoff is known before learning starts. **(ii)** We analyze the naive build-budget-then-learn strategy described above (which we call BudgetFirst) and design a significantly better alternative for stochastic bandits that switches between using the default arm and learning using a version of UCB in a "smoother" fashion. **(iii)** We prove that the new algorithm, which we call Conservative UCB, meets the uniform return constraint (in various senses), while it can achieve significantly less regret than BudgetFirst. In particular, while BudgetFirst is shown to pay a *multiplicative penalty* in the regret for maintaining the return constraint, Conservative UCB only pays an *additive penalty*. We provide both high probability and expectation bounds, consider both high probability and expectation constraints on the return, and also consider the case when the payoff of the default arm is initially unknown. **(iv)** We also

prove a lower bound on the best regret given the constraint and as a result show that the additive penalty is unavoidable; thus Conservative UCB achieves the optimal regret in a worst-case sense. While Unbalanced MOSS of Lattimore (2015a), when specialized to our setting, also achieves the optimal regret (as follows from the analysis of Lattimore (2015a)), as mentioned earlier it does not maintain the constraint uniformly in time (it will explore too much at the beginning of time); it also relies heavily on the knowledge of the mean payoff of the default strategy. **(v)** We also consider the *adversarial setting* where we design an algorithm similar to Conservative UCB: the algorithm uses an underlying "base" adversarial bandit strategy when it finds that the return so far is sufficiently higher than the minimum required return. We prove that the resulting method indeed maintains the return constraint uniformly in time and we also prove a high-probability bound on its regret. We find, however, that the additive penalty in this case is higher than in the stochastic case. Here, the Exp3-γ algorithm of Lattimore (2015a) is an alternative, but again, this algorithm is not able to maintain the return constraint uniformly in time. **(vi)** The theoretical analysis is complemented by synthetic experiments on simple bandit problems whose purpose is to validate that the newly designed algorithm is reasonable and to show that the algorithms' behave as dictated by the theory developed. We also compare our method to Unbalanced MOSS to provide a perspective to see how much is lost due to maintaining the return constraint uniformly over time.

4.1 Conservative Multi-Armed Bandits

The multi-armed bandit problem is a sequential decision-making task in which a learning agent repeatedly chooses an action (called an *arm*) and receives a reward corresponding to that action. We assume there are $K+1$ arms and denote the arm chosen by the agent in round $t \in \{1, 2, \dots\}$ by $I_t \in \{0, \dots, K\}$. There is a reward $X_{t,i}$ associated with each arm i at each round t and the agent receives the reward corresponding to its chosen arm, X_{t,I_t}. The agent does not observe the other rewards $X_{t,j}$ $(j \neq I_t)$.

The learning performance of an agent over a time horizon n is usually measured by its *regret*, which is the difference between its reward and what it could have achieved by consistently choosing the single best arm in hindsight:

$$R_n = \max_{i \in \{0,\dots,K\}} \sum_{t=1}^{n} X_{t,i} - X_{t,I_t}. \tag{4.1}$$

An agent is failing to learn unless its regret grows sub-linearly: $R_n \in o(n)$; good agents achieve $R_n \in O(\sqrt{n})$ or even $R_n \in O(\log n)$.

We also use the notation $T_i(n) = \sum_{t=1}^{n} \mathbb{1}\{I_t = i\}$ for the number of times the agent chooses arm i in the first n time steps.

4.1.1 Conservative Exploration

Let arm 0 correspond to the conservative default action with the other arms $1, \dots, K$ being the alternatives to be explored. We want to be able to choose some $\alpha > 0$ and constrain the learner to earn at least a $1 - \alpha$ fraction of the reward from simply playing arm 0:

$$\sum_{s=1}^{t} X_{s,I_s} \geq (1-\alpha) \sum_{s=1}^{t} X_{s,0} \quad \text{for all } t \in \{1, \dots, n\}. \tag{4.2}$$

It should be clear that small values of α force the learner to be highly conservative, whereas larger α correspond to a weaker constraint.

We introduce a quantity Z_n, called the *budget*, which quantifies how close the constraint (4.2) is to being violated:

$$Z_t = \sum_{s=1}^{t} X_{s,I_s} - (1-\alpha) X_{s,0}; \tag{4.3}$$

the constraint is satisfied if and only if $Z_t \geq 0$ for all $t \in \{1, \dots, n\}$. Note that the constraints must hold uniformly in time.

Our objective is to design algorithms that minimize the regret (4.1) while simultaneously satisfying the constraint (4.2). In the following sections, we will consider two variants of multi-armed bandits: the stochastic setting in Section 4.2 and the adversarial setting in Section 4.3. In each case we will design algorithms that satisfy different versions of the constraint and give regret guarantees.

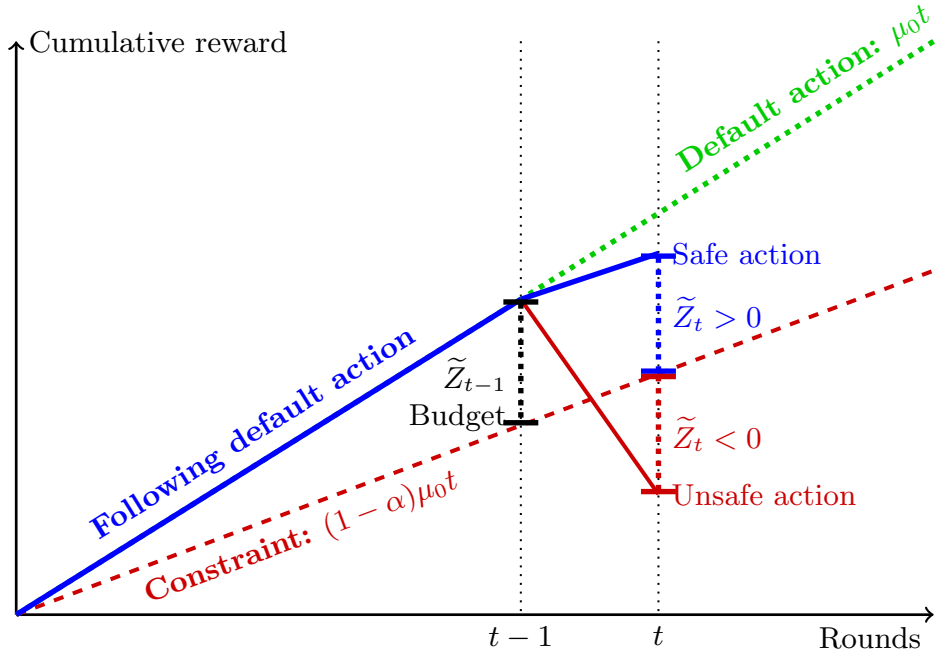

Figure 4.1: Choosing the default arm increases the budget. Then it is safe to explore a non-default arm if it cannot violate the constraint (i.e. make the budget negative).

One may wonder: what if we only care about $Z_n \geq 0$ instead of $Z_t \geq 0$ for all t. Although our algorithms are designed for satisfying the anytime constraint on Z_t our lower bound, which is based on $Z_n \geq 0$ only, shows that in the stochastic setting we cannot improve the regret guarantee even if we only want to satisfy the overall constraint $Z_n \geq 0$.

4.2 The Stochastic Setting

In the stochastic multi-armed bandit setting each arm i and round t has a stochastic reward $X_{t,i} = \mu_i + \eta_{t,i}$, where $\mu_i \in [0, 1]$ is the expected reward of arm i and the $\eta_{t,i}$ are independent random noise variables that we assume have 1-subgaussian distributions. We denote the expected reward of the optimal arm by $\mu^* = \max_i \mu_i$ and the gap between it and the expected reward of the ith arm by $\Delta_i = \mu^* - \mu_i$.

The regret R_n is now a random variable. We can bound it in expectation,

of course, but we are often more interested in high-probability bounds on the weaker notion of *pseudo-regret:*

$$\widetilde{R}_n = n\mu^* - \sum_{t=1}^{n} \mu_{I_t} = \sum_{i=0}^{K} T_i(n)\Delta_i, \tag{4.4}$$

in which the noise in the arms' rewards is ignored and the randomness arises from the agent's choice of arm. The regret R_n and the pseudo-regret $\widetilde{R}_n$ are equal in expectation. High-probability bounds for the latter, however, can capture the risk of exploration without being dominated by the variance in the arms' rewards.

We use the notation $\hat{\mu}_i(n) = \frac{1}{T_i(n)} \sum_{t=1}^{n} \mathbb{1}\{I_t = i\} X_{t,i}$ for the empirical mean of the rewards from arm i observed by the agent in the first n rounds. If $T_i(n) = 0$ then we define $\hat{\mu}_i(n) = 0$. The algorithms for the stochastic setting will estimate the μ_i by $\hat{\mu}_i$ and will construct and act based on high-probability confidence intervals for the estimates.

4.2.1 The Budget Constraint

Just as we substituted regret with pseudo-regret, in the stochastic setting we will use the following form of the constraint (4.2):

$$\sum_{s=1}^{t} \mu_{I_s} \geq (1-\alpha)\mu_0 t \qquad \text{for all } t \in \{1, \ldots, n\}\,; \tag{4.5}$$

the budget then becomes

$$\widetilde{Z}_t = \sum_{s=1}^{t} \mu_{I_s} - (1-\alpha)t\mu_0\,. \tag{4.6}$$

The default arm is always safe to play because it increases the budget by $\mu_0 - (1-\alpha)\mu_0 = \alpha\mu_0$. The budget will decrease for arms i with $\mu_i < (1-\alpha)\mu_0$; the constraint $\widetilde{Z}_n \geq 0$ is then in danger of being violated (Fig. 4.1).

In the following sections we will construct algorithms that satisfy pseudo-regret bounds and the budget constraint (4.5) with high probability $1 - \delta$ (where $\delta > 0$ is a tunable parameter). In Section 4.2.4 we will see how these algorithms can be adapted to satisfy the constraint in expectation and with bounds on their expected regret.

For simplicity, we will initially assume that the algorithms know μ_0, the expected reward of the default arm. This is reasonable in situations where the default action has been used for a long time and is well-characterized. Even so, in Section 4.2.5 we will see that having to learn an unknown μ_0 is not a great hindrance.

4.2.2 BudgetFirst — A Naive Algorithm

Before presenting the new algorithm it is worth remarking on the most obvious naive attempt, which we call the BudgetFirst algorithm. A straightforward modification of UCB leads to an algorithm that accepts a confidence parameter $\delta \in (0, 1)$ and suffers regret at most

$$\widetilde{R}_n = O\left(\sqrt{Kn \log\left(\frac{\log(n)}{\delta}\right)}\right) = R_{\text{worst}} \,. \tag{4.7}$$

Of course this algorithm alone will not satisfy the constraint (4.5), but that can be enforced by naively modifying the algorithm to deterministically choose $I_t = 0$ for the first t_0 rounds where

$$(\forall\, t_0 \leq t \leq n) \quad t\mu_0 - R_{\text{worst}} \geq (1-\alpha)t\mu_0 \,.$$

Subsequently the algorithm plays the high probability version of UCB and the regret guarantee (4.7) ensures the constraint (4.5) is satisfied with high probability. Solving the equation above leads to $t_0 = \tilde{O}(R_{\text{worst}}/\alpha\mu_0)$, and since the regret while choosing the default arm may be $O(1)$ the worst-case regret guarantee of this approach is

$$\widetilde{R}_n = \Omega\left(\frac{1}{\mu_0\alpha}\sqrt{Kn \log\left(\frac{\log(n)}{\delta}\right)}\right) \,.$$

This is significantly worse than the more sophisticated algorithm that is our main contribution and for which the price of satisfying (4.5) is only an additive term rather than a large multiplicative factor.

4.2.3 Conservative UCB

A better strategy is to play the default arm only until the budget (4.6) is large enough to start exploring other arms with a low risk of violating the

constraint. It is safe to keep exploring as long as the budget remains large, whereas if it decreases too much then it must be replenished by playing the default arm. In other words, we intersperse the exploration of a standard bandit algorithm with occasional budget-building phases when required. We show that accumulating a budget does not severely curtail exploration and thus gives small regret.

Conservative UCB (Algorithm 7) is based on UCB with the novel twist of maintaining a positive budget. In each round, UCB calculates upper confidence bounds for each arm; let J_t be the arm that maximizes this calculated confidence bound. Before playing this arm (as UCB would) our algorithm decides whether doing so risks the budget becoming negative. Of course, it does not know the actual budget $\widetilde{Z}_t$ because the μ_i ($i \neq 0$) are unknown; instead, it calculates a lower confidence bound ξ_t based on confidence intervals for the μ_i. More precisely, it calculates a lower confidence bound for what the budget would be if it played arm J_t. If this lower bound is positive then the constraint will not be violated as long as the confidence bounds hold. If so, the algorithm chooses $I_t = J_t$ just as UCB would; otherwise it acts conservatively by choosing $I_t = 0$.

Algorithm 7 Conservative UCB

Input: K, μ_0, δ, $\psi^\delta(\cdot)$
for $t \in 1, 2, \ldots$ **do**
  $\theta_0(t), \lambda_0(t) \leftarrow \mu_0$
  for $i \in 1, \ldots, K$ **do**
    $\Delta_i(t) \leftarrow \sqrt{\psi^\delta(T_i(t-1))/T_i(t-1)}$
    $\theta_i(t) \leftarrow \hat{\mu}_i(t-1) + \Delta_i(t)$
    $\lambda_i(t) \leftarrow \max\{0, \hat{\mu}_i(t-1) - \Delta_i(t)\}$
  end for
  $J_t \leftarrow \operatorname{argmax}_i \theta_i(t)$ {...and find UCB arm.}
  $\xi_t \leftarrow \sum_{s=1}^{t-1} \lambda_{I_s}(t) + \lambda_{J_t}(t) - (1-\alpha)t\mu_0$
  if $\xi_t \geq 0$ **then**
    $I_t \leftarrow J_t$ {...choose UCB arm if safe,}
  else
    $I_t \leftarrow 0$ {...default arm otherwise.}
  end if
end for

Remark 37 (Choosing ψ^δ)**.** The confidence intervals in Algorithm 7 are constructed using the function ψ^δ. Let F be the event that for all rounds $t \in \{1, 2, \ldots\}$ and every action $i \in [K]$, the confidence intervals are valid:

$$|\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{\psi^\delta(T_i(t))}{T_i(t)}}\,.$$

Our goal is to choose $\psi^\delta(\cdot)$ such that

$$\Pr(F) \geq 1 - \delta\,. \tag{4.8}$$

A simple choice is $\psi^\delta(s) = 2\log(Ks^3/\delta)$, for which (4.8) holds by Hoeffding's inequality and union bounds. The following choice achieve better performance in practice:

$$\psi^\delta(s) = \log\max\{3, \log\zeta\} + \log(2e^2\zeta) + \frac{\zeta(1+\log(\zeta))}{(\zeta-1)\log(\zeta)}\log\log(1+s), \tag{4.9}$$

where $\zeta = K/\delta$; it can be seen to achieve (4.8) by more careful analysis motivated by Garivier (2013).

Some remarks on Algorithm 7

- μ_0 is known, so the upper and lower confidence bounds can both be set to μ_0 (line 3). See Section 4.2.5 for a modification that learns an unknown μ_0.
- The max in the definition of the lower confidence bound $\lambda_i(t)$ (line 7) is because we have assumed $\mu_i \geq 0$ and so the lower confidence bound should never be less than 0.
- ξ_t (line 10) is a lower confidence bound on the budget (4.6) if action J_t is chosen. More precisely, it is a lower confidence bound on $\widetilde{Z}_t = \sum_{s=1}^{t-1} \mu_{I_s} + \mu_{J_t} - (1-\alpha)t\mu_0$.
- If the default arm is also the UCB arm ($J_t = 0$) and the confidence intervals all contain the true values, then $\mu^* = \mu_0$ and the algorithm will choose action 0 for all subsequent rounds, incurring no regret.

The following theorem guarantees that Conservative UCB satisfies the constraint while giving a high-probability upper bound on its regret.

Theorem 38. *In any stochastic environment where the arms have expected rewards $\mu_i \in [0,1]$ with 1-subgaussian noise, Algorithm 7 satisfies the following with probability at least $1-\delta$ and for every time horizon n, when ψ^δ is chosen in accordance with Remark 37 and with $L = \psi^\delta(n)$:*

$$\sum_{s=1}^{t} \mu_{I_s} \geq (1-\alpha)\mu_0 t \qquad \text{for all } t \in \{1,\dots,n\}, \tag{4.5}$$

$$\widetilde{R}_n \leq \sum_{i>0:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{6L}{\alpha\mu_0} \sum_{i=1}^{K} \frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}, \tag{4.10}$$

$$\widetilde{R}_n \in O\left(\sqrt{nKL} + \frac{KL}{\alpha\mu_0}\right). \tag{4.11}$$

Proof of Theorem 38. By Remark 37, with probability $\Pr(F) \geq 1-\delta$ the confidence intervals are valid for all t and all arms $i \in \{1,\dots,K\}$:

$$|\hat{\mu}_i(t-1) - \mu_i| \leq \sqrt{\psi^\delta(T_i(t-1))/T_i(t-1)} \leq \sqrt{L/T_i(t-1)};$$

we will henceforth assume that this is the case (i.e. that F holds). By the definition of the confidence intervals and by the construction of Algorithm 7 we immediately satisfy the constraint

$$\sum_{t=1}^{n} \mu_{I_t} \geq (1-\alpha)n\mu_0 \qquad \text{for all } n.$$

We now bound the regret. Let $i > 0$ be the index of a sub-optimal arm and suppose $I_t = i$. Since the confidence intervals are valid,

$$\begin{aligned} \mu^* &\leq \theta_i(t) \leq \hat{\mu}_i(t-1) + \sqrt{L/T_i(t-1)} \\ &\leq \mu_i + 2\sqrt{L/T_i(t-1)}\,, \end{aligned}$$

which implies that arm i has not been chosen too often; in particular we obtain

$$T_i(n) \leq T_i(n-1) + 1 \leq \frac{4L}{\Delta_i^2} + 1. \tag{4.12}$$

and the regret satisfies

$$\widetilde{R}_n = \sum_{i=0}^{K} T_i(n)\Delta_i \leq \sum_{i>0:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + T_0(n)\Delta_0.$$

If $\Delta_0 = 0$ then the theorem holds trivially; we therefore assume that $\Delta_0 > 0$ and find an upper bound for $T_0(n)$.

Let $\tau = \max\{t \leq n \mid I_t = 0\}$ be the last round in which the default arm is played. Since F holds and $\theta_0(t) = \mu_0 < \mu^* < \max_i \theta_i(t)$, it follows that $J_t = 0$ is never the UCB choice; the default arm was only played because $\xi_\tau < 0$:

$$\sum_{i=0}^{K} T_i(\tau-1)\lambda_i(\tau) + \lambda_{J_\tau}(\tau) - (1-\alpha)\mu_0\tau < 0 \tag{4.13}$$

By dropping $\lambda_{J_\tau}(\tau)$, replacing τ with $\sum_{i=0}^{K} T_i(\tau-1)+1$, and rearranging the terms in (4.13), we get

$$\begin{aligned}
\alpha T_0(\tau-1)\mu_0 &< (1-\alpha)\mu_0 + \sum_{i=1}^{K} T_i(\tau-1)\left((1-\alpha)\mu_0 - \lambda_i(\tau)\right) \\
&\leq (1-\alpha)\mu_0 + \sum_{i=1}^{K} T_i(\tau-1)\left((1-\alpha)\mu_0 - \mu_i + \sqrt{\frac{L}{T_i(\tau-1)}}\right) \\
&\leq 1 + \sum_{i=1}^{K} S_i\,. \qquad (4.14)
\end{aligned}$$

where $a_i = (1-\alpha)\mu_0 - \mu_i$ and

$$\begin{aligned}
S_i &= T_i(\tau-1)\cdot\left((1-\alpha)\mu_0 - \mu_i + \sqrt{L/T_i(\tau-1)}\right) \\
&= a_i T_i(\tau-1) + \sqrt{L T_i(\tau-1)}
\end{aligned}$$

is a bound on the decrease in ξ_t in the first $\tau-1$ rounds due to choosing arm i. We will now bound S_i for each $i > 0$.

The first case is $a_i \geq 0$, i.e. $\Delta_i \geq \Delta_0 + \alpha\mu_0$. Then (4.12) gives $T_i(\tau-1) \leq 4L/\Delta_i^2 + 1$ and we get

$$S_i \leq \frac{4La_i}{\Delta_i^2} + \frac{2L}{\Delta_i} + 2 \leq \frac{6L}{\Delta_i} + 2\,. \tag{4.15}$$

The other case is $a_i < 0$, i.e. $\Delta_i < \Delta_0 + \alpha\mu_0$. Then

$$S_i \leq \sqrt{L T_i(\tau-1)} \leq \frac{2L}{\Delta_i} + 1, \tag{4.16}$$

and by using $ax^2 + bx \leq -b^2/4a$ for $a < 0$ we have

$$S_i \leq -\frac{L}{4a_i} = \frac{L}{4(\Delta_0 + \alpha\mu_0 - \Delta_i)}. \tag{4.17}$$

Summarizing (4.15) to (4.17) gives

$$S_i \leq \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2\,.$$

Continuing from (4.14), we get

$$T_0(n) = T_0(\tau - 1) + 1 \leq \frac{2K+2}{\alpha\mu_0} + \frac{1}{\alpha\mu_0}\sum_{i=1}^{K} \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,.$$

We can now upper bound the regret by

$$\widetilde{R}_n \leq \sum_{i>0:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{6L}{\alpha\mu_0}\sum_{i=1}^{K} \frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,. \tag{4.10}$$

We will now show (4.11). To bound the regret due to the non-default arms, Jensen's inequality gives

$$\left(\sum_{i>0} T_i(n)\Delta_i\right)^2 \leq m^2 \sum_{i>0} \frac{T_i(n)}{m}\Delta_i^2,$$

where $m \leq n$ is the number of times non-default arms were chosen. Combining this with $\Delta_i^2 \leq 4L/T_i(n)$ for sub-optimal arms from (4.12) gives

$$\sum_{i>0} T_i(n)\Delta_i \leq 2\sqrt{mKL} \in O(\sqrt{nKL}).$$

To bound the regret due to the default arm, observe that $\max\{\Delta_i, \Delta_0 - \Delta_i\} \geq \Delta_0/2$ and thus $T_0(n)\Delta_0 \in O(KL/\alpha\mu_0)$. Combining these two bounds gives (4.11).

. . . . . □

Standard unconstrained UCB algorithms achieve a regret of order $O(\sqrt{nKL})$; Theorem 38 tells us that the penalty our algorithm pays to satisfy the constraint is an extra additive regret of order $O(KL/\alpha\mu_0)$.

Remark 39. We take a moment to understand how the regret of the algorithm behaves if α is polynomial in $1/n$. Clearly if $\alpha \in O(1/n)$ then we have a constant exploration budget and the problem is trivially hard. In the slightly less extreme case when α is as small as n^{-a} for some $0 < a < 1$, the extra regret penalty is still not negligible: satisfying the constraint costs us $O(n^a)$ more regret in the worst case.

We would argue that the problem-dependent regret penalty (4.10) is more informative than the worst case of $O(n^a)$; our regret increases by

$$\frac{6L}{\alpha\mu_0}\sum_{i=1}^{K}\frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}.$$

Intuitively, even if α is very small, we can still explore as long as the default arm is close-to-optimal (i.e. Δ_0 is small) and most other arms are clearly sub-optimal (i.e. the Δ_i are large). Then the sub-optimal arms are quickly discarded and even the budget-building phases accrue little regret: the regret penalty remains quite small. More precisely, if $\Delta_0 \approx n^{-b_0}$ and $\min_{i>0:\Delta_i>0}\Delta_i \approx n^{-b}$, then the regret penalty is

$$O\left(n^{a+\min\{0,b-b_0\}}\right);$$

small Δ_0 and large Δ_i means $b - b_0 < 0$, giving a smaller penalty than the worst case of $O(n^a)$.

Remark 40. Curious readers may be wondering if $I_t = 0$ is the only conservative choice when the arm proposed by UCB risks violating the constraint. A natural alternative would be to use the lower confidence bound $\lambda_i(t)$ by choosing

$$I_t = \begin{cases} J_t\,, & \text{if } \xi_t \geq 0\,; \\ \operatorname{argmax}_i \lambda_i(t)\,, & \text{otherwise}\,. \end{cases} \tag{4.18}$$

It is easy to see that if F does not occur, then choosing $\operatorname{argmax}_i \lambda_i(t)$ increases the budget at least as much as choosing action 0 while incurring less regret and so this algorithm is preferable to Algorithm 7 in practice. Theoretically speaking, however, it is possible to show that the improvement is by at most a

constant factor so our analysis of the simpler algorithm suffices. The proof of this claim is somewhat tedious so instead we provide two intuitions: Firstly, the upper bound approximately matches the lower bound in the minimax regime, so any improvement must be relatively small in the minimax sense. Secondly, imagine we run the unmodified Algorithm 7 and let t be the first round when $I_t \neq J_t$ and where there exists an $i > 0$ with $\lambda_i(t) \geq \mu_0$. If F does not hold, then the actions chosen by UCB satisfy

$$T_i(t) \in \Omega\left(\min\left\{\frac{L}{\Delta_i^2}, \max_j T_j(t)\right\}\right),$$

which means that arms are being played in approximately the same frequency until they are proving suboptimal (for a similar proof, see Lattimore (2015b)). From this it follows that once $\lambda_{I_t}(t) \geq \mu_0$ for some i it will not be long before either $\lambda_j(t+s) \geq \mu_0$ or $T_j(t+s) \geq 4L/\Delta_i^2$ and in both cases the algorithm will cease playing conservatively. Thus it takes at most a constant proportion more time before the naive algorithm is exclusively choosing the arm chosen by UCB.

Next we discuss how small modifications to Algorithm 7 allow it to handle some variants of the problem while guaranteeing the same order of regret.

4.2.4 Considering the Expected Regret and Budget

One may care about the performance of the algorithm in expectation rather than with high probability, i.e. we want an upper bound on $\mathbb{E}\left[\widetilde{R}_n\right]$ and the constraint (4.5) becomes

$$\mathbb{E}\left[\sum_{s=1}^{t} \mu_{I_s}\right] \geq (1-\alpha)\mu_0 t, \quad \text{for all } t \in \{1, \ldots, n\}. \tag{4.19}$$

We argued in Remark 39 that if $\alpha \in O(1/n)$ then the problem is trivially hard; let us assume therefore that $\alpha \geq c/n$ for some $c > 1$. By running Algorithm 7 with $\delta = 1/n$ and $\alpha' = (\alpha - \delta)/(1-\delta)$ we can achieve (4.19) and a regret bound with the same order as in Theorem 38.

To show (4.19) we have

$$\mathbb{E}\left[\sum_{s=1}^{t} \mu_{I_s}\right] \geq \Pr(F)\,\mathbb{E}\left[\sum_{s=1}^{t} \mu_{I_s} \,\middle|\, F\right] \geq (1-\delta)(1-\alpha')\mu_0 t = (1-\alpha)\mu_0 t\,.$$

As an upper bound, we have $\mathbb{E}[R_n] \leq \mathbb{E}[R_n|F] + \delta n = \mathbb{E}[R_n|F] + 1$. Here $\mathbb{E}[R_n|F]$ can be upper bounded by Theorem 38 with two changes: (i) L becomes $O(\log nK)$ after replacing δ with $1/n$, and (ii) α becomes α'. Since $\alpha'/\alpha \geq 1 - 1/c$ we get essentially the same order of regret bound as in Theorem 38.

4.2.5 Learning an Unknown μ_0

Two modifications to Algorithm 7 allow it to handle the case when μ_0 is unknown. First, just as we do for the non-default arms, we need to set $\theta_0(t)$ and $\lambda_0(t)$ based on confidence intervals. Second, the lower bound on the budget needs to be set as

$$\xi'_t = \sum_{i=1}^{K} T_i(t-1)\lambda_i(t) + \lambda_{J_t}(t) + (T_0(t-1) - (1-\alpha)t)\theta_0(t)\,. \tag{4.20}$$

Theorem 41. *Algorithm 7, modified as above to work without knowing μ_0 but otherwise the same conditions as Theorem 38, satisfies with probability $1-\delta$ and for all time horizons n the constraint* (4.5) *and the regret bound*

$$\widetilde{R}_n \leq \sum_{i:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{7L}{\alpha\mu_0}\sum_{i=1}^{K} \frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,. \tag{4.21}$$

Proof of Theorem 41. We proceed very similarly to the proof of Theorem 38. As we did there, we assume that F holds: the confidence intervals are valid for all rounds and all arms (including the default), which happens with probability $\Pr(F) \geq 1 - \delta$.

To show that the modified algorithm satisfies the constraint (4.5), we write the budget (4.6) as

$$\widetilde{Z}_t = \sum_{i=1}^{K} T_i(t-1)\mu_i + \mu_{J_t} + (T_0(t-1) - (1-\alpha)t)\mu_0$$

when the UCB arm J_t is chosen and show that it is indeed lower-bounded by

$$\xi'_t = \sum_{i=1}^{K} T_i(t-1)\lambda_i(t) + \lambda_{J_t}(t) + (T_0(t-1) - (1-\alpha)t)\theta_0(t)\,. \tag{4.20}$$

This is apparent if $T_0(t-1) < (1-\alpha)t$, since the last term in (4.20) is then negative and $\theta_0(t) \geq \mu_0$. On the other hand, if $T_0(t-1) \geq (1-\alpha)t$ then the constraint is still satisfied:

$$\sum_{s=1}^{t} \mu_{I_s} \geq T_0(t-1)\mu_0 \geq (1-\alpha)\mu_0 t.$$

We now upper-bound the regret. As in the earlier proof, we can show that for any arm $i > 0$ with $\Delta_i > 0$ we have $T_i(n) \leq 4L/\Delta_i^2 + 1$. If this also holds for $i = 0$ or if $\Delta_0 = 0$ then $\widetilde{R}_n \leq \sum_{i:\Delta>0}(4L/\Delta_i + \Delta_i)$ and the theorem holds trivially. From now on we only consider the case when $\Delta_0 > 0$ and $T_0(n) > 4L/\Delta_0^2 + 1$. As before, we will proceed to upper-bound $T_0(n)$.

Let τ be the last round in which $I_\tau = 0$. We can ignore the possibility that $J_\tau = 0$, since then the above bound on $T_i(n)$ would apply even to the default arm, contradicting our assumption above. Thus we can assume that the default arm was played because $\xi'_\tau < 0$:

$$\sum_{i=1}^{K} T_i(\tau-1)\lambda_i(\tau) + \lambda_{J_\tau}(\tau) + \big(T_0(\tau-1) - (1-\alpha)\tau\big)\,\theta_0(\tau) < 0\,,$$

in which we drop $\lambda_{J_\tau}(\tau)$, replace τ with $\sum_{i=0}^{K} T_i(\tau-1)+1$, and rearrange the terms to get

$$\alpha T_0(\tau-1)\theta_0(\tau) < (1-\alpha)\theta_0(\tau) + \sum_{i=1}^{K} T_i(\tau-1)\big((1-\alpha)\theta_0(\tau) - \lambda_i(\tau)\big)\,. \tag{4.22}$$

We lower-bound the left-hand side of (4.22) using $\theta_0(\tau) \geq \mu_0$, whereas we upper-bound the right-hand side using

$$\theta_0(\tau) \leq \mu_0 + \sqrt{\frac{L}{T_0(\tau-1)}} \leq \mu_0 + \frac{\Delta_0}{2}\,,$$

which comes from $T_0(\tau-1) \geq 4L/\Delta_0^2$. Combining these in (4.22) with the lower confidence bound $\lambda_i(\tau) \geq \mu_i - \sqrt{L/T_i(\tau-1)}$ gives

$$\begin{aligned}\alpha\mu_0 T_0(\tau-1) < (1-\alpha)&\left(\mu_0 + \frac{\Delta_0}{2}\right) \\ &+ \sum_{i=1}^{K} T_i(\tau-1)\left((1-\alpha)\left(\mu_0 + \frac{\Delta_0}{2}\right) - \mu_i + \sqrt{\frac{L}{T_i(\tau-1)}}\right)\end{aligned}$$

$$= (1-\alpha)\left(\mu_0 + \frac{\Delta_0}{2}\right) + \sum_{i=1}^{K} S_i$$
$$\leq 1 + \sum_{i=1}^{K} S_i\,, \tag{4.23}$$

where $a_i = (1-\alpha)(\mu_0 + \Delta_0/2) - \mu_i$ and

$$S_i = a_i T_i(\tau - 1) + \sqrt{L T_i(\tau - 1)}$$

is a bound on the decrease in ξ'_t in the first $\tau - 1$ rounds due to choosing arm i. We will now bound S_i for each $i > 0$.

Analogously to the previous proof, we get the bounds

$$S_i \leq \frac{6L}{\Delta_i} + 2, \quad \text{when } a_i \geq 0\,; \tag{4.24}$$
$$S_i \leq \frac{2L}{\Delta_i} + 1\,, \quad \text{otherwise;} \tag{4.25}$$

and in the latter case, using $ax^2 + bx \leq -b^2/4a$ gives

$$S_i \leq -\frac{L}{4a_i} = \frac{L}{4\big((1+\alpha)\Delta_0/2 + \alpha\mu_0 - \Delta_i\big)}\,. \tag{4.26}$$

Summarizing (4.24) to (4.26) gives

$$S_i \leq \frac{6L}{\max\{*\}\,\Delta_i, 24\big((1+\alpha)\Delta_0/2 + \alpha\mu_0 - \Delta_i\big)} + 2$$
$$\leq \frac{7L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2\,.$$

Continuing with (4.23), if $T_0(n) > \frac{4L}{\Delta_0^2} + 1$, we get

$$T_0(n) = T_0(\tau - 1) + 1 \leq \frac{2K+2}{\alpha\mu_0} + \frac{1}{\alpha\mu_0}\sum_{i=1}^{K} \frac{7L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,.$$

We can now upper bound the regret by

$$\widetilde{R}_n \leq \sum_{i:\Delta_i>0}\left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{7L}{\alpha\mu_0}\sum_{i=1}^{K}\frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,. \tag{4.21}$$

. . . . . □

Theorem 41 shows that we get the same order of regret for unknown μ_0.

4.3 The Adversarial Setting

Unlike the stochastic case, in the adversarial multi-armed bandit setting we do not make any assumptions about how the rewards are generated. Instead, we analyze a learner's worst-case performance over all possible sequences of rewards $(X_{t,i})$. In effect, we are treating the environment as an adversary that has intimate knowledge of the learner's strategy and will devise a sequence of rewards that maximizes regret. To preserve some hope of succeeding, however, the learner is allowed to behave randomly: in each round it can randomize its choice of arm I_t using a distribution it constructs; the adversary cannot influence nor predict the result of this random choice.

Our goal is, as before, to satisfy the constraint (4.2) while bounding the regret (4.1) with high probability (the randomness comes from the learner's actions). We assume that the default arm has a fixed reward: $X_{t,0} = \mu_0 \in [0, 1]$ for all t; the other arms' rewards are generated adversarially in $[0, 1]$. The constraint to be satisfied then becomes $\sum_{s=1}^{t} X_{s,I_s} \geq (1 - \alpha)\mu_0 t$ for all t.

Safe-playing strategy: We take any standard any-time high probability algorithm for adversarial bandits and adapt it to play as usual when it is safe to do so, i.e. when $Z_t \geq \sum_{s=1}^{t-1} X_{s,I_s} - (1 - \alpha)\mu_0 t \geq 0$. Otherwise it should play $I_t = 0$. To demonstrate a regret bound, we only require that the bandit algorithm satisfy the following requirement.

Definition 42. An algorithm $\mathcal{A}$ is $\hat{R}_t^\delta$-*admissible* ($\hat{R}_t^\delta$ sub-linear) if for any δ, in the adversarial setting it satisfies

$$\Pr\left(\forall t \in \{1, 2, \dots\}, R_t \leq \hat{R}_t^\delta\right) \geq 1 - \delta.$$

Note that this performance requirement is stronger than the typical high probability bound but is nevertheless achievable. For example, Neu (2015) states the following for the any-time version of their algorithm: given any time horizon n and confidence level δ, $\Pr\left(R_n \leq \hat{R}'_n(\delta)\right) \geq 1 - \delta$ for some sub-linear $\hat{R}'_t(\delta)$. If we let $\hat{R}_t^\delta = \hat{R}'_t(\delta/2t^2)$ then $\Pr\left(R_t \leq \hat{R}_t^\delta\right) \geq 1 - \frac{\delta}{2t^2}$ holds for any fixed t. Since the algorithm does not require n and δ as input, a union bound shows it to be $\hat{R}_t^\delta$-admissible.

Having satisfied ourselves that there are indeed algorithms that meet our requirements, we can prove a regret guarantee for our safe-playing strategy.

Theorem 43. *Any $\hat{R}_t^\delta$-admissible algorithm $\mathcal{A}$, when adapted with our safe-playing strategy, satisfies the constraint (4.2) and has a regret bound of $R_n \leq t_0 + \hat{R}_n^\delta$ with probability at least $1-\delta$ where $t_0 = \max\left\{t \mid \alpha\mu_0 t \leq \hat{R}_t^\delta + \mu_0\right\}$.*

Proof of Theorem 43. It is clear from the description of the safe-playing strategy that it is indeed safe: the constraint (4.2) is always satisfied.

The algorithm plays safe when the following quantity, which is a lower bound on the budget Z_t, is negative:

$$Z_t' = Z_t - X_{t,I_t} = \sum_{s=1}^{t-1} X_{s,I_s} - (1-\alpha)\mu_0 t$$

To upper bound the regret, consider only the rounds in which our safe-playing strategy does not interfere with playing $\mathcal{A}$'s choice of arm. Then with probability $1-\delta$,

$$\max_{i\in\{0,\dots,K\}} \sum_{s=1}^{t} \mathbb{1}\{Z_s' \geq 0\}\,(X_{s,i} - X_{s,I_s}) \leq \hat{R}_{B(t)}^\delta$$

where $B(t) = \sum_{s=1}^{t} \mathbb{1}\{Z_s' \geq 0\}$. Let τ be the last round in which the algorithm plays safe.

$$\begin{aligned}
\mu_0 B(\tau-1) &\leq \max_i \sum_{s=1}^{\tau-1} \mathbb{1}\{Z_s' \geq 0\}\, X_{s,i} \\
&\leq \hat{R}_{B(\tau-1)}^\delta + \sum_{s=1}^{\tau-1} \mathbb{1}\{Z_s' \geq 0\}\, X_{s,I_s} \\
&= \hat{R}_{B(\tau-1)}^\delta + \sum_{s=1}^{\tau-1} X_{s,I_s} - \mu_0(\tau - 1 - B(\tau-1)) \\
&\leq \hat{R}_{B(\tau-1)}^\delta + (1-\alpha)\mu_0\tau - \mu_0(\tau - 1 - B(\tau-1))\,,
\end{aligned}$$

which indicates $\alpha\mu_0\tau \leq \hat{R}_\tau^\delta + \mu_0$ and thus $\tau \leq t_0$. It follows that $R_n \leq t_0 + \hat{R}_n^\delta$.

. . . . . □

Corollary 44. *The any-time high probability algorithm of Neu (2015) adapted with our safe-playing strategy gives $\hat{R}_t^\delta = 7\sqrt{Kt\log K}\log(4t^2/\delta)$ and*

$$R_n \leq 7\sqrt{Kn\log K}\log(4n^2/\delta) + \frac{49K\log K}{\alpha^2\mu_0^2}\log^2\frac{4n^2}{\delta}$$

with probability at least $1 - \delta$.

Corollary 44 shows that a strategy similar to that of Algorithm 7 also works for the adversarial setting. However, we pay a higher regret penalty to satisfy the constraint: $O\left(\frac{KL^2}{(\alpha\mu_0)^2}\right)$ rather than the $O\left(\frac{KL}{\alpha\mu_0}\right)$ we had in the stochastic setting. Whether this is because (i) our algorithm is sub-optimal, (ii) the analysis is not tight, or (iii) there is some intrinsic hardness in the non-stochastic setting is still not clear and remains an interesting open problem.

4.4 Lower Bound on the Regret

We now present a worst-case lower bound where α, μ_0 and n are fixed, but the mean rewards are free to change. For any vector $\mu \in [0,1]^K$, we will write $\mathbb{E}_\mu$ to denote expectations under the environment where all arms have normally-distributed unit-variance rewards and means μ_i (i.e., the fixed value μ_0 is the mean reward of arm 0 and the components of μ are the mean rewards of the other arms). We assume normally distributed noise for simplicity: other subgaussian distributions whose parameter is kept fixed independently of the mean rewards work identically.

Theorem 45. *Suppose for any* $\mu_i \in [0,1]$ *(*$i > 0$*) and* μ_0 *satisfying*

$$\min\{\mu_0, 1-\mu_0\} \geq \max\left\{1/2\sqrt{\alpha}, \sqrt{e+1/2}\right\}\sqrt{K/n},$$

an algorithm satisfies $\mathbb{E}_\mu\left[\sum_{t=1}^n X_{t,I_t}\right] \geq (1-\alpha)\mu_0 n$. *Then there is some* $\mu \in [0,1]^K$ *such that its expected regret satisfies* $\mathbb{E}_\mu[R_n] \geq B$ *where*

$$B = \max\left\{\frac{K}{(16e+8)\alpha\mu_0}, \frac{\sqrt{Kn}}{\sqrt{16e+8}}\right\}. \tag{4.27}$$

Proof of Theorem 45. Pick any algorithm. We want to show that the algorithm's regret on some environment is at least as large as B. If $\mathbb{E}_\mu[R_n] > B$ for some $\mu \in [0,1]^K$, there is nothing to be proven. Hence, without loss of generality, we can assume that the algorithm is *consistent* in the sense that $\mathbb{E}_\mu[R_n] \leq B$ for all $\mu \in [0,1]^K$.

For some $\Delta > 0$, define environment $\mu \in \mathbb{R}^K$ such that $\mu_i = \mu_0 - \Delta$ for all $i \in [K]$. For now, assume that μ_0 and Δ are such that $\mu_i \geq 0$; we will get back to this condition later. Also define environment $\mu^{(i)}$ for each $i = 1, \ldots, K$ by

$$\mu_j^{(i)} = \begin{cases} \mu_0 + \Delta, & \text{for } j = i\,; \\ \mu_0 - \Delta, & \text{otherwise.} \end{cases}$$

In this proof, we use $T_i = T_i(n)$ to denote the number of times arm i was chosen in the first n rounds. We distinguish two cases, based on how large the exploration budget is.

Case 1: $\alpha \geq \dfrac{\sqrt{K}}{\mu_0\sqrt{(16e+8)n}}$**.**

In this case, $B = \frac{\sqrt{Kn}}{\sqrt{16e+8}}$ and we use $\Delta = (4e+2)B/n$. For each $i \in [K]$ define event $A_i = \{T_i \leq 2B/\Delta\}$. First we prove that $\Pr_\mu(A_i) \geq 1/2$:

$$\begin{aligned} \Pr_\mu\left(T_i \leq 2B/\Delta\right) &= 1 - \Pr_\mu\left(T_i > 2B/\Delta\right) \\ &\geq 1 - \frac{\Delta \mathbb{E}_\mu[T_i]}{2B} \geq 1 - \frac{\mathbb{E}_\mu[R_n]}{2B} \\ &\geq \frac{1}{2}\,. \end{aligned}$$

Next we prove that $\Pr_{\mu^{(i)}}(A_i) \leq 1/4e$:

$$\begin{aligned} \Pr_{\mu^{(i)}}\left(T_i \leq 2B/\Delta\right) &= \Pr_{\mu^{(i)}}\left(n - T_i \geq n - 2B/\Delta\right) \\ &\leq \frac{\mathbb{E}_{\mu^{(i)}}[n - T_i]}{n - 2B/\Delta} \leq \frac{B}{\Delta n - 2B} \\ &= \frac{1}{4e}\,. \end{aligned}$$

Note that μ and $\mu^{(i)}$ differ only in the ith component: $\mu_i = \mu_0 - \Delta$ whereas $\mu_i^{(i)} = \mu_0 + \Delta$. Then the KL divergence between the reward distributions of the ith arms is $\mathrm{KL}(\mu_i, \mu_i^{(i)}) = (2\Delta)^2/2 = 2\Delta^2$. Define the *binary relative entropy* to be

$$d(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y};$$

it satisfies $d(x, y) \geq (1/2)\log(1/4y)$ for $x \in [1/2, 1]$ and $y \in (0, 1)$. By a standard change of measure argument (see, e.g., Lemma 1 of Kaufmann et al., 2015b) we get that

$$\mathbb{E}_\mu[T_i] \cdot \mathrm{KL}(\mu_i; \mu_i^{(i)}) \geq d(\Pr_\mu(A_i), \Pr_{\mu^{(i)}}(A_i)) \geq \tfrac{1}{2} \log \tfrac{1}{4(1/4e)} = \tfrac{1}{2}$$

and so $\mathbb{E}_\mu[T_i] \geq 1/4\Delta^2$ for each $i \in [K]$. Hence

$$\mathbb{E}_\mu[R_n] = \Delta \sum_{i \in [K]} \mathbb{E}_\mu[T_i] \geq \frac{K}{4\Delta} = \frac{\sqrt{Kn}}{\sqrt{16e+8}} = B\,.$$

Case 2: $\alpha < \dfrac{\sqrt{K}}{\mu_0\sqrt{(16e+8)n}}$**.**

In this case, $B = \frac{K}{(16e+8)\alpha\mu_0}$ and we use $\Delta = K/4\alpha\mu_0 n$. For each i define the event $A_i = \{T_i \leq 2\alpha\mu_0 n/\Delta\}$. First we prove that $\Pr_\mu(A_i) \geq 1/2$:

$$\begin{aligned}
\Pr_\mu\left\{T_i \leq 2\alpha\mu_0 n/\Delta\right\} &= 1 - \Pr_\mu\left\{T_i > 2\alpha\mu_0 n/\Delta\right\} \\
&\geq 1 - \frac{\Delta\mathbb{E}_\mu[T_i]}{2\alpha\mu_0 n} \geq 1 - \frac{\mathbb{E}_\mu[R_n]}{2\alpha\mu_0 n} \\
&\geq \frac{1}{2}\,,
\end{aligned}$$

where we use the fact that

$$\mathbb{E}_\mu[R_n] = n\mu_0 - \mathbb{E}_\mu\Big[\sum_{t=1}^{n} X_{t,I_t}\Big] \leq n\mu_0 - (1-\alpha)\mu_0 n = \alpha\mu_0 n\,.$$

Next, we show that $\Pr_{\mu^{(i)}}(A_i) < 1/4e$:

$$\begin{aligned}
\Pr_{\mu^{(i)}}\left(T_i \leq 2\alpha\mu_0 n/\Delta\right) &= \Pr_{\mu^{(i)}}\left(n - T_i \geq n - 2\alpha\mu_0 n/\Delta\right) \\
&\leq \frac{\mathbb{E}_{\mu^{(i)}}[n - T_i]}{n - 2\alpha\mu_0 n/\Delta} \leq \frac{B}{\Delta n - 2\alpha\mu_0 n} \\
&= \frac{K}{(4e+2)K - (32e+16)\alpha^2\mu_0^2 n} \\
&< \frac{1}{4e}\,.
\end{aligned}$$

As in the other case, we have $\mathbb{E}_\mu[T_i] > 1/4\Delta^2$ for each $i \in [K]$. Therefore

$$\mathbb{E}_\mu[R_n] = \Delta \sum_{i \in [K]} \mathbb{E}_\mu[T_i] > \frac{K}{4\Delta} = \alpha\mu_0 n,$$

which contradicts the fact that $\mathbb{E}_\mu[R_n] \leq \alpha\mu_0 n$. So there does not exist an algorithm whose worst-case regret is smaller than B.

To summarize, we proved that

$$\mathbb{E}_\mu[R_n] \geq \begin{cases} \dfrac{\sqrt{Kn}}{\sqrt{16e+8}}, & \text{when } \alpha \geq \dfrac{\sqrt{K}}{\mu_0\sqrt{(16e+8)n}} \\ \dfrac{K}{(16e+8)\alpha\mu_0}, & \text{otherwise,} \end{cases}$$

finishing the proof.

. . . . . □

Theorem 45 shows that our algorithm for the stochastic setting is near-optimal (up to a logarithmic factor L) in the worst case. A problem-dependent lower bound for the stochastic setting would be interesting but is left for future work. Also note that in the lower bound we only use $\mathbb{E}_\mu\left[\sum_{t=1}^n X_t\right] \geq (1-\alpha)n\mu_0$ for the last round n, which means that the regret guarantee cannot be improved if we only care about the last-round budget instead of the anytime budget. In practice, however, enforcing the constraint in all rounds will generally lead to significantly worse results because the algorithm cannot explore early on. This is demonstrated empirically in Section 4.5, where we find that the Unbalanced MOSS algorithm performs very well in terms of the expected regret, but does not satisfy the constraint in early rounds.

Remark 46. The theorem above almost follows from the lower bound given by Lattimore (2015a), but in that paper μ_0 is unknown, while here it may be known. This makes our result strictly stronger, as the lower bound is the same up to constant factors.

4.5 Experiments

We evaluate the performance of Conservative UCB compared to UCB and Unbalanced MOSS (Lattimore, 2015a) using simulated data in two regimes. In the first (Fig. 4.2) we fix the horizon and sweep over $\alpha \in [0, 1]$ to show the degradation of the average regret of Conservative UCB relative to UCB as the constraint becomes harsher (α close to zero). In the second regime (Fig. 4.3) we fix $\alpha = 0.1$ and plot the long-term average regret, showing that Conservative UCB is eventually nearly as good as UCB, despite the constraint. Each data point is an average of $N \approx 4000$ i.i.d. samples, which makes error bars too small to see. Results are shown for both versions of Conservative UCB: The first knows the mean μ_0 of the default arm while the second does not and must act more conservatively while learning this value. As predicted

by the theory, the difference in performance between these two versions of the algorithm is relatively small, but note that even when $\alpha = 1$ the algorithm that knows μ_0 is performing better because this knowledge is useful in the unconstrained setting. This is also true of the BudgetFirst algorithm, which is unconstrained when $\alpha = 1$ and exploits its knowledge of μ_0 to eliminate the default arm. This algorithm is so conservative that even when α is nearly zero it must first build a significant budget. We tuned the Unbalanced MOSS algorithm with the following parameters:

$$B_0 = \frac{nK}{\sqrt{nK} + \frac{K}{\alpha\mu_0}} \qquad\qquad B_i = B_K = \sqrt{nK} + \frac{K}{\alpha\mu_0}\,.$$

The quantity B_i determines the regret of the algorithm with respect to arm i up to constant factors, and must be chosen to lie inside the Pareto frontier given by Lattimore (2015a). It should be emphasised that Unbalanced MOSS does *not* constrain the return except for the last round, and has no high-probability guarantees. This freedom allows it to explore early, which gives it a significant advantage over the highly constrained Conservative UCB. Furthermore, it also requires $B_0, \ldots, B_K$ as inputs, which means that μ_0 must be known in advance. The mean rewards in both experiments are $\mu_0 = 0.5$, $\mu_1 = 0.6$, $\mu_2 = \mu_3 = \mu_4 = 0.4$, which means that the default arm is slightly sub-optimal.

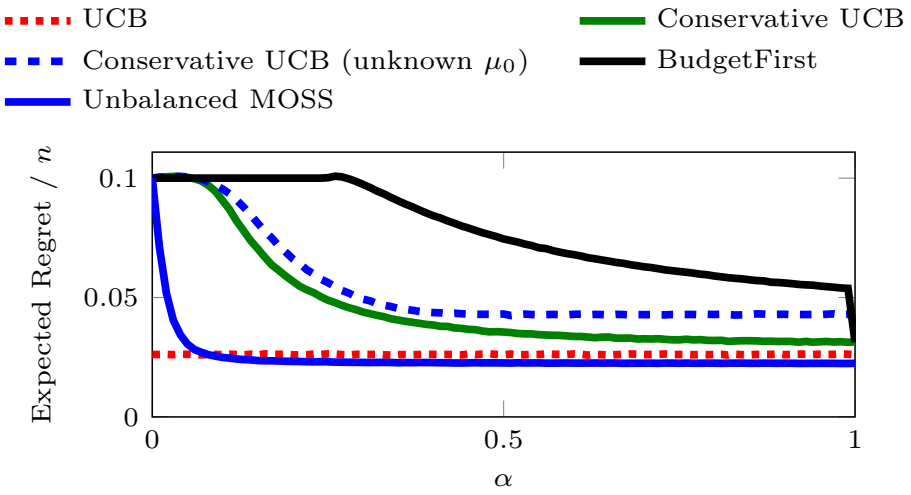

Figure 4.2: Average regret for varying α and $n = 10^4$ and $\delta = 1/n$

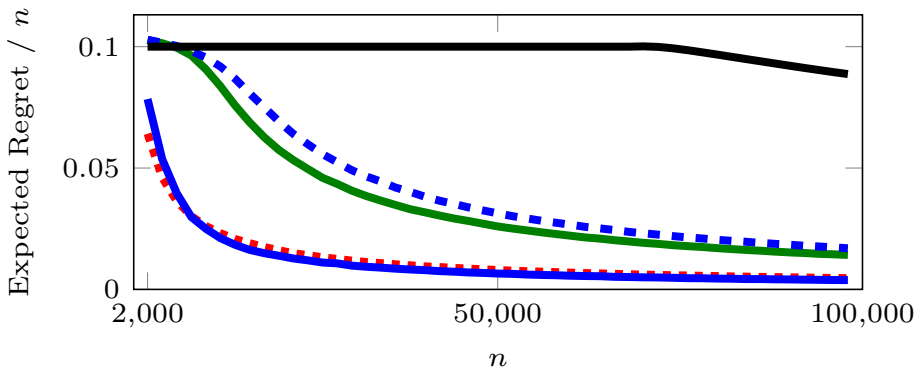

Figure 4.3: Average regret as n varies with $\alpha = 0.1$ and $\delta = 1/n$

4.6 Summary

We introduced a new family of multi-armed bandit frameworks motivated by the requirement of exploring conservatively to maintain revenue and demonstrated various strategies that act effectively under such constraints. We expect that similar strategies generalize to other settings, like contextual bandits and reinforcement learning. We want to emphasize that this is just the beginning of a line of research that has many potential applications. We hope that others will join us in improving the current results, closing open problems, and generalizing the model so it is more widely applicable.

Chapter 5

Conclusions and Future Work

In this thesis we studied three variants of online learning problems with different objectives and presented our recent results. There are several interesting future directions: **(i)** There are still some theoretical open questions in our work, e.g. the gap between the lower and upper bound in Chapter 2, a single algorithm that achieves both asymptotic problem-dependent and worst-case optimality in Chapter 3 and the minimax regret in the adversarial setting in Chapter 4. **(ii)** Another interesting direction is the pure exploration and regret minimization problems in the more general partial monitoring setting (Bartók et al., 2014). To the best of our knowledge there has not been any study on the pure exploration problems in this setting. Regarding regret minimization, Komiyama et al. (2015) introduces an asymptotically problem-dependent optimal algorithm but a minimax optimal algorithm (in terms of both the time horizon and the scaling parameter) is still an open problem. Moreover, the current partial monitoring setting is limited to finite outcome spaces. Extending it to continuous outcome spaces is also an interesting future work. **(iii)** To further push online learning techniques into practical use, we can study other variants of problems based on real applications such as problems with new type of environments, objectives or other type of constraints in the learning process.

Bibliography

N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1610–1618, 2013.

N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: beyond bandits. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 23–35, 2015.

J. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39:967–997, 2014.

R. E. Bechhofer. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

S. Bubeck, T. Wang, and N. Viswanathan. Multiple identifications in multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2013.

S. Buccapatnam, A. Eryilmaz, and N. B. Shroff. Stochastic bandits with side observations on networks. *SIGMETRICS Perform. Eval. Rev.*, 42(1):289–300, June 2014.

S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 142–151, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 521–529, 2014.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pages 255–270, 2002.

E. Even-Dar, M. Kearns, Y. Mansour, and J. Wortman. Regret to the best vs. regret to the average. *Machine Learning*, 72(1-2):21–37, 2008.

R. H. Farrell. Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, 35(1):36–72, 1964.

T. Gabel and M. Riedmiller. Distributed policy search reinforcement learning for job-shop scheduling tasks. *International Journal of Production Research*, 50(1):41–61, 2011.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

V. Gabillon, A. Lazaric, M. Ghavamzadeh, R. Ortner, and P. Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1004–1012, 2016.

J. García and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

A. Garivier. Informational confidence bounds for self-normalized averages and applications. *arXiv preprint arXiv:1309.3376*, 2013.

T. L. Graves and T. L. Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 2 edition, 1993.

M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2014.

K.-S. Jun, K. Jamieson, R. Nowak, and X. Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *The 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

S. Kalyanakrishnan and P. Stone. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of International Conference on Machine Learning (ICML)*, 2010.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2012.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2013.

E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2013.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015a. (to appear).

E. Kaufmann, A. Garivier, and O. Cappé. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2015b. To appear.

T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 613–621, 2014.

T. Kocák, G. Neu, and M. Valko. Online learning with noisy side observations. In *International Conference on Artificial Intelligence and Statistics*, pages 1186–1194, 2016.

J. Komiyama, J. Honda, and H. Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, pages 1783–1791, 2015.

W. M. Koolen. The Pareto regret frontier. In *Advances in Neural Information Processing Systems*, pages 863–871, 2013.

B. H. Korte and J. Vygen. *Combinatorial optimization: theory and algorithms*. Springer, 3 edition, 2006.

T. Lattimore. The Pareto regret frontier for bandits. In *Advances in Neural Information Processing Systems*, 2015a. To appear.

T. Lattimore. Optimally confident UCB : Improved regret for finite-armed bandits. Technical report, 2015b. URL `http://arxiv.org/abs/1507.07880`.

T. Lattimore, A. György, and C. Szepesvári. On learning the optimal waiting time. In P. Auer, A. Clark, T. Zeugmann, and S. Zilles, editors, *Algorithmic Learning Theory*, volume 8776 of *Lecture Notes in Computer Science*, pages 200–214. Springer International Publishing, 2014. ISBN 978-3-319-11661-7.

L. Li, R. Munos, and C. Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 608–616, 2015.

T. Lin, B. Abrahao, R. Kleinberg, J. Lui, and W. Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 901–909, 2014.

Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popović. Towards automatic experimentation of educational knowledge. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2014)*, pages 3349–3358. ACM Press, 2014.

L. Lovász. On the ratio of optimal integral and fractional covers. *Discrete mathematics*, 13(4):383–390, 1975.

S. Magureanu, R. Combes, and A. Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 975–999, 2014.

S. Mannor and O. Shamir. From bandits to experts: on the value of side-observations. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 684–692, 2011.

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

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3150–3158, 2015.

E. Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

R. Raz and S. Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *STOC*, pages 475–484, 1997.

V. Rieser and O. Lemon. Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *ACL-08: HLT*, pages 638–646, 2008.

A. Sani, G. Neu, and A. Lazaric. Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems*, pages 810–818, 2014.

A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, 2003.

P. Slavik. *Approximation Algorithms for Set Cover and Related Problems*. PhD thesis, State University of New York at Buffalo, 1998. AAI9833643.

Y. Sui, A. Gotovos, J. Burdick, and A. Krause. Safe exploration for optimization with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 997–1005, 2015.

V. Vazirani. *Approximation algorithms*. Springer, 2001.

Y. Wu, A. György, and C. Szepesvári. On identifying good options under combinatorially structured feedback in finite noisy environments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1283–1291, 2015a.

Y. Wu, A. György, and C. Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015b.

Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári. Conservative bandits. In *Proceedings of the 33nd International Conference on Machine Learning (ICML-16)*, 2016. To appear.

Y. Zhou, X. Chen, and J. Li. Optimal pac multiple arm identification with applications to crowdsourcing. In *Proceedings of International Conference on Machine Learning (ICML)*, 2014.