Online Learning with Gaussian Payoffs and Side Observations

Yifan Wu, Csaba Szepesvári
Dept. of Computing Science
University of Alberta
{ywu12,szepesva}@ualberta.ca

András György
Dept. of Electrical and Electronic Engineering
Imperial College London
a.gyorgy@imperial.ac.uk

Abstract

We consider a sequential learning problem with Gaussian payoffs and side information: 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. 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).

1 Introduction

Online learning in stochastic environments is a sequential decision problem where in each time step a learner chooses an action from a given finite set, observes some random feedback and receives a random payoff. Several feedback models have been considered in the literature: The simplest is the full information case where the learner observes the payoff of all possible actions at the end of every round. A popular setup is the case of bandit feedback, where the learner only observes its own payoff and receives no information about the payoff of other actions [1]. 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 (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 [2, 3, 4, 5] and the stochastic cases [6, 7]. 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 [5]. 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 [8, 9]. 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 [8, 9], which is in contrast to the standard full information and bandit settings where the feedback is typically assumed to be real-valued. The only exception to this case is the work of [5], which considers graph-structured feedback without the self-loop assumption.

In this paper we consider a generalization of the graph-structured feedback model that can also be viewed as a general partial monitoring model with real-valued feedback. 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 paper 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}, 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 $1/\sqrt{T}$ (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 2, we provide non-asymptotic problem-dependent lower bounds in Section 3, 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., [10, 7]), or finite-time minimax bounds (e.g., [9, 3, 5]). Our bounds can recover all previous bounds up to some universal constant factors not depending on the problem. In Section 4, 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) of partial monitoring problems is characterized by their global/local observability property [9] or, in case of the graph-structured feedback model, by their strong/weak observability property [5]. 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 [6, 7] 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. Finally, we draw conclusions and consider some interesting future directions in Section 5. Due to space constraints, all proofs are deferred to the appendix.

2 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 action, the learner receives a random payoff and also some side observation corresponding to other actions. More precisely, each action $i \in [K] = \{1, \dots, K\}$ 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 $\sigma_{ij} \geq 0$) 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 it knows the variance matrix $\Sigma = (\sigma_{ij}^2)_{i,j \in [K]}$ in advance; (iii) $\theta \in [0, D]$ for some $D > 0$; (iv) $\min_{i \in [K]} \sigma_{ij} \leq \sigma < \infty$, 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 [5]: Each action $i \in [K]$ is associated with an observation set $S_i \subset [K]$ such that $\sigma_{ij} = \sigma_j$ 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 *asymptotically 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.[1] On the other hand, the constant can be quite large: in particular, when the variance of some observations are large relative to the gaps $d_i = \max_i \theta_i - \theta_i$, the situation is rather similar to a partial monitoring setup for a smaller, finite time horizon. In this paper 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.

2.1 Notation

Let $C_T^{\mathbb{N}} = \{c \in \mathbb{N}^K \,:\, c_i \geq 0\,, \sum_{i\in[K]} c_i = T\}$ and $N(T) \in C_T$ 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 paper, $\log$ denotes the natural logarithm and Δ^n denotes the n-dimensional simplex for any positive integer n.

3 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 (the regret of this algorithm 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 [10]. However, this asymptotic notion of consistency is not applicable to finite-horizon problems. Therefore, following [11], 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.

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 [10], 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

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

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

$$\liminf_{T\to\infty} \frac{R_T(\theta)}{\log T} \geq \inf_{c\in C_\theta} \langle c, d(\theta) \rangle \ . \tag{1}$$

Note that the right hand side of (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

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

for other problems. Similar bounds have been provided in [6, 7] 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.

3.1 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

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

$$\text{s.t.} \sum_{a \in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \le \sum_{i \in [K]} \left(I_i(\theta, \theta') \sum_{a \in C_T^{\mathbb{N}}} q(a) a_i \right),$$

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

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

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

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

Here $C_{\theta,B}$ contains all the value of $\mathbb{E}\left[N(T); \theta\right]$ 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 1. *Given any $B > 0$, for any algorithm such that $\sup_{\theta' \in \Theta} R_T(\theta) \le B$, we have, for any environment $\theta \in \Theta$, $R_T(\theta) \ge b(\theta, B)$.*

Remark 2. 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 A Relaxed Lower Bound

Now we introduce a relaxed but more interpretable version of the finite-time lower bound of Theorem 1, which can be shown to match the asymptotic lower bound (1). The idea of deriving the lower bounds 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 $\mathbb{E}\left[N_i(T); \theta\right]$ 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 $\mathbb{E}\left[N_i(T); \theta\right]$ 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 $\mathbb{E}\left[N_i(T); \theta'\right]$ may lead to low finite-time regret in θ'. In the following we make this argument precise to obtain an interpretable lower bound.

3.2.1 Formulation

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

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

where the m_i are defined as follows: For $i \neq i_1$, if $\theta_{i_1} = D$, then $m_i(\theta, B) = 0$. Otherwise let

$$m_i^+(\theta, B) = \max_{\epsilon \in (d_i(\theta), D-\theta_i]} \frac{1}{\epsilon^2} \log \frac{T(\epsilon - d_i(\theta))}{8B},$$
$$m_i^-(\theta, B) = \max_{\epsilon \in (0, \theta_i]} \frac{1}{\epsilon^2} \log \frac{T(\epsilon + d_i(\theta))}{8B},$$

and let $\epsilon_{i,+}$ and $\epsilon_{i,-}$ 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): let

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

where $\epsilon_{i_1(\theta),-}$ and $\epsilon_{i_1(\theta),+}$ 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 $\epsilon_{i,+}$ and $\epsilon_{i,-}$ can be expressed in closed form using the Lambert $W\mathbb{R} \to \mathbb{R}$ function satisfying $W(x)e^{W(x)} = x$: by Lemma 12 and Lemma 13 in Appendix A.2, for any $i \neq i_1(\theta)$,

$$\epsilon_{i,+} = \min\left\{D - \theta_i\,,\, \frac{8\sqrt{e}B}{T} e^{W\left(\frac{d_i(\theta)T}{16\sqrt{e}B}\right)} + d_i(\theta)\right\}\,, \tag{2}$$
$$\epsilon_{i,-} = \min\left\{\theta_i\,,\, \frac{8\sqrt{e}B}{T} e^{W\left(-\frac{d_i(\theta)T}{16\sqrt{e}B}\right)} - d_i(\theta)\right\},$$

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 1:

Corollary 3. *Given $B > 0$, for any algorithm such that $\sup_{\lambda \in \Theta} R_T(\lambda) \leq B$, we have, for any environment $\theta \in \Theta$, $R_T(\theta) \geq b'(\theta, B) = \min_{c \in C'_{\theta,B}} \langle c, d(\theta) \rangle$.*

Next we compare this bound to existing lower bounds.

3.2.2 Comparison to the Asymptotic Lower Bound of (1)

Next we will show that our finite lower bound in Corollary 3 matches the asymptotic lower bound in (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 minimum in (2) is not achieved on the boundary) and has a unique optimal action. Then, for $i \neq i_1(\theta)$, it is easy to show that $\epsilon_{i,+} = \frac{d_i(\theta)}{2W(d_i(\theta)T^{1-\beta}/(16\alpha\sqrt{e}))} + d_i(\theta)$ by (2) and $m_i(\theta, B) = \frac{1}{\epsilon_{i,+}^2} \log \frac{T(\epsilon_{i,+} - d_i(\theta))}{8B}$ 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, $C'_{\theta,B} \to \frac{2\log T}{1-\beta} C_\theta$, under the assumptions of (1), as $T \to \infty$. This implies that Corollary 3 matches the asymptotic lower bound of (1) up to a factor of $(1-\beta)/2$.

3.2.3 Comparison to Minimax Bounds

Now we will show that our θ-dependent finite-time lower bound reproduces the minimax regret bounds of [2] and [5], 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 $\mathcal{W}(\Sigma) = \emptyset$. Σ is weakly observable if $\mathcal{W}(\Sigma) \neq \emptyset$.

Next we will define two key qualities introduced by [2] and [5] 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 4. *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 $\sup_{\theta\in\Theta} b'(\theta, B) \geq \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}\log^{-2/3} K}{51200e^2\alpha^2}$.

Remark 5. In Corollary 4, picking $\alpha = \frac{1}{8\sqrt{e}}$ for strongly observable Σ and $\alpha = \frac{1}{73}$ for weakly observable Σ gives formal worst case 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 $\sup_{\theta\in\Theta} R_T(\theta) \geq \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}}{73\log^{2/3} K}$.

4 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 (1) up to constants. The second algorithm achieves the minimax lower bounds of Corollary 4 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, $d_{i_2(\theta)}(\theta) > 0$.

4.1 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 $\theta_{i,t}$ 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 the round t.

Our first algorithm is presented in Algorithm 1. The main idea, coming from [12], 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 [12] is that we use a sublinear exploration schedule $\beta(n)$ 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 [12]. The expected regret of Algorithm 1 is upper bounded as follows:

Algorithm 1

```
Inputs: $\Sigma$, $\beta(n)$, $\alpha$.
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 $\mathrm{argmin}_{i \in [K]} n_i(t) \in S_{i_t}$.
        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
```

Theorem 6. *For any $\theta \in \Theta$, $\epsilon > 0$, $\alpha > 2$ 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 $m, n \in \mathbb{N}$,*

$$R_T(\theta) \le \left(2K + 2 + \frac{4K}{\alpha - 2}\right) d_{\max}(\theta) + 4K d_{\max}(\theta) \sum_{s=0}^{T} \exp\left(-\frac{\beta(s)\epsilon^2}{2K\sigma^2}\right)$$

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

where $c_i(\theta, \epsilon) = \sup\{c_i(\theta') \, : \, |\theta'_i - \theta_i| \le \epsilon \; \forall i \in [K]\}$.

Further specifying $\beta(n)$ and using the continuity of $c(\theta)$ around θ, it immediately follows that Algorithm 1 achieves asymptotically optimal performance:

Corollary 7. *Suppose the conditions of Theorem 6 hold. Assume, furthermore, that $\beta(n)$ satisfies $\beta(n) = o(n)$ and $\sum_{s=0}^{\infty} \exp\left(-\frac{\beta(s)\epsilon^2}{2K\sigma^2}\right) < \infty$ for any $\epsilon > 0$, then for any θ such that $c(\theta)$ is unique,*

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

Note that any $\beta(n) = an^b$ with $a \in (0, \frac{1}{2}]$, $b \in (0,1)$ satisfies the requirements in Theorem 6 and Corollary 7. Also note that the algorithms presented in [6, 7] do not achieve this asymptotic bound.

4.2 A Minimax Optimal Algorithm

For any $A, A' \subset [K]$, define $c(A, A') = \mathrm{argmax}_{c \in \Delta^{|A|}} \min_{i \in A'} \sum_{j : i \in S_j} c_j$ (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 $|A| \ge 2$, define $A^{\mathcal{S}} = \{i \in A \, : \, \exists j \in A, i \in S_j\}$ and $A^{\mathcal{W}} = A - A^{\mathcal{S}}$. Furthermore, let $g_{i,r}(\delta) = \sigma \sqrt{\frac{2 \log(8K^2 r^3 / \delta)}{n_i(r)}}$ where $n_i(r) = \sum_{s=1}^{r-1} i_{r,i}$ and $\hat{\theta}_{i,r}$ be the empirical estimate of θ_i based on first $n_i(r)$ observations (i.e., the average of the samples).

Our second algorithm, presented in Algorithm 2, 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, it 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 the actions that are 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 2

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

Theorem 8. *For any $\delta \in (0,1)$ and any $\theta \in \Theta$,*

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

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.

Theorem 9 (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 $1-\delta$,*

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

Remark 10. Picking $\delta = 1/T$ gives an $O\left(\log^{3/2} T\right)$ upper bound on the expected regret.

Remark 11. Note that Algortihm 2 is similar to the UCB-LP algorithm of [7], which admits a better problem-dependent upper bound (although does not achieve it with optimal problem-dependent constants), but it does not achieve the minimax bound even under strong observability.

5 Conclusions and Open Problems

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 harness of the problem, and provide algorithms that can adapt to any correspondence between the mean payoffs an the variances (the hardness is that one needs to identify suboptimal actions with good information/cost trade-off).

Acknowledgments

This work was supported by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning (AICML) and NSERC. During this work, A. György was with the Department of Computing Science, University of Alberta.

References

[1] 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.

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

[3] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems 26 (NIPS-2013)*, pages 1610–1618, 2013.

[4] Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 613–621, 2014.

[5] Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: beyond bandits. In *Proceedings of The 28th Conference on Learning Theory (COLT-2015)*, 2015. (to appear).

[6] Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012)*, pages 142–151, 2012.

[7] Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. *SIGMETRICS Perform. Eval. Rev.*, 42(1):289–300, June 2014.

[8] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

[9] G. Bartók, D. Foster, D. Pál, A. Rakhlin, and Cs. Szepesvári. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39:967–997, 2014.

[10] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

[11] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015.

[12] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *Volume 35: Proceedings of The 27th Conference on Learning Theory (COLT-2014)*, 2014.

[13] 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*, 2015. (to appear).

[14] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 521–529, 2014.

A Proofs for Section 3

A.1 Proof of Theorem 1

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., [13, 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

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

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 1.

A.2 Proof of Corollary 3

We start the proof with two technical lemmas on the Lambert W function.

Lemma 12. *Let $a, b > 0$ with $ab < 1$ and $f(x) = \frac{1}{x^2} \log((x + a)b)$ for $x > 0$. Then $f(x) \leq f(x_*)$ for all $x > 0$ where*

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

Lemma 13. *Let $a, b > 0$ and $f(x) = \frac{1}{x^2} \log((x - a)b)$ for $x > a$. Then $f(x) \leq f(x_*)$ for all $x > a$ where*

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

Proof of Lemma 13.

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

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 $x > a$.

□

Proof of Corollary 3. To prove the corollary, it suffices to show $b'(\theta, B) \le b(\theta, B)$.

Define $C'_{\theta,B} = \left\{c \in C_T^{\mathbb{R}} : \sum_{j=1}^K \frac{c_j}{\sigma_{ji}^2} \ge m_i(\theta, B)\,, \forall i \in [K]\right\}$. 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 $\sup_{\theta'\in\Theta} f(\theta, q, \theta') \le B$ 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 14. *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

$$\sum_{a\in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \ge \mathrm{KL}(q(A), q'(A))\,, \tag{3}$$

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) proves the lemma. □

Now we continue the proof of Corollary 3. First consider $i \ne i_1(\theta)$.

If $\sum_{a:a_i\le T/2} q(a) \ge 1/2$, construct $\theta^{(i,+)}$ by replacing θ_i with $\theta_i + \epsilon_{i,+}$. Then $f(\theta, q, \theta^{(i,+)}) \le B$ holds, so there exists $q' \in \Delta^{|C_T^{\mathbb{N}}|}$ such that $\sum_{a\in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,+)})\right\rangle \le B$ and $\sum_{a\in C_T^{\mathbb{N}}} q(a) \log \frac{q(a)}{q'(a)} \le \sum_{j\in[K]} c_j I_j(\theta, \theta^{(i,+)})$. Applying Lemma 14 with $A = \{a : a_i \le T/2\}$ gives

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

where

$$\begin{aligned} q'(A) &= \sum_{a\in C_T^{\mathbb{N}}} \mathbb{I}\left\{\sum_{j\ne i} a_j \ge T/2\right\} q'(a) \le \frac{2}{T} \sum_{a\in C_T^{\mathbb{N}}} q'(a) \sum_{j\ne i} a_j \\ &= \frac{2}{T(\epsilon_{i,+} - d_i(\theta))} \sum_{a\in C_T^{\mathbb{N}}} q'(a) \sum_{j\ne i} a_j(\epsilon_{i,+} - d_i(\theta)) \\ &\le \frac{2}{T(\epsilon_{i,+} - d_i(\theta))} \sum_{a\in C_T^{\mathbb{N}}} q'(a) \left\langle a, d(\theta^{(i,+)})\right\rangle \end{aligned}$$

$$\leq \frac{2B}{T(\epsilon_{i,+} - d_i(\theta))} \,.$$

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

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\epsilon_{i,+}^2} \log \frac{T(\epsilon_{i,+} - d_i(\theta))}{8B} \,. \tag{4}$$

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

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

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 $\theta_i - \epsilon_{i,-}$. 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 14 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

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

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

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\epsilon_{i,-}^2} \log \frac{T(\epsilon_{i,-} + d_i(\theta))}{8B} \,. \tag{5}$$

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 $\theta_i - \epsilon_{i,-}$. 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 14 with $A = \{a : a_i \geq T/2\}$ and

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

gives

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\epsilon_{i,-}^2} \log \frac{T(\epsilon_{i,-} - d_{i_2(\theta)}(\theta))}{8B} \,. \tag{6}$$

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

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

$$> \frac{4B}{T} \frac{T}{2} \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ \sum_{j \neq i} a_j > T/2 \right\} q(a) \geq B \,,$$

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 $\theta_i + \epsilon_{i,+}$. 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 14 with $A = \{a : a_i < T/2\}$ and

$$q'(A) = \sum_{a \in C_T^{\mathbb{N}}} \mathbb{I}\left\{ \sum_{j \neq i} a_j > T/2 \right\} q'(a) \leq \frac{2}{T} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \sum_{j \neq i} a_j$$

$$= \frac{2}{T(\epsilon_{i,+} + d_{i_2(\theta)}(\theta))} \sum_{a \in C_T^{\mathbb{N}}} q'(a) \sum_{j \neq i} a_j (\epsilon_{i,+} + d_{i_2(\theta)}) \leq \frac{2B}{T(\epsilon_{i,+} + d_{i_2(\theta)})}$$

gives

$$\sum_{j \in [K]} \frac{c_j}{\sigma_{ji}^2} \geq \frac{1}{\epsilon_{i,+}^2} \log \frac{T(\epsilon_{i,+} + d_{i_2(\theta)})}{8B} \,. \tag{7}$$

Combining (4) (5) (6) (7) gives $c \in C'_{\theta,B}$, which concludes the proof.

□

A.3 Proof of Corollary 4

Proof of Corollary 4. Define $\epsilon = \frac{8eB}{T}$. 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 $\theta_i = D/2 - \epsilon$ 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

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

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

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

Recall that we defined $C'_{\theta,B} = \left\{ c \in C_T^{\mathbb{R}} : \sum_{j:i \in S_j} c_j \geq \sigma^2 m_i(\theta, B) \,, \forall i \in [K] \right\}$ 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 $c_i \geq \sigma^2 m_i(\theta, B) - a = \frac{\sigma^2}{4\epsilon^2} - a$. Since $d_i(\theta) = \epsilon$ for all $i \in A_\kappa \setminus \{i_1\}$ and $d_i(\theta) = D/2$ for all $i \notin A_\kappa$, we get

$$\langle c, d(\theta) \rangle = \sum_{i \in A_\kappa \setminus \{i_1\}} c_i \epsilon + \frac{aD}{2} \geq (\kappa(\Sigma) - 1) \left(\frac{\sigma^2}{4\epsilon^2} - a \right) \epsilon + \frac{aD}{2}$$

$$\geq \frac{\kappa(\Sigma)}{2}\left(\frac{\sigma^2}{4\epsilon^2} - a\right)\epsilon + \frac{aD}{2} = \frac{\kappa(\Sigma)\sigma^2}{8\epsilon} + \frac{D - \kappa(\Sigma)\epsilon}{2}a$$
$$\geq \frac{\kappa(\Sigma)\sigma^2}{8\epsilon} \tag{8}$$

if $\kappa(\Sigma)\epsilon < D$. 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 $\theta_i = D/2 - \epsilon$ for all $i \neq i_1$. Then $m_i(\theta, B) = \frac{1}{4\epsilon^2}$ 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 $a \geq \sigma^2 m_{i_2}(\theta, B) = \frac{\sigma^2}{4\epsilon^2}$ and

$$\langle c, d(\theta)\rangle = a\epsilon \geq \frac{\sigma^2}{4\epsilon} > \frac{\kappa(\Sigma)\sigma^2}{8\epsilon}\,. \tag{9}$$

Combining (8) and (9) gives the first part of Corollary 4.

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 [5] which based on the following graph-theoretic lemma:

Lemma 15 (Restated from Lemma 8 in [5])**.** *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 15 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 $\theta_i = D/2 - \epsilon$ 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

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

Therefore,

$$\langle c, d(\theta)\rangle \geq \frac{aD}{2} \geq \frac{\rho(\Sigma)\sigma^2 D}{200\epsilon^2\log^2 K} = \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}\log^{-2/3} K}{12800e^2\alpha^2}\,. \tag{10}$$

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 $\theta_{i_1} = D/2$, $\theta_{i_2} = D/2 - \epsilon$ 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

$$\langle c, d(\theta)\rangle \geq \frac{aD}{2} \geq \frac{\sigma^2 m_{i_2}(\theta, B)D}{2} = \frac{D\sigma^2}{8\epsilon^2} = \frac{D^{1/3}(\sigma T)^{2/3}}{512e^2\alpha^2} \cdot \frac{\log^{4/3} K}{\rho(\Sigma)^{2/3}}$$
$$\geq \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}\log^{-2/3} K}{51200e^2\alpha^2}\,. \tag{11}$$

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

Combining (10) and (11) gives the second part of Corollary 4.

□

B Proofs for Section 4.1

B.1 Proof of Theorem 6

Proof of Theorem 6. Define events

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

$$V_t = \left\{\forall i \in [K]\,, |\hat{\theta}_{i,t} - \theta_i| \leq \epsilon\right\}\,,$$

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

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

and U_t^c, V_t^c, W_t^c, Y_t^c be their complements.

$$\begin{aligned}
R_T(\theta) &= \sum_{t=1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\right] \leq K d_{\max}(\theta) + \sum_{t=K+1}^{n} \mathbb{E}\left[d_{i_t}(\theta)\right] \\
&= K d_{\max}(\theta) + \sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\left(\mathbb{I}\left\{U_t^c\right\} + \mathbb{I}\left\{U_t, W_t\right\} + \mathbb{I}\left\{U_t, W_t^c, Y_t\right\}\right.\right. \\
&\left.\left.+\mathbb{I}\left\{U_t, W_t^c, Y_t^c, V_t^c\right\} + \mathbb{I}\left\{U_t, W_t^c, Y_t^c, V_t\right\}\right)\right]\,. \qquad (12)
\end{aligned}$$

Then we will upper bound each quantity in (12) separately.

By Hoeffding's inequality, we have

$$\Pr\left(|\hat{\theta}_{i,t} - \theta_i| > \sqrt{\frac{2\alpha\sigma^2 \log t}{n_i(t)}}\right) \leq 2t^{1-\alpha}\,,$$

where we use a union bound over all possible $n_i(t)$.

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

$$\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t^c\right\}\right] \leq d_{\max}(\theta) \sum_{t=K+1}^{T} \Pr(U_t^c) \leq d_{\max}(\theta) \sum_{t=K+1}^{T} 2Kt^{1-\alpha} \leq \frac{2K d_{\max}(\theta)}{\alpha - 2}\,. \qquad (13)$$

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

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

and

$$\hat{\theta}_{i_1(\hat{\theta}_t),t} - \theta_{i_1(\hat{\theta}_t)} \leq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_{i_1(\hat{\theta}_t)}(t)}} \leq \frac{d_{i_1(\hat{\theta}_t)}(\hat{\theta}_t)}{2} \leq \frac{d_i(\hat{\theta}_t)}{2} \qquad (14)$$

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

$$\theta_i - \hat{\theta}_{i,t} \leq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_i(t)}} \leq \frac{d_i(\hat{\theta}_t)}{2}\,. \qquad (15)$$

Combining (14) and (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

$$\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t, W_t\right\}\right] = 0\,. \tag{16}$$

Consider the next term in (12),

$$\sum_{t=K+1}^{T} \mathbb{E}\left[d_{i_t}(\theta)\mathbb{I}\left\{U_t, W_t^c, Y_t\right\}\right] \leq d_{\max}(\theta)\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{U_t, W_t^c, Y_t\right\}\right]\,. \tag{17}$$

To upper bound (17), we will first prove:

Proposition 16.

$$\sum_{t=K+1}^{T} \mathbb{I}\left\{W_t^c, Y_t\right\} \leq 1 + \beta\left(\sum_{t=K+1}^{T} \mathbb{I}\left\{W_t^c\right\}\right)\,. \tag{18}$$

Proof of (18). According to the algorithm we have $n_e(t) = \sum_{s=K+1}^{t-1} \mathbb{I}\left\{W_s^c\right\}$ for $t > K$, we then proceed by the following proposition:

Proposition 17. *For* $K < t_1 < t_2$*, 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 of Proposition 17. If for such t_1 and t_2, $\min_{i\in[K]} n_i(t_2) = \min_{i\in[K]} n_i(t_1)$, then there must exist 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 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 $j' \neq j$ such that $j' \in S_{i_{s_k}}$ and $n_{j'}(s_k) = n_j(s_k) = \min_{i\in[K]} n_i(s_k)$. Note that each action appears at most once as such j' for $1 \leq k \leq K$ since $n_{j'}(s_k + 1) = n_{j'}(s_k) + 1$, but there are only $K - 1$ actions other than j, which means such j cannot exist. Hence $\min_{i\in[K]} n_i(t_2) \geq \min_{i\in[K]} n_i(t_1) + 1$ is proved. □

Now we define

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

If such t' does not exist, then (18) must hold. If such t' exists, by Proposition 17,

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

Therefore,

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

gives (18). □

Now continue with (17)

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

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

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

Thus we have

$$\sum_{t=K+1}^{T} \mathbb{I}\{U_t, W_t^c, Y_t\}$$

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

and

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

$$\leq 2d_{\max}(\theta) + \frac{2K d_{\max}(\theta)}{\alpha - 2} + d_{\max}(\theta) \sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\{U_t, W_t^c, Y_t^c, V_t^c\} \right]$$

$$+ 2d_{\max}(\theta) \mathbb{E}\left[\beta \left(\sum_{t=K+1}^{n} \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\} \right) \right] \tag{19}$$

by applying (13).

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 from [14] (Lemma 2.1 and 2.2):

Lemma 18. *Let $\{Z_t\}_{t \in \mathbb{N}^+}$ be a sequence of independent random variables from $\mathcal{N}(0, \sigma^2)$. Define $\mathcal{F}_t$ the σ-algebra generated by $\{Z_s\}_{s \leq t}$ and 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. Further define $n(t) = \sum_{s=n_0}^{t-1} B_s$ and ϕ an $\mathcal{F}$-stopping time which satisfies either $n(\phi) \geq r$ or $\phi = T + 1$.*

Then we have

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

Lemma 19. *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 $\epsilon > 0$ and $i \in [K]$:*

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

Proof of Lemma 19. We adapt the proof of Lemma 2.2 from [14]. For $0 \leq s \leq T$, define $\phi_s = t$ if $\Lambda_s = \{t\}$ or $\phi_s = T + 1$ if $\Lambda_s = \emptyset$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{ t \in \Lambda, |\hat{\theta}_{i,t} - \theta_i| > \epsilon \right\} \right] \leq \mathbb{E}\left[\sum_{s=0}^{T} \mathbb{I}\left\{ \phi_s \leq T, |\hat{\theta}_{i,\phi_s} - \theta_i| > \epsilon \right\} \right]$$

$$= \sum_{s=0}^{T} \Pr\left(\phi_s \leq T, |\hat{\theta}_{i,\phi_s} - \theta_i| > \epsilon \right) . \tag{20}$$

Since ϕ_s can be viewed as an $\mathcal{F}$-stopping time and satisfies either $n_i(\phi_s) \geq \lceil \beta(s)/K \rceil$ or $\phi_s = T+1$, if $\lceil \beta(s)/K \rceil \geq 1$ then applying Lemma 18 gives

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

If $\lceil \beta(s)/K \rceil = 0$ then $\Pr\left(\phi_s \leq T, |\hat{\theta}_{i,\phi_s} - \theta_i| > \epsilon\right) < 2 = 2\exp\left(-\frac{\beta(s)\epsilon^2}{2K\sigma^2}\right)$ still holds. Now proceeding from (20) we can get the result of Lemma 19.

□

Now we define $\Lambda = \{t : K+1 \leq t \leq T, U_t, W_t^c, Y_t^c\}$, and $\Lambda_s = \{t : K+1 \leq t \leq T, U_t, W_t^c, n_e(t) = s, \min_{i\in[K]} n_i(t) \geq \beta(s)/K\}$. It can be verified that Λ_s satisfies the conditions in Lemma 19: (i) If $t \in \Lambda$ then there must be some $0 \leq s \leq T$ such that $n_e(t) = s$ and thus $t \in \Lambda_s$. (ii) If $t \in \Lambda_s$ then for $t' > t$, $n_e(t') \geq n_e(t+1) = n_e(t)+1 = s+1$, so $t' \notin \Lambda_s$. Condition (iii) and (iv) are also satisfied from the definition of Λ_s.

Then

$$\begin{aligned}&\sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\left\{U_t, W_t^c, Y_t^c, V_t^c\right\}\right] = \sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\left\{t \in \Lambda, V_t^c\right\}\right] \\ &\leq \sum_{i=1}^{K}\sum_{t=K+1}^{T} \mathbb{E}\left[\mathbb{I}\left\{t \in \Lambda, |\hat{\theta}_{i,t} - \theta_i| > \epsilon\right\}\right] \leq 2K\sum_{s=0}^{T} \exp\left(-\frac{\beta(s)\epsilon^2}{2K\sigma^2}\right) . \qquad (21)\end{aligned}$$

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 $\frac{N_i(t)}{4\alpha\log t} \geq c_i(\hat{\theta}_t)4\alpha\log t$ 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

$$\Theta(\theta, \epsilon) = \{\lambda \in \Theta : \forall i \in [K], |\lambda_i - \theta_i| \leq \epsilon\} ,$$

and

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

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

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

Thus

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

So we have

$$\sum_{t=K+1}^{T} d_{i_t}(\theta)\mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\} \leq 4\alpha\log T\sum_{i\in[K]} c_i(\theta,\epsilon)d_i(\theta) + \sum_{i\in[K]} d_i(\theta) , \qquad (22)$$

and

$$\sum_{t=K+1}^{T} \mathbb{I}\{U_t, W_t^c, Y_t^c, V_t\} \leq 4\alpha\log T\sum_{i\in[K]} c_i(\theta,\epsilon) + K . \qquad (23)$$

Now plugging (23) (21) into (19) and plugging (13) (16) (19) (21) (22) into (12) we get

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

□

C Proofs for Section 4.2

C.1 Proof of Theorem 8

Proof of Theorem 8. For every $r > 0$, define the events

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

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

$$\Pr(\forall r \ge 2, U_r) \ge 1 - \delta\,.$$

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

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

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

$$R_T(\theta) \le \sum_{r=1}^{r_T} \mathbb{I}\{V_r\}\langle i_r, d(\theta)\rangle + \sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\}\langle i_r, d(\theta)\rangle$$
$$\le \sum_{r=1}^{r_T} \mathbb{I}\{V_r\}\|i_r\|_1 D + \sum_{r=1}^{r_T}\mathbb{I}\{V_r^c\}\|i_r\|_1 \max_{i\in A_r} d_i(\theta)\,. \tag{24}$$

We upper bound the two terms in (24) separately. Before proceeding, we introduce the following proposition which lower bounds $n_i(r)$ for $i \in A_r^{\mathcal{W}}$.

Proposition 20. *For any i, r such that $i \in A_r^{\mathcal{W}}$,*

$$n_i(r) \ge \frac{\alpha_{r-1}}{2}\sum_{s=1}^{r-1}\mathbb{I}\{V_s\}\|i_s\|_1 - (\beta_r - 1)K\,, \tag{25}$$

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

Proof of Proposition 20. The proof is done by induction. Let W_r denote the event that for any $1 \le s \le r$ and any $i \in A_s^{\mathcal{W}}$, (25) holds. W_1 holds because $A_1^{\mathcal{W}} = \emptyset$. Now we assume W_r holds and consider W_{r+1}.

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

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

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

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

$$\|i_r\|_1 = \|\lceil c_r \cdot \|c_r\|_0 \rceil\|_1 \le \|c_r\|_0 + \|c_r\|_0 \cdot \|c_r\|_1 = 2\|c_r\|_0 \,. \tag{26}$$

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

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

If such r' exists

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

where the facts $\alpha_r \le 1$, $\|i_{r'}\|_1 \le 2K$ and $\beta_{r'} \le \beta_{r+1} - 1$ 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 20.

□

Based on Proposition 20, $\sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1$ can be upper bounded by the following fact:

Proposition 21. *For any* $r \ge 1$, $\sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 \le \frac{2\gamma(r) + 2K\beta_r}{\alpha_r}$.

Proof of Proposition 21. Let $r' = \max\{s \le r \,:\, V_s\}$. Then

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

Hence

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

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 $\sum_{s=1}^{r} \mathbb{I}\{V_s\} \|i_s\|_1 \le \frac{2\gamma(r) + 2K\beta_r}{\alpha_r}$.

□

Now we are ready to upper bound the first term in (24):

$$\sum_{r=1}^{r_T} \mathbb{I}\{V_r\} \|i_r\|_1 D \le \frac{2\gamma(r_T) + 2K\beta_{r_T}}{\alpha_{r_T}} D = 2\alpha_{r_T}^{-1/3} D^{1/3} (\sigma T)^{2/3} + 2KD \frac{\beta_{r_T}}{\alpha_{r_T}} \,. \tag{27}$$

Next consider the second term in (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 $|\hat{\theta}_{i,r} - \theta_i| \le g_{i,r}(\delta)$ and $\hat{\theta}_{i,r} + g_{i,r}(\delta) \ge \hat{\theta}_{i_1(\theta)} - g_{i_1(\theta),r}(\delta)$. Therefore,

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

So

$$\sum_{r=1}^{r_T} \mathbb{I}\left\{V_r^c\right\} \|i_r\|_1 \max_{i \in A_r} d_i(\theta) \le \sum_{r=1}^{r_T} \mathbb{I}\left\{V_r^c\right\} \|i_r\|_1 \min\left\{D, C(\min_{i\in A_r} n_i(r))^{-1/2}\right\}, \tag{28}$$

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 $i \in A_r^{\mathcal{S}}$,

$$n_i(r) \ge \sum_{s=1}^{r-1} \mathbb{I}\left\{V_s^c\right\} \|c_s\|_0 \, m(A_s, A_s^{\mathcal{S}}) \ge \frac{\eta_{\min}}{2} \sum_{s=1}^{r-1} \mathbb{I}\left\{V_s^c\right\} \|i_s\|_1 . \tag{29}$$

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

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

By Proposition 21,

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

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

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

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

So we have proved that for any $r \le r_T$ such that $t_r \ge T_0 = \frac{125\sigma^2}{D^2}K^3 + 10K^3$ and V_r^c happens, $\min_{i\in A_r} n_i(r) \ge \gamma(r) \ge (\sigma\alpha_{r_T} t_r/D)^{2/3}$. Therefore, following (28) gives

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

$$
\begin{aligned}
&\leq (T_0+2K)D + C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3}\int_{T_0-2K}^{t_{r_T}} x^{-1/3}dx\\
&\leq (T_0+2K)D + \frac{3}{2}C\left(\frac{\sigma\alpha_{r_T}}{D}\right)^{-1/3} T^{2/3}\\
&= \frac{125\sigma^2K^3}{D} + (10K^3+2K)D + \alpha_{r_T}^{-1/3}D^{1/3}(\sigma T)^{2/3}\cdot 6\sqrt{6\log\frac{2KT}{\delta}}\,. \qquad (30)
\end{aligned}
$$

Now plugging (27) and (30) into (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 (24) (28) and (29) we have

$$
\begin{aligned}
R_T(\theta) &\leq \sum_{r=1}^{r_T}\|i_r\|_1 \max_{i\in A_r} d_i(\theta)\\
&\leq \sum_{r=1}^{r_T}(t_{r+1}-t_r)\min\left\{D, C\left(\frac{\eta_{\min}}{2}\right)^{-1/2} t_r^{-1/2}\right\}\\
&\leq 2KD + C\left(\frac{\eta_{\min}}{2}\right)^{-1/2}\int_0^{t_{r_T}} x^{-1/2}dx\\
&\leq 2KD + 8\sigma\sqrt{\frac{T}{\eta_{\min}}\cdot 12\log\frac{2KT}{\delta}}\,.
\end{aligned}
$$

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

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

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

Proof of Proposition 22. Recall that

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

Let $b = c/a$, then

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

□

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

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

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

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

($A^{\mathcal{S}} = A$ 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 15 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 $\frac{1}{\eta_{\min}} \leq \kappa(\Sigma) 50 \log K$.

□

C.2 Proof of Theorem 9

Proof of Theorem 9. Similarly to the proof of Theorem 9, we define high probability events

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

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

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

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 $2g_{i,r}(\delta) + 2g_{i_1(\theta),r}(\delta) \geq d_i(\theta) \geq d_{\min}(\theta)$, where $d_{\min}(\theta)$ denotes $d_{i_2(\theta)}(\theta)$. From $g_{i,r}(\delta) = \sigma\sqrt{\frac{2\log(8K^2r^3/\delta)}{n_i(r)}}$ 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

$$\min_{i \in A_r} n_i(r) \leq \frac{C_r^2}{d_{\min}^2(\theta)}. \tag{31}$$

Then consider the regret:

$$R_T(\theta) \leq \sum_{r=1}^{r_T} \mathbb{I}\{V_r\} \langle i_r, d(\theta) \rangle + \sum_{r=1}^{r_T} \mathbb{I}\{V_r^c\} \langle i_r, d(\theta) \rangle$$

$$\leq \sum_{r=1}^{r_V} \mathbb{I}\left\{V_r\right\} \left\|i_r\right\|_1 d_{\max}(\theta) + \sum_{r=1}^{r_W} \mathbb{I}\left\{V_r^c\right\} \left\|i_r\right\|_1 \max_{i \in A_r} d_i(\theta)\,. \tag{32}$$

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

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

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

by applying Proposition 20. Then we can upper bound the first term in (32) by

$$\sum_{r=1}^{r_V} \mathbb{I}\left\{V_r\right\} \left\|i_r\right\|_1 \leq \frac{2\rho(\Sigma) C_{r_V}^2}{d_{\min}^2(\theta)} + 2\rho(\Sigma)K^2 + 2K\,. \tag{33}$$

Regarding the second term in (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 $\max_{i \in A_r} d_i(\theta) \leq \min\left\{d_{\max}(\theta), C_r \left(\min_{i \in A_r} n_i(r)\right)^{-1/2}\right\}$ gives

$$\begin{aligned}
&\sum_{r=1}^{r_W} \mathbb{I}\left\{V_r^c\right\} \left\|i_r\right\|_1 \max_{i \in A_r} d_i(\theta) \\
&\leq \sum_{r=1}^{r_W} \mathbb{I}\left\{V_r^c\right\} \left\|i_r\right\|_1 \min\left\{d_{\max}(\theta), C_r (\min_{i \in A_r} n_i(r))^{-1/2}\right\} \\
&\leq \sum_{r \geq 1 : t_r < T_0} \left\|i_r\right\|_1 d_{\max}(\theta) + \sum_{r \leq r_W : t_r \geq T_0} \left\|i_r\right\|_1 C_{r_W} \left(\frac{\sigma}{\rho(\Sigma) D}\right)^{-1/3} t_r^{-1/3} \\
&\leq (T_0 + 2K) d_{\max}(\theta) + C_{r_W} \left(\frac{\sigma}{\rho(\Sigma) D}\right)^{-1/3} \sum_{r \leq r_W : t_r \geq T_0} (t_{r+1} - t_r)(t_{r+1} - 2K)^{-1/3} \\
&\leq (T_0 + 2K) d_{\max}(\theta) + C_{r_W} \left(\frac{\sigma}{\rho(\Sigma) D}\right)^{-1/3} \int_{T_0}^{t_{r_W + 1}} (x - 2K)^{-1/3} dx \\
&\leq (T_0 + 2K) d_{\max}(\theta) + C_{r_W} \left(\frac{\sigma}{\rho(\Sigma) D}\right)^{-1/3} \int_{T_0 - 2K}^{t_{r_W}} x^{-1/3} dx \\
&\leq (T_0 + 2K) d_{\max}(\theta) + \frac{3}{2} C_{r_W} \left(\frac{\sigma}{\rho(\Sigma) D}\right)^{-1/3} t_{r_W}^{2/3}\,.
\end{aligned} \tag{34}$$

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

$$t_{r_W}^{2/3} \leq \left(\frac{\sigma}{\rho(\Sigma) D}\right)^{-2/3} \frac{C_{r_W}^2}{d_{\min}^2(\theta)} + T_0^{2/3}\,. \tag{35}$$

Combining (32) (33) (34) and (35) with $C_{r_W} \leq C_{r_T}$ gives

$$\begin{aligned}
R_T(\theta) \leq{}& \frac{1603\rho(\Sigma) D \sigma^2}{d_{\min}^2(\theta)} \left(\log \frac{2Kr_T}{\delta}\right)^{3/2} + 14K^3 D + \frac{125\sigma^2 K^3}{D} \\
&+ 15 \left(\rho(\Sigma) D \sigma^2\right)^{1/3} \left(\frac{125\sigma^2}{D^2} + 10\right) K^2 \left(\log \frac{2Kr_T}{\delta}\right)^{1/2}\,.
\end{aligned} \tag{36}$$

Applying $r_T \leq T$ gives the result of Theorem 9.

Note that using $r_T \leq T$ 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 21, $\sum_{s=1}^{r_T} \mathbb{I}\left\{V_s\right\} \left\|i_s\right\|_1 =$

$O\left(t_{r_T}^{2/3}\right)$, and $t_{r_W} = O\left((\log t_{r_T})^{3/2}\right)$, we have

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

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

□