Asymptotically-Optimal Gaussian Bandits with Side Observations

Alexia Atsidakou [*1] **Orestis Papadigenopoulos** [*2] **Constantine Caramanis** [1] **Sujay Sanghavi** [1 3]
Sanjay Shakkottai [1]

Abstract

We study the problem of Gaussian bandits with general side information, as first introduced by Wu, Szepesvári, and György. In this setting, the play of an arm reveals information about other arms, according to an arbitrary *a priori* known *side information* matrix: each element of this matrix encodes the fidelity of the information that the "row" arm reveals about the "column" arm. In the case of Gaussian noise, this model subsumes standard bandits, full-feedback, and graph-structured feedback as special cases. In this work, we first construct an LP-based asymptotic instance-dependent lower bound on the regret. The LP optimizes the cost (regret) required to reliably estimate the suboptimality gap of each arm. This LP lower bound motivates our main contribution: the first known asymptotically optimal algorithm for this general setting.

1. Introduction

In the stochastic online learning framework, a player sequentially selects from a set of available actions (or "arms") and collects a stochastic reward associated with the chosen action. Under the objective of maximizing the (expected) cumulative reward collected over a number of rounds, the "complexity" of an instance is greatly impacted by the nature of the feedback the player receives at the end of each round. In the *full-feedback* case, for instance, where the player observes the realized rewards of all actions, the most reasonable (and provably optimal) strategy is to greedily select at each round the action with maximum estimated mean reward computed using the observed samples. In the *bandit-feedback* case (Lai & Robbins, 1985; Bubeck & Cesa-Bianchi, 2012), on the other hand, where only the reward of the chosen action is revealed to the player, more sophisticated ideas have to be leveraged in order to align the two conflicting objectives of the problem: *exploration* (learning the best action) and *exploitation* (not playing suboptimal actions).

In a number of real world applications, however, the feedback does not fall into any of the above extreme categories, since an action can potentially leak information regarding, not only its own reward distribution, but also the distributions of other actions. Drawing an example from music recommendation, the fact that a user likes a specific recommended song (corresponding to an action chosen by the platform) can potentially imply that similar songs (e.g., of the same artist or genre) might also be appreciated by the user. In another example, encouraging a certain user of a social network to advertise a purchased product (potentially by offering a discount as part of a promotion campaign), the seller can obtain valuable information about other neighboring users by recording their reactions.

Motivated by such practical scenarios, researchers have focused their attention on online learning models with richer information structures – namely, feedback which interpolates between full and bandit (Mannor & Shamir, 2011; Caron et al., 2012; Lin et al., 2014; Yun et al., 2018; Cortes et al., 2020). While the presence of side observations permits improved regret guarantees compared to those of standard multi-armed bandits, techniques and algorithms that have been proved successful for bandit-feedback (e.g., optimism principle and the UCB method) fail to achieve optimality. Indeed, enriching the feedback model severely perplexes the role of exploration and exploitation. To illustrate, one can think of the following example: $K-1$ arms provide to the player standard bandit feedback. In addition to these, there also exists an "information-revealing" arm with (deterministically) zero reward, such that, once played, the player gets full-feedback on the realized rewards of the round, for all arms. Even in the above simple example, the number of times (if any) this "information-revealing" arm should played in a minimum regret solution is now a complex function of K and the (a priori unknown) suboptimality gaps.

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas, USA [2]Department of Computer Science, University of Texas at Austin, Austin, Texas, USA [3]Amazon Science, USA. Correspondence to: Alexia Atsidakou <atsidakou@utexas.edu>, Orestis Papadigenopoulos <papadig@cs.utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Initiated by the work of (Mannor & Shamir, 2011) in the context of adversarial bandits, the majority of works in this field focuses on the so-called *graph-structured feedback* model (Alon et al., 2015; Wu et al., 2015; Kocák et al., 2016; Arora et al., 2019; Cortes et al., 2020). There, each arm corresponds to a node of a given (directed) graph and every time the player chooses an action, she observes the reward realization of the arm played, but also the realizations of all the adjacent arms in the graph. In the above setting, a series of works (Buccapatnam et al., 2014; Alon et al., 2015; Cohen et al., 2016; Lykouris et al., 2020) has provided algorithms with regret guarantees that no longer depend on the number of arms, but on smaller inherent characteristics of the feedback-graph (e.g., clique or independence number).

One of the most general feedback models that has been introduced in the literature is due to Wu, Szepesvári, and György (Wu et al., 2015). According to their model, playing an arm reveals noisy information about the reward of all other arms according to an arbitrary a priori known *side information* matrix. Specifically, each element of this square matrix encodes the fidelity of the information – expressed in terms of standard deviation of the noise – that the "row" arm reveals about the "column" arm. Notice that, modulo the Gaussian noise assumption made in (Wu et al., 2015), the above model subsumes the full, bandit, and graph-structured feedback as special cases. For the case of Gaussian noise, (Wu et al., 2015) provide finite-time instance-dependent lower bounds on the regret of any algorithm. In terms of upper bounds, the authors only address the special case where each entry of the feedback matrix satisfies $\sigma_{i,j} \in \{\sigma, \infty\}$ for some fixed σ (which is essentially the graph-structured feedback model), but they leave open the question of an algorithm for general feedback matrices.

1.1. Our Contributions

In our work, we provide the first algorithm and regret analysis for the setting of general feedback matrices proposed by (Wu et al., 2015). As we show, the regret guarantee of our algorithm is asymptotically optimal, as it matches our asymptotic instance-dependent lower bound. We now outline the main challenges and technical contributions of this work.

Concentration bounds for a natural ML estimator. In the restricted case where the entries of the feedback matrix are in $\{\sigma, \infty\}$, the natural estimator of the mean reward of an arm is the sample average of the (finite-variance) samples collected. In our algorithm, where the collected samples are subject to different noise levels, we replace the above estimator by the weighted average of the collected samples, where the weight of each sample is the inverse variance of the noise of its source. The above maximum-likelihood estimator suggests that the notion of "number of samples", as a measure of the amount of information collected, needs to be replaced by the *weighted number of samples*.

The use of weighted average as an estimator introduces technical hurdles: in scenarios where the number of samples is a random variable that depends on the trajectory of the algorithm, one needs to apply a union bound over all possible sample numbers in order to decorrelate the estimate from the evolution of the algorithm up to each round. However, in our case where each sample can be collected under K different noise levels, the weighted number of samples of an arm can take exponentially many different values – a fact that invalidates the use of union bound. In Section 3, we show how to overcome the above issue by proving concentration guarantees for our estimator, which extend to the more general case of sub-Gaussian noise.

Asymptotic regret lower bound. In our setting, a minimum regret arm-pulling schedule is a complicated function of the number of arms, the suboptimality gaps, and the feedback matrix. In the imaginary scenario where prior knowledge of the suboptimality gaps is assumed, the underlying (combinatorial) problem is to collect sufficient information in order to distinguish the optimal arm by paying the minimum cost (i.e., total suboptimality gap of the arms played for exploration). Although this problem is computationally hard, it can be closely approximated (up to rounding errors) and relaxed by a linear program (LP). As we prove in Section 4, since the above approximation becomes tighter as the time horizon increases, the exact same LP provides a clear and intuitive asymptotic lower bound on the regret.

Asymptotically optimal algorithm. The LP-based asymptotic regret lower bound that we construct serves as a starting point for the design of our asymptotically optimal algorithm. Since the suboptimality gaps are initially unknown, the role of our algorithm is now twofold: to estimate the LP (including constraints and objective) up to a sufficient degree, while simultaneously to implement its optimal solution in an online manner. In order to achieve the above dual objective, our algorithm interleaves rounds of pure greedy exploitation, when sufficient information (with respect to the estimated LP) has been collected, and exploration rounds. The latter case is further partitioned into rounds where, by greedily collecting high fidelity samples, the algorithm attempts to uniformly estimate the LP, and more refined exploration rounds that are dictated by its solution. In Section 5, we describe our algorithm and prove its asymptotic optimality.

Due to space constraints, all omitted proofs have been moved to the Appendix.

1.2. Related Work

The problem of bandits with graph-structured feedback was introduced by (Mannor & Shamir, 2011) in an adversarial

setting. Their model naturally interpolates between experts, where the learner observes feedback from all actions, and bandits, where the learner receives feedback only from the action selected at each round. Since then, the graph feedback model has been extensively studied in both adversarial and stochastic settings.

In the stochastic case, the work of (Caron et al., 2012) presents a $\Omega(\log(T))$ regret lower bound for the graph feedback setting, as long as not all suboptimal arms have a maximum-reward neighbor. In addition, (Caron et al., 2012) examine the regret of natural UCB variants that include the side observations in the UCB indices. They provide instance-dependent regret guarantees of worst-case order $\mathcal{O}\left(\chi \cdot \frac{\Delta_{\max}}{\Delta^2_{\min}} \cdot \log(T)\right)$, where χ is the *clique-partition number* of the graph (i.e. the minimum number of cliques into which the graph can be partitioned). (Buccapatnam et al., 2014) provide an asymptotic instance-dependent LP lower bound for the graph-structured feedback model. The authors also propose two algorithms which exploit (a relaxation of) the minimum *dominating set* of the graph (i.e., the smallest set of nodes which, in terms of their one-step neighbors, covers all arms) in order to perform more efficient exploration. Their gap-dependent regret guarantees are of the form $\mathcal{O}\left(\sum_{i\in D} \frac{\log(T)}{\Delta_i}\right)$, where D is a special *dominating set* of the graph. Similar guarantees for UCB and Thomson Sampling are provided by (Lykouris et al., 2020) in terms of the *independence number* of the feedback graph (i.e. the size of the maximum independent set).

In principle, it seems that the absence of initial knowledge of the suboptimality gaps prevents the above approaches from achieving optimal regret guarantees. The work of (Wu et al., 2015) is the first to resolve the above issue and achieve optimal instance-dependent regret. Focusing on the Gaussian setting, they present an algorithm that attempts to estimate the gaps as well as satisfy the constraints of the asymptotic instance-dependent LP lower bound due to (Buccapatnam et al., 2014). They also present a minimax optimal algorithm for this setting. In addition, (Wu et al., 2015) generalize the graph-structured feedback model to arbitrary feedback matrices, for which they present finite-time gap-dependent lower bounds.

The graph-structured feedback model has also been extensively studied in an adversarial setting. (Mannor & Shamir, 2011) present algorithms with regret guarantees of the form $\mathcal{O}(\sqrt{\alpha \cdot \log(k) \cdot T})$, for undirected, and $\mathcal{O}(\sqrt{\chi \cdot \log(k) \cdot T})$ for directed graphs, together with $\Omega(\sqrt{\alpha \cdot T})$ lower bounds. (Alon et al., 2013) provide improved results for the directed feedback graph case. In (Alon et al., 2015), the authors investigate the relation between the observability properties of the graph and the optimal achievable regret. (Kocák et al., 2016) generalize the graph feedback to weighted graph and provide $\tilde{\mathcal{O}}(\sqrt{\alpha^* \cdot T})$ guarantees, where α^* is a generalization of the independence number for weighted graphs and the $\tilde{\mathcal{O}}(\cdot)$ notation is used to hide logarithmic terms. Other variants of adversarial online learning with graph-structured feedback have been studied in (Rangi & Franceschetti, 2019; Cortes et al., 2019; Arora et al., 2019; Resler & Mansour, 2019; Cortes et al., 2020).

Finally, additional examples of bandits that incorporate graph-structured feedback include the work of (Liu et al., 2018a;b), where the authors study the Bayesian version of the problem and provide $\mathcal{O}(\sqrt{\alpha \cdot T \cdot \log(K)})$ regret guarantees for time-varying graphs. In (Liu et al., 2018a) the authors employ information directed sampling for this setting. (Li et al., 2020) extend the graph-structured feedback by adding observation probabilities on the edges. (Cohen et al., 2016) study the case where the feedback graph is never fully revealed to the learner. Other variations of the problem include (Yun et al., 2018; Singh et al., 2020; Wang et al., 2020), and the so-called partial monitoring setting (Lin et al., 2014; Bartók et al., 2014; Hanawal et al., 2016).

2. Problem Definition

Model. We consider the variant of the stochastic multi-armed bandit problem where the player is given K arms with (unknown) expected rewards $\mu = (\mu_1, \dots, \mu_K)$ and a feedback matrix $\mathbf{\Sigma} = (\sigma_{i,j})_{i,j\in[K]}$. By playing an action $i \in [K]$ the player observes a noisy sample X_j from the reward of each arm $j \in [K]$, distributed independently as $X_j \sim \mathcal{N}(\mu_j, \sigma^2_{i,j})$, and collects the realized value $X_i \sim \mathcal{N}(\mu_i, \sigma^2_{i,i})$. Therefore, the matrix $\mathbf{\Sigma}$ quantifies the quality of the noisy observations of the expected rewards in terms of standard deviation of the Gaussian noise. At any round t where arm i_t is played, we denote by $X_{j,t}$ the noisy observation for each arm $j \in [K]$, which coincides with the realized collected reward in the case where $j = i_t$. The objective is to minimize the expected cumulative regret over an unknown time horizon T, defined as

$$R_T(\mu) = T \cdot \mu^* - \mathbb{E}\left[\sum_{t\in[T]} X_{i_t,t}\right],$$

where $\mu^* = \max_{i\in[K]} \mu_i$ is the maximum expected reward.

We remark that we do not impose any non-trivial restrictions on the feedback matrix $\mathbf{\Sigma} \in \mathbb{R}^{k\times k}_{\geq 0}$. In particular, we allow $\mathbf{\Sigma}$ to be asymmetric, i.e., $\sigma_{i,j} \neq \sigma_{j,i}$ for $i \neq j$, and we permit that $\sigma_{j,i} < \sigma_{i,i}$, namely, higher quality information about the reward of an action can potentially be obtained by playing a different action. Finally, $\sigma_{i,j} = \infty$ corresponds to the case where pulling arm i reveals no information about the reward of arm j.

Technical notation. We denote by $i^*(\mu)$ the maximum expected reward arm of vector μ and by μ^* its expected re-

ward. In the case of more than one optimal arms, we choose $i^*(\mu)$ to be the optimal arm of smallest index. For any vector μ, the suboptimality gap of arm $i \in [K]$ is defined as $\Delta_i(\mu) = \mu^* - \mu_i$. We define the minimum and maximum suboptimality gaps as $\Delta_{\min}(\mu) = \min_{i:\Delta_i(\mu)>0} \Delta_i(\mu)$ and $\Delta_{\max}(\mu) = \max_{i\in[K]} \Delta_i(\mu)$, respectively. For brevity, we simply use $\Delta_{\max}$ instead of $\Delta_{\max}(\mu)$ when μ is the actual mean reward vector.

For any feedback-matrix $\mathbf{\Sigma}$, we define $\sigma_i^{\min} = \min_{j\in[K]} \sigma_{j,i}$ to be the minimum standard deviation of the noise under which we can obtain information about μ_i. Note that, although the condition $\sigma_i^{\min} < \infty$ for any $i \in [K]$ is essential for an instance to be identifiable (indeed, if $\sigma_i^{\min} = \infty$ for some $i \in [K]$, then it is impossible to estimate μ_i), we do not explicitly make this assumption, given that the dependence on $\{\sigma_i^{\min}\}_{i\in[K]}$ appears naturally in our results. Finally, we define $\bar{\sigma} = \max_{i\in[K]} \sigma_i^{\min}$ to be the maximum $\sigma_i^{\min}$ over all arms.

At any round t, we denote by $N_i(t)$ the number of times arm i has been pulled up to and including round $t-1$. Finally, we define $\zeta_i(t) = \sum_{j\in[K]} N_j(t)/\sigma_{j,i}^2$ to be the *weighted number of samples* corresponding to arm $i \in [K]$ at the beginning of round t.

3. Maximum-Likelihood Estimator and Concentration Bounds

By definition of our model, every time an arm $j \in [K]$ is played at some round τ, for each arm $i \in [K]$, the player observes a sample $X_{i,\tau}$ drawn independently from $\mathcal{N}(\mu_i, \sigma_{j,i}^2)$. From the perspective of a fixed arm $i \in [K]$, the player collects, at each round, a noisy estimate of μ_i under various (Gaussian) noise levels. In particular, the noise variance depends on the played arm, which in turn is a function of the trajectory of the algorithm up to that round. The results of the following sections rely on a natural maximum-likelihood (ML) estimator for the mean of samples from heterogeneous sources of identical means and different standard deviations, given by matrix $\mathbf{\Sigma}$. In this section, we define this estimator and prove useful concentration properties. For brevity, for the rest of this section we fix an arm $i \in [K]$ and drop any reference to it.

The ML estimator for the mean of any arm at the beginning of round t (namely, using $t-1$ samples) is defined as follows:

$$\hat{\mu}(t) = \sum_{\tau=1}^{t-1} \frac{X_\tau}{\sigma_{i_\tau}^2} \Bigg/ \sum_{\tau=1}^{t-1} \frac{1}{\sigma_{i_\tau}^2} = \sum_{j\in[K]} \sum_{\tau=1}^{t-1} \frac{X_\tau \cdot \mathbb{I}\{i_\tau = j\}}{\sigma_j^2} \Bigg/ \zeta(t), \qquad (1)$$

where, again, we use the definition of the weighted number of samples $\zeta(t) = \sum_{j\in[K]} \frac{N_j(t)}{\sigma_j^2}$.

Observe that the above estimator is a weighted average of $t-1$ samples coming from K possible different sources (namely, arms played), where each sample is weighted by the inverse variance of its source. Here, the weighted number of samples $\zeta(t)$ reflects the total "quality" of the information collected up to time t, and generalizes the "number of collected samples" – a critical notion in the standard concentration results used in multi-armed bandits.

In the case where the number of samples from each type of noise distribution is fixed, the following guarantee can be trivially proved:

Fact 3.1. *Assuming that the trajectory of the algorithm (namely, the type of sources) up to and including time $t-1$ is fixed and independent of the observed samples – thus, $\zeta(t)$ is deterministic – the following inequality is true for any $\epsilon > 0$ and $i \in [K]$:*

$$\Pr[|\hat{\mu}(t) - \mu| > \epsilon] < 2 \cdot \exp\left(-\frac{\zeta(t)\cdot\epsilon^2}{2}\right).$$

In scenarios of active sampling, however, where the actions depend on the history of observed rewards, the quantity $\zeta(t)$ is also a random variable depending on the trajectory up to time $t-1$ – a fact that invalidates the use of the above standard concentration result.

Remark. A usual approach in such settings is to apply a union bound over all possible numbers of samples, bypassing in that way the dependence between the trajectory and the observed samples. However, given that in our case $\zeta(t)$ can take $\binom{t+K-2}{K-1}$ different values (since the reward of each arm can be sampled in K different ways at each round), this approach would result in an additional $\mathcal{O}(t^K)$-factor with an undesirable exponential dependence on K in the bound of Fact 3.1.

In order to overcome the above issue, we first prove an auxiliary result, which includes concentration bounds for active sampling related to the estimator in Equation (1).

Lemma 3.1. *Let $\{Z_{t'}\}_{t'\in\mathbb{N}}$ be a sequence of random variables. We denote by $\mathcal{F}_{t'}$ the σ-algebra generated by $\{Z_\tau\}_{\tau\le t'}$ and by $\mathcal{F} = (\mathcal{F}_{t'})_{t'\in\mathbb{N}}$ the corresponding filtration. Each random variable $Z_{t'}$ is drawn independently from a zero-mean sub-Gaussian distribution with variance proxy $\sigma_{t'}^2$, where $\sigma_{t'}$ is an $\mathcal{F}_{t'-1}$-measurable random variable. We define $W_{t'} = \sum_{\tau=1}^{t'} \frac{Z_\tau}{\sigma_\tau^2}$ and $\zeta_{t'} = \sum_{\tau=1}^{t'} \frac{1}{\sigma_\tau^2}$.*

(a) *Let ϕ be an $\mathcal{F}$-stopping time which satisfies either $\zeta_\phi \in I$ for some interval $I = [L, H]$ with $H > L > 0$, or $\phi = t+1$. Then, we have that*

$$\Pr\left[\, |W_\phi| > \sqrt{2\alpha\,\zeta_\phi \log t} \text{ and } \phi \le t \right] \le 2 \cdot t^{-\alpha L/H}.$$

(b) *Let ψ be an $\mathcal{F}$-stopping time which satisfies either $\zeta_\psi \geq r$ for some $r \in \mathbb{R}_{\geq 0}$, or $\psi = t+1$. Then, for any $\epsilon > 0$, we have that*

$$\Pr\left[|W_\psi| > \zeta_\psi\, \epsilon \text{ and } \psi \leq t\right] \leq 2 \cdot \exp\left(-\frac{r \cdot \epsilon^2}{2}\right).$$

Proof sketch. As an auxiliary result, we first prove that the sequence $(\tilde{G}_{t'})_{t' \in \mathbb{N}}$, where $\tilde{G}_{t'} = \exp\left(\lambda W_{t'} - \frac{\lambda^2 \zeta_{t'}}{2}\right) \cdot \mathbb{I}\left(t' \leq t\right)$ for some $\lambda \in \mathbb{R}$, is a super-martingale which satisfies $\mathbb{E}\left[\tilde{G}_{t'}\right] \leq 1$ for any $t' \in \mathbb{N}$. For proving part (a), we focus on bounding the probability that $W_\phi > \sqrt{2\alpha\, \zeta_\phi \log t}$ and $\phi \leq t$, since the other tail bound follows symmetrically. By denoting $G_{t'} = \exp\left(\lambda(W_{t'} - \sqrt{2\alpha\, \zeta_{t'} \log t})\right) \cdot \mathbb{I}\left(t' \leq t\right)$ for some $\lambda > 0$, and using Markov's inequality, we get that

$$\Pr\left[W_\phi > \sqrt{2\alpha\, \zeta_\phi \log t} \text{ and } \phi \leq t\right] \leq \mathbb{E}\left[G_\phi\right].$$

In order to upper-bound $\mathbb{E}\left[G_\phi\right]$, by setting $\tilde{G}_{t'} = \exp\left(\lambda W_{t'} - \frac{\lambda^2 \zeta_{t'}}{2}\right) \cdot \mathbb{I}\left(t' \leq t\right)$, we first rewrite

$$G_{t'} = \tilde{G}_{t'} \cdot \exp\left(\frac{\lambda^2 \zeta_{t'}}{2} - \lambda\sqrt{2\alpha\, \zeta_{t'} \log t}\right).$$

Note that, for $t' = \phi$, the event that $\phi \leq t$ implies that $\zeta_\phi \in I$ and, thus, $L \leq \zeta_\phi \leq H$. Therefore, by setting $\lambda = \frac{1}{H}\sqrt{2\alpha L \log t}$, G_ϕ can be upper-bounded as

$$G_\phi \leq \tilde{G}_\phi \cdot \exp\left(-\frac{\alpha \cdot L}{H} \log t\right).$$

The proof follows by using the fact that $\mathbb{E}\left[\tilde{G}_\phi\right] \leq 1$ for any $\lambda > 0$.

The proof of (b) follows by similar arguments. □

In the next Lemma, we provide a concentration result for our estimator in the case where the weighted number of samples randomly depends on the trajectory of observed realizations. The result holds in the case where we have at least one sample from the source with minimum noise.

Lemma 3.2. *Let $\alpha > 0$. For any $t \geq 2$, for the estimator defined in Equation (1), if $\zeta(t) \geq \frac{1}{(\sigma^{\min})^2}$, where $\sigma^{\min} = \min_{j \in [K]} \sigma_j$, then we have that*

$$\Pr\left[|\,\hat{\mu}(t) - \mu| > \sqrt{\frac{2\alpha \log t}{\zeta(t)}}\right] < 2 \cdot \lceil \log_2(t-1) \rceil \cdot t^{-\alpha/2}.$$

Note that imposing conditions on $\zeta(t)$ or the source noise is necessary in order to obtain any concentration results for the estimator in Equation (1); for instance, we cannot hope for concentration results in the case where all t samples come from a source with $\sigma = \infty$ noise. The proof of the above Lemma is based on an exponentially-spaced discretization of the possible range of values of $\zeta(t)$, combined with Lemma 3.1.

We remark that another possible approach for obtaining concentration guarantees for our estimator is through the idea of the self-normalizing bound (Abbasi-yadkori et al., 2011) combined with the assumptions of Lemma 3.2. However, this approach comes at an expense of a slightly worse dependence on t.

4. Asymptotic Regret Lower Bound

In this section, we provide an asymptotic regret lower bound for policies that are consistent with respect to any environment $(\mu, \mathbf{\Sigma})$. We recall the definition of *consistent* policies:

Definition 4.1. A policy is called *consistent* if, for any environment $(\mu, \mathbf{\Sigma})$, its regret satisfies

$$\lim_{T \to \infty} \frac{R_T(\mu)}{T^p} = 0 \quad \text{for any } p > 0.$$

In this section, we show that, in the asymptotic regime, the problem of collecting sufficient information to distinguish the suboptimality gaps of the arms, while paying the minimum cost in terms of regret accumulated during suboptimal plays, can be closely approximated by an LP. For any reward vector μ, we define the following parameterized set of constraints:

$$C(\mu) = \left\{ c \in \mathbb{R}^K_{\geq 0} : \begin{array}{l} \sum_{j \in [K]} \frac{c_j}{\sigma^2_{j,i}} \geq \frac{2}{\Delta_i^2(\mu)}, \forall i \neq i^*(\mu) \\ \sum_{j \in [K]} \frac{c_j}{\sigma^2_{j,i}} \geq \frac{2}{\Delta^2_{\min}(\mu)}, i = i^*(\mu) \end{array} \right\}. \tag{2}$$

To provide some intuition on the above constraint set, we recall the case of standard bandit feedback, where $\sigma_{i,i} < \infty$ and $\sigma_{i,j} = \infty$ for any $i \neq j$. There, for any suboptimal arm i the corresponding constraint of $C(\mu)$ becomes $c_i \geq \frac{2\,\sigma^2_{i,i}}{\Delta_i^2(\mu)}$. This matches the multiplicative factor in the existing lower bounds for the standard bandit feedback case, where each arm must be played at least $\frac{2\,\sigma^2_{i,i}}{\Delta_i^2(\mu)} \log(T)$ times. In our general feedback setting, for any arm i, in addition to the information collected by playing the arm itself, the constraints also take into account the information collected by any other arm $j \in [K] \setminus \{i\}$, weighted by its inverse variance.

A natural lower bound on the *number* of suboptimal plays of each arm can be constructed using the minimum-cost feasible vector in $C(\mu)$ with respect to the suboptimality

gaps, formally defined as

$$c^*(\mu) = \operatorname*{argmin}_{c \in C(\mu)} \sum_{i \in [K]} c_i \, \Delta_i(\mu). \tag{3}$$

As we show in the following theorem, the optimal solution of the LP in Equation (3) asymptotically characterizes a lower bound on the number of plays of each suboptimal arm relative to $\log(T)$.

Theorem 4.1. *For any environment $(\mu, \mathbf{\Sigma})$, the regret of any consistent policy within T rounds satisfies*

$$\liminf_{T \to \infty} \frac{R_T(\mu)}{\log T} \geq \sum_{i \in [K]} c_i^*(\mu) \, \Delta_i(\mu).$$

In order to prove Theorem 4.1, first, we consider a reward vector μ and identify a suboptimal arm k in μ. For some $\epsilon > 0$ we construct a vector μ' such that

$$\mu'_i = \begin{cases} \mu_i, \text{ if } i \neq k \\ \mu^* + \epsilon, \text{ if } i = k. \end{cases} \tag{4}$$

Then, in the following proposition, we decompose the KL-divergence of two distributions $\mathbb{P}, \mathbb{P}'$, each capturing the interplay of a policy with environments $(\mu, \mathbf{\Sigma})$ and $(\mu', \mathbf{\Sigma})$, respectively.

Proposition 4.1. *Let two Gaussian K-armed bandit instances ν and ν' with the same side information matrix $\mathbf{\Sigma}$ and mean reward vectors μ and μ', respectively. Let $\mathbb{P}$ (resp. $\mathbb{P}'$) be the distribution associated with the interplay of ν (resp. ν') and a policy π within t rounds. If μ and μ' differ only in the reward of arm k, then the KL-divergence of $\mathbb{P}$ with respect to $\mathbb{P}'$ satisfies:*

$$D(\mathbb{P}||\mathbb{P}') = \sum_{i \in [K]} \mathop{\mathbb{E}}_{\nu} [N_i(t)] \, \frac{(\mu_k - \mu'_k)^2}{2 \, \sigma^2_{i,k}}. \tag{5}$$

Theorem 4.1 follows by Proposition 4.1, Definition 4.1 and an application of Bretagnolle-Huber inequality for distributions with mean rewards μ, μ' as defined in Equation (4).

Before we proceed to the presentation of our algorithm, we comment that any given solution $c^*(\mu)$ of the LP in Equation (3) could be turned into an optimal Explore-Then-Commit (ETC) strategy for our problem: by playing each arm $\lceil c_i^*(\mu) \cdot \log T \rceil$ times, thus incurring regret at most $\sum_{i \in [K]} c_i^*(\mu) \Delta_i(\mu) \log T + \sum_{i \in [K]} \Delta_i(\mu)$, by Lemma 3.2 we have collected enough information to distinguish the optimal arm with high probability. Hence, $c^*(\mu)$ readily describes an optimal (up to rounding errors) arm-pulling schedule with respect to the lower bound.

5. Algorithm and Regret Analysis

In this section we provide the first algorithm for the general setting of Gaussian bandits with arbitrary side information matrix $\mathbf{\Sigma}$ and prove its asymptotic optimality.

5.1. Description of the Algorithm and Main Results

The general idea behind our algorithm is the following: recall that in the ideal scenario where the suboptimality gaps are known a priori, the linear program described in Equation (3) would directly provide an optimal (up to vanishing rounding errors) exploration strategy, namely, a minimum-regret arm-pulling schedule collecting the necessary information to distinguish the optimal arm with high probability. The above intuition, however, cannot be readily transformed into an algorithm, since prior knowledge of the suboptimality gaps would trivialize our problem.

At a high level, our algorithm attempts to estimate the LP (including constraints and objective) described in Equation (3) up to some accuracy, while at the same time implement its solution (i.e., play, for each arm, the indicated number of samples) in an online manner. The choice of the desired accuracy, up to which the LP must be estimated, is a key for the success of our algorithm: the LP needs to be estimated well-enough to allow the learner to take near-optimal decisions, yet without suffering a significant estimation overhead. In principle, our algorithm generalizes and extends that of (Wu et al., 2015) for the simple case of graph-structured feedback, i.e., when each entry of $\mathbf{\Sigma}$ satisfies $\sigma_{i,j} \in \{\sigma, \infty\}$ for some $\sigma < \infty$.

Our algorithm is presented in pseudocode in Algorithm 1. The algorithm starts by playing, for each arm $j \in [K]$, the arm $i_j \in [K]$ that provides the most accurate information about j. For the rest of the rounds, the algorithm either exploits the already collected information, or further explores the environment.

Greedy exploitation. Let $\hat{\mu}(t) \in \mathbb{R}^K$ be the vector of estimated means, where the coefficient $\hat{\mu}_i(t)$ for every $i \in [K]$ is given in Equation (1) (note that in Section 3 we have dropped any reference to i for generality). Recall that $C(\hat{\mu}(t))$ is the constraint set of the estimated LP at time t, constructed using the collected samples. In the case where the collected number of samples for each arm provides (up to proper scaling) a feasible solution to $C(\hat{\mu}(t))$ (see Case (A) of Algorithm 1), the algorithm greedily plays the arm with the maximum empirical mean reward.

Uniform and LP-dictated exploration. If the collected samples are not feasible for $C(\hat{\mu}(t))$ (up to proper scaling), then the algorithm enters an exploration phase where it either attempts to refine the estimate of the LP, or to make progress in exploring arms according to what the currently

estimated LP dictates. We refer to each case as *Uniform* and *LP-dictated* exploration, respectively. Here, the player needs to address the delicate issue that the total exploration cost needs to be close to that of the optimal solution of the LP. This is achieved by performing a relatively small number of uniform exploration rounds, which allow the learner to obtain a sufficiently good estimate of the LP.

The algorithm considers the LP to be sufficiently (uniformly) explored if the weighted number of samples for each arm satisfies a uniform lower bound. In particular, our algorithm compares the minimum weighted number of samples over all arms with a sublinear function of the total sample weight collected during all exploration rounds. Formally, Algorithm 1 inspects the weighted number of samples of each arm, and checks whether $\min_{i\in[K]} \zeta_i(t) < \frac{1}{K}\beta\left(n_e(t)\right)$, where $\beta(x) = \frac{x^\gamma}{2\,\bar{\sigma}^2}$, with $\gamma \in (0,1)$ and $\bar{\sigma} = \max_{i\in[K]} \sigma_i^{\min}$, and $n_e(t)$ is the total number of exploration rounds up to time t. In the case where the arms are not uniformly explored (see Case (B) of Algorithm 1), our algorithm attempts to satisfy this uniform exploration bound, by playing the arm which provides the less noisy information about an arm i that has the minimum $\zeta_i(t)$ (breaking ties arbitrarily).

On the other hand, if the arms are sufficiently uniformly explored (see Case (C) of Algorithm 1), our algorithm computes an optimal solution $c^*(\hat{\mu}(t))$ to the estimated LP, and then plays any arm i such that $N_i(t) < 4\alpha \cdot c_i^*(\hat{\mu}(t)) \log t$, namely, which is considered unexplored (again up to proper scaling) according to the LP solution.

Main Results. Before we state the regret guarantee of Algorithm 1, we introduce some useful definitions. For any vector μ, we consider the family of vectors μ' that are guaranteed to be ϵ-close to μ with respect to the ℓ_∞-norm, that is, $|\mu'_i - \mu_i| \le \epsilon$ for all arms $i \in [K]$. Let $c^*(\mu')$ be the optimal solution of the minimization problem in Equation (3) using parameters μ'. The following quantity appears naturally in our regret guarantees:

Definition 5.1. For any vector μ, the worst-case ϵ-approximate LP solution for arm $j \in [K]$ is defined as

$$c_j^*(\mu,\epsilon) = \sup_{\mu':\|\mu'-\mu\|_\infty \le \epsilon} c_j^*(\mu').$$

Note that, by continuity, taking $\epsilon \to 0$ we get $c_j^*(\mu,\epsilon) \to c_j^*(\mu)$ for every arm $j \in [K]$.

We prove the following regret upper bound for Algorithm 1:

Theorem 5.1. *For any $\alpha > 4$, $\gamma \in (0,1)$, and $\epsilon > 0$, the regret of algorithm Algorithm 1 satisfies*

$$\begin{aligned} R_T(\mu) \le & \left(2K + \frac{8K}{\alpha-4} + 2\right)\Delta_{\max} \\ & + 2\,\Delta_{\max} K \sum_{\tau\in[T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right) \\ & + \Delta_{\max}\left(4\alpha \sum_{j\in[K]} c_j^*(\mu,\epsilon)\log T + K\right)^\gamma \\ & + 4\alpha \sum_{j\in[K]} \Delta_j(\mu) c_j^*(\mu,\epsilon)\log T. \end{aligned}$$

We remark that, in the restricting case of graph side-information structure (i.e. each $\sigma_{i,j} = \sigma$ or ∞), the regret upper bound of Algorithm 1 also matches the asymptotically-optimal regret shown in (Wu et al., 2015). Further, notice that, as $\bar{\sigma} = \max_{i\in[K]} \sigma_i^{\min}$ grows to infinity, the dependence on T in the second term of the above bound becomes linear. This is a natural effect, however, since having a large $\bar{\sigma}$ implies the existence of an arm $i \in [K]$ with

Algorithm 1 Asymptotically-Optimal Algorithm for Gaussian Bandits with Side Observations

Input: K arms, Σ, $\beta(x) = \frac{x^\gamma}{2\,\bar{\sigma}^2}$, $\gamma \in (0,1)$, $\alpha > 4$.
Set $n_e(K+1) \leftarrow 0$.
For each arm $j \in [K]$, play arm $i_j \leftarrow \arg\min_{i\in[K]} \sigma_{i,j}^2$.

for $t = K+1, K+2, \ldots$ **do**
  if $\left(\frac{N_1(t)}{4a\log t}, \ldots, \frac{N_K(t)}{4a\log t}\right) \in C(\hat{\mu}(t))$ **then**
    // Case (A): Greedy exploitation (event $\mathcal{A}_t$ in the analysis).
    Play arm $i_t \leftarrow \arg\max_{i\in[K]} \hat{\mu}_i(t)$.
    Set $n_e(t+1) \leftarrow n_e(t)$.
  else if $\min_{i\in[K]} \zeta_i(t) < \frac{1}{K}\beta\left(n_e(t)\right)$ **then**
    // Case (B): Uniform exploration (event $\mathcal{A}_t{}^c, \mathcal{B}_t$ in the analysis).
    Let $i \leftarrow \arg\min_{k\in[K]} \zeta_k(t)$.
    Play $i_t \leftarrow \arg\min_{k\in[K]} \sigma_{k,i}^2$.
    Set $n_e(t+1) \leftarrow n_e(t) + 1$.
  else
    // Case (C): LP-Dictated exploration (event $\mathcal{A}_t{}^c, \mathcal{B}_t{}^c$ in the analysis).
    Compute
    $c^*(\hat{\mu}(t)) \leftarrow \arg\min_{c\in C(\hat{\mu}(t))} \sum_{i\in[K]} c_i\,\Delta_i(\hat{\mu}(t))$.
    Play $i_t \leftarrow i$ for some $i \in [K]$ such that
      $N_i(t) < 4\alpha \cdot c_i^*(\hat{\mu}(t)) \log t$.
    Set $n_e(t+1) \leftarrow n_e(t) + 1$.
  end if
end for

large $\sigma_i^{\min}$, for which there is no possible way of collecting accurate information.

The following corollary bounds the asymptotic regret of Algorithm 1 in terms of the worst-case ϵ-approximate LP solution.

Corollary 5.1. *As the time horizon tends to infinity, the regret of Algorithm 1 satisfies*

$$\limsup_{T\to\infty} \frac{R_T(\mu)}{\log T} \leq 4\alpha \sum_{j\in[K]} \Delta_j(\mu) c_j^*(\mu).$$

Notice that, although the above bound matches the lower bound of Theorem 4.1 for $\epsilon = 0$, we are not allowed to directly use $\epsilon \to 0$ in the bound of Theorem 5.1, since that would make the term $\sum_{\tau\in[T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right)$ linear in T. However, by choosing ϵ to be a decreasing function of T, e.g. $\epsilon \sim \log(T)^{-\gamma/4}$, the suboptimal terms in Theorem 5.1 are vanishing as $T \to \infty$ and the regret bound in Corollary 5.1 matches the asymptotic regret lower bound (up to constant factors).

5.2. Outline of the Regret Analysis

For the rest of this section, we provide an overview of our regret analysis, while the complete proof can be found in Appendix C.

Clearly, the first K rounds contribute at most a $(K-1)\cdot\Delta_{\max}$-additive loss in the regret. For each round $t \geq K+1$, we consider the following events:

$$\mathcal{A}_\mathrm{t} = \left\{ \left(\frac{N_1(t)}{4a\log t}, \ldots, \frac{N_K(t)}{4a\log t} \right) \in C(\hat{\mu}(t)) \right\},$$
$$\mathcal{B}_\mathrm{t} = \left\{ \min_{i\in[K]} \zeta_i(t) < \frac{1}{K}\beta(n_e(t)) \right\}.$$

Notice that the above events immediately characterize how Algorithm 1 operates at each round. Specifically, for $t \geq K+1$, the algorithm enters in Case A when $\mathcal{A}_\mathrm{t}$ holds, in Case B when ${\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}$, and in Case C, when ${\mathcal{A}_\mathrm{t}}^c, {\mathcal{B}_\mathrm{t}}^c$.

We also define the error event:

$$\mathcal{U}_\mathrm{t} = \left\{ |\,\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{2a\log t}{\zeta_i(t)}},\ \ \forall i \in [K] \right\}.$$

As we show in Lemma 3.2 of Section 3, for any t the event $\mathcal{U}_\mathrm{t}$ holds with high probability. Therefore, we can assume that $\mathcal{U}_\mathrm{t}$ holds in every round, since the probability that $\mathcal{U}_\mathrm{t}$ does not hold at some round converges to a small $\mathcal{O}(\frac{K}{\alpha-4}\cdot\Delta_{\max})$-additive loss in the regret.

Rounds satisfying $\mathcal{U}_\mathrm{t}, \mathcal{A}_\mathrm{t}$. When the event $\mathcal{U}_\mathrm{t}$ holds, if the algorithm enters Case A, it can be shown that it chooses the optimal arm during the greedy selection. In particular, the event $\mathcal{U}_\mathrm{t}, \mathcal{A}_\mathrm{t}$ implies that for any arm i that is suboptimal with respect to $\hat{\mu}(t)$, that is, $i \neq i^*(\hat{\mu}(t))$, we have that

$$\mu_i - \hat{\mu}_i(t) \leq \frac{\Delta_i(\hat{\mu}(t))}{2},$$

while for the optimal arm $i^*(\hat{\mu}(t))$, we get

$$\hat{\mu}_{i^*(\hat{\mu}(t))}(t) - \mu_{i^*(\hat{\mu}(t))} \leq \frac{\Delta_{\min}(\hat{\mu}(t))}{2}.$$

Since each true parameter μ_i is $\frac{\Delta_i(\hat{\mu}(t))}{2}$-close to the corresponding estimate at time t, we can conclude that the arm selected by Algorithm 1 in that case is the optimal. Hence, the rounds such that $\mathcal{U}_\mathrm{t}, \mathcal{A}_\mathrm{t}$ holds do not contribute to the regret.

Rounds satisfying $\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}$. We now turn our focus to the rounds where ${\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}$ holds, which implies that Algorithm 1 enters Case B.

Through a counting argument we show that the above can happen at most $\bar{\sigma}^2\,\beta(n_e)+1$ times, where n_e is the total number of exploration rounds, i.e. $n_e = \sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c\right)$. Specifically, we show the following:

Lemma 5.1. *The following inequality holds:*

$$\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right) \leq \frac{1}{2}\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c\right)\right)^\gamma + 1.$$

Notice that the fact that our algorithm plays the most informative arm (i.e., the one with smaller noise) in order to collect a sample for arm $i = \arg\min_{k\in[K]} \zeta_k(t)$ is crucial for the above counting argument to hold. Thus, using Lemma 5.1, we have that

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right) &\leq \sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right) \\ &\leq \frac{\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c\right)\right)^\gamma}{2} + 1. \quad (6)\end{aligned}$$

We now bound the regret accumulated during all exploration rounds (that is, including Cases (2) and (3)). We have that

$$\mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c\right) \leq \mathbb{I}\left({\mathcal{U}_\mathrm{t}}^c\right) + \mathbb{I}\left(\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, {\mathcal{B}_\mathrm{t}}^c\right) + \mathbb{I}\left(\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right). \quad (7)$$

By combining Equation (6) with Equation (7), we can see that in order to upper bound $\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right)$, it suffices to provide a bound on $\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, {\mathcal{B}_\mathrm{t}}^c\right)$. By doing this we are able to conclude that, overall, the number of rounds such that $\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}$ holds is at most order

$$K \sum_{\tau\in[T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right) + \left(\alpha \sum_{j\in[K]} c_j^*(\mu,\epsilon)\log T + K\right)^\gamma.$$

Rounds satisfying $\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c$. It remains to describe how we bound the regret accumulated in rounds where $\{\mathcal{A}_t{}^c, \mathcal{B}_t{}^c\}$ holds, and thus the algorithm enters Case C.

We define the following event which states that, at time t, the error in the mean estimates upper bounded by ϵ:

$$\mathcal{E}_t = \{|\,\hat{\mu}_i(t) - \mu_i| \leq \epsilon,\ \ \forall i \in [K]\}\,. \tag{8}$$

In order to bound the error of case $\{\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c\}$, we further distinguish two cases according to whether $\mathcal{E}_t$ holds:

Recall that, when $\{\mathcal{A}_t{}^c, \mathcal{B}_t{}^c\}$ holds, the algorithm selects an arm j such that $\frac{N_j(t)}{4a\log t} < c_j^*(\hat{\mu}(t))$. For rounds where $\{\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c, \mathcal{E}_t\}$ holds, by Definition 5.1 in combination with the definition of $\mathcal{E}_t$, we prove the following result:

Lemma 5.2. *For every arm $j \in [K]$, it holds that*

$$\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c, \mathcal{E}_t, i_t = j\right) \leq 4\alpha \cdot c_j^*(\mu, \epsilon)\log T.$$

The above results immediately implies that the contribution to the regret of rounds $t \geq K+1$ such that $\{\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c, \mathcal{E}_t\}$ is at most $4\alpha \sum_{j\in[K]} \Delta_j(\mu) c_j^*(\mu,\epsilon)\log T$.

Finally, for the rounds such that $\{\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c, \mathcal{E}_t{}^c\}$ holds, i.e., when the error in some of the estimates is greater than ϵ, by applying a union bound over all arms, we get

$$\begin{aligned}&\mathbb{I}\left(\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c, \mathcal{E}_t{}^c\right)\\ &\leq \sum_{i\in[K]} \mathbb{I}\left(\mathcal{U}_t, \mathcal{A}_t{}^c, \min_{j\in[K]} \zeta_j(t) \geq \frac{\beta(n_e(t))}{K}, |\,\hat{\mu}_i(t) - \mu_i| > \epsilon\right).\end{aligned}$$

Given that $\min_{j\in[K]} \zeta_j(t) \geq \frac{\beta(n_e(t))}{K}$ implies that $\zeta_i(t) \geq \frac{\beta(n_e(t))}{K}$ for any arm $i \in [K]$, using part (c) of Lemma 3.1, we can upper bound the probability of each term in the above summation by $\exp\left(-\frac{\epsilon^2}{2}\frac{1}{K}\beta(\tau)\right)$. Thus, the total contribution to the regret of the rounds such that $\{\mathcal{U}_t, \mathcal{A}_t{}^c, \mathcal{B}_t{}^c, \mathcal{E}_t{}^c\}$ holds can be upper bounded by $\Delta_{\max} K \sum_{\tau\in[T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right)$.

The regret guarantee presented in Theorem 5.1 follows by combining the above losses.

Conclusion and Further Directions

In this work, we revisit the general feedback model introduced by Wu, Szepesvári, and György in (Wu et al., 2015) and provide the first algorithm for the case of arbitrary feedback matrices. To the best of our knowledge, this is one of the most general feedback models that appears in the literature, modulo the Gaussian noise assumption.

A number of questions still remain open in the area of online learning under rich feedback structures. For instance, it would be interesting to explore the existence of algorithms that achieve (minimax) optimal regret in the finite time horizon regime. Another direction would be to examine different noise models, other than Gaussian, or even general sub-Gaussian noise with known variance proxies. We believe that our work could serve as a building block in the above direction.

Acknowledgements

This research is supported in part by NSF Grants 2019844, 2107037 and 2112471, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. From bandits to experts: A tale of domination and independence. *Advances in Neural Information Processing Systems*, 07 2013.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In *COLT*, 2015.

Arora, R., Marinov, T. V., and Mohri, M. Bandits with feedback graphs and switching costs. In *NeurIPS*, 2019.

Bartók, G., Foster, D., Pál, D., Rakhlin, A., and Szepesvári, C. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39: 967–997, 11 2014. doi: 10.1287/moor.2014.0663.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5:1–122, 2012.

Buccapatnam, S., Eryilmaz, A., and Shroff, N. Stochastic bandits with side observations on networks. *ACM SIGMETRICS Performance Evaluation Review*, 42, 06 2014. doi: 10.1145/2591971.2591989.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. Leveraging side observations in stochastic bandits. In *UAI*, 2012.

Cohen, A., Hazan, T., and Koren, T. Online learning with feedback graphs without the graphs. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of

Proceedings of Machine Learning Research, pp. 811–819, New York, New York, USA, 20–22 Jun 2016. PMLR.

Cortes, C., DeSalvo, G., Gentile, C., Mohri, M., and Yang, S. Online learning with sleeping experts and feedback graphs. In *ICML*, 2019.

Cortes, C., DeSalvo, G., Gentile, C., Mohri, M., and Zhang, N. Online learning with dependent stochastic feedback graphs. In *ICML*, 2020.

Hanawal, M. K., Szepesvári, C., and Saligrama, V. Sequential learning without feedback. *CoRR*, abs/1610.05394, 2016.

Kocák, T., Neu, G., and Valko, M. Online learning with noisy side observations. In *AISTATS*, 2016.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Li, S., Chen, W., Wen, Z., and Leung, K.-S. Stochastic online learning with probabilistic graph feedback. *ArXiv*, abs/1903.01083, 2020.

Lin, T., Abrahao, B., Kleinberg, R., Lui, J., and Chen, W. Combinatorial partial monitoring game with linear feedback and its applications. volume 3, 06 2014. doi: 10.13140/RG.2.1.1502.1521.

Liu, F., Buccapatnam, S., and Shroff, N. B. Information directed sampling for stochastic bandits with graph feedback. In *AAAI*, 2018a.

Liu, F., Zheng, Z., and Shroff, N. B. Analysis of thompson sampling for graphical bandits without the graphs. *ArXiv*, abs/1805.08930, 2018b.

Lykouris, T., Tardos, É., and Wali, D. Graph regret bounds for thompson sampling and ucb. In *ALT*, 2020.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 684–692, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Rangi, A. and Franceschetti, M. Online learning with feedback graphs and switching costs. *AISTATS*, 2019.

Resler, A. and Mansour, Y. Adversarial online learning with noise. In *ICML*, 2019.

Singh, R., Liu, F., Liu, X., and Shroff, N. Contextual bandits with side-observations, 06 2020.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Wang, L., Li, B., Zhou, H., Giannakis, G. B., Varshney, L. R., and Zhao, Z. Adversarial linear contextual bandits with graph-structured side observations. *CoRR*, abs/2012.05756, 2020.

Wu, Y., György, A., and Szepesvári, C. Online learning with gaussian payoffs and side observations. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1360–1368, Cambridge, MA, USA, 2015. MIT Press.

Yun, D., Ahn, S., Proutiere, A., Shin, J., and Yi, Y. Multi-armed bandit with additional observations. *ACM SIGMETRICS Performance Evaluation Review*, 46:53–55, 06 2018. doi: 10.1145/3292040.3219639.

A. ML Estimator and Concentration Bounds: Omitted Proofs

A.1. Proof of Lemma 3.1

Lemma 3.1. *Let $\{Z_{t'}\}_{t'\in\mathbb{N}}$ be a sequence of random variables. We denote by $\mathcal{F}_{t'}$ the σ-algebra generated by $\{Z_\tau\}_{\tau\le t'}$ and by $\mathcal{F} = (\mathcal{F}_{t'})_{t'\in\mathbb{N}}$ the corresponding filtration. Each random variable $Z_{t'}$ is drawn independently from a zero-mean sub-Gaussian distribution with variance proxy $\sigma^2_{t'}$, where $\sigma_{t'}$ is an $\mathcal{F}_{t'-1}$-measurable random variable. We define $W_{t'} = \sum_{\tau=1}^{t'} \frac{Z_\tau}{\sigma^2_\tau}$ and $\zeta_{t'} = \sum_{\tau=1}^{t'} \frac{1}{\sigma^2_\tau}$.*

- ***(a)*** *Let ϕ be an $\mathcal{F}$-stopping time which satisfies either $\zeta_\phi \in I$ for some interval $I = [L, H]$ with $H > L > 0$, or $\phi = t+1$. Then, we have that*

$$\Pr\left[\,|W_\phi| > \sqrt{2\alpha\,\zeta_\phi \log t} \text{ and } \phi \le t\right] \le 2\cdot t^{-\alpha L/H}.$$

- ***(b)*** *Let ψ be an $\mathcal{F}$-stopping time which satisfies either $\zeta_\psi \ge r$ for some $r \in \mathbb{R}_{\ge 0}$, or $\psi = t+1$. Then, for any $\epsilon > 0$, we have that*

$$\Pr\left[|W_\psi| > \zeta_\psi\,\epsilon \text{ and } \psi \le t\right] \le 2\cdot\exp\left(-\frac{r\cdot\epsilon^2}{2}\right).$$

Proof. Before we proceed to the main part of the proof, we first show the following auxiliary result:

Proposition A.1. *Let $\tilde{G}_{t'} = \exp\left(\lambda W_{t'} - \frac{\lambda^2\,\zeta_{t'}}{2}\right)\cdot\mathbb{I}\,(t'\le t)$ for some $\lambda\in\mathbb{R}$. Then, $(\tilde{G}_{t'})_{t'\in\mathbb{N}}$ is a super-martingale satisfying $\mathbb{E}\left[\tilde{G}_{t'}\right] \le 1$ for any $t'\in\mathbb{N}$.*

Proof. In order to show that the sequence $(\tilde{G}_{t'})_{t'\in\mathbb{N}}$ is a super-martingale satisfying $\mathbb{E}\left[\tilde{G}_{t'}\right] \le 1$ for any $t'\in\mathbb{N}$, we proceed by induction. For the base case where $t' = t+1$, we have $\mathbb{E}\left[\tilde{G}_{t+1}\mid\mathcal{F}_t\right] = 0 \le \tilde{G}_t$. Let us now fix any time $t'\le t$. Given that $\sigma_{t'}$ is $\mathcal{F}_{t'-1}$-measurable and $Z_{t'}$ is sub-Gaussian with variance proxy $\sigma^2_{t'}$, for any $\lambda'\in\mathbb{R}$ we have that $\mathbb{E}\left[\exp\left(\lambda' Z_{t'} - \frac{\lambda'^2\sigma^2_{t'}}{2}\right)|\mathcal{F}_{t'-1}\right] \le 1$. Thus, by setting $\lambda' = \frac{\lambda}{\sigma^2_{t'}}$, we get that $\mathbb{E}\left[\exp\left(\lambda\frac{Z_{t'}}{\sigma^2_{t'}} - \frac{\lambda^2}{2\sigma^2_{t'}}\right)|\mathcal{F}_{t'-1}\right] \le 1$. Using that, we have

$$\begin{aligned}
\mathbb{E}\left[\tilde{G}_{t'}\mid\mathcal{F}_{t'-1}\right] &= \mathbb{E}\left[\exp\left(\lambda W_{t'} - \frac{\lambda^2\,\zeta_{t'}}{2}\right)|\mathcal{F}_{t'-1}\right]\\
&= \mathbb{E}\left[\exp\left(\sum_{\tau=1}^{t'}\left(\lambda\frac{Z_\tau}{\sigma^2_\tau} - \frac{\lambda^2}{2\sigma^2_\tau}\right)\right)|\mathcal{F}_{t'-1}\right]\\
&= \tilde{G}_{t'-1}\cdot\mathbb{E}\left[\exp\left(\lambda\frac{Z_{t'}}{\sigma^2_{t'}} - \frac{\lambda^2}{2\sigma^2_{t'}}\right)|\mathcal{F}_{t'-1}\right] \le \tilde{G}_{t'-1},
\end{aligned}$$

thus proving that $(\tilde{G}_{t'})_{t'}$ is a super-martingale. In addition, since ϕ satisfies $\phi\le t+1$ almost surely, by Doob's optional stopping theorem, we can conclude that $\mathbb{E}\left[\tilde{G}_{t'}\right]\le 1$ for all $t'\in\mathbb{N}$. □

(a) It suffices to upper-bound the probability that $W_\phi > \sqrt{2\alpha\,\zeta_\phi\log t}$ and $\phi\le t$ by $t^{-\alpha\frac{L}{H}}$. Notice that the other tail bound follows by symmetry and the factor of 2 in the desired bound results from a union bound over the two tails.

By denoting $G_{t'} = \exp\left(\lambda(W_{t'} - \sqrt{2\alpha\,\zeta_{t'}\log t})\right)\cdot\mathbb{I}\,(t'\le t)$ for some $\lambda > 0$, we get

$$\Pr\left[W_\phi > \sqrt{2\alpha\,\zeta_\phi\log t} \text{ and } \phi\le t\right] = \Pr\left[G_\phi \ge 1\right] \le \mathbb{E}\left[G_\phi\right],$$

where the last step above follows by Markov's inequality, given that G_ϕ is by construction a non-negative random variable. Thus, in order to complete the proof, it suffices to upper-bound $\mathbb{E}\left[G_\phi\right]$.

By setting $\tilde{G}_{t'} = \exp\left(\lambda W_{t'} - \frac{\lambda^2 \zeta_{t'}}{2}\right) \cdot \mathbb{I}(t' \leq t)$, we can rewrite

$$G_{t'} = \tilde{G}_{t'} \cdot \exp\left(\frac{\lambda^2 \zeta_{t'}}{2} - \lambda\sqrt{2\alpha\, \zeta_{t'} \log t}\right).$$

Now, for $t' = \phi$, the event that $\phi \leq t$ implies by definition that $\zeta_\phi \in I$ and, thus, $L \leq \zeta_\phi \leq H$. Therefore, by setting $\lambda = \frac{1}{H}\sqrt{2\alpha L \log t}$, we get that

$$G_\phi = \tilde{G}_\phi \cdot \exp\left(\frac{\lambda^2 \zeta_\phi}{2} - \lambda\sqrt{2\alpha\, \zeta_\phi \log t}\right) \leq \tilde{G}_\phi \cdot \exp\left(\frac{\lambda^2 H}{2} - \lambda\sqrt{2\alpha L \log t}\right) = \tilde{G}_\phi \cdot \exp\left(-\frac{\alpha \cdot L}{H} \log t\right).$$

Finally, by using the fact that $\mathbb{E}\left[\tilde{G}_\phi\right] \leq 1$ for any $\lambda > 0$, as proved in Proposition A.1, we have that $\mathbb{E}[G_\phi] \leq \exp\left(-\alpha \frac{L}{H} \log t\right) = t^{-\alpha \frac{L}{H}}$, which concludes the proof.

(b) As in the proof of part (a), we can focus on bounding the probability that $W_\psi > \zeta_\psi\, \epsilon$ and $\psi \leq t$, since the other tail bound follows by symmetry. By denoting $G'_{t'} = \exp\left(\lambda(W_{t'} - \zeta_{t'}\, \epsilon)\right) \cdot \mathbb{I}(t' \leq t)$ for some $\lambda > 0$, we get

$$\Pr\left[W_\psi > \zeta_\psi \text{ and } \psi \leq t\right] = \Pr\left[G'_\psi \geq 1\right] \leq \mathbb{E}\left[G'_\psi\right],$$

where the last step above follows by Markov's inequality.

In order to upper-bound $\mathbb{E}\left[G'_\phi\right]$, by setting $\tilde{G}_{t'} = \exp\left(\lambda W_{t'} - \frac{\lambda^2 \zeta_{t'}}{2}\right) \cdot \mathbb{I}(t' \leq t)$, we can rewrite

$$G'_{t'} = \tilde{G}_{t'} \cdot \exp\left(\frac{\lambda^2 \zeta_{t'}}{2} - \lambda\epsilon\, \zeta_{t'}\right).$$

By setting $\lambda = \epsilon$, we get that $G'_{t'} \leq \tilde{G}_{t'} \exp\left(-\frac{\zeta_{t'} \cdot \epsilon^2}{2}\right)$. Further, if $\psi \leq t$, then by definition we have that $\zeta_\psi \geq r$, which implies that

$$G'_\psi \leq \tilde{G}_\psi \cdot \exp\left(-\frac{r \cdot \epsilon^2}{2}\right).$$

The proof follows by taking expectation in the above expression and using the fact that $\mathbb{E}\left[\tilde{G}_\psi\right] \leq 1$ for any $\lambda \in \mathbb{R}$, as we show in Proposition A.1. □

A.2. Proof of Lemma 3.2

Lemma 3.2. *Let $\alpha > 0$. For any $t \geq 2$, for the estimator defined in Equation (1), if $\zeta(t) \geq \frac{1}{(\sigma^{\min})^2}$, where $\sigma^{\min} = \min_{j \in [K]} \sigma_j$, then we have that*

$$\Pr\left[|\,\hat{\mu}(t) - \mu| > \sqrt{\frac{2\alpha \log t}{\zeta(t)}}\right] < 2 \cdot \lceil \log_2(t-1) \rceil \cdot t^{-\alpha/2}.$$

Proof. Recall that $\zeta(t)$ is defined as $\zeta(t) = \sum_{j \in [K]} N_j(t)/\sigma_j^2$ and that $\sigma^{\min} = \min_{j \in [K]} \sigma_j$. Without loss of generality, we can assume that $\sigma^{\min} > 0$, since, otherwise, by collecting any sample with $\sigma_j = 0$, the concentration inequality follows trivially. Recalling that $\zeta(t)$ involves $t-1$ samples, by definition, by time $t \geq 2$, it is easy to see that

$$\zeta(t) \in \left[\frac{1}{(\sigma^{\min})^2}, \frac{t-1}{(\sigma^{\min})^2}\right] \leq \bigcup_{r \in [\lceil \log_2(t-1) \rceil]} \left[\frac{2^{r-1}}{(\sigma^{\min})^2}, \frac{2^r}{(\sigma^{\min})^2}\right]$$

and, hence, by using the above decomposition and applying union bound we get that

$$\Pr\left[|\hat{\mu}(t)-\mu|>\sqrt{\frac{2a\log t}{\zeta(t)}}\right]=\Pr\left[\bigcup_{r\in[\lceil\log_2(t-1)\rceil]}\left(|\,\hat{\mu}(t)-\mu|>\sqrt{\frac{2a\log t}{\zeta(t)}}\text{ and }\zeta(t)\in\left[\frac{2^{r-1}}{(\sigma^{\min})^2},\frac{2^r}{(\sigma^{\min})^2}\right]\right)\right]$$
$$\leq\sum_{r\in[\lceil\log_2(t-1)\rceil]}\Pr\left[|\,\hat{\mu}(t)-\mu|>\sqrt{\frac{2a\log t}{\zeta(t)}}\text{ and }\zeta(t)\in\left[\frac{2^{r-1}}{(\sigma^{\min})^2},\frac{2^r}{(\sigma^{\min})^2}\right]\right]$$

Notice that, by multiplying both sides with $\zeta(t)$, the inequality $|\hat{\mu}(t)-\mu|>\sqrt{\frac{2a\log t}{\zeta(t)}}$ is equivalent to $\left|\sum_{\tau=1}^{t-1}\frac{X_\tau-\mu}{\sigma^2_{i_\tau}}\right|>\sqrt{2a\,\zeta(t)\log t}$. By setting $Z_{t'}=X_{t'}-\mu$, we can observe that the sequence $\{Z_{t'}\}_{t'\leq t}$ satisfies the conditions of Lemma 3.1. For any fixed $r\in[\lceil\log_2(t-1)\rceil]$ let us define $I(r)=\left[\frac{2^{r-1}}{(\sigma^{\min})^2},\frac{2^r}{(\sigma^{\min})^2}\right]$. Hence, for $L=\frac{2^{r-1}}{(\sigma^{\min})^2}$ and $H=\frac{2^r}{(\sigma^{\min})^2}$, then the event that $\left|\sum_{\tau=1}^{t-1}\frac{X_\tau-\mu}{\sigma^2_{i_\tau}}\right|>\sqrt{2a\,\zeta(t)\log t}$ and $\zeta(t)\in I(r)$ implies that there must exist a stopping-time ϕ where $\zeta_\phi\in I(r)$ as described in Lemma 3.1 for $I=I(r)$, such that $|W_\phi|>\sqrt{2\alpha\,\zeta_\phi\log\phi}$ and $\phi\leq t-1$. Thus, by the result of Lemma 3.1 we get

$$\Pr\left[|\hat{\mu}(t)-\mu|>\sqrt{\frac{2a\log t}{\zeta(t)}}\text{ and }\zeta(t)\in I(r)\right]=\Pr\left[\left|\sum_{\tau=1}^{t-1}\frac{X_\tau-\mu}{\sigma^2_{i_\tau}}\right|>\sqrt{2a\,\zeta(t)\log t}\text{ and }\zeta(t)\in I(r)\right].$$
$$\leq\Pr\left[|W_\phi|>\sqrt{2\alpha\,\zeta_\phi\log\phi}\text{ and }\phi\leq t-1\right]$$
$$<2\cdot t^{-\alpha\frac{2^{r-1}}{2^r}}=2\cdot t^{-\frac{\alpha}{2}}.$$

Therefore, we can conclude that:

$$\Pr\left[|\,\hat{\mu}(t)-\mu|>\sqrt{\frac{2a\log t}{\zeta(t)}}\right]<2\cdot\lceil\log_2(t-1)\rceil\cdot t^{-\alpha/2}.$$

□

B. Asymptotic Regret Lower Bound: Omitted Proofs

B.1. Proof of Proposition 4.1

Proposition 4.1. *Let two Gaussian K-armed bandit instances ν and ν' with the same side information matrix $\mathbf{\Sigma}$ and mean reward vectors μ and μ', respectively. Let $\mathbb{P}$ (resp. $\mathbb{P}'$) be the distribution associated with the interplay of ν (resp. ν') and a policy π within t rounds. If μ and μ' differ only in the reward of arm k, then the KL-divergence of $\mathbb{P}$ with respect to $\mathbb{P}'$ satisfies:*

$$D(\mathbb{P}||\mathbb{P}')=\sum_{i\in[K]}\mathop{\mathbb{E}}_{\nu}[N_i(t)]\,\frac{(\mu_k-\mu'_k)^2}{2\,\sigma^2_{i,k}}.\tag{5}$$

Proof. Recall that we denote by i_τ the arm played at round τ by policy π. Let P_i (resp. P'_i) be the distribution over the random vector of the observations obtained at any round where arm i is played under the instance ν (resp. ν'). By working along the lines of Lemma 15.1 in (Lattimore & Szepesvári, 2020), it can be proved that the KL-divergence between distributions $\mathbb{P}$ and $\mathbb{P}'$ satisfies

$$D(\mathbb{P}||\mathbb{P}')=\sum_{\tau=1}^{t}\mathop{\mathbb{E}}_{\nu}\left[D(P_{i_\tau}||P'_{i_\tau})\right]=\sum_{\tau=1}^{t}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{\nu}\left[D(P_i||P'_i)\,\mathbb{I}\left(i_\tau=i\right)\right]=\sum_{i=1}^{K}D(P_i||P'_i)\mathop{\mathbb{E}}_{\nu}\left[N_i(t)\right].$$

In contrast to Lemma 15.1 in (Lattimore & Szepesvári, 2020), here P_k, P'_k are multivariate distributions of K arms. We denote by $P_{i,j}$ the marginal distribution that corresponds to the reward observation of arm j, when arm i is played. Given

that the noisy reward observations are independent for each arm $j \in [K]$, by using the additivity of the KL-divergence in that case, we have that $D(P_i||P_i') = \sum_{j=1}^{K} D(P_{i,j}, P_{i',j})$, which in turn leads to

$$D(\mathbb{P}||\mathbb{P}') = \sum_{i=1}^{K}\sum_{j=1}^{K} D(P_{i,j}, P_{i',j}) \mathop{\mathbb{E}}_{\nu}[N_i(t)] = \sum_{i=1}^{K} \mathop{\mathbb{E}}_{\nu}[N_i(t)] \frac{(\mu_k - \mu_k')^2}{2\,\sigma_{i,k}^2},$$

where the last equation follows by the fact that for any $i \in [K]$, the distributions $P_{i,k}$ and $P_{i',k}$ are both Gaussian with means μ_k and $\mu_{k'}$, respectively, and the same variance $\sigma_{i,j}^2$. □

B.2. Proof of Theorem 4.1

Theorem 4.1. *For any environment $(\mu, \mathbf{\Sigma})$, the regret of any consistent policy within T rounds satisfies*

$$\liminf_{T\to\infty} \frac{R_T(\mu)}{\log T} \geq \sum_{i\in[K]} c_i^*(\mu)\,\Delta_i(\mu).$$

Proof. Let us fix any consistent policy π. Let μ be a mean reward vector and let k be a suboptimal arm in μ. We define vector μ' such that

$$\mu_i' = \begin{cases} \mu_i, \text{ if } i \neq k \\ \mu^* + \epsilon, \text{ if } i = k \end{cases} \tag{9}$$

for some $\epsilon > 0$. Moreover, let $\mathbb{P}$ and $\mathbb{P}'$ be the measures induced by the interplay of π with the environments $(\mu, \mathbf{\Sigma})$ and $(\mu', \mathbf{\Sigma})$ respectively, and let us denote by $\mathbb{E}$ and $\mathbb{E}'$ the corresponding expectations. For any event $\mathcal{E}$, due to the Bretagnolle-Huber inequality (Tsybakov, 2008), we have that

$$\mathbb{P}[\mathcal{E}] + \mathbb{P}'[\mathcal{E}^c] \geq \frac{1}{2}\exp\left(-D\left(\mathbb{P} \,||\, \mathbb{P}'\right)\right), \tag{10}$$

where $D\left(\mathbb{P} \,||\, \mathbb{P}'\right)$ is the KL-divergence of $\mathbb{P}$ with respect to $\mathbb{P}'$.

Let us define the event $\mathcal{Q} = \left\{N_k(T) > \frac{T}{2}\right\}$, namely, the event that, by round T OP: or by time $T+1$?, arm k has being played more than $T/2$ times. Since k is a suboptimal arm in μ and optimal in μ', for the regret of policy π in the two environments we have that

$$R_T^\pi(\mu) + R_T^\pi(\mu') \geq \frac{T}{2}\mathbb{P}[\mathcal{Q}]\,\Delta_k(\mu) + \frac{T}{2}\mathbb{P}'[\mathcal{Q}^c]\epsilon.$$

By considering the minimum gap in the above two terms we obtain

$$\begin{aligned} R_T^\pi(\mu) + R_T^\pi(\mu') &\geq \min\{\Delta_k(\mu), \epsilon\} \cdot \frac{T}{2} \cdot \left(\mathbb{P}[\mathcal{Q}] + \mathbb{P}'[\mathcal{Q}^c]\right) \\ &\geq \min\{\Delta_k(\mu), \epsilon\} \cdot \frac{T}{4} \cdot \exp\left(-D\left(\mathbb{P} \,||\, \mathbb{P}'\right)\right) \\ &\geq \min\{\Delta_k(\mu), \epsilon\} \cdot \frac{T}{4} \cdot \exp\left(-\sum_{i\in[K]} \mathbb{E}\left[N_i(t)\right] \frac{(\Delta_k(\mu)+\epsilon)^2}{2\,\sigma_{i,k}^2}\right). \end{aligned} \tag{11}$$

where the second inequality follows by Equation (10), while the last inequality follows by Proposition 4.1. By rearranging terms, we obtain

$$\sum_{i\in[K]} \frac{\mathbb{E}\left[N_i(t)\right]}{\sigma_{i,k}^2} \geq \frac{2}{(\Delta_k(\mu)+\epsilon)^2} \log\left(\frac{T \cdot \min\{\Delta_k(\mu), \epsilon\}}{4(R_T^\pi(\mu) + R_T^\pi(\mu'))}\right). \tag{12}$$

We recall that for any consistent policy π, any environment $(\mu, \mathbf{\Sigma})$, and any $p > 0$, we have that

$$\lim_{T\to\infty} \frac{R_T^\pi(\mu)}{T^p} = 0.$$

Thus, dividing Equation (12) by $\log T$ and taking the limit of $T \to \infty$, we get

$$\begin{aligned}\liminf_{T\to\infty} \frac{1}{\log T} \sum_{i\in[K]} \frac{\mathbb{E}\left[N_i(T)\right]}{\sigma_{i,k}^2} &\geq \liminf_{T\to\infty} \frac{1}{\log T} \frac{2}{(\Delta_k(\mu)+\epsilon)^2} \log\left(\frac{T\cdot\min\{\Delta_k(\mu),\epsilon\}}{4(R_T^\pi(\mu)+R_T^\pi(\mu'))}\right) \\ &= \frac{2}{(\Delta_k(\mu)+\epsilon)^2} \liminf_{T\to\infty} \frac{1}{\log T}\left(\log T - \log\left(R_T^\pi(\mu)+R_T^\pi(\mu')\right) + \log\frac{\min\{\Delta_k(\mu),\epsilon\}}{4}\right).\end{aligned} \tag{13}$$

Now, observe that, by Definition 4.1 for consistent policies, we have that

$$\liminf_{T\to\infty} \frac{1}{\log T}\left(\log T - \log\left(R_T^\pi(\mu)+R_T^\pi(\mu')\right) + \log\frac{\min\{\Delta_k(\mu),\epsilon\}}{4}\right) = 1 - \liminf_{T\to\infty} \frac{\log\left(R_T^\pi(\mu)+R_T^\pi(\mu')\right)}{\log T} = 1,$$

and, therefore, the bound in Equation (13) becomes

$$\liminf_{t\to\infty} \frac{\sum_{i\in[K]} \frac{\mathbb{E}'[N_i(t)]}{\sigma_{i,k}^2}}{\log t} \geq \frac{2}{(\Delta_k(\mu)+\epsilon)^2}.$$

Notice that the above analysis holds for any suboptimal arm and $\epsilon > 0$ and thus is sufficient to conclude the theorem. To illustrate:

Let us consider the reward vector μ and the following constraint set indicated by the above analysis:

$$C(\mu) = \left\{ c \in \mathbb{R}_{\geq 0}^K : \sum_{j\in[K]} \frac{c_j}{\sigma_{j,i}^2} \geq \frac{2}{\Delta_i^2(\mu)} \;\; \forall i \neq i^*(\mu) \right\}.$$

Suppose that the expected number of plays of policy π divided by $\log t$ do not satisfy the above constraints, i.e. $\left(\frac{\mathbb{E}[N_1(t)]}{\log t}, ..., \frac{\mathbb{E}[N_K(t)]}{\log t}\right) \notin C(\mu)$. Then, w.l.o.g. we assume that the constraint of $C(\mu)$ that is not satisfied is the one that corresponds to the k-th suboptimal arm, that is:

$$\sum_{j\in[K]} \frac{\mathbb{E}[N_j(t)]}{\sigma_{j,k}^2 \cdot \log t} < \frac{2}{\Delta_k^2(\mu)}.$$

Then, we can write that $\sum_{j\in[K]} \frac{\mathbb{E}[N_j(t)]}{\sigma_{j,k}^2} = \frac{2\log t}{(\Delta_k(\mu)+\epsilon')^2}$ for some $\epsilon' > 0$. If we consider environments μ and μ' as defined in Equation (9) for some $\epsilon < \epsilon'$, by using Equation (11) we get that:

$$\begin{aligned}R_t^\pi(\mu) + R_t^\pi(\mu') &\geq \frac{t\cdot\min\{\Delta_k(\mu),\epsilon\}}{4} \exp\left(-\frac{\sum_{i\in[K]}\mathbb{E}\left[N_i(t)\right]}{2\,\sigma_{i,k}^2}(\Delta_k(\mu)+\epsilon)^2\right) \\ &= \frac{t\cdot\min\{\Delta_k(\mu),\epsilon\}}{4} \exp\left(-\frac{2\log t}{2(\Delta_k(\mu)+\epsilon')^2}(\Delta_k(\mu)+\epsilon)^2\right) \\ &= \frac{t\cdot\min\{\Delta_k(\mu),\epsilon\}}{4} \cdot t^{-\left(\frac{\Delta_k(\mu)+\epsilon}{\Delta_k(\mu)+\epsilon'}\right)^2}\end{aligned}$$

which is a contradiction to the fact that policy π is consistent.

Finally, by construction of the objective of the LP, we conclude that the optimal solution of the LP characterizes an asymptotic lower bound on the regret of any consistent policy.

□

C. Algorithm and Regret Analysis: Omitted Proofs

C.1. Proof of Theorem 5.1

We prove the main result of Section 5:

Theorem 5.1. *For any $\alpha > 4$, $\gamma \in (0,1)$, and $\epsilon > 0$, the regret of algorithm Algorithm 1 satisfies*

$$\begin{aligned} R_T(\mu) \leq & \left(2K + \frac{8K}{\alpha - 4} + 2\right) \Delta_{\max} \\ & + 2\,\Delta_{\max} K \sum_{\tau \in [T]} \exp\left(-\frac{\tau^{\gamma}\epsilon^2}{4K\,\bar{\sigma}^2}\right) \\ & + \Delta_{\max}\left(4\alpha \sum_{j \in [K]} c_j^*(\mu,\epsilon)\log T + K\right)^{\gamma} \\ & + 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu,\epsilon)\log T. \end{aligned}$$

Proof. As a first step, we define the following events which correspond to the conditions examined in Cases A and B of Algorithm 1, respectively:

$$\mathcal{A}_{\mathrm{t}} = \left\{\left(\frac{N_1(t)}{4a\log t}, ..., \frac{N_K(t)}{4a\log t}\right) \in C(\hat{\mu}(t))\right\} \text{ and } \mathcal{B}_{\mathrm{t}} = \left\{\min_{i\in[K]} \zeta_i(t) < \frac{1}{K}\beta(n_e(t))\right\}.$$

Moreover, we consider the following error events for the ML estimator $\hat{\mu}(t)$:

$$\mathcal{U}_{\mathrm{t}} = \left\{|\,\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{2a\log t}{\zeta_i(t)}},\ \ \forall i \in [K]\right\} \text{ and } \mathcal{E}_{\mathrm{t}} = \left\{\|\,\hat{\mu}(t) - \mu\|_{\infty} \leq \epsilon\right\}.$$

Using the above definitions, we can provide an upper bound on the regret accumulated by Algorithm 1 within T rounds through the following decomposition:

$$\begin{aligned} R_T(\mu) &= T\mu^* - \mathbb{E}\left[\sum_{t\in[T]} X_{i_t,t}\right] = \mathbb{E}\left[\sum_{t\in[T]} \Delta_t(\mu)\right] \\ &\leq K\,\Delta_{\max} + \sum_{t=K+1}^{T} \mathbb{E}\left[\Delta_t(\mu)\left(\mathbb{I}\left({\mathcal{U}_{\mathrm{t}}}^c\right) + \mathbb{I}\left(\mathcal{U}_{\mathrm{t}}, \mathcal{A}_{\mathrm{t}}\right) + \mathbb{I}\left(\mathcal{U}_{\mathrm{t}}, {\mathcal{A}_{\mathrm{t}}}^c, \mathcal{B}_{\mathrm{t}}\right) + \mathbb{I}\left(\mathcal{U}_{\mathrm{t}}, \mathcal{A}_{\mathrm{t}}{}^c, \mathcal{B}_{\mathrm{t}}{}^c, \mathcal{E}_{\mathrm{t}}\right) + \mathbb{I}\left(\mathcal{U}_{\mathrm{t}}, \mathcal{A}_{\mathrm{t}}{}^c, \mathcal{B}_{\mathrm{t}}{}^c, \mathcal{E}_{\mathrm{t}}{}^c\right)\right)\right]. \end{aligned} \tag{14}$$

In the rest of this proof, we bound each of the above terms separately.

Regret due to $\mathcal{U}_{\mathrm{t}}{}^c$. The regret accumulated due to the event $\mathcal{U}_{\mathrm{t}}{}^c$ over T rounds can be upper-bounded by via a union bound over all arms and, then, by applying the concentration result proved in Lemma 3.2. In particular, we obtain that

$$\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{U}_{\mathrm{t}}}^c\right) \leq \sum_{t=K+1}^{T} \sum_{i\in[K]} \Pr\left[|\,\hat{\mu}_i(t) - \mu_i| > \sqrt{\frac{2a\log t}{\zeta_i(t)}}\right] \leq K \sum_{t=K+1}^{T} 2\cdot\lceil\log_2(t-1)\rceil\cdot t^{-\alpha/2} \leq \frac{4K}{\alpha-4}. \tag{15}$$

Therefore, for the rest of the cases we can assume the the event $\mathcal{U}_{\mathrm{t}}$ holds for every $t \geq K$.

Regret due to $\mathcal{U}_{\mathrm{t}}, \mathcal{A}_{\mathrm{t}}$. When the events $\mathcal{U}_{\mathrm{t}}$ and $\mathcal{A}_{\mathrm{t}}$ hold, we show that the greedy arm selection in Case A of Algorithm 1 leads to the selection of the optimal arm. By using the definitions of the events $\mathcal{U}_{\mathrm{t}}$ and $\mathcal{A}_{\mathrm{t}}$, we have that the error in the mean estimate of any arm $i \in [K]$ which is suboptimal in the vector of estimations $\hat{\mu}(t)$ can be upper-bounded as

$$|\,\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{2a\log t}{\zeta_i(t)}} \leq \sqrt{\frac{2a\log t}{\frac{8\alpha\log t}{\Delta_i^2(\hat{\mu}(t))}}} = \frac{\Delta_i(\hat{\mu}(t))}{2},\ \ \forall i \neq i^*(\hat{\mu}(t)).$$

In particular, this implies that for any arm $i \neq i^*(\hat{\mu}(t))$ (i.e., which is suboptimal with respect to estimates the $\hat{\mu}(t)$), we have that

$$\mu_i - \hat{\mu}_i(t) \leq \frac{\Delta_i(\hat{\mu}(t))}{2}. \tag{16}$$

Similarly, by using the definitions of $\mathcal{A}_\mathrm{t}$ and $C(\hat{\mu}_t)$, for the optimal arm of $\hat{\mu}(t)$, $i^*(\hat{\mu}(t))$, we have

$$\hat{\mu}_{i^*(\hat{\mu}(t))}(t) - \mu_{i^*(\hat{\mu}(t))} \leq \frac{\Delta_{\min}(\hat{\mu}(t))}{2}. \tag{17}$$

By adding Equation (16) and Equation (17) we get

$$\mu_i - \hat{\mu}_i(t) + \hat{\mu}_{i^*(\hat{\mu}(t))}(t) - \mu_{i^*(\hat{\mu}(t))} \leq \frac{\Delta_i(\hat{\mu}(t))}{2} + \frac{\Delta_{\min}(\hat{\mu}(t))}{2} \leq \Delta_i(\hat{\mu}(t)).$$

Now, since $\Delta_i(\hat{\mu}(t)) = \hat{\mu}_{i^*(\hat{\mu}(t))}(t) - \hat{\mu}_i(t)$ by definition, eliminating the identical terms in the above expression yields $\mu_i \leq \mu_{i^*(\hat{\mu}(t))}$.

By repeating the above analysis for each arm $i \neq i^*(\hat{\mu}(t))$, we can conclude that the mean reward of $i^*(\hat{\mu}(t))$ satisfies $\mu_{i^*(\hat{\mu}(t))} \geq \max_{i\in[K]} \mu_i$, which in turn implies that $i^*(\hat{\mu}(t))$ also corresponds to the optimal arm of vector μ. Thus, the rounds where the events $\mathcal{U}_\mathrm{t}$ and $\mathcal{A}_\mathrm{t}$ hold do not contribute to the regret, namely,

$$\Delta_t(\mu)\, \mathbb{I}\left(\mathcal{U}_\mathrm{t}, \mathcal{A}_\mathrm{t}\right) = 0. \tag{18}$$

Regret due to $\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}$. We focus on the regret accumulated during the rounds where $\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c$ and $\mathcal{B}_\mathrm{t}$ hold. These correspond to the the rounds where Algorithm 1 attempts to ensure uniform exploration for all arms in terms of their weighted numbers of samples.

In order to upper-bound the regret accumulated in the rounds where the events $\mathcal{U}_\mathrm{t}, {\mathcal{A}_\mathrm{t}}^c$ and $\mathcal{B}_\mathrm{t}$ hold, we relate the number of rounds where ${\mathcal{A}_\mathrm{t}}^c$ and $\mathcal{B}_\mathrm{t}$ hold with the total number of "exploration rounds", i.e., the number of times the algorithm enters Case B or C (which happens only if ${\mathcal{A}_\mathrm{t}}^c$ holds). We achieve the above via a counting argument, based on the following proposition:

Proposition C.1. *Let $t_2 > t_1 > K$. If $\sum_{t=t_1}^{t_2-1} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right) \geq K$, then $\min_{i\in[K]} \zeta_i(t_2) \geq \min_{i\in[K]} \zeta_i(t_1) + \frac{1}{\bar{\sigma}^2}$.*

In simple terms, Proposition C.1 states that if ${\mathcal{A}_\mathrm{t}}^c$ and $\mathcal{B}_\mathrm{t}$ hold (which corresponds to Case B of Algorithm 1) at least K times within an interval $[t_1, t_2 - 1]$, then the minimum weighted number of samples, $\zeta_i(t)$, over all arms $i \in [K]$ after these rounds, must have increased by at least $1/\bar{\sigma}^2$.

Proof of Proposition C.1. The proposition follows from a pigeonhole argument. Let ℓ_1 be the first round in the interval $[t_1, t_2 - 1]$, where ${\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}$ are satisfied and, hence, Algorithm 1 enters Case B. Let $j = \operatorname{argmin}_{k\in[K]} \zeta_k(\ell_1)$ be the arm with minimum weighted number of samples at the beginning of round ℓ_1, which was chosen in Case B of Algorithm 1. At round ℓ_1, Algorithm 1 plays by construction an arm $i_{\ell_1} = \arg\min_{k\in[K]} \sigma^2_{k,j}$ and, hence, the weighted number of samples of arm j at round $\ell_1 + 1$ must satisfy

$$\zeta_j(\ell_1 + 1) = \zeta_j(\ell_1) + \max_{k\in[K]} \frac{1}{\sigma^2_{k,j}} = \zeta_j(\ell_1) + \frac{1}{(\sigma_j^{\min})^2} \geq \zeta_j(\ell_1) + \frac{1}{\bar{\sigma}^2} \geq \min_{i\in[K]} \zeta_i(t_1) + \frac{1}{\bar{\sigma}^2}, \tag{19}$$

where the last inequality follows by the fact that $\zeta_j(\ell_1) = \min_{i\in[K]} \zeta_i(\ell_1) \geq \min_{i\in[K]} \zeta_i(t_1)$, since the weighted number of samples of any arm can only increase as time progresses.

Observe that, if at any round $\ell \in [\ell_1 + 1, t_2 - 1]$ arm i is again a minimizer of $\min_{k\in[K]} \zeta_k(\ell)$, Equation (19) implies that, at round t_2, it has to be that $\min_{i\in[K]} \zeta_i(t_2) \geq \zeta_j(\ell_1) \geq \min_{i\in[K]} \zeta_i(t_1) + 1/\bar{\sigma}^2$, in which case the proposition follows directly. Thus, we can assume that for any round $\ell \in [\ell_1 + 1, t_2 - 1]$ only arms in $[K] \setminus \{j\}$ are minimizers of $\min_{k\in[K]} \zeta_k(\ell)$.

By repeatedly applying the above argument, it follows that each arm in $[K]$ can be the minimizer of $\min_{k\in[K]} \zeta_k(\ell)$ for at most one round $\ell \in [t_1, t_2 - 1]$. However, since $\sum_{t=t_1}^{t_2-1} \mathbb{I}\left({\mathcal{A}_\mathrm{t}}^c, \mathcal{B}_\mathrm{t}\right) \geq K$, by assumption, this can only imply that each arm

$j \in [K]$ is the minimizer of $\min_{k\in[K]} \zeta_k(\ell)$ for at least one round $\ell \in [t_1, t_2 - 1]$. In this case, the weighted number of samples for each arm $i \in [K]$ (including the minimizer of round t_1) has been increased by at least $1/\bar{\sigma}^2$ during the interval $[t_1, t_2 - 1]$, and thus

$$\min_{i\in[K]} \zeta_i(t_2) \geq \min_{i\in[K]} \left(\zeta_i(t_1) + \frac{1}{\bar{\sigma}^2}\right) = \min_{i\in[K]} \zeta_i(t_1) + \frac{1}{\bar{\sigma}^2},$$

which concludes the proof. □

Lemma 5.1. *The following inequality holds:*

$$\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c, \mathcal{B}_t\right) \leq \frac{1}{2}\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c\right)\right)^{\gamma} + 1.$$

Proof. Let $K + 1 \leq t' \leq T$ be the maximum round where the events ${\mathcal{A}_t}^c, \mathcal{B}_t$ hold, that is,

$$t' = \max_{t\in\mathbb{N}} \{t \mid \mathbb{I}\left({\mathcal{A}_t}^c, \mathcal{B}_t\right) = 1 \text{ and } K + 1 \leq t \leq T\}.$$

Due to Proposition C.1, we have that the minimum weighted number of samples of any arm $i \in [K]$ at time t' satisfies

$$\min_{i\in[K]} \zeta_i(t') \geq \frac{1}{\bar{\sigma}^2 K} \sum_{t=K+1}^{t'-1} \mathbb{I}\left({\mathcal{A}_t}^c, \mathcal{B}_t\right),$$

since the above minimum increases by at least $1/\sigma^2$ every K times the events ${\mathcal{A}_t}^c, \mathcal{B}_t$ occur.

By rearranging terms in the above inequality and using the fact that, when ${\mathcal{A}_t}^c, \mathcal{B}_t$ hold, we have that $\min_{i\in[K]} \zeta_i(t) < \frac{1}{K}\beta\left(n_e(t)\right)$, we obtain the following upper bound on the number of times the events ${\mathcal{A}_t}^c, \mathcal{B}_t$ can occur over T rounds:

$$\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c, \mathcal{B}_t\right) \leq \sum_{t=K+1}^{t'-1} \mathbb{I}\left({\mathcal{A}_t}^c, \mathcal{B}_t\right) + 1 \leq K\,\bar{\sigma}^2 \min_{i\in[K]} \zeta_i(t') + 1 < \bar{\sigma}^2\,\beta\left(n_e(t')\right) + 1. \tag{20}$$

Now using that $n_e(t') = \sum_{t=K+1}^{t'-1} \mathbb{I}\left({\mathcal{A}_t}^c\right)$, the definition of $\beta(x) = {x^\gamma}/{2}$, and that $t' \leq T$, we have

$$\begin{aligned}(20) = \bar{\sigma}^2\,\beta\left(\sum_{t=K+1}^{t'-1} \mathbb{I}\left({\mathcal{A}_t}^c\right)\right) + 1 &\leq \bar{\sigma}^2\,\beta\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c\right)\right) + 1 \\ &= \frac{1}{2}\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c\right)\right)^{\gamma} + 1,\end{aligned}$$

which concludes the proof. □

Equipped with Lemma 5.1 and using a decomposition of the event ${\mathcal{A}_t}^c$, we can bound the term of the regret depending on $\mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, \mathcal{B}_t\right)$ as follows:

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, \mathcal{B}_t\right) &\leq \sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c, \mathcal{B}_t\right) \\ &\leq \frac{1}{2}\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{A}_t}^c\right)\right)^{\gamma} + 1 \\ &\leq \frac{1}{2}\left(\sum_{t=K+1}^{T} \mathbb{I}\left({\mathcal{U}_t}^c, {\mathcal{A}_t}^c\right) + \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, \mathcal{B}_t\right) + \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\right) + \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\right)\right)^{\gamma} + 1 \\ &\leq \frac{1}{2}\sum_{t=K+1}^{T}\left(\mathbb{I}\left({\mathcal{U}_t}^c\right) + \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, \mathcal{B}_t\right) + \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\right)\right) + \frac{1}{2}\left(\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\right)\right)^{\gamma} + 1,\end{aligned}$$

where the last inequality comes from the fact that the function $f(x) = x^\gamma$ is subadditive for any $\gamma \in (0,1)$, and $x^\gamma \leq x$ for any integer x.

By rearranging and grouping terms we obtain the following bound:

$$\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c\right) \leq \sum_{t=K+1}^{T} \left(\mathbb{I}\left({\mathcal{U}_t}^c\right) + \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\right)\right) + \left(\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\right)\right)^\gamma + 2. \tag{21}$$

Thus, in order to bound the regret in the case where $\mathcal{U}_t, {\mathcal{A}_t}^c$ and ${\mathcal{B}_t}^c$ hold, we need to bound the regret accumulated due to the terms $\mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\right)$ and $\mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\right)$.

Regret due to $\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t$. In rounds where the events ${\mathcal{A}_t}^c$ and ${\mathcal{B}_t}^c$ hold and, thus, the constraints of the estimated LP, $C(\hat{\mu}(t))$, are not satisfied, Algorithm 1 plays any arm $j \in [K]$ that violates its corresponding constraint, i.e. $\frac{N_j(t)}{4a \log t} < c_j^*(\hat{\mu}(t))$. In addition, we recall that the event $\mathcal{E}_t$ implies that $\| \hat{\mu}(t) - \mu \|_\infty \leq \epsilon$, namely, all mean estimates are ϵ-close to the true mean rewards. In that case, for any arm $j \in [K]$ we can construct an ϵ-approximate LP, using the worst-case ϵ-approximate LP solution introduced in Definition 5.1:

$$c_j^*(\mu, \epsilon) = \sup_{\mu' : \|\mu' - \mu\|_\infty \leq \epsilon} c_j^*(\mu').$$

Lemma 5.2. *For every arm $j \in [K]$, it holds that*

$$\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t, i_t = j\right) \leq 4\alpha \cdot c_j^*(\mu, \epsilon) \log T.$$

Proof. Let t' be the last round such that $t' \leq T$ where $\{\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\}$ holds and the algorithm plays arm j. Then we have that:

$$\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t, i_t = j\right) = \sum_{t=K+1}^{t'} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t, i_t = j\right) \leq N_j(t') \leq 4\alpha c_j^*(\hat{\mu}(t')) \log t' \leq 4\alpha c_j^*(\mu, \epsilon) \log T,$$

where the second inequality comes from the fact that $\frac{N_j(t')}{4a \log t'} < c_j^*(\hat{\mu}(t'))$, by definition of j. The last follows from $t' \leq T$ and Definition 5.1 combined with the fact that $\| \hat{\mu}(t') - \mu \|_\infty \leq \epsilon$. □

By applying the above lemma for any arm $j \in [K]$, we can bound the term of the regret depending on $\mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\right)$ as follows:

$$\begin{aligned}\sum_{t=K+1}^{T} \Delta_t(\mu)\, \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t\right) &\leq \sum_{j \in [K]} \Delta_j(\mu) \sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, \mathcal{E}_t, i_t = j\right) \\ &\leq 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu, \epsilon) \log T. \end{aligned} \tag{22}$$

Regret due to $\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c$. We can bound the term of the regret due to the events $\{\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\}$ as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=K+1}^{T} \Delta_{i_t}(\mu)\, \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\right)\right] &\leq \Delta_{\max}\, \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, {\mathcal{E}_t}^c\right)\right] \\ &= \Delta_{\max}\, \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c,\ \exists j \in [K] : |\hat{\mu}_j(t) - \mu_j| > \epsilon\right)\right] \\ &\leq \Delta_{\max}\, \mathbb{E}\left[\sum_{j \in [K]} \sum_{t=K+1}^{T} \mathbb{I}\left(\mathcal{U}_t, {\mathcal{A}_t}^c, {\mathcal{B}_t}^c, |\hat{\mu}_j(t) - \mu_j| > \epsilon\right)\right], \end{aligned} \tag{23}$$

where the last inequality follows by a union bound on any possible arm $j \in [K]$ satisfying $|\hat{\mu}_j(t) - \mu_j| > \epsilon$.

For an easier manipulation of the bound in Equation (23), let us construct the following sets:

$$\Lambda = \left\{ t \in [T] : \mathcal{U}_t, {\mathcal{A}_t}^c, \min_{i \in [K]} \zeta_i(t) \geq \frac{1}{K} \beta\left(n_e(t)\right) \right\}$$

and the more refined

$$\Lambda(\tau) = \left\{ t \in [T] : \mathcal{U}_t, {\mathcal{A}_t}^c, \min_{i \in [K]} \zeta_i(t) \geq \frac{1}{K} \beta\left(\tau\right), n_e(t) = \tau \right\}.$$

Note that if $t \in \Lambda(\tau)$ then $\zeta_j(t) \geq \frac{1}{K}\beta\left(n_e(t)\right)$ and that $|\Lambda(\tau)| \leq 1$ for any τ, by construction. In addition, $\Lambda \subseteq \cup_{\tau \in [T]} \Lambda(\tau)$. Then, let ϕ_τ be a stopping time such that $\phi_\tau = t$ if $\Lambda(\tau) = \{t\}$ and $\phi_\tau = T + 1$ otherwise. We can write that:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t \in [T]} \mathbb{I}\left(t \in \Lambda, |\, \hat{\mu}_i(t) - \mu_i| > \epsilon\right) \right] &\leq \mathbb{E}\left[\sum_{\tau \in [T]} \mathbb{I}\left(\phi_\tau \leq T, |\, \hat{\mu}_i(t) - \mu_i| > \epsilon\right) \right] \\
&= \sum_{\tau \in [T]} \Pr\left[\phi_\tau \leq T, |\, \hat{\mu}_i(t) - \mu_i| > \epsilon\right] \\
&\leq \sum_{\tau \in [T]} 2 \exp\left(-\frac{\epsilon^2}{2} \frac{1}{K} \beta(\tau) \right),
\end{aligned}$$

where in the last inequality we apply the result of Lemma 3.1.

Therefore, Equation (23) can be bounded as follows:

$$\begin{aligned}
\textit{Equation (23)} &= \Delta_{\max} \mathbb{E}\left[\sum_{i \in [K]} \sum_{t=K+1}^{T} \mathbb{I}\left(t \in \Lambda, |\, \hat{\mu}_i(t) - \mu_i| > \epsilon\, \right) \right] \\
&\leq \Delta_{\max} \sum_{i \in [K]} \sum_{\tau \in [T]} \exp\left(-\frac{\beta(\tau)\epsilon^2}{2K} \right) \qquad (24) \\
&= \Delta_{\max} \sum_{i \in [K]} \sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\, \bar{\sigma}^2} \right). \qquad (25)
\end{aligned}$$

where in the last inequality we used the definition of $\beta(\cdot)$ function.

Combining everything. By combining the bounds in Equations (14), (15), (18), (21), (22) and (24), we can conclude that the regret of Algorithm 1 can be upper bounded as follows:

$$\begin{aligned}
R_T(\mu) \leq K\, \Delta_{\max} + \Delta_{\max} \frac{8K}{\alpha - 4} + \Delta_{\max} \left(4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu, \epsilon) \log T \right)^\gamma \\
+ 2\, \Delta_{\max} K \sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\, \bar{\sigma}^2} \right) + 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu, \epsilon) \log T + 2.
\end{aligned}$$

□

C.2. Proof of Corollary 5.1

Corollary 5.1. *As the time horizon tends to infinity, the regret of Algorithm 1 satisfies*

$$\limsup_{T \to \infty} \frac{R_T(\mu)}{\log T} \leq 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu).$$

Proof. Recall that by Theorem 5.1:

$$R_T(\mu) \leq \left(2K + \frac{8K}{\alpha - 4} + 2\right) \Delta_{\max} + 2\,\Delta_{\max} K \sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right) + \Delta_{\max} \left(4\alpha \sum_{j \in [K]} c_j^*(\mu, \epsilon) \log T + K\right)^\gamma + 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu, \epsilon) \log T.$$

Observe that for any constant $\epsilon, \bar{\sigma} > 0$ and $\gamma \in (0, 1)$ we have that $\lim_{T \to \infty} \sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right) < \infty$. We emphasize that, while $\bar{\sigma}$ is a fixed (possibly unbounded) parameter of the problem instance, ϵ is only a component of the analysis. Moreover, notice that we are not allowed to directly use $\epsilon \to 0$ in the above regret upper bound, since that would make the term $\sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right)$ linear in T. However, if we choose ϵ to be a decreasing function of T, then the suboptimal terms in Theorem 5.1 are vanishing as $T \to \infty$. In particular, choosing $\epsilon = (4K\,\bar{\sigma}^2)^{\frac{1}{2}} \log(T)^{-\gamma/4}$ gives

$$\begin{aligned}
\sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma \epsilon^2}{4K\,\bar{\sigma}^2}\right) &= \sum_{\tau \in [T]} \exp\left(-\frac{\tau^\gamma}{(\log T)^{\gamma/2}}\right) \\
&= \sum_{\tau \in [\lceil (\log T)^{1/2} \rceil]} \exp\left(-\frac{\tau^\gamma}{(\log T)^{\gamma/2}}\right) + \sum_{\tau \in [T] \setminus [\lceil (\log T)^{1/2} \rceil]} \exp\left(-\frac{\tau^\gamma}{(\log T)^{\gamma/2}}\right) \\
&\leq \lceil (\log T)^{1/2} \rceil + \sum_{n \in \left[\frac{T}{(\log T)^{1/2}}\right]} \exp\left(-n^\gamma\right)
\end{aligned}$$

where in the above expression the first term is $o(\log(T)$ and the second term is bounded above by $1/\left(e^{\frac{1}{\gamma}} - 1\right)$ for any T. Furthermore, the term

$$\left(4\alpha \sum_{j \in [K]} c_j^*(\mu, \epsilon) \log T + K\right)^\gamma \in o\left(4\alpha \sum_{j \in [K]} c_j^*(\mu, \epsilon) \log T + K\right)$$

and thus when it is divided by $\log(T)$ it vanishes asymptotically for $T \to \infty$. Finally, for the value $\epsilon = (4K\,\bar{\sigma}^2)^{\frac{1}{2}} \log(T)^{-\gamma/4}$ selected above, we have that $\epsilon \to 0$ as $T \to \infty$, thus the leading term becomes

$$\begin{aligned}
\limsup_{T \to \infty} \frac{4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu, \epsilon) \log T}{\log T} &= \limsup_{T \to \infty}\ 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu, \epsilon) \\
&= 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu)
\end{aligned}$$

by Definition 5.1 of the ϵ-approximate LP solution. Therefore, we conclude that

$$\limsup_{T \to \infty} \frac{R_T(\mu)}{\log T} = 4\alpha \sum_{j \in [K]} \Delta_j(\mu) c_j^*(\mu).$$

□