Mostly Exploration-Free Algorithms for Contextual Bandits

Hamsa Bastani
IBM Thomas J. Watson Research and Wharton School, hamsab@wharton.upenn.edu

Mohsen Bayati
Stanford Graduate School of Business, bayati@stanford.edu

Khashayar Khosravi
Stanford University Electrical Engineering, khosravi@stanford.edu

The contextual bandit literature has traditionally focused on algorithms that address the exploration-exploitation tradeoff. In particular, greedy algorithms that exploit current estimates without any exploration may be sub-optimal in general. However, exploration-free greedy algorithms are desirable in practical settings where exploration may be costly or unethical (e.g., clinical trials). Surprisingly, we find that a simple greedy algorithm can be rate-optimal if there is sufficient randomness in the observed contexts. We prove that this is always the case for a two-armed bandit under a general class of context distributions that satisfy a condition we term *covariate diversity*. Furthermore, even absent this condition, we show that a greedy algorithm can be rate-optimal with nonzero probability. Thus, standard bandit algorithms may unnecessarily explore. Motivated by these results, we introduce Greedy-First, a new algorithm that uses only observed contexts and rewards to determine whether to follow a greedy algorithm or to explore. We prove that this algorithm is rate-optimal without any additional assumptions on the context distribution or the number of arms. Extensive simulations demonstrate that Greedy-First successfully reduces experimentation and outperforms existing (exploration-based) contextual bandit algorithms such as Thompson sampling or UCB.

Key words: sequential decision-making, contextual bandit, greedy algorithm, exploration-exploitation

1. Introduction

Service providers across a variety of domains are increasingly interested in personalizing decisions based on customer characteristics. For instance, a website may wish to tailor content based on an Internet user's web history (Li et al. 2010), or a medical decision-maker may wish to choose treatments for patients based on their medical records (Kim et al. 2011). In these examples, the costs and benefits of each decision depend on the individual customer or patient, as well as their specific context (web history or medical records respectively). Thus, in order to make optimal decisions, the decision-maker must learn a model predicting individual-specific rewards for each decision based on the individual's observed contextual information. This problem is often formulated as a contextual bandit (Auer 2003, Langford and Zhang 2008, Li et al. 2010), which generalizes the classical multi-armed bandit problem (Lai and Robbins 1985).

In this setting, the decision-maker has access to K possible decisions (arms) with uncertain rewards. Each arm i is associated with an unknown parameter $\beta_i \in \mathbb{R}^d$ that is predictive of its individual-specific rewards. At each time t, the decision-maker observes an individual with an associated context vector $X_t \in \mathbb{R}^d$. Upon choosing arm i, she realizes a (linear) reward of

$$X_t^\top \beta_i + \varepsilon_{i,t} ,$$

where $\varepsilon_{i,t}$ are independent subGaussian noise. The decision-maker's goal is to maximize the cumulative reward over T different individuals by gradually learning the arm parameters. Devising an optimal policy for this setting is often computationally intractable, and thus, the literature has focused on effective heuristics that are asymptotically optimal, including UCB (Dani et al. 2008, Abbasi-Yadkori et al. 2011), Thompson sampling (Agrawal and Goyal 2013, Russo and Van Roy 2014a), information-directed sampling (Russo and Van Roy 2014b), and algorithms inspired by ϵ-greedy methods (Goldenshluger and Zeevi 2013, Bastani and Bayati 2015).

The key ingredient in designing these algorithms is addressing the *exploration-exploitation trade-off*. On one hand, the decision-maker must explore or sample each decision for random individuals to improve her estimate of the unknown arm parameters $\{\beta_i\}_{i=1}^K$; this information can be used to improve decisions for future individuals. Yet, on the other hand, the decision-maker also wishes to exploit her current estimates $\{\hat{\beta}_i\}_{i=1}^K$ to make the estimated best decision for the current individual in order to maximize cumulative reward. The decision-maker must therefore carefully balance both exploration and exploitation to achieve good performance. In general, algorithms that fail to explore sufficiently may fail to learn the true arm parameters, yielding poor performance.

However, exploration may be prohibitively costly or infeasible in a variety of practical environments (Bird et al. 2016). In medical decision-making, choosing a treatment that is not the estimated-best choice for a random patient may be unethical; in marketing applications, testing out an inappropriate ad on a potential customer may result in the costly, permanent loss of the customer. Such concerns may deter decision-makers from deploying bandit algorithms in practice.

In this paper, we analyze the performance of *exploration-free* greedy algorithms. Surprisingly, we find that a simple greedy algorithm can achieve the same state-of-the-art asymptotic performance guarantees as standard bandit algorithms *if* there is sufficient randomness in the observed contexts (thereby creating natural exploration). In particular, we prove that the greedy algorithm is near-optimal for a two-armed bandit when the context distribution satisfies a condition we term *covariate diversity*, a property that is satisfied by a natural class of continuous and discrete context distributions. Furthermore, even absent this condition, we find that a greedy approach provably converges to the optimal policy (for all time) with some constant probability that depends on the problem parameters. Thus, exploration may not be necessary at all in a general class of problem instances, and may only sometimes be necessary in other problem instances. Simulations confirm our theoretical results and suggest that the greedy algorithm in fact outperforms popular bandit algorithms when exploration is provably unnecessary.

Unfortunately, one may not know a priori when a greedy algorithm will converge, since its convergence depends on unknown problem parameters. For instance, the decision-maker may not know if the context distribution satisfies covariate diversity; if covariate diversity is not satisfied, the greedy algorithm may be undesirable since it may achieve linear regret some fraction of the time (i.e., it fails to converge to the optimal policy with some probability). To address this concern, we present Greedy-First, a new algorithm that seeks to reduce exploration when possible by starting with a greedy approach, and incorporating exploration only when it is confident that the greedy algorithm is failing with high probability. In particular, we formulate a simple hypothesis test using observed contexts and rewards to verify (with high probability) if the greedy arm parameter estimates are converging at the asymptotically optimal rate. If not, our algorithm transitions to a standard exploration-based contextual bandit algorithm.

Greedy-First satisfies the same asymptotic guarantees as standard contextual bandit algorithms without our additional assumptions on covariate diversity or any restriction on the number of arms. More importantly, Greedy-First does not perform any exploration (i.e., remains greedy) with high probability if the covariate diversity condition is met. Furthermore, even when covariate diversity is not met, Greedy-First provably reduces the expected amount of exploration compared to standard bandit algorithms. This occurs because the vanilla greedy algorithm provably converges to the optimal policy with some probability even for problem instances without covariate diversity; however, it achieves linear regret on average since it may fail some fraction of the time. Greedy-First leverages this observation by following a purely greedy algorithm until it detects that this approach has failed. Thus, on average, the Greedy-First policy explores less than standard algorithms that always explore. Simulations confirm our theoretical results, and demonstrate that Greedy-First outperforms existing contextual bandit algorithms even when covariate diversity is not met.

Finally, Greedy-First provides decision-makers with a natural interpretation for exploration. The hypothesis test for adopting exploration only triggers when an arm has not received sufficiently diverse samples; at this point, the decision-maker can choose to explore that arm by assigning it random individuals, or to discard it based on current estimates and continue with a greedy approach. In this way, Greedy-First reduces the opaque nature of experimentation, which we believe can be valuable for aiding the adoption of bandit algorithms in practice.

1.1. Related Literature

There has been significant interest in operational methods for personalizing service decisions as a function of observed user covariates (see, e.g., Ban and Rudin 2014, Bertsimas and Kallus 2014, Chen et al. 2015, Kallus 2016). We take a sequential decision-making approach with *bandit feedback*, i.e., the decision-maker only observes feedback for her chosen decision and does not observe counterfactual feedback from other decisions she could have made. This obstacle inspires the exploration-exploitation tradeoff in multi-armed bandit problems.

Our work falls within the framework of contextual bandits (or a linear bandit with changing action space), which has been extensively studied in the operations, computer science, and statistics literature (we refer the reader to Chapter 4 of Bubeck and Cesa-Bianchi (2012) for an informative review). This setting was first introduced by Auer (2003) through the LinRel algorithm and was subsequently improved through the OFUL algorithm by Dani et al. (2008) and the LinUCB algorithm by Chu et al. (2011). More recently, Abbasi-Yadkori et al. (2011) prove an upper bound of $\mathcal{O}(d\sqrt{T})$ regret in the low-dimensional setting. (We note that they also prove a "problem-dependent" bound of $\mathcal{O}(d \log T/\Delta)$ if one assumes a constant gap Δ between arm rewards; this bound does not apply to the contextual bandit since there is no such gap between arm rewards.)

As mentioned earlier, this literature typically allows for arbitrary (adversarial) covariate sequences. We consider the case where contexts are generated i.i.d., which is more suited for certain applications (e.g., clinical trials on treatments for a non-infectious disease). In this setting one can achieve exponentially better regret bounds in T. In particular, Goldenshluger and Zeevi (2013) present the OLS Bandit algorithm and prove a corresponding upper bound of $\mathcal{O}(d^3 \log T)$ on its cumulative regret. They also prove a lower bound of $\mathcal{O}(\log T)$ regret for this problem (i.e., the contextual bandit with i.i.d. contexts and linear payoffs).

However, this substantial literature requires exploration. Greedy policies are desirable in practical settings where exploration may be costly or unethical. Notable exceptions are Woodroofe (1979) and Sarkar (1991), who consider a Bayesian one armed bandit with a single i.i.d. covariate and a parametric reward with a known prior. They show that a greedy policy based on dynamic programming achieves optimal discounted reward as the discount factor converges to 1. Wang et al. (2005a,b) extend this result with a single covariate and two arms. Mersereau et al. (2009) consider the Bayesian setting where there is some known structure between arm rewards; in this case, they prove that the greedy algorithm is asymptotically optimal with respect to the known prior. These papers differ from our setting in two ways: (i) our arm parameters are unknown and deterministic, and (ii) we use the more standard notion of minimax regret since the prior is usually unknown. In this setting, Goldenshluger and Zeevi (2013) explicitly acknowledge that "we were not able to prove that the myopic policy is rate optimal in our setting." Our work addresses this issue by showing that the greedy (or myopic) policy is optimal under some additional assumptions on the context distribution.

Our approach is also related to recent literature on designing conservative bandit algorithms (Wu et al. 2016, Kazerouni et al. 2016) that operate within a safety margin, i.e., the regret is constrained to stay below a certain threshold that is determined by a baseline policy. These papers propose algorithms that restrict the amount of exploration (similar to the present work) in order to satisfy their safety constraint. Wu et al. (2016) studies the classical multi-armed bandit, and Kazerouni et al. (2016) generalizes these results to the contextual linear bandit.

Finally, we note that there are technical parallels between our work and the analysis of the greedy policy and its variants in the dynamic pricing literature (Lattimore and Munos 2014, Broder and Rusmevichientong 2012). In particular, the most commonly-studied dynamic pricing problem (without covariates) can be viewed as a linear bandit problem without changing action space and with a modified reward function (den Boer and Zwart 2013, Keskin and Zeevi 2014a). When there are no covariates, the greedy algorithm has been shown to be undesirable since it provably converges to a suboptimal price (a fixed point known as the "uninformative price") with nonzero probability (den Boer and Zwart 2013, Keskin and Zeevi 2014a, 2015). Thus, bandit-like algorithms have been proposed, which always explore in order to guarantee convergence to the optimal price (den Boer and Zwart 2013, Keskin and Zeevi 2014b,a, den Boer and Zwart 2015). Applying an approach like Greedy-First (i.e., using a hypothesis test to only explore when the greedy policy has failed) may be relevant in this setting as well, in order to reduce exploration. However, such an analysis is outside the scope of this paper.

More recently, some have studied dynamic pricing with changing demand covariates (Cohen et al. 2016, Javanmard and Nazerzadeh 2016) or a changing demand environment (Keskin and Zeevi 2015). In this case, there is no sub-optimal fixed point since the demand curves are continually changing. Thus, the greedy algorithm has been shown to be asymptotically optimal (Keskin and Zeevi 2015, Qiang and Bayati 2016) in this setting. Our work significantly differs from this line of analysis since we need to learn multiple reward functions simultaneously. In dynamic pricing, the decision-maker always receives feedback from the true demand function, and thus must only avoid a single problematic fixed point (i.e., the uninformative price). Consequently, the greedy policy is always asymptotically optimal for dynamic pricing with demand covariates. In the contextual bandit, we only receive feedback from a decision if we choose it; thus, the greedy algorithm may suffer from entirely dropping an arm, i.e., never sampling a particular decision due to poor initial estimates. As a result, the asymptotic optimality of the greedy algorithm in the contextual bandit is not always guaranteed. We derive a novel condition (covariate diversity), under which we prove that the greedy policy always achieves asymptotic optimality.

2. Problem Formulation

We consider a K-armed contextual bandit for T time steps, where T is unknown. Each arm i is associated with an unknown parameter $\beta_i \in \mathbb{R}^d$. For any integer n, let $[n]$ denote the set $\{1, ..., n\}$. At each time t, we observe a new individual with context vector $X_t \in \mathbb{R}^d$. We assume that $\{X_t\}_{t\geq 0}$ is a sequence of i.i.d. samples from some unknown probability density $p_X(\mathbf{x})$. If we pull arm $i \in [K]$, we observe a stochastic linear reward[1]

$$Y_t = X_t^\top \beta_i + \varepsilon_{i,t} \,,$$

where $\varepsilon_{i,t}$ are independent σ-subgaussian random variables (see Definition 1 below).

Definition 1. A random variable Z is σ-subgaussian if for all $\tau > 0$ we have $\mathbb{E}[e^{\tau Z}] \leq e^{\tau^2\sigma^2/2}$.

We seek to construct a sequential decision-making policy π that learns the arm parameters $\{\beta_i\}_{i=1}^K$ over time in order to maximize expected reward for each individual.

We measure the performance of π by its *cumulative expected regret*, which is the standard metric in the analysis of bandit algorithms (Lai and Robbins 1985, Auer 2003). In particular, we compare ourselves to an oracle policy π^*, which knows the arm parameters $\{\beta_i\}_{i=1}^K$ in advance. Upon observing context X_t, the oracle will always choose the best expected arm $\pi_t^* = \max_{j\in[K]}(X_t^\top \beta_j)$. Thus, if we choose an arm $i \in [K]$ at time t, we incur *instantaneous expected regret*

$$r_t \;\equiv\; \mathbb{E}_{X_t \sim p_X}\left[\max_{j\in[K]}(X_t^\top \beta_j) - X_t^\top \beta_i\right] \,,$$

which is simply the expected difference in reward between the oracle's choice and our choice. We seek to minimize the cumulative expected regret $R_T := \sum_{t=1}^T r_t$. In other words, we seek to mimic the oracle's performance by gradually learning the arm parameters.

Additional Notation: Let B_R^d be the ball of radius R around the origin in $\mathbb{R}^d$ and let the volume of a set $S \subset \mathbb{R}^d$ be $\mathrm{vol}(S) \equiv \int_S \mathrm{d}\mathbf{x}$.

2.1. Assumptions

We now describe the assumptions required for our regret analysis. Some assumptions will be relaxed in later sections of the paper as noted below.

Our first assumption is that the contexts as well as the arm parameters $\{\beta_i\}_{i=1}^K$ are bounded. This ensures that the maximum regret at any time step t is bounded. This is a standard assumption made in the bandit literature (see e.g., Dani et al. 2008).

[1] We consider a generalization of the linear reward function in Appendix E.1.

Assumption 1 **(Parameter Set)**. *There exists a positive constant* $x_{\max}$ *such that the context probability density* p_X *has no support outside the ball of radius* $x_{\max}$*, i.e.,* $\|X_t\|_2 \leq x_{\max}$ *for all* t*. There also exists a constant* $b_{\max}$ *such that* $\|\beta_i\|_2 \leq b_{\max}$ *for all* $i \in [K]$*.*

Second, we assume that the context probability density p_X satisfies a margin condition, which comes from the classfication literature (Tsybakov 2004). We do not require this assumption to prove convergence of the greedy algorithm, but the rate of convergence differs depending on whether it holds. In particular, Goldenshluger and Zeevi (2009) prove matching upper and lower bounds demonstrating that all bandit algorithms achieve $\mathcal{O}(\log T)$ regret when the margin condition holds, but they can achieve up to $\mathcal{O}(\sqrt{T})$ regret when this condition is violated. We show analogous results hold for the simple greedy algorithm as well (see Appendix E.2 for an analysis when Assumption 2 is violated). This is because the margin condition rules out unusual context distributions that become unbounded near the decision boundary (which has zero measure), thereby making learning difficult. Our regret bounds for the greedy algorithm match those of standard bandit algorithms in the literature with and without the margin condition.

Assumption 2 **(Margin Condition)**. *There exists a constant* $C_0 > 0$ *such that for each* $\kappa > 0$*:*

$$\forall \ i \neq j: \qquad \mathbb{P}_X\left[0 < |X^\top(\beta_i - \beta_j)| \leq \kappa\right] \leq C_0 \kappa \, .$$

Thus far, we have made generic assumptions that are standard in the bandit literature. Our third assumption introduces the covariate diversity condition, which is essential for proving that the greedy algorithm always converges to the optimal policy. This condition guarantees that no matter what our arm parameter estimates are at time t, there is a diverse set of possible contexts (supported by the context probability density p_X) under which each arm may be chosen.

Assumption 3 **(Covariate Diversity)**. *There exists a positive constant* λ_0 *such that for each vector* $\mathbf{u} \in \mathbb{R}^d$ *we have*

$$\lambda_{\min}\Big(\mathbb{E}_X\left[XX^\top \mathbb{I}\{X^\top \mathbf{u} \geq 0\}\right]\Big) \geq \lambda_0 \, .$$

Assumption 3 holds for a general class of distributions. For instance, if the context probability density p_X is bounded below by a nonzero constant in an open set around the origin, then it would satisfy covariate diversity. This includes common distributions such as the uniform or truncated gaussian distributions. Furthermore, discrete distributions such as the classic Rademacher distribution on binary random variables also satisfy covariate diversity.

2.2. Example Distributions satisfying Assumptions

While Assumptions 1-2 are generic, Assumption 3 does not seem straightforward to verify. The following lemma provides sufficient conditions (that are easier to check) that guarantee Assumption 3; the proof is provided in Appendix A.

LEMMA 1. *If there exists a set $W \subset \mathbb{R}^d$ that satisfies conditions (a), (b), and (c) given below, then p_X satisfies Assumption 3.*

(a) W is symmetric around the origin; i.e., if $\mathbf{x} \in W$ then $-\mathbf{x} \in W$.

(b) There exist positive constants $a, b \in \mathbb{R}$ such that for all $\mathbf{x} \in W$, $a \cdot p_X(-\mathbf{x}) \leq b \cdot p_X(\mathbf{x})$.

(c) There exists a positive constant λ such that $\int_W \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x})\mathrm{d}\mathbf{x} \succeq \lambda I_d$. For discrete distributions, the integral is replaced with a sum.

We now use Lemma 1 to demonstrate that covariate diversity holds for a wide range of continuous and discrete context distributions, and we explicitly provide the corresponding constants. It is straightfoward to verify that these examples also satisfy Assumptions 1 and 2.

1. **Uniform Distribution.** Consider the uniform distribution over an arbitrary bounded set V that contains the origin. Then, there exists some $R > 0$ such that $B_R^d \subset V$. Taking $W = B_R^d$, we note that conditions (a) and (b) of Lemma 1 follow immediately. We now check condition (c) by first stating the following lemma (see Appendix A for proof):

   LEMMA 2. $\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left[\frac{R^2}{d+2}\mathrm{vol}(B_R^d)\right] I_d$ *for any $R > 0$.*

   By definition, $p_X(\mathbf{x}) = 1/\mathrm{vol}(V)$ for all $\mathbf{x} \in V$, and $\mathrm{vol}(B_R^d) = R^d \mathrm{vol}(B_{x_{\max}}^d)/x_{\max}^d$. Applying Lemma 2, we see that condition (c) of Lemma 1 holds with constant $\lambda = R^{d+2}/[(d+2)x_{\max}^d]$.

2. **Truncated Multivariate Gaussian Distribution.** Let p_X be a multivariate Gaussian distribution $\mathrm{N}(\mathbf{0}_d, \Sigma)$, truncated to 0 for all $\|\mathbf{x}\|_2 \geq x_{\max}$. The density after renormalization is

$$p_X(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x}\right)}{\int_{B_{x_{\max}}^d} \exp\left(-\frac{1}{2}\mathbf{z}^\top \Sigma^{-1}\mathbf{z}\right)\mathrm{d}\mathbf{z}} \mathbb{I}(\mathbf{x} \in B_{x_{\max}}^d).$$

   Taking $W = B_{x_{\max}}^d$, conditions (a) and (b) of Lemma 1 follow immediately. Condition (c) of Lemma 1 holds with constant $\lambda = \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \frac{x_{\max}^2}{d+2} \mathrm{vol}(B_{x_{\max}}^d)$, as shown in Lemma 6 in Appendix A.

3. **Gibbs Distributions with Positive Covariance.** Consider the set $\{\pm 1\}^d \subset \mathbb{R}^d$ equipped with a discrete probability density p_X, which satisfies

$$p_X(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{1 \leq i,j \leq d} J_{ij} x_i x_j\right) \cdot \delta(\mathbf{x}),$$

   for any $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \{\pm 1\}^d$. Here, $\delta(\mathbf{x})$ is the Dirac delta function, $J_{ij} \in \mathbb{R}$ are (deterministic) parameters, and Z is a normalization term known as the *partition function* in the statistical physics literature. We define $W = \{\pm 1\}^d$, satisfying conditions (a) and (b) of Lemma 1. Furthermore, condition (c) follows by definition since the covariance of the distribution is positive-definite. This class of distributions includes the well-known Rademacher distribution (by setting all $J_{ij} = 0$).

Finally, note that any product of these distributions would also satisfy our assumptions.

3. Greedy Bandit

Notation. Let the *design matrix* $\mathbf{X}$ be the $T \times d$ matrix whose rows are X_t. Similarly, for $i \in [K]$, let Y_i be the length T vector of potential outcomes $X_t^\top \beta_i + \varepsilon_{i,t}$. Since we only obtain feedback when arm i is played, entries of Y_i may be missing. For any $t \in [T]$, let $\mathcal{S}_{i,t} = \{j \mid \pi_j = i\} \cap [t]$ be the set of times when arm i was played within the first t time steps. We estimate β_i at time t based on the observations $\mathbf{X}(\mathcal{S}_{i,t}), Y(\mathcal{S}_{i,t})$ using ordinary least squares (OLS) regression (defined below). We denote this estimator $\hat{\beta}_{\mathbf{X}(\mathcal{S}_{i,t}), Y(\mathcal{S}_{i,t})}$, or $\hat{\beta}(\mathcal{S}_{i,t})$ for short.

Definition 2 (OLS Estimator). For any $\mathbf{X}_0 \in \mathbb{R}^{n \times d}$ and $Y_0 \in \mathbb{R}^{n \times 1}$, the OLS estimator is $\hat{\beta}_{\mathbf{X}_0, Y_0} \equiv \arg\min_\beta \|Y_0 - \mathbf{X}_0 \beta\|_2^2$, which is equal to $(\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top Y_0$ when $\mathbf{X}_0^\top \mathbf{X}_0$ is invertible.

We now describe the greedy algorithm and provide performance guarantees when covariate diversity holds (see Appendix E for extensions to more general rewards and more general margin conditions), and probabilistic performance guarantees when covariate diversity does not hold.

3.1. Algorithm

At each time step, we observe a new context X_t and use the current arm estimates $\hat{\beta}(\mathcal{S}_{i,t-1})$ to play the arm with the highest estimated reward, i.e., $\pi_t = \arg\max_{i \in [K]} X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$. Upon playing arm π_t, a reward $Y_t = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$ is observed. We then update our estimate for arm π_t (we need not update the arm parameter estimates for other arms as $\hat{\beta}(\mathcal{S}_{i,t-1}) = \hat{\beta}(\mathcal{S}_{i,t})$ for $i \neq \pi_t$). The update formula is given by

$$\hat{\beta}(\mathcal{S}_{\pi_t,t}) = \left[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\right]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{Y}(\mathcal{S}_{\pi_t,t})\,.$$

We do not update the parameter of arm π_t if $\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ is not invertible. The pseudo-code for the algorithm is given in Algorithm 1.

3.2. Performance of Greedy Bandit with Covariate Diversity

We establish an upper bound of $\mathcal{O}(\log T)$ on the cumulative expected regret of the Greedy Bandit for the two-armed contextual bandit when covariate diversity is satisfied.

Algorithm 1 Greedy Bandit

Initialize $\hat{\beta}(\mathcal{S}_{i,0}) = 0$ for $i \in [K]$
for $t \in [T]$ **do**
  Observe $X_t \sim p_X$
  $\pi_t \leftarrow \arg\max_i X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$ (break ties randomly)
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
  Play arm π_t, observe $Y_t = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$
  Update the arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t}) \leftarrow \Big[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\Big]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{Y}(\mathcal{S}_{\pi_t,t})$
end for

THEOREM 1. *If $K = 2$ and Assumptions 1-3 are satisfied, the cumulative expected regret of the Greedy Bandit at time T is at most*

$$\begin{aligned} R_T(\pi) &\le \frac{C_0 \tilde{C} x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} \log T + \tilde{C}\left(\frac{C_0 x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} + \frac{b_{\max} x_{\max}^3 d}{\lambda_0} + 2 x_{\max} b_{\max}\right) \\ &\le C_{GB} \log T = \mathcal{O}(\log T) \ , \end{aligned} \tag{1}$$

where the constant C_0 is defined in Assumption 2 and

$$\bar{C} = \left(\frac{1}{3} + \frac{7}{2}(\log d)^{-0.5} + \frac{38}{3}(\log d)^{-1} + \frac{67}{4}(\log d)^{-1.5}\right) \in (1/3, 52)\,. \tag{2}$$

REMARK 1. Goldenshluger and Zeevi (2013) established a lower bound of $\mathcal{O}(\log T)$ for any algorithm in a two-armed contextual bandit. While they do not make Assumption 3, the distribution used in their proof satisfies Assumption 3; thus the result applies to our setting. Combined with our upper bound, this demonstrates that the Greedy Bandit achieves rate-optimal cumulative regret.

3.3. Proof of Theorem 1

Notation. Let $\mathcal{R}_i = \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \beta_i \ge \max_{j \ne i} \mathbf{x}^\top \beta_j\right\}$ denote the true set of contexts where arm i is optimal. Then, let $\hat{\mathcal{R}}_{i,t} = \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \ge \max_{j \ne i} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{j,t-1})\right\}$ denote the estimated set of contexts at time t where arm i appears optimal; in other words, if the context $X_t \in \hat{\mathcal{R}}_{i,t}$, then the greedy policy will choose arm i at time t. (For brevity, we assume without loss of generality that ties are broken randomly and thus, $\{\mathcal{R}_i\}_{i=1}^K$ and $\{\hat{\mathcal{R}}_{i,t}\}_{i=1}^K$ partition the context space $\mathcal{X}$.)

For any $t \in [T]$, let $\mathcal{H}_{t-1} = \sigma\left(\mathbf{X}_{1:t}, \pi_{1:t-1}, Y_1(\mathcal{S}_{1,t-1}), Y_2(\mathcal{S}_{2,t-1}), \ldots, Y_K(\mathcal{S}_{K,t-1})\right)$ denote the σ-algebra containing all observed information up to time t before taking an action; thus, our policy π_t is $\mathcal{H}_{t-1}$-measurable. Let $\mathbb{E}_k$ be the conditional expectation with respect to the filtration $\mathcal{H}_k$.

Define $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ as the sample covariance matrix for observations from arm i up to time t. We may compare this to the expected covariance matrix for arm i under the greedy policy, defined as $\tilde{\Sigma}_{i,t} = \sum_{k=1}^t \mathbb{E}_{k-1}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{i,k}^\pi]\right]$.

Proof Strategy. Intuitively, covariate diversity (Assumption 3) guarantees that there is sufficient randomness in the observed contexts, which creates natural "exploration." In particular, no matter what our current arm parameter estimates $\{\hat{\beta}(\mathcal{S}_{1,t}), \hat{\beta}(\mathcal{S}_{2,t})\}$ are at time t, each arm will be chosen by the greedy policy with at least some constant probability (with respect to p_X) depending on the observed context. We formalize this intuition in the following lemma.

Lemma 3. *Given Assumptions 1 and 3, the following holds for any $u \in \mathbb{R}^d$:*

$$\mathbb{P}_X[X^T u \geq 0] \geq \frac{\lambda_0}{x_{\max}^2}\,.$$

Proof. For any observed context x, note that $xx^T \leq x_{\max}^2 I_d$ by Assumption 1. Re-stating Assumption 3 for each $u \in \mathbb{R}^d$, we can write

$$\lambda_0 I_d \ \leq \ \int xx^T \mathbb{I}(x^T u \geq 0) p_X(x) \mathrm{d}x \ \leq \ x_{\max}^2 I_d \int \mathbb{I}(x^T u \geq 0) p_X(x) \mathrm{d}x \ = \ x_{\max}^2 \mathbb{P}_X[X^T u \geq 0] I_d,$$

since the indicator function and p_X are both nonnegative. □

Taking $u = \hat{\beta}(\mathcal{S}_{1,t}) - \hat{\beta}(\mathcal{S}_{2,t})$, Lemma 3 implies that arm 1 will be pulled with probability at least $\lambda_0/x_{\max}^2$ at each time t; the claim holds analogously for arm 2. Thus, each arm will be played at least $\lambda_0 T/x_{\max}^2 = \mathcal{O}(T)$ times in expectation. However, this is not sufficient to guarantee that each arm parameter estimate $\hat{\beta}_i$ converges to the true parameter β_i. In Lemma 4, we establish a sufficient condition for convergence.

First, we show that covariate diversity guarantees that the minimum eigenvalue of each arm's expected covariance matrix $\tilde{\Sigma}_{i,t}$ under the greedy policy grows linearly with t. This result implies that not only does each arm receive a sufficient number of observations under the greedy policy, but also that these observations are sufficiently diverse (in expectation). Next, we apply a standard matrix concentration inequality (see Lemma 8 in Appendix B) to show that the minimum eigenvalue of each arm's sample covariance matrix $\hat{\Sigma}(\mathcal{S}_{i,t})$ also grows linearly with t. This will guarantee the convergence of our regression estimates for each arm parameter.

Lemma 4. *Take $C_1 = \lambda_0/(40x_{\max}^2)$. Given Assumptions 1 and 3, the following holds for the minimum eigenvalue of the empirical covariance matrix of each arm $i \in [2]$:*

$$\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda_0 t/4\right] \geq 1 - \exp(\log d - C_1 t)\,.$$

Proof. Without loss of generality, take $i = 1$. For any $k \leq t$, let $\mathbf{u}_k = \hat{\beta}(\mathcal{S}_{1,k}) - \hat{\beta}(\mathcal{S}_{2,k})$; by the greedy policy, we pull arm 1 if $X_k^\top \mathbf{u}_{k-1} > 0$ and arm 2 if $X_k^\top \mathbf{u}_{k-1} < 0$ (ties are broken randomly using a fair coin flip W_k). Thus, the estimated set of optimal contexts for arm 1 is

$$\hat{\mathcal{R}}_{1,k} = \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \mathbf{u}_{k-1} > 0\right\} \cup \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \mathbf{u}_{k-1} = 0, W_k = 0\right\}.$$

First, we seek to bound the minimum eigenvalue of the expected covariance matrix $\tilde{\Sigma}_{1,t} = \sum_{k=1}^{t} \mathbb{E}_{k-1}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{1,k}]\right]$. Expanding one term in the sum, we can write

$$\begin{aligned}
\mathbb{E}_{k-1}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{1,k}]\right] &= \mathbb{E}_{k-1}\left[X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0)\right]\right. \\
&= \mathbb{E}_X\left[X X^\top \left(\mathbb{I}[X^\top \mathbf{u}_{k-1} > 0] + \frac{1}{2}\mathbb{I}[X^\top \mathbf{u}_{k-1} = 0]\right)\right] \\
&\geq \lambda_0/2\,,
\end{aligned}$$

where the last line follows from Assumption 3. Since the minimum eigenvalue function $\lambda_{\min}(\cdot)$ is concave over positive semi-definite matrices, we can write

$$\begin{aligned}
\lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right) &= \lambda_{\min}\left(\sum_{k=1}^{t} \mathbb{E}_{k-1} X X^\top \mathbb{I}[X \in \hat{\mathcal{R}}_{1,k}]\right) \\
&\geq \sum_{k=1}^{t} \lambda_{\min}\left(\mathbb{E}_{k-1} X X^\top \mathbb{I}[X \in \hat{\mathcal{R}}_{1,k}]\right) \geq \frac{\lambda_0 t}{2}\,.
\end{aligned}$$

Next, we seek to use matrix concentration inequalities (Lemma 8 in Appendix B) to bound the minimum eigenvalue of the sample covariance matrix $\hat{\Sigma}(\mathcal{S}_{1,t})$. To apply the concentration inequality, we also need to show an upper bound on the maximum eigenvalue of $X_k X_k^\top$; this follows trivially from Assumption 1 using the Cauchy-Schwarz inequality:

$$\lambda_{\max}(X_k X_k^\top) = \max_{\mathbf{u}} \frac{\|X_k X_k^\top \mathbf{u}\|_2}{\|\mathbf{u}\|_2} \leq \frac{\|X_k\|_2^2 \|\mathbf{u}\|_2}{\|\mathbf{u}\|_2} \leq x_{\max}^2.$$

We can now apply Lemma 8, taking the finite adapted sequence $\{X_k\}$ to be $\left\{X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{1,k}]\right\}$, so that $Y = \hat{\Sigma}(\mathcal{S}_{1,t})$ and $W = \tilde{\Sigma}_{1,t}$. We also take $R = x_{\max}^2$ and $\gamma = 1/2$. Thus, we have

$$\begin{aligned}
\mathbb{P}_X\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{1,t})\right) \leq \frac{\lambda_0 t}{4} \;\text{ and }\; \lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right) \geq \frac{\lambda_0 t}{2}\right] &\leq d\left(\frac{e^{-0.5}}{0.5^{0.5}}\right)^{\frac{\lambda_0}{4x_{\max}^2} t} \\
&\leq \exp\left(\log d - \frac{0.1\lambda_0}{4x_{\max}^2} t\right),
\end{aligned}$$

using the fact $-0.5 - 0.5\log(0.5) \leq -0.1$. As we showed earlier, $\mathbb{P}_X\left(\lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right) \geq \frac{\lambda_0 t}{2}\right) = 1$. This proves the result. $\square$

Next, Lemma 5 guarantees with high probability that each arm's parameter estimate has small ℓ_2 error with respect to the true parameter if the minimum eigenvalue of the sample covariance matrix $\hat{\Sigma}(\mathcal{S}_{i,t})$ has a positive lower bound. Note that we cannot directly use results on the convergence of the OLS estimator since the set of samples $\mathcal{S}_{i,t}$ from arm i at time t are not i.i.d.[2] Instead, we use a Bernstein concentration inequality to guarantee convergence with adaptive observations.

[2] We use the arm estimate $\hat{\beta}(\mathcal{S}_{i,t-1})$ to decide whether to play arm i at time t; thus, the samples in $\mathcal{S}_{i,t}$ are correlated.

LEMMA 5. *Taking $C_2 = \lambda^2/(2d\sigma^2 x_{\max}^2)$ and $n \geq |S_{i,t}|$, we have for all $\lambda, \chi > 0$,*

$$\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \quad \textit{and} \quad \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \leq 2d \exp\left(-C_2 t^2 \chi^2/n\right).$$

Proof of Lemma 5. We begin by noting that if the event $\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t$ holds, then

$$\begin{aligned}
\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 &= \| \left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right)^{-1} \mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \\
&\leq \| \left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right)^{-1} \|_2 \|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \ \leq \ \frac{1}{\lambda t}\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2.
\end{aligned}$$

As a result, we can write

$$\begin{aligned}
&\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \quad \text{and} \quad \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\
&= \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \ \mid \ \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\
&\leq \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda \ \mid \ \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\
&\leq \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda\right] \\
&\leq \sum_{r=1}^{d} \mathbb{P}\left[|\varepsilon(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})^{(r)}| \geq \frac{\lambda t \cdot \chi}{\sqrt{d}}\right],
\end{aligned}$$

where $\mathbf{X}^{(r)}$ denotes the r^{th} column of $\mathbf{X}$. We can expand

$$\varepsilon(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})^{(r)} = \sum_{j=1}^{t} \varepsilon_j X_{j,r} \mathbb{I}\left[j \in \mathcal{S}_{i,j}\right].$$

For simplicity, define $D_j = \varepsilon_j X_{j,r} \mathbb{I}\left[j \in \mathcal{S}_{i,j}\right]$. First, note that D_j is $(x_{\max}\sigma)$-subgaussian, since ε_j is σ-subgaussian and $|X_{j,r}| \leq x_{\max}$. Next, note that $X_{j,r}$ and $\mathbb{I}\left[j \in \mathcal{S}_{i,j}\right]$ are both $\mathcal{H}_{j-1}$ measurable; taking the expectation gives $\mathbb{E}[D_j \mid \mathcal{H}_{j-1}] = X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{i,j}\right] \mathbb{E}[\varepsilon_j \mid \mathcal{H}_{j-1}] = 0$. Thus, the sequence $\{D_j\}_{j=1}^{t}$ is a martingale difference sequence adapted to the filtration $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_t$. Applying a standard Bernstein concentration inequality (see Lemma 7 in Appendix B), we can write

$$\mathbb{P}\left[\Big|\sum_{j=1}^{t} D_j\Big| \geq \frac{\lambda t \cdot \chi}{\sqrt{d}}\right] \leq 2 \exp\left(-\frac{t^2 \lambda^2 \chi^2}{2d\sigma^2 x_{\max}^2 n}\right),$$

where n is an upper bound on the number of nonzero terms in above sum, i.e., an upper bound on $|\mathcal{S}_{i,t}|$. This yields the desired result. □

To summarize, Lemma 4 provides a lower bound (with high probability) on the minimum eigenvalue of the sample covariance matrix. Lemma 5 states that if such a bound holds on the minimum eigenvalue of the sample covariance matrix, then the estimated parameter $\hat{\beta}(\mathcal{S}_{i,t})$ is close to the true β_i (with high probability). Having established convergence of the arm parameters under the Greedy Bandit, we use a standard peeling argument (see Appendix C) to bound the cumulative regret of the Greedy Bandit and prove Theorem 1.

3.4. Performance of Greedy Bandit without Covariate Diversity

In this section, we build on more general versions of two concentration inequalities stated in Lemmas 5 and 8 to derive probabilistic guarantees on the performance of Greedy Bandit when covariate diversity does not hold or when there are more than two arms, i.e., $K > 2$. Having these probabilistic bounds helps us to confirm that Greedy Bandit algorithm can effectively reduce the experimentation in some settings.

Assumptions. We now remove our additional assumptions: namely, that there are only two arms and Assumption 3 on the diversity of covariates. Instead, we will make a weaker assumption that with a slight modification, is typically made in the contextual bandit literature (see Goldenshluger and Zeevi (2013) and Bastani and Bayati (2015)). In particular, this assumption allows for multiple arms (some of which may be uniformly sub-optimal), and relaxes the assumption on the covariates, e.g., allowing for intercept terms.

Assumption 4 **(Positive-Definiteness)**. *Let $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ be mutually exclusive sets that include all K arms. Sub-optimal arms $i \in \mathcal{K}_{sub}$ satisfy $\mathbf{x}^\top \beta_i < \max_{j \neq i} \mathbf{x}^\top \beta_j - h$ for some $h > 0$ and every $\mathbf{x} \in \mathcal{X}$. On the other hand, each optimal arm $i \in \mathcal{K}_{opt}$, has a corresponding set $U_i = \{\mathbf{x} \mid \mathbf{x}^\top \beta_i > \max_{j \neq i} \mathbf{x}^\top \beta_j + h\}$. Define $\Sigma_i \equiv \mathbb{E}[XX^\top \mathbb{I}(X \in U_i)]$ for all $i \in \mathcal{K}_{opt}$. Then, there exists $\lambda_1 > 0$ such that for all $i \in \mathcal{K}_{opt}$, $\lambda_{\min}(\Sigma_i) \geq \lambda_1 > 0$.*

Remark 2. This assumption is slightly different from two assumptions made in analysis of OLS Bandit in Goldenshluger and Zeevi (2013) and Bastani and Bayati (2015) (see Assumptions 5 and 6 in Appendix D for the comparison). In general, having these two assumptions always implies Assumption 4. The other direction is not always true, but for bounded distributions p_X it is actually true. In other words, Assumptions 1 and 4 imply Assumptions 5 and 6. For completeness, we included these two assumptions in Appendix D.

Having this less restrictive assumption, we now move to state and prove our probabilistic guarantees on the performance of Greedy Bandit for a larger class of distributions than those satisfy the covariate diversity assumption. In particular, problem instances that we consider can violate covariate diversity (Assumption 3) or have more than two arms ($K > 2$). In our analysis we consider a slightly different version of Greedy Bandit algorithm where m random samples are drawn in the beginning for each arm. This might seem to be contradictory to our initial goal of reducing the experimentation, but it is worth noting that: 1) for effectively estimating arm parameters using any estimation method, we might roughly need d samples for each arm (for instance, in the case of OLS, at least d samples are needed to ensure the invertibility of covariance matrix), and 2) value of m here is fixed and can be taken as small as $m = d$, comparing to other algorithms that require indefinite exploration. These bounds highly depend on the distribution of covariates p_X and can even guarantee good performance when $m = d$. From here, we assume m to be an arbitrary positive integer, while later we discuss how these bounds scale when we vary m.

To establish these bounds, we derive lower bounds on the probability of the event that all arms estimates remain within $\theta_1 = h/(2x_{\max})$ euclidean distance from their true value for all time periods. In fact, as we observe later, having this condition ensures that Greedy Bandit algorithm succeeds and achieves logarithmic regret. These probability bounds can be derived using Lemma 5 once we prove suitable lower bounds on the minimum eigenvalue of covariance matrices. In particular, for time periods $1 \le t \le p-1$ we consider a lower bound equal to a constant number δ, while for $t \ge p$ we let this to be a linear bound equal to $(1-\gamma)\lambda_1 t - m|\mathcal{K}_{sub}|$. The probability that one of these events does not occur is exactly described in the first and fourth terms in the function $L(\gamma, \delta, p)$ below, respectively. The remaining terms are the probability that an arm estimate deviate more than θ_1 from its true value computed over different time period intervals. Parameters γ, δ, and p can be chosen arbitrarily and we optimize over their choice.

Theorem 2. *Suppose Assumptions 1, 2, and 4 hold. Assume that Greedy Bandit algorithm with m rounds (per arm) of forced sampling in the beginning is executed. Then, Greedy Bandit algorithm succeeds and achieves logarithmic regret with probability at least equal to*

$$S^{gb}(m, K, \sigma, x_{\max}, \lambda_1, h) := 1 - \inf_{\gamma \in (0,1), \delta > 0, p \ge Km+1} L(\gamma, \delta, p)\,, \tag{3}$$

where the function $L(\gamma, \delta, p)$ is defined as

$$\begin{aligned} L(\gamma,\delta,p) := {} & \underbrace{1 - \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \ge \delta\right]^K}_{\mathbb{P}(\exists i : \lambda_{\min}(\hat{\Sigma}_{i,Km}) \le \delta)} + \underbrace{2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \ge \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\}}_{\mathbb{P}(\exists i : \|\hat{\beta}(\mathcal{S}_{i,Km}) - \beta_i\|_2 \ge \theta_1)} \\ & + \underbrace{\sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\}}_{\mathbb{P}(\exists t \in [Km+1, p-1] : \|\hat{\beta}(\mathcal{S}_{\pi_t,t}) - \beta_i\|_2 \ge \theta_1)} + \underbrace{\frac{d \exp\left(-D_1(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_1(\gamma))}}_{\mathbb{P}(\exists t \ge p : \lambda_{\min}(\hat{\Sigma}_{\pi_t,t}) \le (1-\gamma)\lambda_1 t - m|\mathcal{K}_{sub}|)} \\ & + \underbrace{\frac{2d \exp\left(-D_2(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_2(\gamma))}}_{\mathbb{P}(\exists t \ge p : \|\hat{\beta}(\mathcal{S}_{\pi_t,t}) - \beta_i\|_2 \ge \theta_1)}\,. \end{aligned} \tag{4}$$

Here $\mathbf{X}_{1:m}$ denotes the matrix obtained by drawing m random samples from distribution p_X and constants are

$$D_1(\gamma) = \frac{\lambda_1(\gamma + (1-\gamma)\log(1-\gamma))}{x_{\max}^2}\,, \tag{5}$$

$$D_2(\gamma) = \frac{\lambda_1^2 h^2 (1-\gamma)^2}{8d\sigma^2 x_{\max}^4}\,. \tag{6}$$

The proof of this theorem is provided in Appendix G. Key steps of the proof are as follows:

1. Under assumption that the smallest eigenvalue of all covariance matrices are above some arbitrary number $\delta > 0$, derive a lower bound on the probability that after Km rounds, all the $\hat{\beta}$s are in balls with the radius $\theta_1 = h/(2x_{\max})$ around their true values (Lemma 10).
2. Derive a lower bound on the probability that these estimates remain on the same ball after $p \geq Km+1$ rounds for an arbitrary p (Lemma 11).
3. Use concentration result in Lemma 8 to derive a lower bound on the probability that the minimum eigenvalue of the covariance matrix of all arms in $\mathcal{K}_{opt}$ are above $(1-\gamma)\lambda_1(t - m|\mathcal{K}_{sub}|)$ for any $t \geq p$. Here, p should be chosen in such a way that sum of these probability terms converges to a positive number which lies in $[0,1]$ (Lemma 12).
4. Derive a lower bound on the probability that the estimates ultimately remain inside the ball with radius θ_1. Note that this further ensures that no sub-optimal arm is played for any $t \geq Km$ (Lemma 13).
5. Summing up these probability terms implies Theorem 2.

Note that the concentration result in Lemma 8 that was used in the third step above can also be used to derive a lower bound on the probability term $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]^K$ appeared in the function S^{gb}. However, this concentration inequality is designed to work for a general and arbitrary distribution and usually leads to error probabilities above 1 when m is very close to d. As we demonstrate later via examples, the bound in Theorem 2 works even for $m = d$ for some distributions. Nevertheless, we apply the concentration result in Lemma 8 to get the following corollary which provides a simpler and more readable version of Theorem 2. The proof is provided in Appendix G.

Corollary 1. *Under the similar assumptions made in Theorem 2, the Greedy Bandit algorithm executed with m-rounds of forced sampling in the beginning enjoys the logarithmic regret with probability at least equal to:*

$$1 - \frac{3Kd\exp(-D_{\min}|\mathcal{K}_{opt}|m)}{1-\exp(-D_{\min})},$$

where function $D_{\min}$ is defined as

$$D_{\min} = \min\left\{\frac{0.153\lambda_1}{x_{\max}^2}, \frac{\lambda_1^2 h^2}{32 d\sigma^2 x_{\max}^4}\right\}. \tag{7}$$

The following Proposition reveals some of the properties of the function S^{gb}. In particular, this proposition shows that S^{gb} is non-increasing with respect to σ and provides its limit when σ goes to zero. The proof of this proposition is provided in Appendix G

Proposition 1. *The function* $S^{gb}(m, K, \sigma, x_{\max}, \lambda_1, h)$ *defined in Equation* (3) *is non-increasing with respect to* σ*, and is non-decreasing with respect to* m*,* λ_1 *and* h*. Furthermore, if one keeps the total random sampling rounds* Km*, and the number of samples allocated to sub-optimal arms* $|\mathcal{K}_{sub}|m$ *fixed, then it is non-increasing with respect to* K*. Finally, the limit of this function when* σ *goes to zero is equal to*

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K.$$

The derived probability bound on the success of Greedy Bandit algorithm suggests that this algorithm can be effective and optimal at least for a part of the probability space. Therefore, it is interesting to seek for a policy that follows the Greedy Bandit algorithm and uses data to check whether the greedy updates are actually converging or not. Having this in mind, we are ready to introduce Greedy-First algorithm.

4. Greedy-First Algorithm

As discussed earlier, we made two additional assumptions than typically required for the convergence of bandit algorithms. In particular, our theoretical guarantees in Theorem 1 do not hold for the Greedy Bandit if (i) there are more than two arms, i.e., $K > 2$, or (ii) there is insufficient diversity of covariates (Assumption 3). The second condition rules out some standard settings. For instance, the arm rewards cannot have an intercept term (since the addition of a column of ones to all covariates would violate Assumption 3). Alternatively, an arm cannot be uniformly sub-optimal across all covariates. While there are many examples that satisfy these conditions (see §2.2), the decision-maker may not know a priori whether her particular setting satisfies these assumptions. Thus, we introduce the Greedy-First algorithm that uses observed data to determine whether a greedy algorithm will converge.

4.1. Algorithm

Greedy-First algorithm receives as the input λ_0 and t_0, and starts by taking greedy actions according to Greedy Bandit algorithm up to time t_0. After time t_0 in each iteration it checks whether the minimum eigenvalue of sample covariance matrices of all K arms are greater than or equal to $\lambda_0 t/4$. If this condition is satisfied, it means that the covariates for each arm are diverse enough to guarantee the convergence of greedy estimates, with a high probability. On the other hand, if this condition is not met, the algorithm switches to OLS Bandit algorithm which is introduced by Goldenshluger and Zeevi (2013) for two arms and extended to the general setting by Bastani and Bayati (2015). For completeness we have presented OLS Bandit in Appendix D. Note that, once the algorithm switches from Greedy Bandit to OLS Bandit, it keeps taking actions according to OLS Bandit and it does not switch any longer.

Algorithm 2 Greedy-First Bandit

Input parameters: λ_0, t_0
Initialize $\hat{\beta}(\mathcal{S}_{i,0})$ at random for $i \in [K]$
Initialize switch to $R = 0$
for $t \in [T]$ **do**
  if $R \neq 0$ **then** break
  end if
  Observe $X_t \sim p_X$
  $\pi_t \leftarrow \arg\max_i X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$ (break ties randomly)
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
  Play arm π_t, observe $Y_t = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$
  Update arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t}) = \left[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\right]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{Y}(\mathcal{S}_{\pi_t,t})$
  Compute covariance matrices $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ for $i \in [K]$
  if $t > t_0$ and $\min_{i \in [K]} \lambda_{\min}\left(\hat{\Sigma}_(\mathcal{S}_{i,t})\right) < \frac{\lambda_0 t}{4}$ **then**
    Set $R = t$
  end if
end for
Execute OLS Bandit for $t \in [R+1, T]$

In practice, λ_0 may be an unknown constant. Thus, we suggest the following heuristic routine to estimate this parameter. For t_0 time steps the decision maker executes Greedy Bandit. Then, the parameter λ_0 can be estimated using the observed data via $\hat{\lambda}_0 = \frac{1}{2t_0} \min_{i \in [K]} \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t_0})\right)$. If $\hat{\lambda}_0 = 0$, this suggests that one of the arms may not be receiving sufficient samples and thus, Greedy-First will switch to OLS Bandit immediately. Otherwise, the decision-maker executes Greedy-First Bandit for $t \in [t_0+1, T]$ with $\lambda_0 = \hat{\lambda}_0$. The pseudo-code version of this heuristic is shown in Appendix D. Note that if t_0 is small, then this technique might incorrectly switch to OLS Bandit even when a greedy algorithm may converge; thus, choosing $t_0 \gg Kd$ is advisable. Note that the regret guarantees for Greedy-First are always valid, but upon choosing a very large λ_0 Greedy-First might wrongfully switch even when Greedy itself converges. On the other hand, if λ_0 is small Greedy-First may switch late that leads to a larger regret.

Remark 3. Greedy-First can switch to any contextual bandit algorithm (e.g., OFUL by Abbasi-Yadkori et al. (2011) or Thompson sampling by Agrawal and Goyal (2013), Russo and Van Roy (2014b)) instead of the OLS Bandit. Then, the above assumptions would be replaced with analogous assumptions required by that algorithm. Our proof naturally generalizes to adopt the assumptions and regret guarantees of the new algorithm when the Greedy Bandit fails.

4.2. Regret Analysis of Greedy-First

We now establish an upper bound of $\mathcal{O}(\log T)$ on the expected cumulative regret of the Greedy-First algorithm under Assumptions 1, 2, 5, and 6 (or equivalently Assumptions 1, 2, and 4), which will be proved in Appendix C.

Theorem 3. *The cumulative expected regret of the Greedy-First Bandit at time T is at most*

$$C\log T + 2t_0 x_{\max} b_{\max}\,,$$

where $C = (K-1)C_{GB} + C_{OB}$, C_{GB} is the constant defined in Theorem 1, and C_{OB} is the constant in the upper bound of the regret of the OLS Bandit algorithm.

Furthermore, if Assumption 3 is satisfied (with the specified parameter λ_0) and $K = 2$, then the Greedy-First algorithm will purely execute the greedy policy (and will not switch to the OLS Bandit algorithm) with probability at least $1-\delta$, where

$$\delta = 2d\exp[-t_0 C_1]/C_1\,,$$

and $C_1 = \lambda_0/40x_{\max}^2$. Note that δ can be made arbitrarily small since t_0 is an input parameter to the algorithm.

The key insight to this result is that the proof of Theorem 1 only requires Assumption 3 in the proof of Lemma 4. The remaining steps of the proof hold without the assumption. Thus, if the conclusion of Lemma 4, $\min_{i\in[K]}\lambda_{\min}(\hat{\Sigma}(\mathcal{S}_{i,t})) \geq \frac{\lambda_0 t}{4}$ holds at every $t \in [t_0+1, T]$, then we are guaranteed at most $\mathcal{O}(\log T)$ regret by Theorem 1, regardless of whether Assumption 3 holds. Notice that the constant C_{GB} in Theorem 1 is replaced by $(K-1)C_{GB}$, as we are considering K arms. This can also be easily implied from Lemma 9. Note that this condition must hold for Greedy-First to continue pursuing a purely greedy policy, and that happens with high probability when Assumption 3 is true. If this condition fails and Greedy-First switches to a OLS Bandit policy, then we can simply adopt the regret guarantees previously proven for this algorithm.

Proof of Theorem 3. First, we will show that Greedy-First achieves asymptotically optimal regret. Note that the expected regret during the first t_0 rounds is upper bounded by $2x_{\max}b_{\max}t_0$. For the period $[t_0+1, T]$ we consider two cases: (i) the algorithm pursues a purely greedy strategy, i.e., $R = 0$, or (ii) the algorithm switches to the OLS Bandit algorithm, i.e., $R \in [t_0+1, T]$.

Case 1: By construction, we know that

$$\min_{i\in[K]}\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \frac{\lambda_0 t}{4}\,,$$

for all $t > t_0$. This is true according to the fact that Greedy-First only switches when the minimum of minimum eigenvalues of the covariance matrices is less than $\lambda_0 t/4$. Therefore, if the algorithm does not switch, it basically implies that the minimum eigenvalue of all covariance matrices are greater that or equal to $\lambda_0 t/4$ for all values of $t > t_0$. Then, the conclusion of Lemma 4 holds in this time range ($\mathcal{F}_{i,t}^{\lambda}$ holds for all $i \in [K]$). Consequently, even if Assumption 3 does not hold and $K \neq 2$, Lemma 9 holds and proves an upper bound on the expected regret r_t. This implies that the regret bound of Theorem 1, after multiplying by $(K-1)$, holds for the Greedy-First. Therefore, Greedy-First is guaranteed to achieve $(K-1)C_{GB}\log(T-t_0)$ regret in the period $[t_0+1,T]$, for some constant C_{GB} that depends only on p_X, b and σ. Hence, the regret in this case is upper bounded by $2x_{\max}b_{\max}t_0 + (K-1)C_{GB}\log T$.

Case 2: Once again, by construction, we know that $\min_{i\in[K]}\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda_0 t/4$ for all $t \in [t_0+1,R]$. Then, using the same argument as above, Theorem 1 guarantees that we achieve at most $(K-1)C_{GB}\log(R-t_0)$ regret, for some constant C_{GB} over the interval $[t_0+1,R]$. Next, Theorem 2 of Bastani and Bayati (2015) guarantees that the OLS Bandit will achieve at most $C_{OB}\log(T-R)$ regret under Assumptions 1, 2, 5, 6 and when we apply this algorithm to $t \in [R+1,T]$. Thus, the total regret is at most $2x_{\max}b_{\max}t_0 + ((K-1)C_{GB}+C_{OB})\log T$. Note that although here R is a random variable depending on when algorithm switches, the upper bound $2x_{\max}b_{\max}t_0 + ((K-1)C_{GB}+C_{OB})\log T$ holds uniformly, regardless of what R actually is.

Thus, the Greedy-First algorithm always achieves $\mathcal{O}(\log T)$ cumulative expected regret. Next, we prove that when Assumption 3 holds and $K=2$, the Greedy-First algorithm maintains a purely greedy policy with high probability. In particular, Lemma 4 states that if the specified λ_0 satisfies

$$\lambda_{\min}\left(\mathbb{E}_X\left[XX^\top \mathbb{I}(X^\top \mathbf{u} \geq 0)\right]\right) \geq \lambda_0$$

for each vector $\mathbf{u}\in\mathbb{R}^d$, then at each time t and $C_1 = \lambda_0/40x_{\max}^2$,

$$\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \frac{\lambda_0 t}{4}\right] \geq 1-\exp\left[\log d - C_1 t\right] .$$

Thus, by using a union bound over all $K=2$ arms, the probability that the algorithm switches to the OLS Bandit algorithm is at most

$$\begin{aligned} K\sum_{t=t_0+1}^{T}\exp\left[\log d - C_1 t\right] &\leq 2\int_{t_0}^{\infty}\exp\left[\log d - C_1 t\right]\mathrm{d}t \\ &= \frac{2d}{C_1}\exp\left[-t_0 C_1\right] . \end{aligned}$$

This concludes the proof. $\square$

4.3. Probabilistic Guarantees for Greedy-First Algorithm

In §3.4 we used two concentration inequalities in Lemmas 5 and 8 to derive probabilistic guarantees on the performance of Greedy Bandit algorithm. As the main motivation behind introducing Greedy-First algorithm was to reduce the experimentation whenever the greedy policy was converging itself, it is natural to ask how effective Greedy-First is in achieving this goal. In other words, is it possible to derive a lower bound on the probability that Greedy-First algorithm becomes purely greedy and does not switch to the exploratory algorithm? We answered this question in Theorem 3 for the situation that $K = 2$ and the covariate-diversity assumption held. Can we replicate the approach in §3.4 to derive probabilistic bounds on the performance of Greedy-First, when instead of covariate diversity and $K = 2$, only Assumption 4 holds? We give an affirmative answer to this question and provide similar probabilistic bounds in this section. Based on Greedy-First's switching policy, we should only hope to have meaningful probabilistic bounds when all arms are optimal and no sub-optimal arm exists, i.e., $\mathcal{K}_{opt} = [K], \mathcal{K}_{sub} = \emptyset$. The reason is simple; in order to have a logarithmic regret, sub-optimal arms need to be assigned at a rate that is not larger than logarithmic in terms of t and therefore the minimum eigenvalue of the covariance matrices of such arms cannot grow better than logarithmic. This would cause Greedy-First algorithm to switch with probability 1. Therefore, we only consider the case $\mathcal{K}_{sub} = \emptyset$ in the analysis of Greedy-First. We are now ready to state our result.

Theorem 4. *Let Assumptions 1, 2, and 4 hold and suppose that $\mathcal{K}_{sub} = \emptyset$. Assume that the Greedy-First algorithm is executed with inputs λ_0, t_0,(assuming that Greedy Bandit algorithm being executed with m rounds of random sampling in the beginning for each arm), then with probability at least*

$$S^{gf}(m, K, \sigma, x_{\max}, \lambda_1, h) = 1 - \inf_{\gamma \le 1-\lambda_0/(4\lambda_1), \delta > 0, Km+1 \le p \le t_0} L'(\gamma, \delta, p)\,, \tag{8}$$

it becomes purely greedy, does not switch to the pre-selected exploratory algorithm, and achieves a logarithmic regret. Function L' is closely related to function L and is defined as

$$L'(\gamma, \delta, p) = L(\gamma, \delta, p) + \underbrace{(K-1)\frac{d\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))}}_{\mathbb{P}(\exists t \ge p, i \ne \pi_t : \lambda_{\min}(\hat{\Sigma}_{i,t}) \le (1-\gamma)\lambda_1 t)}\,. \tag{9}$$

Proof steps are similar to such steps for Greedy Bandit algorithm. Let us demonstrate why the bounds of Greedy-First and Greedy Bandit are different. Recall from the third step of the proof from §3.4 that we use concentration result in Lemma 8 to derive a lower bound on the probability that the minimum eigenvalue of the covariance matrix of all arms in $\mathcal{K}_{opt}$ are above $(1-\gamma)\lambda_1 t$ for any $t \ge p$ (note that we assumed $\mathcal{K}_{sub} = \emptyset$). For Greedy Bandit it is only required to include this probability for the *played arm*, while for Greedy-First we need to make sure that *all arms* have their minimum eigenvalue above $(1-\gamma)\lambda_1 t$. This would cause the difference in L and L' arising from the union bound that we need to use over all K arms for establishing our probability bound. The further additional constraint on p is due to the fact that the lower bound of $(1-\gamma)\lambda_1 t$ only holds after p rounds which we want it to be less than or equal to t_0, since algorithm might switch before reaching to p otherwise. Similarly, the lower bound of $(1-\gamma)\lambda_1 t$ is only good if it is bigger than the Greedy-First threshold $\lambda_0 t/4$, as the algorithm would switch otherwise.

Similar to Corollary 1 we can turn the result of Theorem 4 into a more readable and easier to understand corollary as following:

COROLLARY 2. *Under the similar assumptions made in Theorem 4, if the Greedy-First algorithm is executed with input parameters $\lambda_0 \le 8\lambda_1$ and $t_0 \ge Km$, then Greedy-First becomes purely greedy, does not switch to the pre-defined exploratory algorithm and enjoys a logarithmic regret with probability at least*

$$1 - \frac{3Kd\exp(-D_{\min}Km)}{1-\exp(-D_{\min})},$$

where function $D_{\min}$ is defined in Equation (7)*.*

Similar to Proposition 1, analyzing L' would lead us to the following proposition. The proof of this proposition is deferred to Appendix G.

PROPOSITION 2. *The function $S^{gf}(m, K, \sigma, x_{\max}, \lambda_1, h)$ defined in Equation* (8) *is non-increasing with respect to σ, and is non-decreasing with respect to λ_1 and h. Furthermore, if one keeps the total number of random sampling rounds Km fixed, then it is non-increasing with respect to K (recall that for Greedy-First we assumed $|\mathcal{K}_{opt}| = 0$). Finally, the limit of this function when σ goes to zero is equal to*

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K - \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))},$$

where $\gamma^* = 1 - \lambda_0/(4\lambda_1)$.

Let us elaborate more on how to use the bounds derived in Theorems 2 and 4 through an example.

EXAMPLE 1. Let $K = 3$ and $d = 2$. Suppose that arm parameters are given by $\beta_1 = (1,0), \beta_2 = (-1/2, \sqrt{3}/2)$ and $\beta_3 = (-1/2, -\sqrt{3}/2)$. Furthermore, suppose that the distribution of covariates p_X is the uniform distribution on the unit ball $B_1^2 = \{\mathbf{x} \in \mathbb{R}^2 \mid \|x\| \le 1\}$ meaning that $x_{\max} = 1$. Constants h and λ_1 should be chosen so that they satisfy Assumption 4. Note that for this specific distribution, value of h can be taken any positive number less than $\sqrt{3}$, however for larger h the value of λ_1 would decrease. For having a good trade-off we take $h = 0.3$, and after some calculations we obtain $\lambda_1 \approx 0.025$. Having these parameters, we can find the infimum of functions $L(\gamma, \delta, p)$ and $L'(\gamma, \delta, p)$ to find the lower bound on the probability of success of Greedy Bandit and Greedy-First algorithms, via Equations (3) and (8) respectively. In order to better observe these bounds, functions S^{gb} and S^{gf} are depicted for several values of m and σ.

One important thing to note is that bounds that we derived in this section and §3.4 are very conservative bounds and in practice, both Greedy Bandit and Greedy-First work with presumably larger probability. Most probability bounds that are used for proving these two theorems can be sharpened further, but for keeping proofs simpler and more readable we decided to keep this current version. As an example, as we observed in Example 1, one can further optimize over λ_1 and h. In the next section, we verify via simulations that both Greedy Bandit and Greedy-First are successful with a higher probability than the probabilistic bounds that we deployed in this section and §3.4.

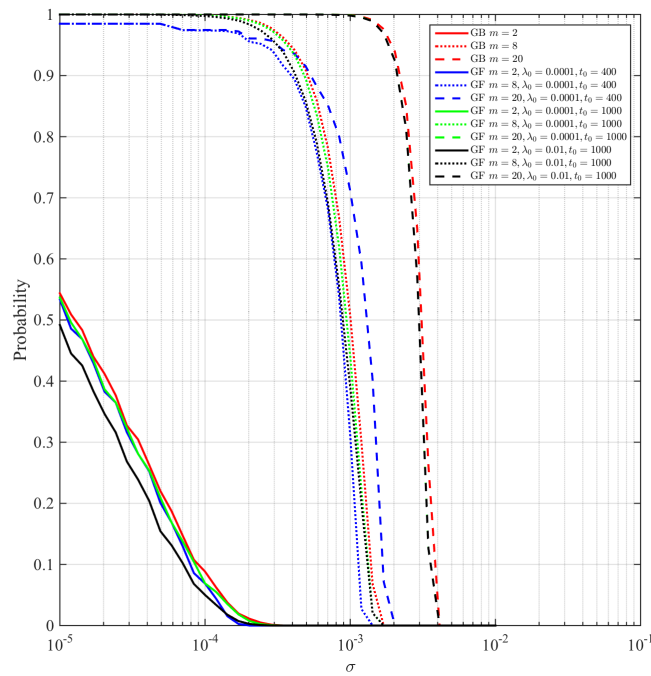


Figure 1 Lower (theoretical) bound on the probability of success for Greedy Bandit and Greedy-First. For $m = 20, t_0 = 1000$, the performance of Greedy-First for $\lambda_0 \in \{0.01, 0.0001\}$ is the same and indistinguishable.

5. Simulations

In this section we compare the Greedy Bandit and Greedy-First algorithms with the state-of-the-art algorithms on synthetic data and a data with real covariates. The list of other algorithms is: (i) OFUL algorithm of Abbasi-Yadkori et al. (2011) that is an advanced adaptation of the upper confidence bound (UCB) approach of Lai and Robbins (1985). (ii) Two adaptations of Thompson Sampling Thompson (1933) by Agrawal and Goyal (2013) and Russo and Van Roy (2014a) that will be referred to by *prior-free TS* and *prior-dependent TS* respectively. (iii) OLS Bandit algorithm of Goldenshluger and Zeevi (2013).

Remark 4. Our naming of Thompson sampling algorithms is based on the fact that the implementation by Russo and Van Roy (2014a) assumes the knowledge of prior distribution of parameters β_i while Agrawal and Goyal (2013) does not. But both algorithms assume the knowledge of noise variance. They also require the type of distribution for parameters β_i to be known which would be needed to calculate posterior distribution of the parameters given past observations. In particular, we follow the authors of these papers and implement the version of TS that assumes these variables are gaussian[3]. Finally, we note that the OFUL algorithm also relies on the knowledge of noise variance.

First, we simulate all algorithms on a synthetically-generated data. Since TS assumes β_i's are random and considers, as performance metric, expected regret with respect to the randomness of covariates, noise, and β_i (also known as Bayes risk), we also look at the same metric for all algorithms. In particular, we generate 1000 problem instances where each time the true parameters β_i are sampled (independently) from a distribution that will be described in following. Then, we graph the average and 95% confidence interval for the cumulative regret of each algorithm for all values of $T \in [10000]$. We take $K = 2$ and $d = 3$. In Appendix F.1 we show simulations with $K > 2$ and other values of d.

To understand the effect of "knowing the prior distribution and noise variance", we consider two regimes for the parameters:

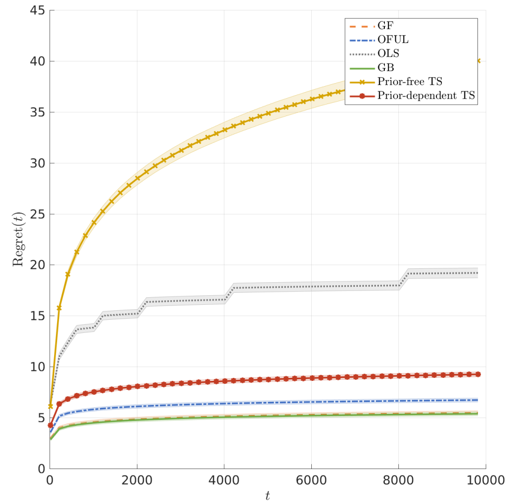
- **Correct prior:** Here we assume that OFUL and both versions of TS know the noise variance. We assume that β_i are sampled independently from $\mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ Also, prior-dependent TS has access to the true prior distribution of β_i and also the noise variance $\sigma^2 = 0.25$.
- **Incorrect prior:** Here we assume that OFUL and both versions of TS do not know the noise variance. They start from an uninformed point and sequentially estimate the noise variance given past data. We also assume that prior-dependent TS is initialized with an incorrect prior for β_i's. Precisely, it assumes β_i's are sampled from $10 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$. Finally, we assume the true β_i's are generated from a mixture of gaussians, while both versions of TS think that they are gaussian. Precisely, each β_i is sampled from $0.5 \times \mathrm{N}(\mathbf{1}_d, \mathbf{I}_d)$ with probability 0.5 or from $0.5 \times \mathrm{N}(-\mathbf{1}_d, \mathbf{I}_d)$ with probability 0.5. Here $\mathbf{1}_d \in \mathbb{R}^{d\times 1}$ is a vector with all entries equal to 1.

For each of the above regimes, we consider two scenarios: (i) when the covariate diversity condition holds and (ii) when it does not hold. For (i) we assume that they are sampled from $0.5 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ and are truncated to have ℓ_∞ norm at most 1, and for (ii) we add an intercept term to the model in (i). Overall, we will have four types of regret plots that are shown in Figure 2. The result shows that while Greedy Bandit has the best performance when the covariate diversity condition holds, Greedy-First is the top performer in all regimes.

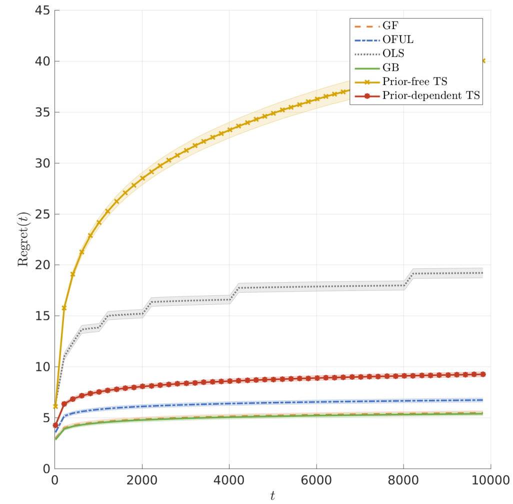
5.1. Additional Simulations

We performed several additional simulations to check the robustness of our results that are presented in Appendix F. Here we provide a brief summary of them.

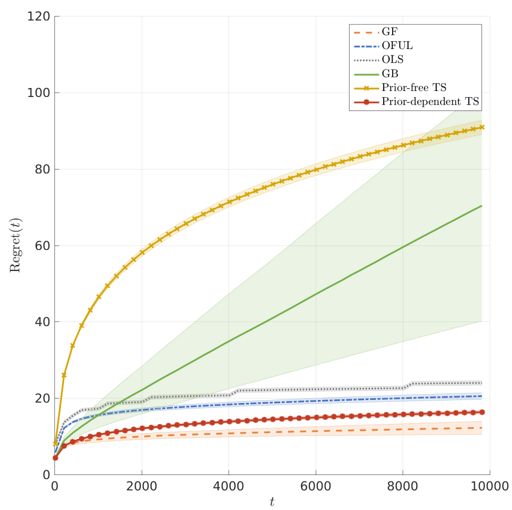
[3] We emphasize that the analyses of Agrawal and Goyal (2013), Russo and Van Roy (2014a) does not require the parameters to be gaussian. The gaussian assumption is only made for designing the algorithms.



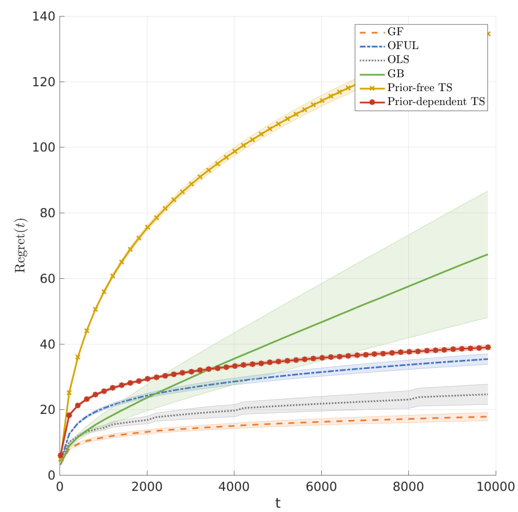
(a) Correct prior and covariate diversity.



(b) Incorrect prior and covariate diversity.



(c) Correct prior and no covariate diversity.



(d) Incorrect prior and no covariate diversity.

Figure 2 Expected regret of all algorithms on synthetic data in four different regimes for the covariate diversity condition and whether OFUL and TS are provided with correct or incorrect information on true prior distribution of the parameters. Out of 1000 runs of each simulation Greedy-First never switched in (a) and (b) and switched only 69 times in (c) and 139 times in (d).

• **More than two arms:** In Appendix F.1 we provide two types of simulations. First, in Figure 3, we show that Greedy Bandit behavior for $K > 2$ arms depends on the dimension of covariates. In particular, Figure 3 shows that performance of Greedy Bandit undergoes a dramatic improvement as dimension d increases. Next, in Figure 4, we compare performance of all algorithms in two cases $K = 5, d = 3$ and $K = 5, d = 7$ and observe that Greedy-First is the overall winner again.

- **Real data covariates:** Figure 6 in Appendix F.3 shows results of the situation in Figure 2(a) when covariates are sampled from a real data set. The results have similar qualitative behavior.

- **Sensitivity to parameters** Each of the algorithms requires a set of input parameters. We chose these parameters based on the recommendations in Abbasi-Yadkori (2012), Russo and Van Roy (2014a), Bastani and Bayati (2015). In particular, OFUL uses $\lambda = 1$ and $\delta = 0.01$. Both versions of Thompson sampling use $\delta = 0.01$, and OLS Bandit uses $h = 5$ and $q = 1$. Greedy Bandit is parameter free and Greedy-First uses the same parameters as OLS Bandit when it switches to that. The only remaining parameter will be t_0 in Greedy-First that will be set at $8Kd$. Figure 5 in Appendix F.1, shows that the choice of parameters h, q, and t_0 will have small impact on performance of Greedy-First.

6. Conclusions and Discussions

We prove that a greedy policy can achieve the optimal asymptotic regret for a two-armed contextual bandit as long as the contexts are i.i.d. and the distribution of contexts p_X are diverse enough. Greedy algorithms may be significantly preferable when exploration can be costly (e.g., result in lost customers for online advertising or A/B testing) or unethical (e.g., personalized medicine or clinical trials). Furthermore, our algorithm is entirely parameter-free, which makes it desirable in settings where tuning is difficult or where there is limited knowledge of problem parameters. Despite its simplicity, our simulations suggest that for the two-armed contextual bandit algorithm, Greedy Bandit can outperform OFUL, Thompson Sampling, and OLS Bandit algorithms.

However, in some scenarios, it is possible to have more than two arms or that the distribution of contexts p_X is not diverse enough. In these cases, the desirable greedy algorithm might sometimes fail. We provide probabilistic guarantees on the performance of Greedy Bandit and proved that with some positive probability this algorithm would succeed without any extra experimentation. To make this more practical, we propose the Greedy-First algorithm which uses the empirical distributions of contexts to determine whether the Greedy Bandit might successfully proceed or not. This algorithm maintains greedy policy and only switches to an exploratory algorithm, such as OLS Bandit, if the minimum eigenvalue of empirical covariance matrices do not grow fast enough. The established probabilistic bound on the performance of Greedy Bandit algorithm are then used to prove that the Greedy-First algorithm is effective in reducing the experimentation. Simulations suggest that in many problem instances, Greedy-First outperforms OFUL, Thompson Sampling, OLS Bandit, and even Greedy Bandit algorithms. It is also worth noting that, Greedy-First achieves a very similar regret to than of Greedy Bandit for a two-armed contextual bandit problem with diverse covariates, as expected.

References

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670, 2010.

Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3:397–422, 2003.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, pages 817–824, 2008.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. pages 355–366, 2008.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014a.

Dan Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *NIPS*, pages 1583–1591, 2014b.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1): 230–261, 2013.

Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates, 2015. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2661896`.

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. 2016.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Working Paper*, 2014.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*, 2014.

Xi Chen, Zachary Owen, Clark Pixton, and David Simchi-Levi. A statistical learning approach to personalization in revenue management. *Available at SSRN 2579462*, 2015.

Nathan Kallus. Learning to personalize from observational data. *arXiv preprint arXiv:1608.08925*, 2016.

Sbastien Bubeck and Nicol Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. ISSN 1935-8237.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, pages 208–214, 2011.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Jyotirmoy Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, pages 1978–2002, 1991.

Chih-Chun Wang, S. R. Kulkarni, and H. V. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005a.

Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Arbitrary side observations in bandit problems. *Advances in Applied Mathematics*, 34(4):903 – 938, 2005b.

Adam J Mersereau, Paat Rusmevichientong, and John N Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, 2009.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvari. Conservative bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1254–1262. PMLR, 2016.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits, 2016. `https://arxiv.org/abs/1611.06426`.

Tor Lattimore and Remi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems 27*, pages 550–558. 2014.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Oper. Res.*, 60(4):965–980, July 2012.

Arnoud V den Boer and Bert Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783, 2013.

N Bora Keskin and Assaf Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014a.

N Bora Keskin and Assaf Zeevi. On incomplete learning and certainty-equivalence control. *preprint*, 2015.

N. Bora Keskin and Assaf Zeevi. Chasing demand: Learning and earning in a changing environment. 2014b. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2389750`.

Arnoud V. den Boer and Bert Zwart. Dynamic pricing and learning with finite inventories. *Operations Research*, 63(4):965–978, 2015.

Maxime C Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. 2016. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737045`.

Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. 2016. https://arxiv.org/abs/1609.07574.

Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates, 2016. https://arxiv.org/abs/1604.07463.

Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

Alexander Goldenshluger and Assaf Zeevi. Woodroofes one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4):1603–1633, 2009.

W. R. Thompson. On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25:285–294, 1933.

Yasin Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems.* PhD thesis, 2012.

Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* Working Publication, 2016.

Joel A Tropp. User-friendly tail bounds for matrix martingales. Technical report, DTIC Document, 2011.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. 2015. https://arxiv.org/abs/1511.05720.

Appendix A: Properties of Covariate Diversity

LEMMA 1 *If there exists a set $W \subset \mathbb{R}^d$ that satisfies conditions (a), (b), and (c) given below, then p_X satisfies Assumption 3.*

(a) W is symmetric around the origin; i.e., if $\mathbf{x} \in W$ then $-\mathbf{x} \in W$.

(b) There exist positive constants $a, b \in \mathbb{R}$ such that for all $\mathbf{x} \in W$, $a \cdot p_X(-\mathbf{x}) \leq b \cdot p_X(\mathbf{x})$.

(c) There exists a positive constant λ such that $\int_W \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \succeq \lambda I_d$. For discrete distributions, the integral is replaced with a sum.

Proof of Lemma 1. Since for all $\mathbf{u} \in \mathbb{R}^d$ at least one of $\mathbf{x}^\top \mathbf{u} \geq 0$ or $-\mathbf{x}^\top \mathbf{u} \geq 0$ holds, and using conditions (a), (b), and (c) of Lemma 1 we have:

$$\begin{aligned}
\int \mathbf{x}\mathbf{x}^\top \mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) \mathrm{d}\mathbf{x} &\succeq \int_W \mathbf{x}\mathbf{x}^\top \mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{1}{2} \int_W \mathbf{x}\mathbf{x}^\top \Big[\mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) + \mathbb{I}(-\mathbf{x}^\top \mathbf{u} \geq 0) p_X(-\mathbf{x}) \Big] \mathrm{d}\mathbf{x} \\
&\succeq \frac{1}{2} \int_W \mathbf{x}\mathbf{x}^\top \Big[\mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) + \frac{a}{b} \mathbb{I}(\mathbf{x}^\top \mathbf{u} \leq 0) \Big] p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\succeq \frac{a}{2b} \int_W \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\succeq \frac{a\lambda}{2b} I_d \,.
\end{aligned}$$

Here, the first inequality follows from the fact that $\mathbf{x}\mathbf{x}^\top$ is positive semi-definite, the first equality follows from condition (a) and a change of variable ($\mathbf{x} \to -\mathbf{x}$), the second inequality is by condition (b), the third inequality uses $a \leq b$ which follows from condition (b), and the last inequality uses condition (c). $\square$

We now state the proofs of lemmas that were used in §2.2.

LEMMA 2 *For any* $R > 0$ *we have* $\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left[\frac{R^2}{d+2}\mathrm{vol}(B_R^d)\right] I_d$.

Proof. First note that B_R^d is symmetric with respect to each axis, therefore the off-diagonal entries in $\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x}$ are zero. In particular, the (i,j) entry of the integral is equal to $\int_{B_R^d} x_i x_j \mathrm{d}\mathbf{x}$ which is zero when $i \neq j$ using a change of variable $x_i \to -x_i$ that has the identity as its Jacobian and keeps the domain of integral unchanged but changes the sign of $x_i x_j$. Also, by symmetry, all diagonal entry terms are equal. In other words,

$$\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left(\int_{B_R^d} x_1^2 \mathrm{d}\mathbf{x}\right) I_d\,. \tag{10}$$

Now for computing the right hand side integral, we introduce the spherical coordinate system as

$$\begin{aligned}
x_1 &= r\cos\theta_1,\\
x_2 &= r\sin\theta_1\cos\theta_2,\\
&\ \vdots\\
x_{d-1} &= r\sin\theta_1\sin\theta_2\ldots\sin\theta_{d-2}\cos\theta_{d-1},\\
x_d &= r\sin\theta_1\sin\theta_2\ldots\sin\theta_{d-2}\sin\theta_{d-1},
\end{aligned}$$

and the determinant of its Jacobian is given by

$$\det J(r,\boldsymbol{\theta}) = \det\left[\frac{\partial \mathbf{x}}{\partial r \partial \boldsymbol{\theta}}\right] = r^{d-1}\sin^{d-2}\theta_1\sin^{d-3}\theta_2\ldots\sin\theta_{d-2}.$$

Now, using symmetry, and summing up equation (10) with x_i^2 used instead of x_1^2 for all $i \in [d]$, we obtain

$$\begin{aligned}
d\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} &= \int_{B_R^d} \left(x_1^2 + x_2^2 + \ldots + x_d^2\right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_d\\
&= \int_{\theta_1,\ldots,\theta_{d-1}} \int_{r=0}^{R} r^{d+1}\sin^{d-2}\theta_1\sin^{d-3}\theta_2\ldots\sin\theta_{d-2}\,\mathrm{d}r\,\mathrm{d}\theta_1\ldots\mathrm{d}\theta_{d-1}\,.
\end{aligned}$$

Comparing this to

$$\mathrm{vol}(B_R^d) = \int_{\theta_1,\ldots,\theta_{d-1}} \int_{r=0}^{R} r^{d-1}\sin^{d-2}\theta_1\sin^{d-3}\theta_2\ldots\sin\theta_{d-2}\,\mathrm{d}r\,\mathrm{d}\theta_1\ldots\mathrm{d}\theta_{d-1}\,,$$

we obtain that

$$\begin{aligned}
\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} &= \left[\frac{\int_0^R r^{d+1}\mathrm{d}r}{d\int_0^R r^{d-1}\mathrm{d}r}\mathrm{vol}(B_R^d)\right] I_d\\
&= \left[\frac{R^2}{d+2}\mathrm{vol}(B_R^d)\right] I_d\,.
\end{aligned}$$

LEMMA 6. *The following inequality holds*

$$\int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,trunc}(\mathbf{x})\mathrm{d}\mathbf{x} \succeq \lambda_{uni}\mathbf{I}_d\,,$$

where $\lambda_{uni} \equiv \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \frac{x_{\max}^2}{d+2} \mathrm{vol}(B^d_{x_{\max}})$. □

Proof of Lemma 6. We can lower-bound the density $p_{X,\text{trunc}}$ by the uniform density as follows. Note that we have $\mathbf{x}^\top\Sigma^{-1}\mathbf{x} \le \|\mathbf{x}\|_2^2\lambda_{\max}(\Sigma^{-1})$ and as a result for any $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 \le x_{\max}$ we have

$$p_{X,\text{trunc}}(\mathbf{x}) \ge p_X(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right) \ge \frac{\exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right)}{(2\pi)^{d/2}|\Sigma|^{d/2}} = p_{X,\text{uniform-lb}}\,.$$

Using this we can derive a lower bound on the desired covariance as following

$$\begin{aligned}
\int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,\text{trunc}}(\mathbf{x})\mathrm{d}\mathbf{x} &\succeq \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,\text{uniform-lb}}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \frac{x_{\max}^2}{d+2} \mathrm{vol}(B^d_{x_{\max}}) I_d \\
&= \lambda_{\text{uni}} I_d\,,
\end{aligned}$$

where we used Lemma 2 in the third line. This concludes the proof. □

Appendix B: Useful Concentration Results

LEMMA 7 **(Bernstein Concentration)**. *Let* $\{D_k, \mathcal{H}_k\}_{k=1}^\infty$ *be a martingale difference sequence, and let* D_k *be* σ_k*-subgaussian. Then, for all* $t > 0$ *we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right| \ge t\right] \le 2\exp\left\{-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2}\right\}.$$

Proof. See Theorem 2.3 of Wainwright (2016) and let $b_k = 0$ and $\nu_k = \sigma_k$ for all k. □

LEMMA 8 **(Theorem 3.1 of Tropp (2011))**. *Consider a finite adapted sequence* $\{X_k\}$ *of positive semi-definite matrices with dimension* d*, and suppose that* $\lambda_{\max}(X_k) \le R$ *almost surely. Define the series* $Y \equiv \sum_k X_k$ *and* $W \equiv \sum_k \mathbb{E}_{k-1}X_k$*. Then for all* $\mu \ge 0, \gamma \in [0,1)$ *we have:*

$$\mathbb{P}\left[\lambda_{\min}(Y) \le (1-\gamma)\mu \quad \textit{and} \quad \lambda_{\min}(W) \ge \mu\right] \le d\left(\frac{e^{-\gamma}}{(1-\gamma)^{1-\gamma}}\right)^{\mu/R}.$$

Appendix C: Proof of Theorem 1

We first prove a lemma on the instantaneous regret of the Greedy Bandit using a standard peeling argument. The proof here is adapted from Bastani and Bayati (2015) with a few modifications; we present it here for completeness.

Notation. We define the following events to simplify notation. For any $\lambda, \chi > 0$, let

$$\mathcal{F}_{i,t}^{\lambda} = \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\right\} \tag{11}$$

$$\mathcal{G}_{i,t}^{\chi} = \left\{\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 < \chi\right\}. \tag{12}$$

Lemma 9. *The instantaneous expected regret of the Greedy Bandit at time $t \geq 2$ satisfies*

$$r_t(\pi) \leq \frac{4(K-1)C_0 \bar{C} x_{\max}^2 (\log d)^{3/2}}{C_2} \frac{1}{t-1} + 4(K-1) b_{\max} x_{\max} \left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right),$$

where $C_2 = \lambda^2/(2d\sigma^2 x_{\max}^2)$, C_0 is defined in Assumption 2, and $\bar{C}$ is defined in Theorem 1.

Proof. We can decompose the regret as $r_t(\pi) = \mathbb{E}[\text{Regret}_t(\pi)] = \sum_{i=1}^K \mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_i] \cdot \mathbb{P}(X_t \in \mathcal{R}_i)$. Now we can expand each term as

$$\mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] = \mathbb{E}\left[X_t^\top(\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l\right],$$

where the expectation is taken with respect to filtration $\mathcal{H}_{t-1}$. For each $1 \leq i, l \leq K$ satisfying $i \neq l$, let us define the region where arm i is superior over arm l

$$\hat{\mathcal{R}}_{i \geq l, t} := \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,t-1})\right\},$$

Note that we may incur a nonzero regret if $X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) > X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ or if $X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) = X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ and the tie-breaking random variable W_t indicates an action other than l as the action to be taken. It is worth mentioning that in the case $X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) = X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ we do not incur any regret if W_t indicates arm l as the action to be taken. Nevertheless, as regret is a non-negative quantity, we can write

$$\begin{aligned}
\mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] &\leq \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) \geq X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})) X_t^\top(\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l\right] \\
&\leq \sum_{i \neq l} \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&= \sum_{i \neq l} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \geq l,t}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\leq \sum_{i \neq l} \Bigg\{\mathbb{E}\left[\mathbb{I}(\hat{\mathcal{R}}_{i \geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\qquad + \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \geq l,t}, \overline{\mathcal{F}_{l,t-1}^{\lambda_0/4}}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \overline{\mathcal{F}^{\lambda_0/4}_{i,t-1}}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \Bigg\} \\
&\leq \sum_{i\neq l} \Bigg\{ \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\quad + 2b_{\max}x_{\max}\left(\mathbb{P}(\overline{\mathcal{F}^{\lambda_0/4}_{l,t-1}}) + \mathbb{P}(\overline{\mathcal{F}^{\lambda_0/4}_{i,t-1}})\right) \Bigg\} \\
&\leq \sum_{i\neq l} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\quad + 4(K-1)b_{\max}x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}^{\lambda_0/4}_{i,t-1}})
\end{aligned} \tag{13}
$$

where in the second line we used a union bound, in the sixth line we used the fact that $\mathcal{F}^{\lambda_0/4}_{i,t-1}$ and $\mathcal{F}^{\lambda_0/4}_{l,t-1}$ are independent of the event $X_t \in \mathcal{R}_l$ which only depends on X_t, and also a Cauchy-Schwarz inequality showing $X_t^\top(\beta_l - \beta_i) \leq 2b_{\max}x_{\max}$. Therefore, we need to bound the first term in above. Fix i and note that when we include events $\mathcal{F}^{\lambda_0/4}_{i,t-1}$ and $\mathcal{F}^{\lambda_0/4}_{l,t-1}$, we can use Lemma 5 which proves sharp concentrations for $\hat{\beta}(\mathcal{S}_{l,t-1})$ and $\hat{\beta}(\mathcal{S}_{i,t-1})$. Let us now define the following set

$$
I^h = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top(\beta_l - \beta_i) \in (2\delta x_{\max} h, 2\delta x_{\max}(h+1)]\},
$$

where $\delta = 1/\sqrt{(t-1)C_2}$. Note that since $X_t^\top(\beta_l - \beta_i)$ is bounded above by $2b_{\max}x_{\max}$, the set I^h only needs to be defined for $h \leq h^{\max} = \lceil b_{\max}/\delta \rceil$. We can now expand the first term in Equation (13) for i, by conditioning on $X_t \in I^h$ as following

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&= \sum_{h=0}^{h^{\max}} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t \in I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1) \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t \in I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1) \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]] \\
&\leq \sum_{h=0}^{h^{\max}} 4C_0\delta^2 x_{\max}^2 (h+1)^2 \mathbb{P}\left[X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I_h\right],
\end{aligned} \tag{14}
$$

where in the first inequality we used the fact that conditioning on $X_t \in I^h$, $X_t^\top(\beta_l - \beta_i)$ is bounded above by $2\delta x_{\max}(h+1)$, in the second inequality we used the fact that the event $X_t \in I^h$ is a subset of the event $X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]$, and in the last inequality we used the margin condition given in Assumption 2. Now we reach to the final part of the proof, where conditioning on $\mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}$, and $X_t \in I^h$ we want to bound the probability that we pull a wrong arm. Note that conditioning on $X_t \in I^h$, the event $X_t^\top\left(\hat{\beta}(\mathcal{S}_{i,t-1}) - \hat{\beta}(\mathcal{S}_{l,t-1})\right) \geq 0$ happens only when at least one of the following two events: i) $X_t^\top(\beta_l - \hat\beta(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h$ or ii) $X_t^\top(\hat\beta(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h$ happens. This is true according to

$$\begin{aligned}
0 &\leq X_t^\top \left(\hat\beta(\mathcal{S}_{i,t-1}) - \hat\beta(\mathcal{S}_{l,t-1})\right) \\
&= X_t^\top(\hat\beta(\mathcal{S}_{i,t-1}) - \beta_i) + X_t^\top(\beta_i - \beta_l) + X_t^\top(\beta_l - \hat\beta(\mathcal{S}_{l,t-1})) \\
&\leq X_t^\top(\hat\beta(\mathcal{S}_{i,t-1}) - \beta_i) - 2\delta x_{\max} h + X_t^\top(\beta_l - \hat\beta(\mathcal{S}_{l,t-1}))\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad\leq \mathbb{P}\left[X_t^\top(\beta_l - \hat\beta(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad+ \mathbb{P}\left[X_t^\top(\hat\beta(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad\leq \mathbb{P}\left[X_t^\top(\beta_l - \hat\beta(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\qquad\qquad+ \mathbb{P}\left[X_t^\top(\hat\beta(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad\leq \mathbb{P}\left[\|\beta_l - \hat\beta(\mathcal{S}_{l,t-1})\|_2 \geq \delta h, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] + \mathbb{P}\left[\|\hat\beta(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \geq \delta h, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right],
\end{aligned}$$

where in the third line we used $P(A, B \mid C) \leq P(A \mid C)$, in the fourth line we used Cauchy-Schwarz inequality. Now using the notation described in Equation (12) this can be rewritten as

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{G}_{l,t-1}^{\delta h}}, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] + \mathbb{P}\left[\overline{\mathcal{G}_{i,t-1}^{\delta h}}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\leq \mathbb{P}\left[\overline{\mathcal{G}_{l,t-1}^{\delta h}}, \mathcal{F}_{l,t-1}^{\lambda_0/4}\right] + \mathbb{P}\left[\overline{\mathcal{G}_{i,t-1}^{\delta h}}, \mathcal{F}_{i,t-1}^{\lambda_0/4}\right] \\
&\leq 4d \exp\left(-C_2(t-1)(\delta h)^2\right) \\
&= 4d \exp(-h^2),
\end{aligned}$$

in the fifth line we used the fact that both $\mathcal{R}_l$ and I^h only depend on X_t which is independent of $\hat\beta(\mathcal{S}_{q,t-1})$ for all q, and in the sixth line we used Lemma 5. We can also bound this probability by 1, which is better than $4d\exp(-h^2)$ for small values of h. Hence, using $\sum_{l=1}^K \mathbb{P}[\mathcal{R}_l] = 1$ we can write the regret as

$$\begin{aligned}
&\mathbb{E}[\text{Regret}_t(\pi)] = \sum_{l=1}^K \mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] \cdot \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\leq \sum_{l=1}^K \left(\sum_{i\neq l}\sum_{h=0}^{h^{\max}} \left[4C_0\delta^2 x_{\max}^2 (h+1)^2 \min\{1, 4d\exp(-h^2)\}\right] + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}})\right) \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\leq 4(K-1) C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h^{\max}} (h+1)^2 \min\{1, 4d\exp(-h^2)\}\right) + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \\
&\leq 4(K-1)\left(C_0\delta^2 x_{\max}^2 \left(\sum_{h=0}^{h_0} (h+1)^2 + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^2 \exp(-h^2)\right) + b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}})\right),
\end{aligned} \tag{15}$$

where we take $h_0 = \lfloor\sqrt{\log 4d}\rfloor + 1$. Note that functions $f(x) = x^2\exp(-x^2)$ and $g(x) = x\exp(-x^2)$ are both decreasing for $x \geq 1$ and therefore

$$\begin{aligned}\sum_{h=h_0+1}^{h^{\max}}(h+1)^2\exp(-h^2) &= \sum_{h=h_0+1}^{h^{\max}}(h^2+2h+1)\exp(-h^2)\\ &= \sum_{h=h_0+1}^{h^{\max}} h^2\exp(-h^2) + 2\sum_{h=h_0+1}^{h^{\max}} h\exp(-h^2) + \sum_{h=h_0+1}^{h^{\max}}\exp(-h^2)\\ &\leq \int_{h_0}^{\infty} h^2\exp(-h^2)\mathrm{d}h + \int_{h_0}^{\infty} 2h\exp(-h^2)\mathrm{d}h + \int_{h_0}^{\infty}\exp(-h^2)\mathrm{d}h. \qquad (16)\end{aligned}$$

Computing the above terms using integration by parts yields

$$\begin{aligned}&\sum_{h=0}^{h_0}(h+1)^2 + 4d\sum_{h=h_0+1}^{h^{\max}}(h+1)^2\exp(-h^2)\\ &= \frac{(h_0+1)(h_0+2)(2h_0+3)}{6} + d(2h_0+7)\exp(-h_0^2)\\ &\leq \frac{1}{3}h_0^3 + \frac{3}{2}h_0^2 + \frac{13}{6}h_0 + 1 + d(2h_0+7)\frac{1}{4d}\\ &\leq \frac{1}{3}\left(\sqrt{\log 4d}+1\right)^3 + \frac{3}{2}\left(\sqrt{\log 4d}+1\right)^2 + \frac{8}{3}\left(\sqrt{\log 4d}+1\right) + \frac{11}{4}\\ &\leq \left(\sqrt{\log d}+2\right)^3 + \frac{3}{2}\left(\sqrt{\log d}+2\right)^2 + \frac{8}{3}\left(\sqrt{\log d}+2\right) + \frac{11}{4}\\ &= \frac{1}{3}(\log d)^{3/2} + \frac{7}{2}\log d + \frac{38}{3}(\log d)^{1/2} + \frac{67}{4}\\ &\leq (\log d)^{3/2}\left((\frac{1}{3} + \frac{7}{2}(\log d)^{-0.5} + \frac{38}{3}(\log d)^{-1} + \frac{67}{4}(\log d)^{-1.5}\right)\\ &\leq (\log d)^{3/2}\bar{C}\end{aligned}$$

where $\bar{C}$ is defined as (2). By replacing this in (15) and substituting $\delta = 1/\sqrt{(t-1)C_2}$ we get

$$r_t(\pi) = \mathbb{E}[\mathrm{Regret}_t(\pi)] \leq \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}\frac{1}{t-1} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right)$$

as desired. □

Having this lemma proved, it is now fairly straightforward to prove Theorem 1.

Proof of Theorem 1. The expected cumulative regret is the sum of expected regret for times up to time T. As the regret term at time $t=1$ is upper bounded by $2x_{\max}b_{\max}$ and as $K=2$, by using Lemma 4 and Lemma 9 we can write

$$\begin{aligned}R_T(\pi) &= \sum_{t=1}^{T} r_t(\pi)\\ &\leq 2x_{\max}b_{\max} + \sum_{t=2}^{T}\left[\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}\frac{1}{t-1} + 4b_{\max}x_{\max}d\exp(-C_1(t-1))\right]\\ &= 2x_{\max}b_{\max} + \sum_{t=1}^{T-1}\left[\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}\frac{1}{t} + 4b_{\max}x_{\max}d\exp(-C_1 t)\right]\end{aligned}$$

$$
\begin{aligned}
&\leq 2x_{\max}b_{\max} + \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}(1+\int_1^T \frac{1}{t}\mathrm{d}t) + 4b_{\max}x_{\max}d\int_1^\infty \exp(-C_1 t)\mathrm{d}t \\
&= 2x_{\max}b_{\max} + \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}(1+\log T) + \frac{4b_{\max}x_{\max}d}{C_1} \\
&= \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2}\log T + \left(2x_{\max}b_{\max} + \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2} + \frac{160b_{\max}x_{\max}^3 d}{\lambda_0}\right) \\
&= \mathcal{O}(\log T),
\end{aligned}
$$

finishing up the proof. □

Appendix D: Regret analysis of Greedy-First

Before we start with the proofs, we present the pseudo-code for OLS-Bandit and also the the heuristic for Greedy-First that were not presented in §4 due to space limitations. The OLS bandit algorithm is introduced by Goldenshluger and Zeevi (2013) and generalized by Bastani and Bayati (2015). Here, we describe the more general version that applies to more than two arms where some arms may be uniformly sub-optimal. For more details, we defer to the aforementioned papers. As mentioned earlier, in addition to Assumptions 1 and 2, OLS bandit needs two additional assumptions as follows:

Assumption 5 **(Arm optimality)**. *. Let $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ be mutually exclusive sets that include all K arms. Sub-optimal arms $i \in \mathcal{K}_{sub}$ satisfy $X^\top\beta_i < \max_{j\neq i} X^\top\beta_j - h$ for some $h>0$ and every $X \in \mathcal{X}$. On the other hand, each optimal arm $i\in\mathcal{K}_{opt}$, has a corresponding set $U_i = \{X \mid X^\top\beta_i > \max_{j\neq i} X^\top\beta_j + h\}$ We assume there exists $p_* > 0$ such that $\min_{i\in\mathcal{K}_{opt}} \Pr[U_i] \geq p^*$.*

Assumption 6 **(Conditional Positive-Definiteness)**. *Define $\Sigma_i \equiv \mathbb{E}[XX^\top \mid X \in U_i]$ for all $i \in \mathcal{K}_{opt}$. Then, there exists $\lambda_1 > 0$ such that for all $i\in\mathcal{K}_{opt}, \lambda_{\min}(\Sigma_i) \geq \lambda_1 > 0$.*

The algorithm requires definition of *forced-sample sets*. In particular, let us prescribe a set of times when we forced-sample arm i (regardless of the observed covariates X_t):

$$
\mathcal{T}_i \equiv \left\{ (2^n - 1)\cdot Kq + j \;\middle|\; n \in \{0,1,2,...\} \text{ and } j \in \{q(i-1)+1, q(i-1)+2, ..., iq\} \right\}. \tag{17}
$$

Thus, the set of forced samples from arm i up to time t is $\mathcal{T}_{i,t} \equiv \mathcal{T}_i \cap [t] = \mathcal{O}(q\log t)$.

We also need to define *all-sample sets* $\mathcal{S}_{i,t} = \left\{t' \mid \pi_{t'} = i \text{ and } 1 \leq t' \leq t\right\}$ that are the set of times we play arm i up to time t. Note that by definition $\mathcal{T}_{i,t} \subset \mathcal{S}_{i,t}$. The algorithm proceeds as follows. During any forced sampling time $t \in \mathcal{T}_i$, the corresponding arm (arm i) is played regardless of observed covariates X_t. However, for other times, the algorithm uses two different estimations of arm parameters in order to make decision. First, it estimates arm parameters via OLS applied only on the forced samples set and discards each arm that is sub-optimal by a margin at least equal to

$h/2$. Then, it applies OLS to all-sample sets and picks the arm with the highest estimated reward among the remaining arms. Algorithm 3 explains the pseudo-code for OLS Bandit.

Algorithm 3 OLS Bandit

Input parameters: q, h
Initialize $\hat{\beta}(\mathcal{T}_{i,0})$ and $\hat{\beta}(\mathcal{S}_{i,0})$ by 0 for all i in $[K]$
Use q to construct force-sample sets $\mathcal{T}_i$ using Eq. (17) for all i in $[K]$
for $t \in [T]$ **do**
  Observe $X_t \in \mathcal{P}_X$
  if $t \in \mathcal{T}_i$ for any i **then**
    $\pi_t \leftarrow i$
  else
    $\hat{\mathcal{K}} = \left\{ i \in K \mid X_t^T \hat{\beta}(\mathcal{T}_{i,t-1}) \geq \max_{j \in K} X_t^T \hat{\beta}(\mathcal{T}_{j,t-1}) - h/2 \right\}$
    $\pi_t \leftarrow \arg\max_{i \in \hat{\mathcal{K}}} X_t^T \hat{\beta}(\mathcal{S}_{i,t-1})$
  end if
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
  Play arm π_t, observe $Y_t = X_t^T \beta_{\pi_t} + \varepsilon_{i,t}$
end for

The pseudo-code for Heuristic Greedy-First bandit is as follows.

Algorithm 4 Heuristic Greedy-First Bandit

Input parameters: t_0
Execute Greedy Bandit for $t \in [t_0]$
Set $\hat{\lambda}_0 = \frac{1}{2t_0} \min_{i \in [K]} \lambda_{\min} \left(\hat{\Sigma}(\mathcal{S}_{i,t_0}) \right)$
if $\hat{\lambda}_0 \neq 0$ **then**
  Execute Greedy-First Bandit for $t \in [t_0 + 1, T]$ with $\lambda_0 = \hat{\lambda}_0$
else
  Execute OLS Bandit for $t \in [t_0 + 1, T]$
end if

Appendix E: Extensions to nonlinear rewards and α-margin boundary conditions

E.1. Extension to nonlinear reward functions

The Greedy Bandit algorithm can be simply modified to be used for a larger family of reward functions. Recall that a function $\psi : I \to \mathbb{R}$ is called *Lipschitz*, if there exists a constant M such that for all $x, y \in I$,

$$|\psi(x) - \psi(y)| \leq M|x - y|.$$

Note that although many interesting functions such as exponential reward are generally not Lipschitz over $I = \mathbb{R}$, they become Lipschitz once the interval $I = (a, b)$ is bounded. In our case, $X_t^\top \beta_i$ always belongs to $[-b_{\max} x_{\max}, b_{\max} x_{\max}]$, according to Assumption 1 and a Cauchy-Schwarz inequality. Therefore, the function ψ needs to be Lipschitz only on the interval $I = [-b_{\max} x_{\max}, b_{\max} x_{\max}]$.

In particular, any differentiable function $\psi(z)$ with a bounded derivative $\psi'(z) \in [-D, D]$ over the interval $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$ is Lipschitz with the same constant D. This can be implied using the Mean-Value Theorem as

$$|\psi(x) - \psi(y)| = |\psi'(z)(x-y)| \leq D|x-y|.$$

In above, z is between x and y and therefore it belongs to $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$ meaning that $|\psi'(z)| \leq D$. Therefore, this class includes many interesting reward functions ψ, such as $\psi(z) = z^n$ and $\psi(z) = \exp(z)$.

Our aim is to provide regret bounds for the bandit problem in which the linear reward is replaced with

$$Y_t = \psi\left(X_t^\top \beta_i + \varepsilon_{i,t}\right). \tag{18}$$

Here ψ is a function that is strictly increasing over $\mathbb{R}$ and Lipschitz over the interval $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$ and the noise terms $\varepsilon_{i,t}$ are i.i.d. samples from a distribution P_ε that is σ-subgaussian. Similar to the case of linear reward, we exclude the dependency of regret with respect to the noise terms. More precisely, suppose that at time t, the best action is indexed by i while we pull the arm $j \neq i$. The expected regret is equal to

$$\mathbb{E}\left[\psi\left(X_t^\top \beta_i\right) - \psi\left(X_t^\top \beta_j\right)\right], \tag{19}$$

where the expectation is taken with respect to the policy π and covariate X_t. In Corollary 3 we provide regret bounds for the case that ψ is strictly increasing and Lipschitz over $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$. In order to state our result, we first need to modify the Greedy Bandit Algorithm that was previously explained in Algorithm 1, for a general function ψ as following.

Algorithm 5 Modified Greedy Bandit

Input parameters: Function ψ
Initialize $\hat{\beta}(\mathcal{S}_{i,0})$ at random for $i \in [K]$
for $t \in [T]$ **do**
  Observe $X_t \sim p_X$
  $\pi_t \leftarrow \arg\max_i \psi\left(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})\right)$ (break ties randomly)
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
  Play arm π_t, observe $Y_t = \psi\left(X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}\right)$
  Let $Z_t = \psi^{-1}(Y_t)$
  Update arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t}) = \left[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\right]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top Z(\mathcal{S}_{\pi_t,t})$
end for

Note that for a strictly increasing function ψ, the value $\psi(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}))$ is maximized for the same arm that maximizes $X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$. Furthermore, the transformation $Z_t = \psi^{-1}(Y_t)$ is valid, since any strictly increasing function on reals is invertible. The following corollary, provides regret bounds for a reward function ψ that is Lipschitz over $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$.

Corollary 3. *Let ψ be a strictly increasing function on $\mathbb{R}$. In addition, let $I = [-b_{\max}x_{\max}, b_{\max}x_{\max}]$ and suppose that ψ is Lipschitz over the interval I, with the constant M. Furthermore, suppose that $K=2$ and Assumptions 1, 2, and 3 hold. Then the following inequality on the regret of Modified Greedy Bandit Algorithm holds*

$$R_T^{\psi}(\pi) \le M R_T(\pi),$$

where $R_T(\pi)$ given in (1), *is the regret of Greedy Bandit when the reward is linear, and $R_T^{\psi}(\pi)$ is the regret of Modified Greedy Bandit when the reward comes from the model in Equation* (18). *According to Theorem 1, this further implies that*

$$R_T^{\psi}(\pi) \le M C_{GB} \log T,$$

with C_{GB} defined in Theorem 1. □

Proof. The regret proof of the Greedy Bandit algorithm (designed for the linear reward) can be readily extended to the Modified Greedy Bandit algorithm with two slight modifications that we need to: 1) verify that the concentration results for $\hat{\beta}_i$ holds, similar to what we had in Lemma 5, and 2) derive an upper bound on the expected regret that is incurred in the presence of ψ, based on the expected regret of the linear case.

1. As ψ is strictly increasing, any decision based on ψ is equivalent to a similar one on the linear case. In other words, $\psi(X_t^\top \hat{\beta}(\mathcal{S}_{i,t})) \ge \psi(X_t^\top \hat{\beta}(\mathcal{S}_{j,t}))$ is equivalent to $X_t^\top \hat{\beta}(\mathcal{S}_{i,t}) \ge X_t^\top \hat{\beta}(\mathcal{S}_{j,t})$, which means that the policies coincide. Therefore, using the Assumption 2 and 3 the result of Lemma 4 holds. As we apply the ordinary least squares to minimize $\|\mathbf{X}\beta - Z\|_2^2$, and as

   $$Z_t = \psi^{-1}(Y_t) = \psi^{-1}(\psi(\mathbf{X}_t^\top \beta_i + \varepsilon_t)) = X_t^\top \beta_i + \varepsilon_t,$$

   the concentration result in Lemma 5 holds.
2. For proving an upper bound on the regret of Modified Greedy Bandit algorithm, we need to derive an upper bound on $r_t^{\psi}(\pi)$ which is the expected reward of Modified Greedy Bandit algorithm at time t. As ψ is Lipschitz over $I = [-b_{\max}x_{\max}, b_{\max}x_{\max}]$ with the constant M, and as $X_t^\top \beta_i$ and $X_t^\top \beta_j$ belong to the interval I, we can write

   $$|\psi(X_t^\top \beta_i) - \psi(X_t^\top \beta_j)| \le M|X_t^\top \beta_i - X_t^\top \beta_j|,$$

   which by taking the expectation with respect to the policy and X turns into

   $$r_t^{\psi}(\pi) = \mathbb{E}\left[|\psi(X_t^\top \beta_i) - \psi(X_t^\top \beta_j)|\right] \le M\mathbb{E}\left[|X_t^\top \beta_i - X_t^\top \beta_j|\right] = M r_t(\pi),$$

where we used our previous observation that the optimal policy under the reward ψ matches with that of the linear case. Summing up all the terms $t = 1, 2, \ldots, T$ we obtain

$$R_T^{\psi}(\pi) \leq M R_T(\pi),$$

which implies the first result. The second result is a straightforward application of Theorem 1. $\square$

E.2. Regret bounds for more general margin conditions

While the assumed margin condition in Assumption 2 holds for many well-known distributions, one can construct a distribution with a growing density near the decision boundary that violates Assumption 2. Therefore, it is interesting to see how regret bounds would change if we assume other type of margin conditions. Similar to what proposed in Weed et al. (2015), we assume that the distribution of contexts p_X satisfies a more general α-margin condition as following.

Assumption 7 (**α-Margin Condition**). *For $\alpha \geq 0$, we say that the distribution p_X satisfies the α-margin condition, if there exists a constant $C_0' > 0$ such that for each $\kappa' > 0$:*

$$\forall\ i \neq j: \quad \mathbb{P}_X\left[0 < |X^\top(\beta_i - \beta_j)| \leq \kappa'\right] \leq C_0' \kappa'^{\alpha}.$$

Although it is straightforward to verify that any distribution p_X satisfies the 0-margin condition, it is easy to construct a distribution violating the α-margin condition, for an arbitrary $\alpha > 0$. In addition, if p_X satisfies the α-margin condition, then for any $\alpha' < \alpha$ it also satisfies the α'-margin condition. In the case that there exist some gap between arm rewards, meaning the existence of $\kappa_0 > 0$ such that

$$\forall\ i \neq j: \quad \mathbb{P}_X\left[0 < |X^\top(\beta_i - \beta_j)| \leq \kappa_0\right] = 0,$$

the distribution p_X satisfies the α-margin condition for all $\alpha \geq 0$.

Having this definition in mind, we can prove the following result on the regret of Greedy Bandit algorithm when p_X satisfies the α-margin condition:

Corollary 4. *Let $K = 2$ and suppose that p_X satisfies the α-margin condition. Furthermore, assume that Assumptions 1 and 3 hold, then we have the following asymptotic bound on the expected cumulative regret of Greedy Bandit algorithm*

$$R_T(\pi) = \begin{cases} \mathcal{O}\left(T^{(1-\alpha)/2}\right) & \text{if } 0 \leq \alpha < 1, \\ \mathcal{O}\left(\log T\right) & \text{if } \alpha = 1, \\ \mathcal{O}(1) & \text{if } \alpha > 1, \end{cases} \tag{20}$$

This result shows that if the distribution p_X satisfies the α-margin condition for $\alpha > 1$, then the Greedy Bandit algorithm is capable of learning the parameters β_i while incurring a constant regret in expectation.

Proof. This corollary can be easily implied from Lemma 9 and Theorem 1 with a very slight modification. Note that all the arguments in Lemma 9 hold and the only difference is where we want to bound the probability $\mathbb{P}[X_t \in I^h]$ in Equation (14). In this Equation, if we use the α-margin bound as

$$\mathbb{P}[X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]] \le C_0' \left(2\delta x_{\max}(h+1)\right)^\alpha ,$$

we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\ge l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\quad \le \sum_{h=0}^{h^{\max}} 2^{1+\alpha} C_0' \delta^{1+\alpha} x_{\max}^{1+\alpha}(h+1)^{1+\alpha} + \mathbb{P}\left[X_t \in \hat{\mathcal{R}}_{i\ge l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right],
\end{aligned}$$

which turns the regret bound in Equation (15) into

$$\begin{aligned}
r_t(\pi) \le (K-1)&\left[C_0' 2^{1+\alpha}\delta^{1+\alpha} x_{\max}^{1+\alpha}\Big(\sum_{h=0}^{h_0}(h+1)^{1+\alpha} + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^{1+\alpha}\exp(-h^2)\Big)\right] \\
&+ 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}^{\lambda_0}_{i,t-1}}),
\end{aligned} \tag{21}$$

Now we claim that the above summation has an upper bound that only depends on d and α. If we prove this claim, the dependency of the regret bound with respect to t can only come from the term $\delta^{1+\alpha}$ and therefore we can prove the desired asymptotic bounds. For proving this claim, consider the summation above and let $h_1 = \lceil \sqrt{3+\alpha} \rceil$. Note that for each $h \ge h_2 = \max(h_0, h_1)$ using $h^2 \ge (3+\alpha)h \ge (3+\alpha)\log h$ we have

$$(h+1)^{1+\alpha}\exp(-h^2) \le (2h)^{1+\alpha}\exp(-h^2) \le 2^{1+\alpha}\exp(-h^2 + (1+\alpha)\log h) \le \frac{2^{1+\alpha}}{h^2}.$$

Furthermore, all the terms corresponding to $h \le h_2 = \max(h_0, h_1)$ have an upper bound equal to $(h+1)^{1+\alpha}$ (remember that for $h \ge h_0 + 1$ we have $4d\exp(-h^2) \le 1$). Therefore, the summation in (21) is bounded above by

$$\sum_{h=0}^{h_0}(h+1)^{1+\alpha} + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^{1+\alpha}\exp(-h^2) \le \sum_{h=0}^{h_2}(h+1)^{1+\alpha} + \sum_{h=h_2+1}^{\infty}\frac{1}{h^2} \le (1+h_2)^{2+\alpha} + \frac{\pi^2}{6} = g(d,\alpha)$$

for some function g. This is true according to the fact that h_2 is the maximum of h_0, that only depends on d, and h_1 that only depends on α. Now replacing $\delta = 1/\sqrt{(t-1)C_2}$ in the Equation (21) and putting together all the constants we reach to

$$r_t(\pi) = (K-1) g_1(d, \alpha, C_0', x_{\max}, \sigma, \lambda_0)(t-1)^{-(1+\alpha)/2} + 4(K-1) b_{\max} x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}^{\lambda_0}_{i,t}}]\right)$$

for some function g_1.

The last part of the proof is summing up the instantaneous regret terms for $t = 1, 2, \ldots, T$. Note that $K = 2$, and using Lemma 4 for $i = 1, 2$, we can bound the probabilities $\mathbb{P}[\overline{\mathcal{F}^{\lambda_0}_{i,t-1}}]$ by $d \exp(-C_1(t-1))$ and therefore

$$
\begin{aligned}
R_T(\pi) &\le 2x_{\max} b_{\max} + \sum_{t=2}^{T} g_1(d, \alpha, C_0', x_{\max}, \sigma, \lambda_0)(t-1)^{-(1+\alpha)/2} + 4 b_{\max} x_{\max} d \exp(-C_1(t-1)) \\
&\le 2x_{\max} b_{\max} + \sum_{t=1}^{T-1} g_1(d, \alpha, C_0', x_{\max}, \sigma, \lambda_0) t^{-(1+\alpha)/2} + 4 b_{\max} x_{\max} d \exp(-C_1 t) \\
&\le 2x_{\max} b_{\max} + g_1(d, \alpha, C_0', x_{\max}, \sigma, \lambda_0) \left[1 + \left(\int_{t=1}^{T} t^{-(1+\alpha)/2} \mathrm{d}t\right)\right] + 4 d b_{\max} x_{\max} \int_0^{\infty} \exp(-C_1 t) \mathrm{d}t \\
&= 2x_{\max} b_{\max} + g_1(d, \alpha, C_0', x_{\max}, \sigma, \lambda_0) \left[1 + \left(\int_{t=1}^{T} t^{-(1+\alpha)/2} \mathrm{d}t\right)\right] + \frac{4 b_{\max} x_{\max} d}{C_1}.
\end{aligned}
$$

Now note that the integral of $t^{-(1+\alpha)/2}$ over the interval $[1, T]$ satisfies

$$
\int_{t=1}^{T} t^{-(1+\alpha)/2} \le \begin{cases} \frac{T^{(1-\alpha)/2}}{(1-\alpha)/2} & \text{if } 0 \le \alpha < 1, \\ \log T & \text{if } \alpha = 1, \\ \frac{1}{(\alpha-1)/2} & \text{if } \alpha > 1, \end{cases}
$$

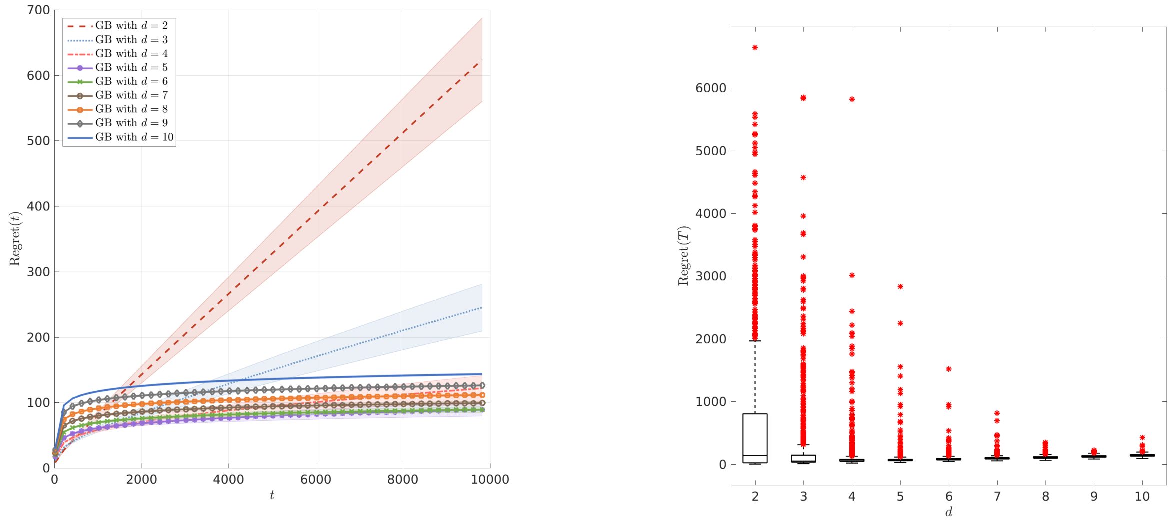
which yields the desired result. □

Appendix F: Additional Simulations

F.1. More than Two Arms ($K > 2$)

For investigating the performance of the Greedy-Bandit algorithm in presence of more than two arms, we run Greedy Bandit algorithm for $K = 5$ and $d = 2, 3, \ldots, 10$ while keeping the distribution of covariates as $0.5 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated at 1. We assume that β_i is again drawn from $\mathrm{N}(0_d, \mathbf{I}_d)$. For having a fair comparison, we scale the noise variance by d so as to keep the signal-to-noise ratio fixed (i.e., $\sigma = 0.25\sqrt{d}$). For small values of d, it is likely that Greedy Bandit algorithm drops an arm due to the poor estimations and as a result its regret becomes linear. However, for large values of d this issue is resolved and Greedy Bandit starts to perform very well.

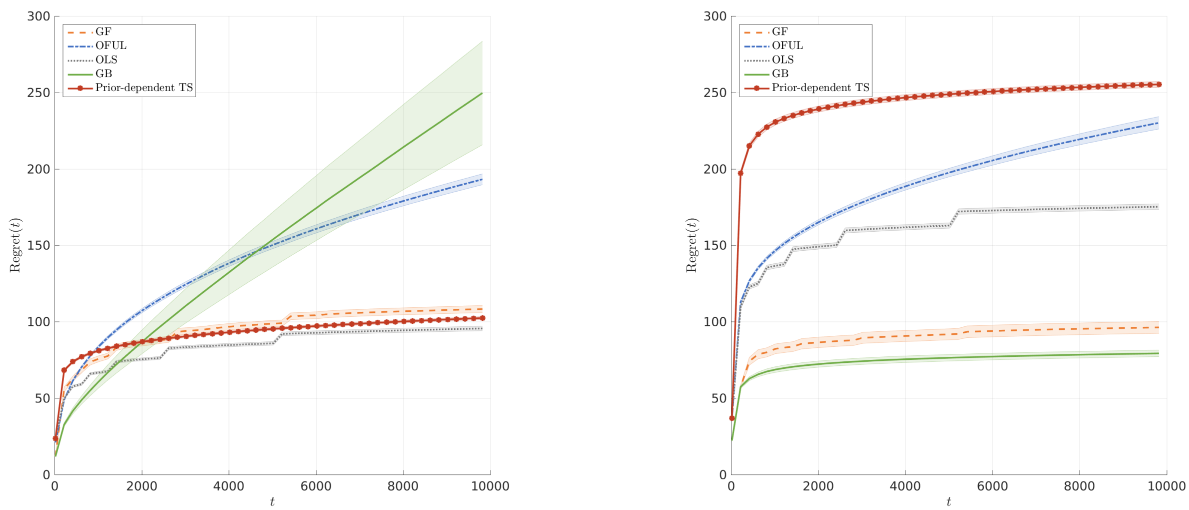
We then repeat the simulations of §5 for $K = 5$ and $d \in \{3, 7\}$ while keeping the other parameters as in §5. In other words, we assume that β_i is drawn from $\mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$. Also, X is drawn from $0.5 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated to have its ℓ_∞ norm at most one. We create 1000 problem instances and plot the average cumulative regret of algorithms for $T \in \{1, 2, \ldots, 10000\}$. We use the correct prior regime for OFUL and TS. The results, as shown in Figure 4, demonstrate that Greedy-First nearly ties with Greedy Bandit as the winner when $d = 7$. However for $d = 3$ that Greedy Bandit performs poorly, while Greedy-First performs very close to the best algorithms.



(a) Regret for $t = 1, \ldots, 10000$.

(b) Distribution of regret at $T = 10000$.

Figure 3 These figures show a sharp change in the performance of Greedy Bandit for $K = 5$ arms as d increases.



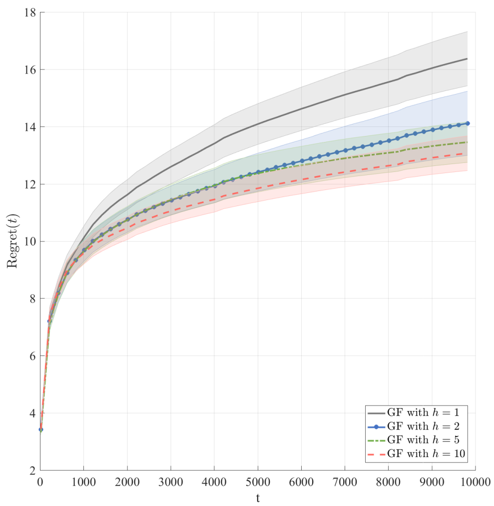
(a) $K = 5, d = 3$

(b) $K = 5, d = 7$

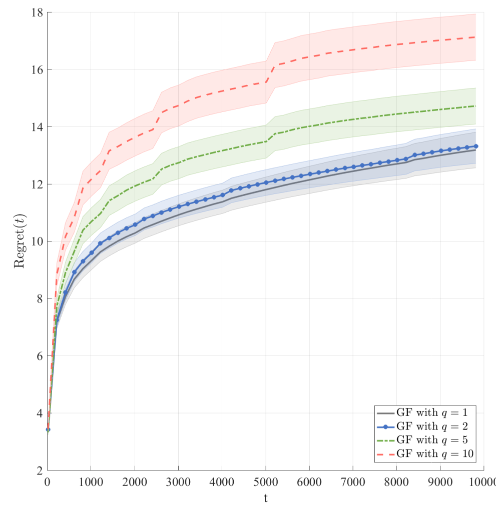
Figure 4 Simulations for $K > 2$ arms.

F.2. Sensitivity to parameters

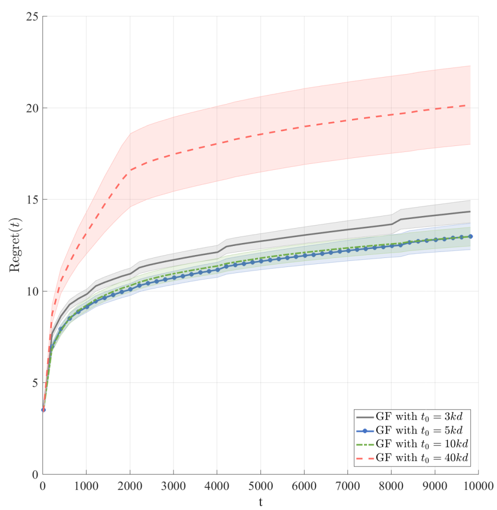
In this section, we will perform a sensitivity analysis to demonstrate that the choice of parameters h, q, and t_0 has a small impact on performance of Greedy First. The sensitivity analysis is performed with the same problem parameters as in Figure 2 for the case that covariate diversity does not hold. As it can be observed from Figure 5, the choice of parameters h, q, and t_0 does have a very small impact on the performance of the Greedy-First algorithm, which verifies the robustness of Greedy-First algorithm to the choice of parameters.



(a) Sensitivity with respect to h.



(b) Sensitivity with respect to q.

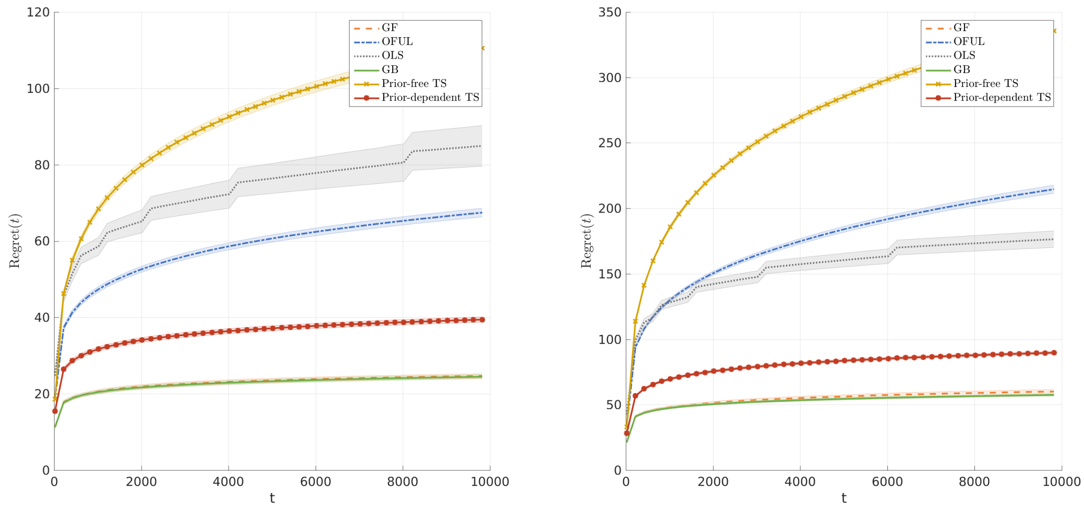


(c) Sensitivity with respect to t_0.

Figure 5 Sensitivity analysis for the expected regret of Greedy-First algorithm with respect to the input parameters h, q, and t_0.

F.3. Real Data Covariates

To test the robustness of our results, we repeat simulations of §5 with a single change that the covariates are not generated by us but instead, they are sampled from a large real data set. In particular, the covariates are obtained from a large data set of hotel listings on Expedia that is publicly available at https://www.kaggle.com/c/expedia-personalized-sort. The covariates are Star Rating, Average Reviews, Brand, Position, Weekend Indicator, Location Popularity, and an indicator variable for Summer for 267832 user bookings. For any one of 1000 simulations of the bandit algorithms, we sample $T = 10000$ covariate vectors from this dataset at random. We use the correct prior regime of §5 and run the simulations for $K = 2$ and $K = 3$. Figure 6 shows the results which has the same qualitative behavior as in Figure 2(a) that the covariate diversity holds.



(a) $K = 2$ (b) $K = 3$

Figure 6 Performance of various algorithms on the Expedia data.

Appendix G: Missing Proofs of §3.4 and §4.3

Proof of Proposition 1. We first start by proving monotonicity results:

- Let $\sigma_1 < \sigma_2$. Note that only the second, the third, and the last term of $L(\gamma, \delta, p)$, defined in Equation (4), depend on σ. As for any positive number χ, the function $\exp(-\chi/\sigma^2)$ is increasing with respect to σ, second and third terms are increasing with respect to σ. Furthermore, the last term can be expressed as

$$\frac{2d \exp\left(-D_2(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_2(\gamma))} = 2d \sum_{t=p-m|\mathcal{K}_{sub}|}^{\infty} \exp\left(-\frac{\lambda_1^2 h^2 (1-\gamma)^2}{8d\sigma^2 x_{\max}^4} t\right).$$

Each term in above sum is increasing with respect to σ. Therefore, the function L is increasing with respect to σ. As S^{gb} is one minus the infimum of L taken over the possible parameter space of γ, δ, and p, that is also non-increasing with respect to σ, yielding the desired result.

• Let $m_1 < m_2$ and suppose that we use the superscript $L^{(i)}$ for the function $L(\cdot,\cdot,\cdot)$ when $m = m_i, i = 1,2$. We claim that for all $\gamma \in (0,1), \delta > 0$, and $p \geq Km_1 + 1$, conditioning on $L^{(1)}(\gamma,\delta,p) \leq 1$ we have $L^{(1)}(\gamma,\delta,p) \geq L^{(2)}(\gamma,\delta,p + K(m_2 - m_1))$. Note that the region for which $L^{(1)}(\gamma,\delta,p) > 1$ does not matter as it leads to a negative probability of success in the formula $S^{\text{gb}} = 1 - \inf_{\gamma,\delta,p} L(\gamma,\delta,p)$, and we can only restrict our attention to the region for which $L^{(1)}(\gamma,\delta,p) \leq 1$. To prove the claim let $\theta_i = \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m_i}^\top \mathbf{X}_{1:m_i}) \geq \delta\right],\ i = 1,2$ and define $f(\theta) = 1 - \theta^K + QK\theta$ for the constant $Q = 2d\exp\left(-(h^2\delta)/(8d\sigma^2 x_{\max}^2)\right)$. Note that $f(\theta_i)$ is equal to to the first two terms of $L^{(i)}(\gamma,\delta,p)$ in Equation (4). As we later going to replace $\theta = \theta_i$ we only restrict our attention to $\theta \geq 0$. The derivative of f is equal to $f'(\theta) = -K\theta^{K-1} + QK$ which is negative when $\theta^{K-1} > Q$. Note that if $\theta^{K-1} \leq Q$ and if we drop the third, fourth, and fifth term in L (see Equation (4)) that are all positive, we obtain $L^{(i)}(\gamma,\delta,p) > 1 - \theta^K + QK\theta > 1 - \theta^K + Q\theta \geq 1$, leaving us in the unimportant regime. Therefore, on the important regime the derivative is negative and f is decreasing. It is not very difficult to see that $\theta_1 \leq \theta_2$. Returning to our original claim, if we calculate $L^{(1)}(\gamma,\delta,p) - L^{(2)}(\gamma,\delta,p + K(m_2 - m_1))$ it is easy to observe that the third term cancels out and we end up with

$$
\begin{aligned}
L^{(1)}(\gamma,\delta,p) - L^{(2)}(\gamma,\delta,p + K(m_2 - m_1)) &= f(\theta_1) - f(\theta_2) \\
&+ \frac{\exp\left(-D_1(\gamma)(p - m_1|\mathcal{K}_{sub}|)\right) - \exp\left(-D_1(\gamma)(p - m_2|\mathcal{K}_{sub}| + K(m_2 - m_1))\right)}{1 - \exp(-D_1(\gamma))} \\
&+ \frac{\exp\left(-D_2(\gamma)(p - m_1|\mathcal{K}_{sub}|)\right) - \exp\left(-D_2(\gamma)(p - m_2|\mathcal{K}_{sub}| + K(m_2 - m_1))\right)}{1 - \exp(-D_2(\gamma))} \\
&\geq 0\,,
\end{aligned}
$$

where we used the inequality $(p - m_1|\mathcal{K}_{sub}|) - (p - m_2|\mathcal{K}_{sub}| + K(m_2 - m_1)) = |\mathcal{K}_{opt}|(m_2 - m_1) \geq 0$. This proves our claim. Note that whenever when p varies in the range $[Km_1 + 1, \infty)$, the quantity $p + K(m_2 - m_1)$ covers the range $[Km_2 + 1, \infty)$. Therefore, we can write that

$$
\begin{aligned}
S^{\text{gb}}(m_1, K, \sigma, x_{\max}, \lambda_1, h) &= 1 - \inf_{\gamma \in (0,1), \delta, p \geq Km_1 + 1} L^{(1)}(\gamma,\delta,p) \leq 1 - \inf_{\gamma \in (0,1), \delta, p \geq Km_1 + 1} L^{(1)}(\gamma,\delta,p + K(m_2 - m_1)) \\
&= 1 - \inf_{\gamma \in (0,1), \delta, p' \geq Km_2 + 1} L^{(2)}(\gamma,\delta,p') = S^{\text{gb}}(m_2, K, \sigma, x_{\max}, \lambda_1, h),
\end{aligned}
$$

as desired.

• Let $h_1 < h_2$. In this case it is very easy to check that the first, fourth and fifth terms in L (see Equation (4)) do not depend on h. Dependency of second and third terms are in the form $\exp(-Qh^2)$ for some constant Q, which is decreasing with respect h. Therefore, if we use the superscript $L^{(i)}$ for the function $L(\cdot,\cdot,\cdot)$ when $h = h_i, i = 1,2$, we have that $L^{(1)}(\gamma,\delta,p) \geq L^{(2)}(\gamma,\delta,p)$ which implies

$$
\begin{aligned}
S^{\text{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h_1) &= 1 - \inf_{\gamma \in (0,1), \delta, p \geq Km + 1} L^{(1)}(\gamma,\delta,p) \leq 1 - \inf_{\gamma \in (0,1), \delta, p \geq Km + 1} L^{(2)}(\gamma,\delta,p) \\
&= 1 - \inf_{\gamma \in (0,1), \delta, p' \geq Km + 1} L^{(2)}(\gamma,\delta,p') = S^{\text{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h_2),
\end{aligned}
$$

as desired.

• Similar to the previous part, it is easy to observe that the first, second, and third term in L, defined in Equation (4) do not depend on λ_1. The dependency of last two terms with respect to λ_1 is of the form $\exp(-Q_1\lambda_1)$ and $\exp(-Q_2\lambda_1^2)$ which both are decreasing functions of λ_1. The rest of argument is similar to the previous part and by replicating it with reach to the conclusion that S^{gb} is non-increasing with respect to λ_1.

• Let us suppose that $K_1 m_1 = K_2 m_2, |\mathcal{K}_{1_{sub}}| m_1 = |\mathcal{K}_{2_{sub}}| m_2$, and $K_1 < K_2$. Similar to before, we use superscript $L^{(i)}$ to denote the function $L(\cdot,\cdot,\cdot)$ when $m = m_i, K = K_i, \mathcal{K}_{sub} = \mathcal{K}_{i_{sub}}$. Then it is easy to check that the last three terms in $L^{(1)}$ and $L^{(2)}$ are the same. Therefore, for comparing $S^{\mathrm{gb}}(m_1, K_1, \sigma, x_{\max}, \lambda_1)$ and $S^{\mathrm{gb}}(m_2, K_2, \sigma, x_{\max}, \lambda_1)$ one only needs to compare the first two terms. Letting $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m_i}^\top \mathbf{X}_{1:m_i}) \geq \delta\right] = \theta_i,\ i = 1, 2$ and $Q = 2d \exp\left(-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right)$ we have

$$L^{(1)}(\gamma,\delta,p) - L^{(2)}(\gamma,\delta,p) = \theta_2^{K_2} - \theta_1^{K_1} + QK_1\theta_1 - QK_2\theta_2.$$

Similar to the proof of second part, it is not very hard to prove that on the reasonable regime for the parameters the function $g(\theta) = -\theta^{K_1} + QK_1\theta$ is decreasing and therefore

$$L^{(1)}(\gamma,\delta,p) - L^{(2)}(\gamma,\delta,p) = \theta_2^{K_2} - \theta_1^{K_1} + QK_1\theta_1 - QK_2\theta_2 \leq \theta_2^{K_2} - \theta_2^{K_1} + QK_1\theta_2 - QK_2\theta_2 < 0,$$

as $\theta_1 \geq \theta_2 \in [0, 1]$ and $K_2 > K_1$. Taking the infimum implies the desired result.

Now let us derive the limit of L when $\sigma \to 0$. For each $\sigma < (1/Km)^2$, define $\gamma(\sigma) = 1/2,\ \delta(\sigma) = \sqrt{\sigma}$, and $p(\sigma) = \lceil 1/\sqrt{\sigma} \rceil$. Then, by computing the function L for these specific choices of parameters and upper bounding the summation in Equation (4) with its maximum times the number of terms we get

$$\begin{aligned} L(\gamma(\sigma),\delta(\sigma),p(\sigma)) \leq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \sqrt{\sigma}\right]\right)^K + 2Kd\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \sqrt{\sigma}\right] \exp\left(-Q_1/\sigma^{3/2}\right) \\ + 2d/\sqrt{\sigma} \exp\left(-Q_2/\sqrt{\sigma}\right) + d\frac{\exp\left(-Q_3/\sqrt{\sigma}\right)}{1 - \exp(-Q_3)} + 2d\frac{\exp\left(-Q_4/\sigma^{5/2}\right)}{1 - \exp\left(-Q_4/\sigma^2\right)} := J(\sigma), \end{aligned}$$

for positive constants Q_1, Q_2, Q_3, and Q_4 that do not depend on σ. Note that for $\sigma > 0$,

$$\inf_{\gamma \in (0,1), \delta > 0, p \geq Km+1} L(\gamma,\delta,p) \leq J(\sigma).$$

Therefore, by taking limit with respect to σ we get

$$\begin{aligned} \lim_{\sigma \downarrow 0} S^{\mathrm{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h) &= 1 - \lim_{\sigma \downarrow 0} L(\gamma,\delta,p) \\ &\geq \lim_{\sigma \downarrow 0} (1 - J(\sigma)) = 1 - \left\{1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]\right)^K\right\} \\ &= \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K, \end{aligned}$$

proving one side of the result. For achieving the desired result we need to prove that $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K \geq \lim_{\sigma\downarrow 0} S^{\mathrm{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h)$ which is the easier way. Note that the function L always satisfies

$$L(\gamma,\delta,p) \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]\right)^K.$$

As a result, for any $\sigma > 0$ we have

$$S^{\mathrm{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h) \leq 1 - \left(1 - \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]\right)^K = \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K.$$

By taking limits we reach to the desired conclusion. □

Proof of Proposition 2. We omit proofs regarding to the monotonicity results as they are very similar to those provided in Proposition 1.

For deriving the limit when $\sigma \to 0$, define $\gamma(\sigma) = \gamma^*$, $\delta(\sigma) = \sqrt{\sigma}$, and $p(\sigma) = t_0$. Then, by computing the function L' for these specific values we have

$$\begin{aligned} L'(\gamma(\sigma),\delta(\sigma),p(\sigma)) \leq 1 &- \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \sqrt{\sigma}\right]\right)^K \\ &+ 2Kd\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \sqrt{\sigma}\right] \exp\left(-Q_1'/\sigma^{3/2}\right) \\ &+ 2dt_0 \exp\left\{-\frac{Q_2'}{\sigma}\right\} + \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))} + 2d\frac{\exp\left(-Q_3' t_0/\sigma^2\right)}{1-\exp\left(-Q_3'/\sigma^2\right)} := J'(\sigma), \end{aligned}$$

for positive constants Q_1', Q_2', and Q_3' that do not depend on σ. Note that for $\sigma > 0$,

$$\inf_{\gamma \leq \gamma^*, \delta > 0, Km+1 \leq p \leq t_0} L'(\gamma,\delta,p) \leq J'(\sigma).$$

Therefore, by taking limit with respect to σ we get

$$\begin{aligned} \lim_{\sigma\downarrow 0} S^{\mathrm{gf}}(m, K, \sigma, x_{\max}, \lambda_1, h) &= 1 - \lim_{\sigma\downarrow 0}) L'(\gamma,\delta,p) \\ &\geq \lim_{\sigma\downarrow 0}(1 - J'(\sigma)) \\ &= 1 - \left\{1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]\right)^K + \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))}\right\} \\ &= \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K - \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))}, \end{aligned}$$

proving one side of the result. For achieving the desired result we need to prove that the other side of this inequality. Note that the function L' always satisfies

$$L'(\gamma,\delta,p) \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K + \frac{Kd\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))}. \tag{22}$$

Note that the function $D_1(\gamma)$ is increasing with respect to γ. This is easy to verify as the first derivative of $D_1(\gamma)$ with respect to γ is equal to

$$\frac{\partial D_1}{\partial \gamma} = \frac{\lambda_1}{x_{\max}^2}\left\{1 - \log(1-\gamma) - 1\right\} = -\frac{\lambda_1}{x_{\max}^2}\log(1-\gamma),$$

which is increasing for $\gamma \in [0,1)$. Therefore, by using $p \le t_0$ and $\gamma \le \gamma^*$ we have

$$\frac{Kd\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))} \ge \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))}.$$

Substituting this in Equation (22) implies that

$$\begin{aligned} S^{\mathrm{gf}}(m,K,\sigma,x_{\max},\lambda_1,h) &\le 1-\left\{\left(1-\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top\mathbf{X}_{1:m})>0\right]\right)^K + \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))}\right\} \\ &= \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top\mathbf{X}_{1:m})>0\right]^K - \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))}. \end{aligned}$$

By taking limits we reach to the desired conclusion. □

Proofs of Theorems 2 and 4

Let us first start by introducing two new notations and recalling some others. For each $\delta > 0$ define

$$\begin{aligned} \mathcal{H}_i^\delta &:= \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top\mathbf{X}(\mathcal{S}_{i,Km})\right)\ge\delta\right\} \\ \mathcal{J}_{i,t}^\lambda &= \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top\mathbf{X}(\mathcal{S}_{i,t})\right)\ge\lambda t - m|\mathcal{K}_{sub}|\right\}, \end{aligned}$$

and recall that

$$\begin{aligned} \mathcal{F}_{i,t}^\lambda &= \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top\mathbf{X}(\mathcal{S}_{i,t})\right)\ge\lambda t\right\} \\ \mathcal{G}_{i,t}^\chi &= \left\{\|\hat\beta(\mathcal{S}_{i,t})-\beta_i\|_2<\chi\right\}. \end{aligned}$$

Note that whenever $|\mathcal{K}_{sub}|=0$, the sets $\mathcal{J}$ and $\mathcal{F}$ coincide. We first start by proving some lemmas that will be used later to prove Theorems 2 and 4.

Lemma 10. *Let $i \in [K]$ be arbitrary. Then*

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] \le 2d\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^\top\mathbf{X}_{1:m}\right)\ge\delta\right\}\exp\left\{-\frac{\theta_1^2\delta}{2d\sigma^2}\right\}$$

Remark 5. Note that Lemma 5 provides an upper bound on the same probability event described above. However, those results are addressing the case that samples are highly correlated due to greedy decisions. In the first Km rounds that m rounds of random sampling are executed for each arm, samples are independent and we can use sharper tail bounds. This would help us to get better probability guarantees for the Greedy Bandit algorithm.

Proof. Note that we can write

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] = \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top\mathbf{X}(\mathcal{S}_{i,Km})\right)\ge\delta, \|\hat\beta(\mathcal{S}_{Km,t})-\beta_i\|_2\ge\theta_1\right]. \tag{23}$$

Note that if $\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right) \geq \delta > 0$, this means that the covariance matrix is invertible. Therefore, we can write

$$\begin{aligned}\hat{\beta}(\mathcal{S}_{Km,t}) - \beta_i &= \left[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})^\top Y(\mathcal{S}_{i,Km}) - \beta_i \\ &= \left[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})^\top \left[\mathbf{X}(\mathcal{S}_{i,Km})\beta_i + \varepsilon(\mathcal{S}_{i,Km})\right] - \beta_i \\ &= \left[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})^\top \varepsilon(\mathcal{S}_{i,Km}).\end{aligned}$$

To avoid clutter, we drop the term $\mathcal{S}_{i,Km}$ in equations. By letting $M = \left[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})$ the probability in Equation (23) turns into

$$\begin{aligned}\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] &= \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \|M\varepsilon\|_2 \geq \theta_1\right] \\ &= \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \sum_{j=1}^d |m_j^\top \varepsilon| \geq \theta_1\right] \\ &\leq \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \exists j \in [d], |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d}\right] \\ &\leq \sum_{j=1}^d \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d}\right] \\ &= \sum_{j=1}^d \mathbb{P}_{\mathbf{X}}\mathbb{P}_{\varepsilon|\mathbf{X}}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d} \mid \mathbf{X} = \mathbf{X}_0\right], \end{aligned} \tag{24}$$

where in the second inequality we used a union bound. Note that in above $\mathbb{P}_{\mathbf{X}}$ means the probability distribution over the matrix $\mathbf{X}$, which can also be thought as the multi-dimensional probability distribution of p_X, or alternatively p_X^m. Now fixing $\mathbf{X} = \mathbf{X}_0$, the matrix M only depends on $\mathbf{X}_0$ and we can use the well-known Chernoff bound for subgaussian random variables to achieve

$$\begin{aligned}\mathbb{P}[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d} \mid \mathbf{X} = \mathbf{X}_0] &= \mathbb{I}\left[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta\right] \mathbb{P}[|m_j^\top \varepsilon| \geq \theta_1/\sqrt{d} \mid \mathbf{X} = \mathbf{X}_0] \\ &\leq 2\mathbb{I}\left[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta\right] \exp\left\{-\frac{\theta_1^2}{2d\sigma^2\|m_j\|_2^2}\right\}\end{aligned}$$

Now note that when $\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta$ we have

$$\max_{j\in[d]} \|m_j\|_2^2 = \max\left(\operatorname{diag}\left(MM^\top\right)\right) = \max\left(\operatorname{diag}\left(\mathbf{X}^\top \mathbf{X}^{-1}\right)\right) \leq \lambda_{\max}\left(\mathbf{X}^\top \mathbf{X}^{-1}\right) = \frac{1}{\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right)} \leq \frac{1}{\delta},$$

Hence,

$$\mathbb{P}_{\varepsilon|\mathbf{X}}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d} \mid \mathbf{X} = \mathbf{X}_0\right] \leq 2\mathbb{I}\left[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta\right] \exp\left\{-\frac{\theta_1^2 \delta}{2d\sigma^2}\right\}.$$

Putting this back in Equation (24) gives

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] \leq 2d\mathbb{P}_{\mathbf{X}}\left[\left(\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right)\right) \geq \delta\right] \exp\left\{-\frac{\theta_1^2 \delta}{2d\sigma^2}\right\} = 2d\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}\right) \geq \delta\right\} \exp\left\{-\frac{\theta_1^2 \delta}{2d\sigma^2}\right\},$$

as desired. In above we use the fact that $\mathbb{P}_{\mathbf{X}}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta\right]$ is equal to $\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}\right) \geq \delta\right\}$ as they both describe the probability that the minimum eigenvalue of a matrix derived from m random samples from p_X is not smaller than δ. $\square$

Lemma 11. *For an arbitrary $Km+1 \le t \le p-1$ and $i \in [K]$ we have*

$$\mathbb{P}\left[\mathcal{H}_i^{\delta} \cap \overline{\mathcal{G}_{i,t}^{\theta_1}}\right] \le 2d \exp\left\{-\frac{\theta_1^2 \delta^2}{2d(t-(K-1)m)\sigma^2 x_{\max}^2}\right\}$$

Proof. This is an immediate consequence of Lemma 5. Replace $\chi = \theta_1, \lambda = \delta/t$ and note that $|\mathcal{S}_{i,t}| \le t-(K-1)m$ always holds as $(K-1)m$ rounds of random sampling for arms other than i exist in algorithm. $\square$

The next step is proving that if all arm estimates are within the ball of radius θ_1 around their true values, the minimum eigenvalue of arms in $\mathcal{K}_{opt}$ grow linearly, while sub-optimal arms are not picked by Greedy Bandit algorithm. The proof is a general extension of Lemma 4.

Lemma 12. *For each $t \ge p, i \in \mathcal{K}_{opt}$*

$$\mathbb{P}\left[\ \overline{\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}} \cap \left(\cap_{l=1}^{K} \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}\right)\right] \le d\exp\left(-D_1(\gamma)(t-m|\mathcal{K}_{sub}|)\right).$$

Furthermore, for each $t \ge Km+1$ and $i \in \mathcal{K}_{sub}$ conditioning on the event $\cap_{l=1}^{K}\mathcal{G}_{l,t-1}^{\theta_1}$, arm i would not be played at time t under greedy policy.

Proof. The idea is again using concentration inequality in Lemma 8. Let $i \in \mathcal{K}_{opt}$ and recall that

$$\tilde{\Sigma}_{i,t} = \sum_{k=1}^{t} \mathbb{E}_{k-1}\left(X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right]\right)$$
$$\hat{\Sigma}_{i,t} = \sum_{k=1}^{t} X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right],$$

denote the expected and sample covariance matrices of arm i at time t respectively. The aim is deriving an upper bound on the probability that minimum eigenvalue of $\hat{\Sigma}_{i,t}$ is less than the threshold $t\lambda_1(1-\gamma) - m|\mathcal{K}_{sub}|$. Note that $\hat{\Sigma}_{i,t}$ consists of two different types of terms: 1) random sampling rounds $1 \le k \le Km$ and 2) greedy action rounds $Km+1 \le k \le t$. We analyze these two types separately as following:

- $k \le Km$. Note that during the first Km periods, each arm receives m random samples from the distribution p_X and therefore using concavity of the function $\lambda_{\min}(\cdot)$ we have

$$\begin{aligned}
\lambda_{\min}\left(\sum_{k=1}^{Km} \mathbb{E}_{k-1}\left(X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right]\right)\right) &\ge m\lambda_{\min}\mathbb{E}\left(XX^\top\right) \\
&\ge m\lambda_{\min}\left(\sum_{j \in \mathcal{K}_{opt}} \mathbb{E}\left(XX^\top \mathbb{I}\left(X^\top \beta_j > \max_{l \ne j} X^\top \beta_l + h\right)\right)\right) \\
&\ge m|\mathcal{K}_{opt}|\lambda_1,
\end{aligned}$$

where X is a random sample from distribution p_X.

- $k \geq Km+1$. If $\mathcal{G}_{l,j}^{\theta_1}$ holds for all $l \in [K]$, then

$$\mathbb{E}_{k-1}\left[X_k X_k^\top \mathbb{I}\left(X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right)\right] \succeq \mathbb{E}\left[XX^\top \mathbb{I}\left(X^\top \hat{\beta}(\mathcal{S}_{i,k}) > \max_{l \neq i} X^\top \hat{\beta}(\mathcal{S}_{l,k})\right)\right] \succeq \lambda_1 \mathbf{I}.$$

The reason is very simple; basically having $\cap_{l=1}^K \mathcal{G}_{l,j}^{\theta_1}$ means that $\|\hat{\beta}(\mathcal{S}_{l,k}) - \beta_l\| < \theta_1$ and therefore for each $\mathbf{x}$ satisfying $\mathbf{x}^\top \beta_i \geq \max_{l \neq i} \mathbf{x}^\top \beta_l + h$, using two Cauchy-Schwarz inequalities we can write

$$\mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,j}) - \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,j}) > \mathbf{x}^\top(\beta_i - \beta_l) - 2x_{\max}\theta_1 = \mathbf{x}^\top(\beta_i - \beta_l) - h \geq 0,$$

for each $l \neq i$. Therefore, by taking a maximum over l we obtain $\mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,j}) - \max_{i \neq l} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,j}) > 0$. Hence,

$$\mathbb{E}_{k-1}\left[X_k X_k^\top \mathbb{I}\left(X_k^\top \hat{\beta}(\mathcal{S}_{i,k}) > \max_{l \neq i} X_k^\top \hat{\beta}(\mathcal{S}_{l,j})\right)\right] \succeq \mathbb{E}\left[XX^\top \mathbb{I}\left(X^\top \beta_i > \max_{l \neq i} X^\top \beta_l + h\right)\right] \succeq \lambda_1 \mathbf{I},$$

using Assumption 4, which holds for all optimal arms, i.e, $i \in \mathcal{K}_{opt}$.

Putting these two results together and using concavity of $\lambda_{\min}(\cdot)$ over positive semi-definite matrices we have

$$\begin{aligned}
\lambda_{\min}\left(\tilde{\Sigma}_{i,t}\right) &= \lambda_{\min}\left(\sum_{k=1}^t \mathbb{E}_{k-1}\left(X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right]\right)\right) \\
&\geq \sum_{k=1}^{Km} \lambda_{\min}\left(\mathbb{E}_{k-1}\left(X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right]\right)\right) + \sum_{k=Km+1}^{t} \lambda_{\min}\left(\mathbb{E}_{k-1}\left(X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right]\right)\right) \\
&\geq m|\mathcal{K}_{opt}|\lambda_1 + (t - Km)\lambda_1 = (t - m|\mathcal{K}_{sub}|)\lambda_1.
\end{aligned}$$

Now the rest of the argument is similar to Lemma 4. Note that in the proof of Lemma 4, we simply put $\gamma = 0.5$, however if use an arbitrary $\gamma \in (0,1)$ together with $X_k X_k^\top \preceq x_{\max}^2 \mathbf{I}$, which is the result of Cauchy-Schwarz inequality, then Lemma 8 implies that

$$\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}_{i,t}\right) \leq (t - m|\mathcal{K}_{sub}|)\lambda_1(1-\gamma) \text{ and } \lambda_{\min}\left(\tilde{\Sigma}_{i,t}\right) \geq (t - m|\mathcal{K}_{sub}|)\lambda_1\right] \leq d\exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right).$$

The second event inside the probability event can be removed, as it always holds under $\left(\cap_{l=1}^K \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}\right)$. The first event also can be translated to $\overline{\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}}$ and therefore for all $i \in \mathcal{K}_{opt}$ we have

$$\mathbb{P}\left[\overline{\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}} \cap \left(\cap_{l=1}^K \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}\right)\right] \leq d\exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right),$$

as desired.

For a sub-optimal arm $i \in \mathcal{K}_{sub}$ using Assumption 4, for each $\mathbf{x} \in \mathcal{X}$ there exist $l \in [K]$ such that $\mathbf{x}^\top \beta_i \leq \mathbf{x}^\top \beta_l - h$ and as a result conditioning on $\cap_{l=1}^K \mathcal{G}_{l,t-1}^{\theta_1}$ by using a Cauchy-Schwarz inequality we have

$$\mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,t-1}) - \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) > \mathbf{x}^\top(\beta_l - \beta_i) - 2x_{\max}\theta_1 = \mathbf{x}^\top(\beta_l - \beta_i) - h > 0.$$

This implies that $i \notin \arg\max_{l \in [K]} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ and therefore arm i is not played for $\mathbf{x}$ at time t (Note that once Km rounds of random sampling are finished the algorithm executes greedy algorithm). As this result holds for all choices of $\mathbf{x} \in \mathcal{X}$, arm i becomes sub-optimal at time t, as desired. $\square$

Here, we state the final Lemma, which bounds the probability that the event $\overline{\mathcal{G}_{i,t}^{\theta_1}}$ occurs whenever $\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}$ holds for any $t \geq p$.

LEMMA 13. *For each $t \geq p, i \in [K]$*

$$\mathbb{P}\left[\,\overline{\mathcal{G}_{i,t}^{\theta_1}} \cap \mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}\right] \leq 2d \exp\left(-D_2(\gamma)(t - m|\mathcal{K}_{sub}|)\right) .$$

Proof. This is again obvious using Lemma 5. □

Now we are ready to prove Theorems 2 and 4. As the proofs of these two theorems are very similar we state and prove a lemma that implies both theorems.

LEMMA 14. *Let Assumption and 4 hold. Suppose that Greedy Bandit algorithm with m-rounds of forced sampling in the beginning is executed. Let $\gamma \in (0,1), \delta > 0, p \geq Km + 1$. Suppose that $\mathcal{W}$ is an event which can be decomposed as $\mathcal{W} = \cap_{t \geq p} \mathcal{W}_t$, then event*

$$\left(\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \mathcal{W}$$

holds with probability at least

$$1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K + 2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\}$$
$$+ \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} + \sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \overline{\mathcal{W}_t}\right)\right] .$$

In above, $\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})$ denotes the minimum eigenvalue of a matrix obtained from m random samples from the distribution p_X and constants are defined in Equations (11) *and* (12)*.*

Proof. One important property to note is the following result on the events:

$$\left\{\left(\cap_{i=1}^{K} \mathcal{G}_{i,t-1}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K} \overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\right\} = \left\{\left(\cap_{i=1}^{K} \mathcal{G}_{i,t-1}^{\theta_1}\right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right\} . \tag{25}$$

The reason is that the estimates for arms other than arm π_t do not change at time t, meaning that for each $i \neq \pi_t, \mathcal{G}_{i,t-1}^{\theta_1} = \mathcal{G}_{i,t}^{\theta_1}$. Therefore, the above equality is obvious. This observation comes handy when we want to avoid using a union bound over different arms for the probability of undesired event. For deriving a lower bound on the probability of desired event we have

$$\mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \mathcal{W}\right] = 1 - \mathbb{P}\left[\left(\cup_{i=1}^{K} \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}}\right) \cup \overline{\mathcal{W}}\right].$$

Therefore, we can write

$$\mathbb{P}\left[\left(\cup_{i=1}^{K} \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}}\right) \cup \overline{\mathcal{W}}\right] \leq \mathbb{P}\left[\cup_{i=1}^{K} \overline{\mathcal{H}_i^{\delta}}\right] + \mathbb{P}\left[\left(\cap_{i=1}^{K} \mathcal{H}_i^{\delta}\right) \cap \left[\left(\cup_{i=1}^{K} \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}}\right) \cup \overline{\mathcal{W}}\right]\right] .$$

The first term is equal to $1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K$. The reason is simple; probability of each $\mathcal{H}_i^{\delta}, i \in [K]$ is given by $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]$ and these events are all independent due to the random sampling. Therefore, the probability that at least one of them does not happen is given by the mentioned expression. In addition, the probability of the second event can be upper bounded by

$$\begin{aligned}
&\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left[\left(\cup_{i=1}^{K}\cup_{t\geq Km}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\cup\overline{\mathcal{W}}\right]\right]\\
&\leq\sum_{l=1}^{K}\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\overline{\mathcal{G}_{l,Km}^{\theta_1}}\right]+\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\mathcal{G}_{i,Km}^{\theta_1}\right)\cap\left[\left(\cup_{i=1}^{K}\cup_{t\geq Km}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\cup\overline{\mathcal{W}}\right]\right]\\
&\leq\sum_{l=1}^{K}\mathbb{P}\left[\mathcal{H}_l^{\delta}\cap\overline{\mathcal{G}_{l,Km}^{\theta_1}}\right]+\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\mathcal{G}_{i,Km}^{\theta_1}\right)\cap\left[\left(\cup_{i=1}^{K}\cup_{t\geq Km}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\cup\overline{\mathcal{W}}\right]\right]\\
&\leq 2Kd\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^{\top}\mathbf{X}_{1:m}\right)\geq\delta\right\}\exp\left\{-\frac{\theta_1^2\delta}{2d\sigma^2}\right\}+\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\mathcal{G}_{i,Km}^{\theta_1}\right)\cap\left[\left(\cup_{i=1}^{K}\cup_{t\geq Km}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\cup\overline{\mathcal{W}}\right]\right],
\end{aligned}$$

where we used Lemma 10 together with a union bound. For finding an upper bound on the the second probability, we treat terms $t\in[Km+1,p-1]$ and $t\geq p$ differently. Basically, for the first interval we have guarantees when $\cap_{i=1}^{K}\mathcal{H}_i^{\delta}$ holds (Lemma 11) and for the second interval the guarantee comes from having the event $\cap_{l=1}^{K}\cap_{j=Km}^{t-1}\mathcal{G}_{l,j}^{\theta_1}$ (Lemma 12). Following this path leads to

$$\begin{aligned}
&\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\mathcal{G}_{i,Km}^{\theta_1}\right)\cap\left[\left(\cup_{i=1}^{K}\cup_{t\geq Km}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\cup\overline{\mathcal{W}}\right]\right]\\
&\leq\sum_{t=Km+1}^{p-1}\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\cap_{k=Km}^{t-1}\mathcal{G}_{i,k}^{\theta_1}\right)\cap\left(\cup_{i=1}^{K}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\right]\\
&\quad+\sum_{t\geq p}\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\cap_{k=Km}^{t-1}\mathcal{G}_{i,k}^{\theta_1}\right)\cap\left(\cup_{i=1}^{K}\overline{\mathcal{G}_{i,t}^{\theta_1}}\cup\overline{\mathcal{W}_t}\right)\right]\\
&\leq\sum_{t=Km+1}^{p-1}\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\mathcal{G}_{i,t-1}^{\theta_1}\right)\cap\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right]+\sum_{t\geq p}\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\left(\cap_{i=1}^{K}\cap_{k=Km}^{t-1}\mathcal{G}_{i,k}^{\theta_1}\right)\cap\left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\cup\overline{\mathcal{W}_t}\right)\right]\\
&\leq\sum_{t=Km+1}^{p-1}\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right]+\sum_{t\geq p}\mathbb{P}\left[\left(\cap_{i=1}^{K}\cap_{k=Km}^{t-1}\mathcal{G}_{i,k}^{\theta_1}\right)\cap\left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\cup\overline{\mathcal{W}_t}\right)\right].
\end{aligned}$$

using Equation (25) and carefully breaking down the event $\left[\left(\cup_{i=1}^{K}\cup_{t\geq Km}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\cup\overline{\mathcal{W}}\right]$. Note that by using the second part of Lemma 12, if the event $\cap_{i=1}^{K}\mathcal{G}_{i,t-1}^{\theta_1}$ holds, then π is equal to one of the elements in $\mathcal{K}_{opt}$ and sub-optimal arms in $\mathcal{K}_{sub}$ will not be pulled. Therefore, with further reduction the first term is upper bounded by

$$\begin{aligned}
\sum_{t=Km+1}^{p-1}\sum_{l\in\mathcal{K}_{opt}}\mathbb{P}\left[\pi_t=l\right]\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right)\cap\overline{\mathcal{G}_{l,t}^{\theta_1}}\right]&\leq\sum_{t=Km+1}^{p-1}\sum_{l\in\mathcal{K}_{opt}}\mathbb{P}\left[\pi_t=l\right]2d\exp\left\{-\frac{\theta_1^2\delta^2}{2d(t-(K-1)m)\sigma^2x_{\max}^2}\right\}\\
&\leq\sum_{t=Km+1}^{p-1}2d\exp\left\{-\frac{\theta_1^2\delta^2}{2d(t-(K-1)m)\sigma^2x_{\max}^2}\right\},
\end{aligned}$$

using uniform upper bound provided in Lemma 11 and $\sum_{l\in\mathcal{K}_{opt}}\mathbb{P}\left[\pi_t=l\right]=1$. This concludes the proof. □

Proof of Theorem 2 The proof consists of using Lemma 14. Basically, if we know that the events $\mathcal{G}_{i,t}^{\theta_1}$ for $i \in [K]$ and $t \geq Km$ all hold, we have derived a lower bound on the probability that greedy succeeds. The reason is pretty simple here, if the distance of true parameters β_i and $\hat{\beta}_i$ is at most θ_1 for each t, we can easily ensure that the minimum eigenvalue of covariance matrices of optimal arms are growing linearly, and sub-optimal arms remain sub-optimal for all $t \geq Km+1$ using Lemma 12. Therefore, we can prove the optimality of Greedy Bandit algorithm and also establish its logarithmic regret. Therefore, in this case we need not use any $\mathcal{W}$ in Lemma 14, we simply put $\mathcal{W}_t = \mathcal{W} = \Omega$, where Ω is the whole probability space. Then we have

$$\mathbb{P}\left[\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right] \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K + 2Kd\, \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\}$$
$$+ \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} + \sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right].$$

The upper bound on the last term can be derived as following

$$\begin{aligned}
&\sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K} \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right)\right] \\
&\quad = \sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K} \overline{\mathcal{G}_{l,t}^{\theta_1}}\right)\right] \\
&\quad \leq \sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \left\{\mathbb{P}\left[\overline{\mathcal{J}_{l,t}^{\lambda_1(1-\gamma)}} \cap \left(\cap_{i=1}^{K} \cap_{j=Km}^{t-1} \mathcal{G}_{i,j}^{\theta_1}\right)\right] + \mathbb{P}\left[\overline{\mathcal{G}_{l,t}^{\theta_1}} \cap \mathcal{J}_{l,t}^{\lambda_1(1-\gamma)}\right]\right\},
\end{aligned}$$

which by using Lemmas 12 and 13 can be upper bounded by

$$\begin{aligned}
&\sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \left\{d \exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right) + 2d \exp\left(-D_2(\gamma)(t - m|\mathcal{K}_{sub}|)\right)\right\} \\
&\qquad = \sum_{t \geq p} \exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right) + \sum_{t \geq p} 2d \exp\left(-D_2(\gamma)(t - m|\mathcal{K}_{sub}|)\right) \\
&\qquad = \frac{d \exp\left(-D_1(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_1(\gamma))} + \frac{2d \exp\left(-D_2(\gamma)(p - |\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_2(\gamma))}.
\end{aligned}$$

Summing up all these term yields the desired upper bound. Now note that this upper bound is algorithm-independent and holds for all values of $\gamma \in (0,1), \delta \geq 0$, and $p \geq Km$ and therefore we can take the supremum over these values for our desired event (or infimum over undesired event). This concludes the proof. □

For proving Theorem 4 the steps are very similar, the only difference is that the desired event happens if all events $\mathcal{G}_{i,t}^{\theta_1}$, $i \in [K], t \geq Km$ hold, and in addition to that, events $\mathcal{F}_{i,t}^{\lambda}, i \in [K], t \geq t_0$ all need to hold for some $\lambda > \lambda_0/4$. Recall that in Theorem 4, $\mathcal{K}_{sub} = \emptyset$ and therefore we can use the notations $\mathcal{J}$ and $\mathcal{F}$ interchangeably. For Greedy-First, we define $\mathcal{W} = \cap_{i \in [K]} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda}$ for some λ. This basically, means we need to take $\mathcal{W}_t = \cap_{i \in [K]} \mathcal{F}_{i,t}^{\lambda}$ for some λ.

Proof of Theorem 4 The proof is very similar to proof of Theorem 2. For arbitrary γ, δ, p with we want to derive a bound on the probability of the event

$$\mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \left(\cap_{i=1}^{K} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}\right)\right] .$$

Note that if $p \leq t_0$ and $\gamma \leq 1 - \lambda_0/(4\lambda_1)$, then having events $\mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}, i \in [K], t \geq p$ implies that the events $\mathcal{F}_{i,t}^{\lambda_0/4}, i \in [K], t \geq t_0$ all hold. In other words, Greedy-First does not switch to the exploratory algorithm and is able to achieve logarithmic regret. Let us substitute $\mathcal{W}_t = \cap_{i=1}^{K} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}$ which implies that $\mathcal{W} = \cap_{i=1}^{K} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}$. Lemma 14 can be used to establish a lower bound on the probability of this event as

$$\begin{aligned}
\mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \left(\cap_{i=1}^{K} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}\right)\right] \geq\ & 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K \\
& + 2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\} \\
& + \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} \\
& + \sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \left(\overline{\cap_{i=1}^{K} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}}\right)\right)\right] .
\end{aligned}$$

Hence, we only need to derive an upper bound on the last term. By expanding this based on the value of π_t we have

$$\begin{aligned}
& \sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \left(\overline{\cap_{i=1}^{K} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}}\right)\right)\right] \\
& \quad = \sum_{t \geq p} \sum_{l=1}^{K} \mathbb{P}[\pi_t = l] \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{l,t}^{\theta_1}} \cup \left(\cup_{i=1}^{K} \overline{\mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}}\right)\right)\right] \\
& \quad \leq \sum_{t \geq p} \sum_{l=1}^{K} \mathbb{P}[\pi_t = l] \left\{\sum_{w=1}^{K} \left(\mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{j=Km}^{t-1} \mathcal{G}_{i,j}^{\theta_1}\right) \cap \overline{\mathcal{F}_{w,t}^{\lambda_1(1-\gamma)}}\right]\right) + \mathbb{P}\left[\ \overline{\mathcal{G}_{l,t}^{\theta_1}} \cap \mathcal{F}_{l,t}^{\lambda_1(1-\gamma)}\right]\right\},
\end{aligned}$$

using a union bound and the fact that the space $\overline{\mathcal{F}_{l,t}^{\lambda_1(1-\gamma)}}$ has already been included in the first term, so its complement can be included in the second term. Now, using Lemmas 12 and 13 this can be upper bounded by

$$\begin{aligned}
\sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \left\{Kd \exp(-D_1(\gamma)t) + 2d \exp(-D_2(\gamma)t)\right\} &= \sum_{t \geq p} Kd \exp(-D_1(\gamma)t) + \sum_{t \geq p} 2d \exp(-D_2(\gamma)t) \\
&= \frac{Kd \exp(-D_1(\gamma)p)}{1 - \exp(-D_1(\gamma))} + \frac{2d \exp(-D_2(\gamma)p)}{1 - \exp(-D_2(\gamma))} .
\end{aligned}$$

As mentioned earlier, we can take supremum on parameters p, γ, δ as long as they satisfy $p \leq t_0, \gamma \leq 1 - \lambda_0/(4\lambda_1)$, and $\delta > 0$. They would lead to the same result only with the difference that the infimum over L should be replaced by L' and these two functions satisfy

$$L'(\gamma, \delta, p) = L(\gamma, \delta, p) + (K-1) \frac{d \exp(-D_1(\gamma)p)}{1 - \exp(-D_1(\gamma))},$$

which yields the desired result. $\square$

Proof of Corollary 1. We want to use the result of Theorem 2. In this theorem, let us substitute $\gamma = 0.5, p = Km+1$, and $\delta = 0.5\lambda_1 m|\mathcal{K}_{opt}|$. After this substitution, Theorem 2 implies that the Greedy Bandit algorithm succeeds with probability at least

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right]^K - 2Kd\,\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right] \exp\left\{-\frac{0.5h^2\lambda_1 m|\mathcal{K}_{opt}|}{8d\sigma^2 x_{\max}^2}\right\}$$
$$- \frac{d\exp\{-D_1(0.5)(Km+1-m|\mathcal{K}_{sub}|)\}}{1-\exp\{-D_1(0.5)\}}$$
$$- \frac{2d\exp\{-D_2(0.5)(Km+1-m|\mathcal{K}_{sub}|)\}}{1-\exp\{-D_2(0.5)\}}.$$

For deriving a lower bound on the first term let us use the concentration inequality in Lemma 8. Note that here the samples are drawn i.i.d. from the same distribution p_X. Therefore, by applying this Lemma we have

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \leq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right) \text{ and } \mathbb{E}[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})] \geq \lambda_1 m|\mathcal{K}_{opt}|\right] \leq d\left(\frac{e^{-0.5}}{0.5^{0.5}}\right)^{\lambda_1 m|\mathcal{K}_{opt}|/x_{\max}^2}$$
$$= d\exp\left\{-\frac{\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}(-0.5-0.5\log(0.5))\right\} \geq d\exp\left(-0.153\frac{\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\right).$$

Note that the second event, i.e. $\mathbb{E}[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})] \geq \lambda_1 m|\mathcal{K}_{opt}|$ happens with probability one. This is true according to

$$\mathbb{E}[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})] = \mathbb{E}[\lambda_{\min}(\sum_{l=1}^m X_l X_l^\top)] \geq \mathbb{E}[\sum_{l=1}^m \lambda_{\min}(X_l X_l^\top)] = \sum_{l=1}^m \mathbb{E}[\lambda_{\min}(X_l X_l^\top)] = m\mathbb{E}[\lambda_{\min}(XX^\top)],$$

where $X \sim p_X$ and the inequality is true according to the Jensen's inequality for the concave function $\lambda_{\min}(\cdot)$. Now note that, this expectation can be bounded by

$$\begin{aligned}\mathbb{E}[\lambda_{\min}(XX^\top)] &\geq \mathbb{E}\left[\lambda_{\min}\left(\sum_{i=1}^K XX^\top \mathbb{I}(X^\top\beta_i \geq \max_{j\neq i} X^\top\beta_j + h)\right)\right] \\ &\geq \sum_{i=1}^K \mathbb{E}\left[\lambda_{\min}\left(XX^\top \mathbb{I}(X^\top\beta_i \geq \max_{j\neq i} X^\top\beta_j + h)\right)\right] \\ &\geq |\mathcal{K}_{opt}|\lambda_1,\end{aligned}$$

according to Assumption 4 and another use of Jensen's inequality for the function $\lambda_{\min}(\cdot)$. Note that this part of proof was very similar to Lemma 12. Thus, with a slight modification we get

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right] \geq 1 - d\exp\left(-0.153\frac{\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\right).$$

After using this inequality together with the inequality $(1-x)^K \geq 1-Kx$, and after replacing values of $D_1(0.5)$ and $D_2(0.5)$, the lower bound on the probability of success of Greedy Bandit reduces to

$$1 - Kd\exp\left(\frac{-0.153\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\right) - 2Kd\exp\left(-\frac{h^2\lambda_1 m|\mathcal{K}_{opt}|}{16d\sigma^2 x_{\max}^2}\right)$$
$$- d\sum_{l=(K-|\mathcal{K}_{sub}|)m+1}^{\infty} \exp\left(\frac{-0.153\lambda_1}{x_{\max}^2}l\right) - 2d\sum_{l=(K-|\mathcal{K}_{sub}|)m+1}^{\infty} \exp\left(-\frac{\lambda_1^2 h^2}{32d\sigma^2 x_{\max}^4}l\right).$$

In above we used the expansion $1/(1-x)=\sum_{l=0}^{\infty}x^l$. In order to finish the proof note that by a Cauchy-Schwarz inequality $\lambda_1 \leq x_{\max}^2$. Furthermore, $K-|\mathcal{K}_{sub}|=|\mathcal{K}_{opt}|$ and therefore the above bound is greater than or equal to

$$1-Kd\sum_{l=m|\mathcal{K}_{opt}|}^{\infty}\exp\left(\frac{-0.153\lambda_1}{x_{\max}^2}l\right)-2Kd\sum_{l=m|\mathcal{K}_{opt}|}^{\infty}\exp\left(-\frac{\lambda_1^2h^2}{32d\sigma^2x_{\max}^4}l\right)\geq 1-\frac{3Kd\exp(-D_{\min}m|\mathcal{K}_{opt}|)}{1-\exp(-D_{\min})},$$

as desired. $\square$

Proof of Corollary 2. Proof of this corollary is very similar to the previous corollary. Extra conditions of the corollary ensure that both $\gamma=0.5, p=Km+1$ lie on their accepted region. For avoiding clutter, we skip the proof. $\square$