Mostly Exploration-Free Algorithms for Contextual Bandits

Hamsa Bastani
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 (achieves asymptotically optimal regret) if there is sufficient randomness in the observed contexts (covariates). 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 positive 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 exploration and outperforms existing (exploration-based) contextual bandit algorithms such as Thompson sampling or upper confidence bound (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 (Thompson 1933, 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

$$X_t^\top \beta_i + \varepsilon_{i,t}\,, \tag{1}$$

where $\varepsilon_{i,t}$ are idiosyncratic shocks. One can also consider nonlinear rewards given by generalized linear models (e.g., logistic, probit, and Poisson regression); in this case, (1) is replaced with

$$\mu(X_t^\top \beta_i) + \varepsilon_{i,t}\,, \tag{2}$$

where μ is a suitable *inverse link function* (Filippi et al. 2010, Li et al. 2017). 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 2014b), information-directed sampling (Russo and Van Roy 2014a), 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 specific 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*; this property requires that the covariance matrix of the observed contexts conditioned on any half space is positive definite. We show that covariate diversity is satisfied by a natural class of continuous and discrete context distributions. Furthermore, even absent covariate diversity, we show that a greedy approach provably converges to the optimal policy with some probability that depends on the problem parameters. Our results hold for arm rewards given by both linear and generalized linear models. Thus, exploration may not be necessary at all in a general class of problem instances, and is only sometimes be necessary in other problem instances.

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 positive 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 forced 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 a positive fraction of the time. Greedy-First leverages this observation by following a purely greedy algorithm until it detects that this approach has failed. Thus, in any bandit problem, the Greedy-First policy explores less on average 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

We study sequential decision-making algorithms under the classic *linear contextual bandit* framework, which has been extensively studied in the computer science, operations, and statistics literature (see Chapter 4 of Bubeck and Cesa-Bianchi (2012) for an informative review). A key feature of this setting is the presence of *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 bandit problems.

The contextual bandit 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) proved an upper bound of $\mathcal{O}(d\sqrt{T})$ regret after T time periods when contexts are d-dimensional. While this literature often allows for arbitrary (adversarial) context sequences, we consider the more restricted setting where contexts are generated i.i.d. from some unknown distribution. This additional structure is well-suited to certain applications (e.g., clinical trials on treatments for a non-infectious disease) and allows for improved regret bounds in T (see Goldenshluger and Zeevi 2013, who prove an upper bound of $\mathcal{O}(d^3 \log T)$ regret), and more importantly, allows us to delve into the performance of exploration-free policies which have not been analyzed previously.

Recent work has applied contextual bandit techniques for personalization in a variety of applications such as healthcare (Bastani and Bayati 2015, Tewari and Murphy 2017, Mintz et al. 2017, Kallus and Zhou 2018, Chick et al. 2018, Zhou et al. 2019), recommendation systems (Chu et al. 2011, Kallus and Udell 2016, Agrawal et al. 2017, Bastani et al. 2018), and dynamic pricing (Cohen et al. 2016, Qiang and Bayati 2016, Javanmard and Nazerzadeh 2019, Ban and Keskin 2018, Bastani et al. 2019). However, this substantial literature requires exploration. Exploration-free greedy policies are desirable in practical settings where exploration may be costly or unethical.

Greedy Algorithms. A related literature studies greedy (but not exploration-free) algorithms in discounted Bayesian multi-armed bandit problems. The seminal paper by Gittins (1979) showed that greedily applying an index policy is optimal for a classical multi-armed bandit in Bayesian regret (with a known prior over the unknown parameters). Woodroofe (1979) and Sarkar (1991)

extend this result to a Bayesian one armed bandit with a single i.i.d. covariate when the discount factor approaches 1, and Wang et al. (2005a,b) generalize this result with a single covariate and two arms. Mersereau et al. (2009) further model known structure between arm rewards. However, these policies are not exploration-free; in particular, the Gittins index of an arm is not simply the arm parameter estimate, but includes an additional factor that implicitly captures the value of exploration for under-sampled arms. Recent work has shown a sharp equivalence between the UCB policy (which incorporates exploration) and the Gittins index policy as the discount factor approaches one (Russo 2019). In contrast, we consider a greedy policy with respect to *unbiased* arm parameter estimates, i.e., without incorporating any exploration. It is surprising that such a policy can be effective; in fact, we show that it is not rate optimal in general, but is rate optimal for the linear contextual bandit if there is sufficient randomness in the context distribution.

It is also worth noting that, unlike the literature above, we consider undiscounted minimax regret with unknown and deterministic arm parameters. Gutin and Farias (2016) show that the Gittins analysis does not succeed in minimizing Bayesian regret over all sufficiently large horizons, and propose "optimistic" Gittins indices (which incorporate additional exploration) to solve the undiscounted Bayesian multi-armed bandit.

There are also technical parallels between our work and the analysis of greedy policies in the dynamic pricing literature (Lattimore and Munos 2014, Broder and Rusmevichientong 2012). When there is no context, the greedy algorithm provably converges to a suboptimal price with nonzero probability (den Boer and Zwart 2013, Keskin and Zeevi 2014, 2015). However, in the presence of contexts, Qiang and Bayati (2016) show that changes in the demand environment can induce natural exploration for an exploration-free greedy algorithm, thereby ensuring asymptotically optimal performance. Our work significantly differs from this line of analysis since we need to learn multiple reward functions (for each arm) simultaneously. Specifically, in dynamic pricing, the decision-maker always receives feedback from the true demand function; in contrast, in the contextual bandit, we only receive feedback from a decision if we choose it, thereby complicating the analysis. As a result, the greedy policy is always rate-optimal in the setting of Qiang and Bayati (2016), but only rate-optimal in the presence of covariate diversity in our setting.

Covariate Diversity. The adaptive control theory literature has studied "persistent excitation": for linear models, if the sample path of the system satisfies this condition, then the minimum eigenvalue of the covariance matrix grows at a suitable rate, implying that the parameter estimates converge over time (Narendra and Annaswamy 1987, Nguyen 2018). Thus, if persistent excitation holds for each arm, we will eventually recover the true arm rewards. However, the problem remains to derive policies that ensure that such a condition holds for each (optimal) arm; classical bandit algorithms achieve this goal with high probability by incorporating exploration for under-sampled

arms. Importantly, a greedy policy that does not incorporate exploration may not satisfy this condition, e.g., the greedy policy may "drop" an arm. The covariate diversity assumption ensures that there is sufficient randomness in the observed contexts, thereby exogenously ensuring that persistent excitation holds for each arm regardless of the sample path taken by the bandit algorithm.

Conservative Bandits. 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. This literature proposes algorithms that restrict the amount of exploration (similar to the present work) in order to satisfy a 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.

Additional Related Work. Since the first draft of this paper appeared online, there have been two follow-up papers that cite our work and provide additional theoretical and empirical validation for our results. Kannan et al. (2018) consider the case where an adversary selects the observed contexts, but these contexts are then perturbed by white noise; they find that the greedy algorithm can be rate optimal in this setting even for small perturbations. Bietti et al. (2018) perform an extensive empirical study of contextual bandit algorithms on 524 datasets that are publicly available on the OpenML platform. These datasets arise from a variety of applications including medicine, natural language, and sensors. Bietti et al. (2018) find that the greedy algorithm outperforms a wide range of bandit algorithms in cumulative regret on more that 400 datasets. This study provides strong empirical validation of our theoretical findings.

1.2. Main Contributions and Organization of the Paper

We begin by studying conditions under which the greedy algorithm performs well. In §2, we introduce the *covariate diversity* condition (Assumption 3), and show that it holds for a general class of continuous and discrete context distributions. In §3, we show that when covariate diversity holds, the greedy policy is asymptotically optimal for a two-armed contextual bandit with linear rewards (Theorem 1); this result is extended to rewards given by generalized linear models in Proposition 1. For problem instances with more than two arms or where covariate diversity does not hold, we prove that the greedy algorithm is asymptotically optimal with some probability, and we provide a lower bound on this probability (Theorem 2).

Building on these results, in §4, we introduce the Greedy-First algorithm that uses observed contexts and rewards to determine whether the greedy algorithm is failing or not via a hypothesis test. If the test detects that the greedy steps are not receiving sufficient exploration, the algorithm switches to a standard exploration-based algorithm. We show that Greedy-First achieves rate optimal regret bounds without our additional assumptions on covariate diversity or number of arms.

More importantly, we prove that Greedy-First remains purely greedy (while achieving asymptotically optimal regret) for almost all problem instances for which a pure greedy algorithm is sufficient (Theorem 3). Finally, for problem instances with more than two arms or where covariate diversity does not hold, we prove that Greedy-First remains exploration-free and rate optimal with some probability, and we provide a lower bound on this probability (Theorem 4). This result implies that Greedy-First reduces exploration on average compared to standard bandit algorithms.

Finally, in §5, we run several simulations on synthetic and real datasets to verify our theoretical results. We find that the greedy algorithm outperforms standard bandit algorithms when covariate diversity holds, but can perform poorly when this assumption does not hold. However, Greedy-First outperforms standard bandit algorithms even in the absence of covariate diversity, while remaining competitive with the greedy algorithm in the presence of covariate diversity. Thus, Greedy-First provides a desirable compromise between avoiding exploration and learning the true policy.

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 distribution that admits probability density $p_X(\mathbf{x})$ with respect to the Lebesgue measure. If we pull arm $i \in [K]$, we observe a stochastic linear reward (in §3.4, we discuss how our results can be extended to generalized linear models)

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

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*

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

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 closed ℓ_2 ball of radius R around the origin in $\mathbb{R}^d$ defined as $B_R^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$, and let the volume of a set $S \subset \mathbb{R}^d$ be $\text{vol}(S) \equiv \int_S \mathrm{d}\mathbf{x}$.

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).

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 classification 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 can obtain analogous results for the simple greedy algorithm as well (see Appendix E.2 for details). 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.

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

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

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 $\mathbf{u} \in \mathbb{R}^d$ the minimum eigenvalue of $\mathbb{E}_X\left[XX^\top \mathbb{I}\{X^\top \mathbf{u} \geq 0\}\right]$ is at least λ_0, i.e.,*

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

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.

REMARK 1. As discussed in the related literature, the adaptive control theory literature has studied "persistent excitation," which is reminiscent of the covariate diversity condition without the indicator function $\mathbb{I}\{X^\top \mathbf{u} \geq 0\}$. If persistent excitation holds for each arm in a given sample path, then the minimum eigenvalue of the corresponding covariance matrix grows at a suitable rate, and the arm parameter estimate converges over time. However, a greedy policy that does not incorporate exploration may not satisfy this condition, e.g., the greedy policy may "drop" an arm. Assumption 3 ensures that there is sufficient randomness in the observed contexts, thereby exogenously ensuring that persistent excitation holds for each arm (see Lemma 4), regardless of the sample path taken by the bandit algorithm.

2.2. Examples of Distributions Satisfying Assumptions 1-3

While Assumptions 1-2 are generic, it is not straightforward to verify Assumption 3. The following lemma provides sufficient conditions (that are easier to check) that guarantee Assumption 3.

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 straightforward to verify that these examples (and any product of their distributions) 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

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

   Taking $W = B^d_{x_{\max}}$, conditions (a) and (b) of Lemma 1 follow immediately. Condition (c) of Lemma 1 holds with constant

$$\lambda = \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x^2_{\max}}{2\lambda_{\min}(\Sigma)}\right) \frac{x^2_{\max}}{d+2} \mathrm{vol}(B^d_{x_{\max}}) \,,$$

   as shown in Lemma 7 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

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

   for any $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \{\pm 1\}^d$. Here, $J_{ij} \in \mathbb{R}$ 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 $J_{ij} = 0$).

A special case under which the conditions in Lemma 1 hold is when W is the entire support of the distribution P_X; this is the case in the Gaussian and Gibbs distributions, where $W = B^d_{x_{\max}}$ and $W = \{\pm 1\}^d$ respectively. Now, let $X^{(1)}$ be a random vector that satisfies this special case and has mean 0. Let $X^{(2)}$ be another vector that is independent of $X^{(1)}$ and satisfies the general form of Lemma 1. Then it is easy to see that $X = (X^{(1)}, X^{(2)})$ also satisfies the conditions in Lemma 1: parts (a) and (b) clearly hold; to see why (c) holds, note that the cross diagonal entries in $XX^\top$ are zero since $X^{(1)}$ has mean 0. This construction illustrates how covariate diversity works for distributions that contain a mixture of discrete and continuous components.

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 use the notation $\mathbf{X}(\mathcal{S}_{i,t}), Y(\mathcal{S}_{i,t})$, and $\varepsilon(\mathcal{S}_{i,t})$ to refer to the design matrix, the outcome vector, and vector of idiosyncratic shocks respectively, for observations restricted to time periods in $\mathcal{S}_{i,t}$. We estimate β_i at time t based on $\mathbf{X}(\mathcal{S}_{i,t})$ and $Y(\mathcal{S}_{i,t})$, using ordinary least squares (OLS) regression that is 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 its performance guarantees under covariate diversity.

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 $Y_{\pi_t,t} = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$ is observed. We then update our estimate for arm π_t but 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 (see Remark 2 below for alternative choices). The pseudo-code for the algorithm is given in Algorithm 1.

Algorithm 1 Greedy Bandit

Initialize $\hat{\beta}(\mathcal{S}_{i,0}) = 0 \in \mathbb{R}^d$ 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 $Y_{\pi_t,t} = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$
  If $\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ is invertible, update the arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t})$ via

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

end for

REMARK 2. In Algorithm 1, we only update the arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t})$ from its (arbitrary) initial value of 0 when the covariance matrix $\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ is invertible. However, one can alternatively update the parameter using ridge regression or a pseudo inverse to improve empirical performance. Our theoretical analysis is unaffected by this choice — as we will show in Lemma 4, no matter what estimator $\hat{\beta}(\mathcal{S}_{i,t})$ we use, covariate diversity ensures that the probability that these covariance matrices are singular is upper bounded by $\exp(\log d - C_1 t)$, thereby contributing at most an additive constant factor to the cumulative regret (the second term in Lemma 6).

3.2. Performance of Greedy Bandit with Covariate Diversity

We now establish a finite-sample upper bound on the cumulative expected regret of the Greedy Bandit for the two-armed contextual bandit when covariate diversity is satisfied.

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

$$R_T(\pi) \leq \frac{128 C_0 \bar{C} x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} \log T + \bar{C} \left(\frac{128 C_0 x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} + \frac{160 b_{\max} x_{\max}^3 d}{\lambda_0} + 2 x_{\max} b_{\max} \right) \tag{3}$$

$$\leq C_{GB} \log T = \mathcal{O}(\log T) \,,$$

where the constant C_0 is defined in Assumption 2 and

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

We prove an analogous result for the greedy algorithm in the case where arm rewards are given by generalized linear models (see §3.4 and Proposition 1 for details).

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 their result applies to our setting. Combined with our upper bound (Theorem 1), we conclude that the Greedy Bandit is rate optimal.

REMARK 3. Our upper bound in Theorem 1 scales as $\mathcal{O}(d^3 (\log d)^{3/2} \log T)$ in the context dimension d. This is because the term $x_{\max}^2 / \lambda_0$ scales as $\mathcal{O}(d)$ for standard distributions satisfying covariate diversity (e.g., truncated multivariate gaussian or uniform distribution). Thus, our upper bound for the Greedy Bandit is slightly worse (by a factor of d) than the upper bound of $\mathcal{O}(d^2 (\log d)^{3/2} \log T)$ established in Bastani and Bayati (2015) for the OLS Bandit.

3.3. Proof of Theorem 1

Notation. Let $\mathcal{R}_i = \left\{ \mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \beta_i \geq \max_{j \neq i} \mathbf{x}^\top \beta_j \right\}$ denote the true set of contexts where arm i is optimal. Then, let $\hat{\mathcal{R}}_{i,t}^\pi = \left\{ \mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq \max_{j \neq i} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{j,t-1}) \right\}$ denote the estimated set of contexts at time t where arm i appears optimal; in other words, if the context $X_t \in \hat{\mathcal{R}}_{i,t}^\pi$, then the greedy policy will choose arm i at time t. (since we assume without loss of generality that ties are broken randomly as selected by π and thus, $\{\mathcal{R}_i\}_{i=1}^K$ and $\{\hat{\mathcal{R}}_{i,t}^\pi\}_{i=1}^K$ 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. Furthermore, let $\mathcal{H}_{t-1}^- = \sigma\left(\mathbf{X}_{1:t-1}, \pi_{1:t-1}, Y_1(\mathcal{S}_{1,t-1}), Y_2(\mathcal{S}_{2,t-1}), \ldots, Y_K(\mathcal{S}_{K,t-1})\right)$ which is the σ-algebra containing all observed information *before* time t.

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

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* $\mathbf{u} \in \mathbb{R}^d$*:*

$$\mathbb{P}_X[\mathbf{x}^\top \mathbf{u} \geq 0] \geq \frac{\lambda_0}{x_{\max}^2}.$$

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

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

since the indicator function and p_X are both nonnegative. □

Taking $\mathbf{u} = \hat{\beta}(\mathcal{S}_{1,t}) - \hat{\beta}(\mathcal{S}_{2,t})$, 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 9 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* $C_1 = \lambda_0 / (40 x_{\max}^2)$*. 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 $\tilde{\Sigma}_{1,t} = \sum_{k=1}^{t} \mathbb{E}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{1,k}] \mid \mathcal{H}_{k-1}^-\right]$. Expanding one term in the sum, we can write

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

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

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

Next, we seek to use matrix concentration inequalities (Lemma 9 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 9, 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. (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.). Instead, we use a Bernstein concentration inequality to guarantee convergence with adaptive observations. In the following lemma, note that n is any deterministic upper bound on the total number of times that arm i is pulled until time t. In the proof of Lemma 6, we will take $n = t$; however, we state the lemma for general n for later use in our probabilistic guarantees.

LEMMA 5. *Taking* $C_2 = \lambda^2/(2d\sigma^2 x_{\max}^2)$ *and* $n \geq |\mathcal{S}_{i,t}|$, *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

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

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 8 in Appendix B), we can write

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

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, one can use a standard peeling argument (as in Goldenshluger and Zeevi (2013)) to bound the instantaneous expected regret of the Greedy Bandit algorithm.

LEMMA 6. *Define* $\mathcal{F}_{i,t}^{\lambda} = \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^{\top}\mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\right\}$. *Then, the instantaneous expected regret of the Greedy Bandit at time* $t \geq 2$ *satisfies*

$$r_t(\pi) \leq \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t-1} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right),$$

where $C_3 = \lambda_0^2/(32d\sigma^2x_{\max}^2)$, C_0 *is defined in Assumption 2, and* $\bar{C}$ *is defined in Theorem 1.*

Note that $\mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]$ can be upper bounded using Lemma 4. Substituting this in the upper bound derived on $r_t(\pi)$ in Lemma 6, and using $R_T(\pi) = \sum_{t=1}^{T} r_t(\pi)$ finishes the proof of Theorem 1.

3.4. Generalized Linear Rewards

In this section, we discuss how our results generalize when the arm rewards are given by a generalized linear model (GLM). Now, upon playing arm i after observing context X_t, the decision-maker realizes a reward $Y_{i,t}$ with expectation $\mathbb{E}[Y_{i,t}] = \mu(X_t^{\top}\beta_i)$, where μ is the inverse link function. For instance, in logistic regression, this would correspond to a binary reward $Y_{i,t}$ with $\mu(z) = 1/(1+\exp(-z))$; in Poisson regression, this would correspond to an integer-valued reward $Y_{i,t}$ with $\mu(z) = \exp(z)$; in linear regression, this would correspond to $\mu(z) = z$.

In order to describe the greedy policy in this setting, we give a brief overview of the exponential family, generalized linear model, and maximum likelihood estimation.

Exponential family. A univariate probability distribution belongs to the *canonical exponential family* if its density with respect to a reference measure (e.g., Lebesgue measure) is given by

$$p_{\theta}(z) = \exp\left[z\theta - A(\theta) + B(z)\right], \tag{5}$$

where θ is the underlying real-valued parameter, $A(\cdot)$ and $B(\cdot)$ are real-valued functions, and $A(\cdot)$ is assumed to be twice continuously differentiable. For simplicity, we assume the reference measure is the Lebesgue measure. It is well known that if Z is distributed according to the above canonical exponential family, then it satisfies $\mathbb{E}[Z] = A'(\theta)$ and $\mathrm{Var}[Z] = A''(\theta)$, where A' and A'' denote the first and second derivatives of the function A with respect to θ, and A is strictly convex (see e.g., Lehmann and Casella 1998).

Generalized linear model (GLM). The natural connection between exponential families and GLMs is provided by assuming that the density of $Y_{i,t}$ for the context X_t and arm i is given by $g_{\beta_i}(Y_{i,t} \mid X_t) = p_{X_t^{\top}\beta_i}(Y_{i,t})$. where p is defined in (5). In other words, the reward upon playing arm i for context X_t is $Y_{i,t}$ with density

$$\exp\left[Y_{i,t}X_t^{\top}\beta_i - A(X_t^{\top}\beta_i) + B(Y_{i,t})\right].$$

Using the aforementioned properties of the exponential family, $\mathbb{E}[Y_{i,t}] = A'(X_t^{\top}\beta_i)$, i.e., the link function $\mu = A'$. This implies that μ is continuously differentiable and its derivative is A''. Thus, μ is strictly increasing since A is strictly convex.

Maximum likelihood estimation. Suppose that we have n samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from a distribution with density $g_\beta(Y \mid X)$. The maximum likelihood estimator of β based on this sample is given by

$$\arg\max_\beta \sum_{\ell=1}^{n} \log g_\beta(Y_\ell \mid X_\ell) = \arg\max_\beta \sum_{\ell=1}^{n} \left[Y_\ell X_\ell^\top \beta - A(X_\ell^\top \beta) + B(Y_\ell) \right] . \tag{6}$$

Since A is strictly convex (so $-A$ is strictly concave), the solution to (6) can be obtained efficiently (see e.g., McCullagh and Nelder 1989). It is not hard to see that whenever $\mathbf{X}^\top \mathbf{X}$ is positive definite, this solution is unique (see Appendix E.1 for a proof). We denote this unique solution by $h_\mu(\mathbf{X}, \mathbf{Y})$.

Now we are ready to generalize the Greedy Bandit algorithm when the arm rewards are given by a GLM. Using similar notation as in the linear reward case, given the estimates $\left\{ \hat{\beta}(\mathcal{S}_{i,t-1}) \right\}_{i \in [K]}$ at time t, the greedy policy plays the arm that maximizes expected estimated reward, i.e.,

$$\pi_t = \arg\max_{i \in [K]} \mu \left(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \right) .$$

Since μ is a strictly increasing function, this translates to $\pi_t = \arg\max_{i \in [K]} X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$.

Algorithm 2 Greedy Bandit for Generalized Linear Models

Input parameters: inverse link function μ
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)
  Play arm π_t, observe $Y_{i,t} = \mu(X_t^\top \beta_{\pi_t}) + \varepsilon_{\pi_t,t}$
  Update $\hat{\beta}(\mathcal{S}_{\pi_t,t}) \leftarrow h_\mu\left(\mathbf{X}(\mathcal{S}_{\pi_t,t}), \mathbf{Y}(\mathcal{S}_{\pi_t,t})\right)$, where $h_\mu(\mathbf{X}, \mathbf{Y})$ is the solution to the maximum likelihood estimation in Equation (6)
end for

Next, we state the following result (proved in Appendix E.1) that Algorithm 2 achieves logarithmic regret when $K = 2$ and the covariate diversity assumption holds.

PROPOSITION 1. *Consider arm rewards given by a GLM with σ-subgaussian noise $\varepsilon_{i,t} = Y_{i,t} - \mu(X_t^\top \beta_i)$. Define $m_\theta = \min\{\mu'(z) : z \in [-(b_{\max} + \theta)x_{\max}, (b_{\max} + \theta)x_{\max}]\}$. If $K = 2$ and Assumptions 1-3 are satisfied, the cumulative expected regret of Algorithm 2 at time T is at most*

$$R_T(\pi) \le \frac{128 C_0 \bar{C}_\mu L_\mu x_{\max}^4 \sigma^2 d}{\lambda_0^2} \log T + \bar{C}_\mu L_\mu \left(128 \frac{C_0 x_{\max}^4 \sigma^2 d}{\lambda_0^2} + 160 \frac{b_{\max} x_{\max}^3 d}{\lambda_0} + 2 x_{\max} b_{\max} \right) = \mathcal{O}(\log T) ,$$

where the constant C_0 is defined in Assumption 2, L_μ is the Lipschitz constant[1] *of the function $\mu(\cdot)$ on the interval $[-x_{\max} b_{\max}, x_{\max} b_{\max}]$, and $\bar{C}_\mu$ is defined as* $\bar{C}_\mu = \frac{1}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right)^3 + \frac{3}{2}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right)^2 + \frac{8}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right) + \frac{1}{m_{b_{\max}}^3}\left(\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right)\frac{m_{b_{\max}}}{2} + \frac{1}{4}\right) + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2 m_{b_{\max}}}$.

[1] Exists by continuity of $\mu' = A''$.

3.5. Performance of Greedy Bandit without Covariate Diversity

Thus far, we have shown that the greedy algorithm is rate optimal when there are only two arms and in the presence of covariate diversity in the observed context distribution. However, when these additional assumptions do not hold, the greedy algorithm may fail to converge to the true arm parameters and achieve linear regret. We now show that a greedy approach achieves rate optimal performance with *some probability* even when these assumptions do not hold. This result will motivate the design of the Greedy-First algorithm in §4.

Assumptions. For the rest of the paper, we allow the number of arms $K > 2$, and remove Assumption 3 on covariate diversity. Instead, we will make the following weaker Assumption 4, which is typically made in the contextual bandit literature (see e.g., Goldenshluger and Zeevi 2013, Bastani and Bayati 2015), which allows for multiple arms, and relaxes the assumption on observed contexts (e.g., allowing for intercept terms in the arm parameters).

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 $\Sigma_i \equiv \mathbb{E}\left[XX^\top \mathbb{I}(X \in U_i)\right]$ 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 4. This assumption is slightly different as stated than the assumptions made in prior literature; however, these assumptions are equivalent for bounded context distributions p_X (Assumption 1). We discuss the comparison in Appendix D for completeness.

Algorithm. We consider a small modification of the Greedy Bandit (Algorithm 1), by initializing each arm parameter estimate with $m > 0$ random samples. Note that OLS requires at least d samples for an arm parameter estimate to be well-defined, and Algorithm 1 does not update the arm parameter estimates from the initial ad-hoc value of 0 until this stage is reached (i.e., the covariance matrix $\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ for a given arm i becomes invertible); thus, all actions up to that point are essentially random. Consequently, we argue that initializing each arm parameter with $m = d$ samples at the beginning is qualitatively no different than Algorithm 1. We consider general values of m to study how the probabilistic guarantees of the greedy algorithm vary with the number of initial samples.

REMARK 5. We note that there is a class of explore-then-exploit bandit algorithms that follow a similar strategy of randomly sampling each arm for a length of time and using those estimates for the remaining horizon (Bubeck and Cesa-Bianchi 2012). However, (i) m is a function of the horizon length T in these algorithms (typically $m = \sqrt{T}$) while we consider m to be a (small) constant with respect to T, and (ii) these algorithms do not follow a greedy strategy since they do not update the parameter estimates after the initialization phase.

Result. The following theorem shows that the Greedy Bandit converges to the correct policy and achieves rate optimal performance with at least some problem-specific probability.

THEOREM 2. *Under Assumptions 1, 2, and 4, Greedy Bandit achieves logarithmic cumulative regret with probability at least*

$$S^{gb}(m, K, \sigma, x_{\max}, \lambda_1, h) := 1 - \inf_{\gamma \in (0,1), \delta > 0, p \geq Km+1} L(\gamma, \delta, p)\,, \tag{7}$$

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

$$\begin{aligned} L(\gamma, \delta, p) := 1 &- \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]^K + 2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\} \\ &+ \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} + \frac{d\exp\left(-D_1(\gamma)(p-m|\mathcal{K}_{sub}|)\right)}{1-\exp(-D_1(\gamma))} \\ &+ \frac{2d\exp\left(-D_2(\gamma)(p-m|\mathcal{K}_{sub}|)\right)}{1-\exp(-D_2(\gamma))}\,. \end{aligned} \tag{8}$$

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

$$D_1(\gamma) = \frac{\lambda_1(\gamma + (1-\gamma)\log(1-\gamma))}{x_{\max}^2}\,, \tag{9}$$

$$D_2(\gamma) = \frac{\lambda_1^2 h^2 (1-\gamma)^2}{8d\sigma^2 x_{\max}^4}\,. \tag{10}$$

Proof Strategy. The proof of Theorem 2 is provided in Appendix G. We observe that if all arm parameter estimates remain within a Euclidean distance of $\theta_1 = h/(2x_{\max})$ from their true values for all time periods $t > m$, then the Greedy Bandit converges to the correct policy and is rate optimal. We derive lower bounds on the probability that this event occurs using Lemma 5, after proving suitable lower bounds on the minimum eigenvalue of the covariance matrices. The key steps are as follows:

1. Assuming that the minimum eigenvalue of the sample covariance matrix for each arm is above some threshold value $\delta > 0$, we derive a lower bound on the probability that after initialization, each arm parameter estimates lie within a ball of radius $\theta_1 = h/(2x_{\max})$ centered around the true arm parameter.
2. Next, we derive a lower bound on the probability that these estimates remain within this ball after $p \geq Km + 1$ rounds for some choice of p.
3. We use the concentration result in Lemma 9 to derive a lower bound on the probability that the minimum eigenvalue of the sample covariance matrix of each arm in $\mathcal{K}_{opt}$ is above $(1-\gamma)\lambda_1(t - m|\mathcal{K}_{sub}|)$ for any $t \geq p$.

4. We derive a lower bound on the probability that the estimates ultimately remain inside the ball with radius θ_1. This ensures that no sub-optimal arm is played for any $t \geq Km$.
5. Summing up these probability terms implies Theorem 2. The parameters γ, δ, and p can be chosen arbitrarily and we optimize over their choice.

The following Proposition 2 illustrates some of the properties of the function S^{gb} in Theorem 2 with respect to problem-specific parameters. The proof is provided in Appendix G.

PROPOSITION 2. *The function $S^{gb}(m, K, \sigma, x_{\max}, \lambda_1, h)$ defined in Equation (7) is non-increasing with respect to σ and K; it is non-decreasing with respect to m, λ_1 and h. Furthermore, the limit of this function when σ goes to zero is*

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

In other words, the greedy algorithm is more likely to succeed when there is less noise and when there are fewer arms; it is also more likely to succeed with additional initialization samples, when the optimal arms each have a larger probability of being the best arm under p_X, and when the sub-optimal arms are worse than the optimal arms by a larger margin. Intuitively, these conditions make it easier for the Greedy Bandit to avoid "dropping a good arm" early on, which would result in its convergence to the wrong policy. As the noise goes to zero, the greedy algorithm always succeeds as long as the sample covariance matrix for each of the K arms is positive definite after the initialization periods.

In Corollary 1, we simplify the expression in Theorem 2 for better readability. However, the simplified expression leads to poor tail bounds when m is close to d, while the general expression in Theorem 2 works when $m = d$ as demonstrated later in §4.3 (see Figure 1).

COROLLARY 1. *Under the assumptions of Theorem 2, Greedy Bandit achieves logarithmic cumulative regret with probability at least*

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

where function $D_{\min}$ is defined as $D_{\min} = \min\left\{\frac{0.153\lambda_1}{x_{\max}^2}, \frac{\lambda_1^2 h^2}{32d\sigma^2 x_{\max}^4}\right\}$.

To summarize, these probabilistic guarantees on the success of Greedy Bandit suggest that a greedy approach can be effective and rate optimal in general with at least some probability. Therefore, in the next section, we introduce the Greedy-First algorithm which executes a greedy strategy and only resorts to forced exploration when the observed data suggests that the greedy updates are not converging. This helps eliminate unnecessary exploration with high probability.

4. Greedy-First Algorithm

As noted in Theorem 1, the optimality of the Greedy Bandit requires that there are only two arms and that the context distribution satisfies covariate diversity. The latter condition rules out some standard settings, e.g., the arm rewards cannot have an intercept term (since the addition of a one to every context vector would violate Assumption 3). While there are many examples that satisfy these conditions (see §2.2), the decision-maker may not know a priori whether a greedy algorithm is appropriate for her particular setting. Thus, we introduce the Greedy-First algorithm (Algorithm 3), which is rate optimal without these additional assumptions, but seeks to use the greedy algorithm without forced exploration when possible.

4.1. Algorithm

Algorithm 3 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 $Y_{i,t} = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$
  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 $t \in [R+1, T]$

The Greedy-First algorithm has two inputs λ_0 and t_0. It starts by following the greedy algorithm up to time t_0, after which it iteratively checks whether all the arm parameter estimates are converging to their true values at a suitable rate. A sufficient statistic for checking this is simply the minimum eigenvalue of the sample covariance matrix of each arm; if this value is above the threshold of $\lambda_0 t/4$, then greedy estimates are converging with high probability. On the other hand, if this condition is not met, the algorithm switches to a standard bandit algorithm with forced exploration. We choose the OLS Bandit algorithm (introduced by Goldenshluger and Zeevi (2013) for two arms and extended to the general setting by Bastani and Bayati (2015)), which is provided in Appendix D for completeness.

REMARK 6. 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 (2014a)) instead of the OLS Bandit. Then, the assumptions used in the theoretical analysis 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 Greedy Bandit fails.

In practice, λ_0 may be an unknown constant. Thus, we suggest the following heuristic routine to estimate this parameter:

1. Execute Greedy Bandit for t_0 time steps.
2. Estimate λ_0 using the observed data via $\hat{\lambda}_0 = \frac{1}{2t_0} \min_{i\in[K]} \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t_0})\right)$.
3. If $\hat{\lambda}_0 = 0$, this suggests that one of the arms is not receiving sufficient samples, and thus, Greedy-First will switch to OLS Bandit immediately. Otherwise, execute Greedy-First for $t \in [t_0+1, T]$ with $\lambda_0 = \hat{\lambda}_0$.

The pseudo-code for this heuristic is given in Appendix D. The regret guarantees of Greedy-First (given in the next section) are always valid, but the choice of the input parameters may affect the empirical performance of Greedy-First and the probability with which it remains exploration-free. For example, if t_0 is too small, then Greedy-First may incorrectly switch to OLS Bandit even when a greedy algorithm will converge; thus, choosing $t_0 \gg Kd$ is advisable.

4.2. Regret Analysis of Greedy-First

As noted in §3.5, we replace the more restrictive assumption on covariate diversity (Assumption 3) with a more standard assumption made in the bandit literature (Assumption 2). Theorem 3 establishes an upper bound of $\mathcal{O}(\log T)$ on the expected cumulative regret of Greedy-First. Furthermore, we establish that Greedy-First remains purely greedy with high probability when there are only two arms and covariate diversity is satisfied.

THEOREM 3. *The cumulative expected regret of Greedy-First 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 coefficient of $\log(T)$ 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.

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: (1) the algorithm pursues a purely greedy strategy, i.e., $R=0$, or (2) the algorithm switches to the OLS Bandit algorithm, i.e., $R \in [t_0+1, T]$.

Case 1: By construction, we know that $\min_{i\in[K]} \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda_0 t/4$, for all $t > t_0$. This is because Greedy-First only switches when the minimum eigenvalue of the sample covariance matrix for some arm is less than $\lambda_0 t/4$. Therefore, if the algorithm does not switch, it implies that the minimum eigenvalue of each arm's sample covariance matrix is 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 6 holds and provides 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 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 $t \in [t_0+1, R]$ before the switch. Then, using the same argument as in Case 1, 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, under Assumptions 1, 2 and 2, the OLS Bandit's cumulative regret in the interval $t \in [R+1, T]$ is upper bounded by $C_{OB}\log(T-R)$ for some constant C_{OB}. Thus, the total regret is at most $2x_{\max}b_{\max}t_0 + ((K-1)C_{GB} + C_{OB})\log T$. Note that although the switching time R is a random variable, the upper bound on the cumulative regret $2x_{\max}b_{\max}t_0 + ((K-1)C_{GB} + C_{OB})\log T$ holds uniformly regardless of the value of R.

Thus, the Greedy-First algorithm always achieves $\mathcal{O}(\log T)$ cumulative 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,

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

where $C_1 = \lambda_0/40x_{\max}^2$. 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

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

This concludes the proof. □

4.3. Probabilistic Guarantees for Greedy-First Algorithm

The key value proposition of Greedy-First is to reduce forced exploration when possible. Theorem 2 established that Greedy-First eliminates forced exploration entirely with high probability when there are only two arms and when covariate diversity holds. However, a natural question might be the extent to which Greedy-First reduces forced exploration in general problem instances.

To answer this question, we leverage the probabilistic guarantees we derived for the greedy algorithm in §3.5. Note that unlike the greedy algorithm, Greedy-First always achieves rate optimal regret. We now study the probability with which Greedy-First is purely greedy under an arbitrary number of arms K and the less restrictive Assumption 2. However, we impose that all K arms are optimal for some set of contexts under p_X, i.e., $\mathcal{K}_{opt} = [K], \mathcal{K}_{sub} = \emptyset$. This is because Greedy-First *always* switches to the OLS Bandit when an arm is sub-optimal across all contexts. In order for any algorithm to achieve logarithmic cumulative regret, sub-optimal arms must be assigned fewer samples over time and thus, the minimum eigenvalue of the sample covariance matrices of those arms cannot grow sufficiently fast; as a result, the Greedy-First algorithm will switch with probability 1. This may be practically desirable as the decision-maker can decide whether to "drop" the arm and proceed greedily or to use an exploration-based algorithm when the switch triggers.

THEOREM 4. *Let Assumptions 1, 2, and 4 hold and suppose that $\mathcal{K}_{sub} = \emptyset$. Then, with probability at least*

$$S^{gf}(m, K, \sigma, x_{\max}, \lambda_1, h) = 1 - \inf_{\gamma \leq 1-\lambda_0/(4\lambda_1), \delta > 0, Km+1 \leq p \leq t_0} L'(\gamma, \delta, p)\,, \tag{11}$$

Greedy-First remains purely greedy (does not switch to an exploration-based bandit algorithm) and achieves logarithmic cumulative regret. The function L' is closely related to the function L from Theorem 2, and is defined as

$$L'(\gamma, \delta, p) = L(\gamma, \delta, p) + (K-1)\frac{d\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))}\,. \tag{12}$$

The proof of Theorem 4 is provided in Appendix G. The steps followed are similar to that of the proof of Theorem 2. In the third step of the proof strategy of Theorem 2 (see §3.5), we used concentration results to derive a lower bound on the probability that the minimum eigenvalue of the sample covariance matrix of all arms in $\mathcal{K}_{opt}$ are above $(1-\gamma)\lambda_1 t$ for any $t \geq p$ (note that we are assuming $\mathcal{K}_{sub} = \emptyset$ in this section). For Greedy Bandit, this result was only required for the *played arm*; in contrast, for Greedy-First to remain greedy, *all arms* are required to have the minimum eigenvalues of their sample covariance matrices above $(1-\gamma)\lambda_1 t$. This causes the difference in L and L' since we need a union bound over all K arms. The additional constraints on p ensure that the Greedy-First algorithm does not switch,

The following Proposition 3 illustrates some of the properties of the function S^{gf} in Theorem 4 with respect to problem-specific parameters. The proof is provided in Appendix G.

PROPOSITION 3. *The function* $S^{gf}(m, K, \sigma, x_{\max}, \lambda_1, h)$ *defined in Equation* (11) *is non-increasing with respect to* σ *and* K*; it is non-decreasing with respect to* λ_1 *and* h*. Furthermore, the limit of this function when* σ *goes to zero is*

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

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

These relationships mirror those in Proposition 2, i.e., Greedy-First is more likely to remain exploration-free when Greedy Bandit is more likely to succeed. In particular, Greedy-First is more likely to avoid exploration entirely when there is less noise and when there are fewer arms; it is also more likely to avoid exploration with additional initialization samples and when the optimal arms each have a larger probability of being the best arm under p_X. Intuitively, these conditions make it easier for the greedy algorithm to avoid "dropping" an arm, so the minimum eigenvalue of each arm's sample covariance matrix grows at a suitable rate over time, allowing Greedy-First to remain greedy.

In Corollary 2, we simplify the expression in Theorem 4 for better readability. However, the simplified expression leads to poor tail bounds when m is close to d, while the general expression in Theorem 4 works when $m = d$ as demonstrated in Figure 1.

COROLLARY 2. *Under the assumptions made in Theorem 4, Greedy-First remains purely greedy and achieves logarithmic cumulative regret with probability at least*

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

where the function $D_{\min}$ *is defined in Corollary 1.*

We now illustrate the probabilistic bounds given in Theorems 2 and 4 through a simple 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 $B_1^2 = \{\mathbf{x} \in \mathbb{R}^2 \mid \|x\| \leq 1\}$, implying $x_{\max} = 1$. The constants h and λ_1 are chosen to satisfy Assumption 4; here, we choose $h = 0.3$, and $\lambda_1 \approx 0.025$. We then numerically plot our lower bounds on the probability of success of the Greedy Bandit (Theorem 2) and on the probability that Greedy-First remains greedy (Theorem 4) via Equations (7) and (11) respectively. Figure 1 depicts these probabilities as a function of the noise σ for several values of initialization samples m.

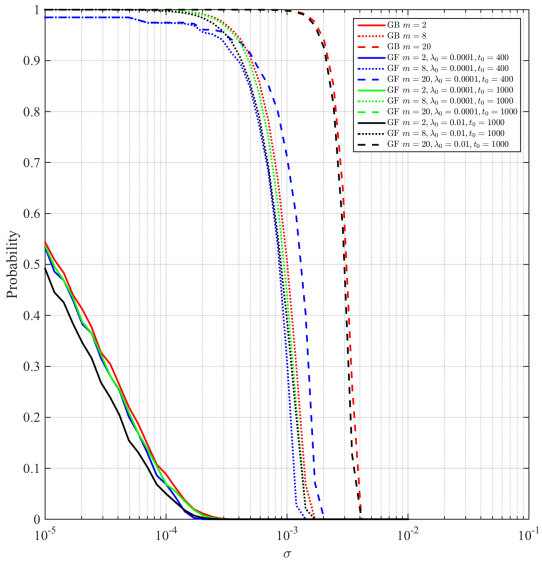


Figure 1 **Lower (theoretical) bound on the probability of success for Greedy Bandit and Greedy-First. For $m = 20, t_0 = 1000$, the performance of Greedy-First for $\lambda_0 \in \{0.01, 0.0001\}$ are similar and indistinguishable.**

We note that our lower bounds are very conservative, and in practice, both Greedy Bandit and Greedy-First succeed and remain exploration-free respectively with much larger probability. For instance, as observed in Example 1, one can optimize over the choice of 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 our lower bounds may suggest.

5. Simulations

We now validate our theoretical findings on synthetic and real datasets.

5.1. Synthetic Data

Linear Reward. We compare Greedy Bandit and Greedy-First with state-of-the-art contextual bandit algorithms. These include:

1. *OFUL* by Abbasi-Yadkori et al. (2011), which builds on the original upper confidence bound (UCB) approach of Lai and Robbins (1985),
2. *Prior-dependent TS* by Russo and Van Roy (2014b), which builds on the original Thompson sampling approach of Thompson (1933),
3. *Prior-free TS* by Agrawal and Goyal (2013), which builds on the original Thompson sampling approach of Thompson (1933), and
4. *OLS Bandit* by Goldenshluger and Zeevi (2013), which builds on ϵ-greedy methods.

REMARK 7. Prior-dependent TS requires knowledge of the prior distribution of arm parameters β_i, while prior-free TS does not. All algorithms above require knowledge of an upper bound on the noise variance σ.

Following the setup of (Russo and Van Roy 2014b), we consider Bayes regret over randomly-generated arm parameters. In particular, for each scenario, we generate 1000 problem instances and sample the true arm parameters $\{\beta_i\}_{i=1}^K$ independently. At each time step within each instance, new context vectors are drawn i.i.d. from a fixed context distribution p_X. We then plot the average Bayes regret across all these instances, along with the 95% confidence interval, as a function of time t with a horizon length $T = 10,000$. We take $K = 2$ and $d = 3$ (see Appendix F for simulations with other values of K and d). The noise variance $\sigma^2 = 0.25$.

We consider four different scenarios, varying (i) whether covariate diversity holds, and (ii) whether algorithms have knowledge of the true prior. The first condition allows us to explore how the performance of Greedy Bandit and Greedy-First compare against benchmark bandit algorithms when conditions are favorable / unfavorable for the greedy approach. The second condition helps us understand how knowledge of the prior distribution and noise variance affects the performance of benchmark algorithms relative to Greedy Bandit and Greedy-First (which do not require this knowledge). When the correct prior is provided, we assume that OFUL and both versions of TS know the noise variance.

Context vectors: For scenarios where covariate diversity holds, we sample the context vectors from a truncated Gaussian distribution, i.e., $0.5 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated to have ℓ_∞ norm at most 1. For scenarios where covariate diversity does not hold, we generate the context vectors the same way but we add an intercept term.

Arm parameters and prior: For scenarios where the algorithms have knowledge of the true prior, we sample the arm parameters $\{\beta_i\}$ independently from $\mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$, and provide all algorithms with knowledge of σ, and prior-dependent TS with the additional knowledge of the true prior distribution of arm parameters. For scenarios where the algorithms do not have knowledge of the true prior, we sample the arm parameters $\{\beta_i\}$ independently from a mixture of Gaussians, i.e., they are sampled from the distribution $0.5 \times \mathrm{N}(\mathbf{1}_d, \mathbf{I}_d)$ with probability 0.5 and from the distribution $0.5 \times \mathrm{N}(-\mathbf{1}_d, \mathbf{I}_d)$ with probability 0.5. However, prior-dependent TS is given the following incorrect prior distribution over the arm parameters: $10 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$. The OLS Bandit parameters are set to $h = 5, q = 1$, and $t_0 = 4Kd$ for Greedy-First. None of the algorithms in this scenario are given knowledge of σ; rather, this parameter is sequentially estimated over time using past data within the algorithm.

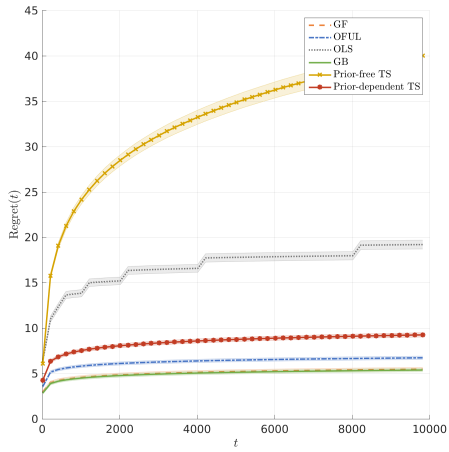
Results. Figure 2 shows the cumulative Bayes regret of all the algorithms for the four different scenarios discussed above (with and without covariate diversity, with and without the true prior). When covariate diversity holds (a-b), the Greedy Bandit is the clear frontrunner, and Greedy-First

achieves the same performance since it never switches to OLS Bandit. However, when covariate diversity does not hold (c-d), we see that the Greedy Bandit performs very poorly (achieving linear regret), but Greedy-First is the clear frontrunner. This is because the greedy algorithm succeeds a significant fraction of the time (Theorem 2), but fails on other instances. Thus, always following the greedy algorithm yields poor performance, but a standard bandit algorithm like the OLS Bandit explores unnecessarily in the instances where a greedy algorithm would have sufficed. Greedy-First leverages this observation by only exploring (switching to OLS Bandit) when the greedy algorithm has failed (with high probability), thereby outperforming both Greedy Bandit and OLS Bandit. Thus, Greedy-First provides a desirable compromise between avoiding exploration and learning the true policy.

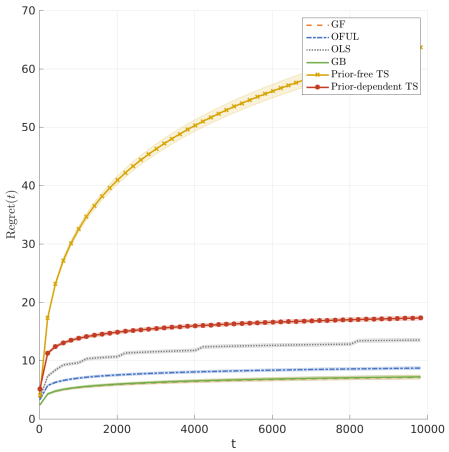
Logistic Reward. We now move beyond linear rewards and explore how the performance of Greedy Bandit (Algorithm 2) compares to other bandit algorithms for GLM rewards when covariate diversity holds. We compare to the state-of-the-art GLM-UCB algorithm (Filippi et al. 2010), which is designed to handle GLM reward functions unlike the bandit algorithms from the previous section. Our reward is logistic, i.e, $Y_{it} = 1$ with probability $1/[1+\exp(-X_t^\top \beta_i)]$ and is 0 otherwise.

We again consider Bayes regret over randomly-generated arm parameters. For each scenario, we generate 10 problem instances (due to the computational burden of solving a maximum likelihood estimation step in each iteration) and sample the true arm parameters $\{\beta_i\}_{i=1}^K$ independently. At each time step within each instance, new context vectors are drawn i.i.d. from a fixed context distribution p_X. We then plot the average Bayes regret across all these instances, along with the 95% confidence interval, as a function of time t with a horizon length $T = 2,000$. Once again, we sample the context vectors from a truncated Gaussian distribution, i.e., $0.5 \times \mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated to have ℓ_2 norm at most $x_{\max}$. Note that this context distribution satisfies covariate diversity. We take $K = 2$, and we sample the arm parameters $\{\beta_i\}$ independently from $\mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$. We consider two different scenarios for d and $x_{\max}$. In the first scenario, we take $d = 3, x_{\max} = 1$; in the second scenario, we take $d = 10, x_{\max} = 5$.

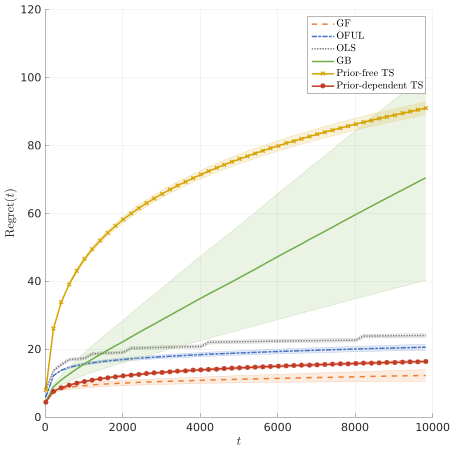
Results: Figure 3 shows the cumulative Bayes regret of the Greedy Bandit and GLM-UCB algorithms for the two different scenarios discussed above. As is evident from these results, the Greedy Bandit far outperforms GLM-UCB. We suspect that this is due to the conservative construction of confidence sets in GLM-UCB, particularly for large values of d and $x_{\max}$. In particular, the radius of the confidence set in GLM-UCB is proportional to $(\inf_{z \in C} \mu'(z))^{-1}$ where $C = \{z \mid z \in [-x_{\max} b_{\max}, x_{\max} b_{\max}]\}$. Hence, the radius of the confidence set scales as $\exp(x_{\max} b_{\max})$, which is exponentially large in $x_{\max}$. This can be seen from the difference in Figure 3 (a) and (b); in (b),$x_{\max}$ is much larger, causing GLM-UCB's performance to severely degrade. Although the



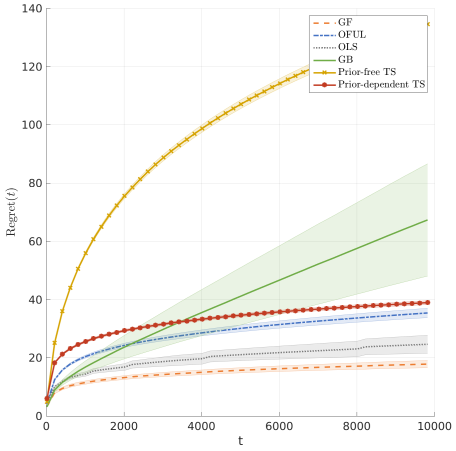
(a) Correct prior and covariate diversity.



(b) Incorrect prior and covariate diversity.



(c) Correct prior and no covariate diversity.

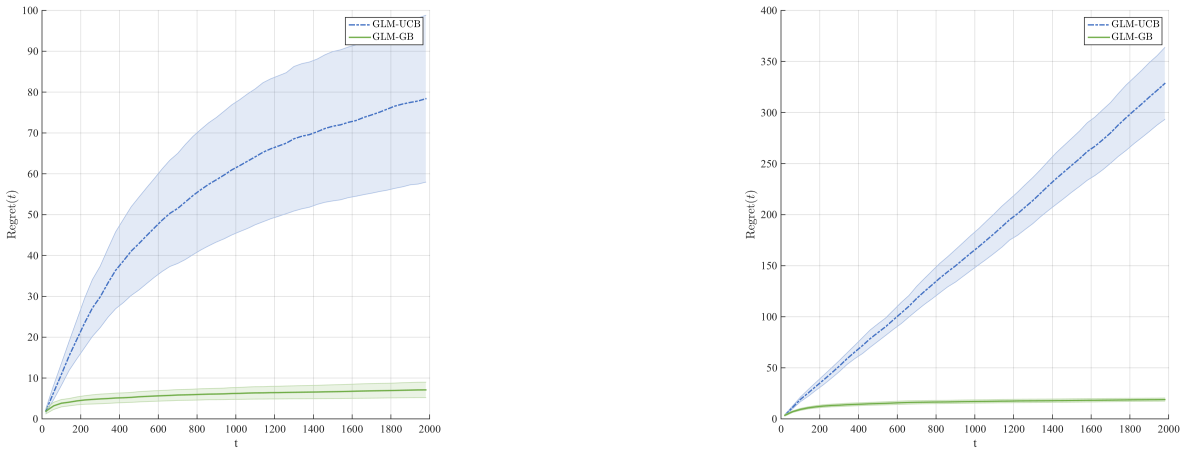


(d) Incorrect prior and no covariate diversity.

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).**

same quantity appears in the theoretical analysis of Greedy Bandit for GLM (Proposition 1), the empirical performance of Greedy Bandit appears much better.

Additional Simulations. We explore the performance of Greedy Bandit as a function of K and d; we find that the performance of Greedy Bandit improves dramatically as the dimension d increases, while it degrades with the number of arms K (as predicted by Proposition 2). We also study the dependence of the performance of Greedy-First on the input parameters t_0 (which determines when to switch) and h, q (which are inputs to OLS Bandit after switching); we find that



(a) $d = 3, x_{\max} = 1$ (b) $d = 10, x_{\max} = 5$

Figure 3 Expected regret of GLM-GB and GLM-UCB on synthetic data for logistic reward

the performance of Greedy-First is quite robust to the choice of inputs. Note that Greedy Bandit is entirely parameter-free. These simulations can be found in Appendix F.

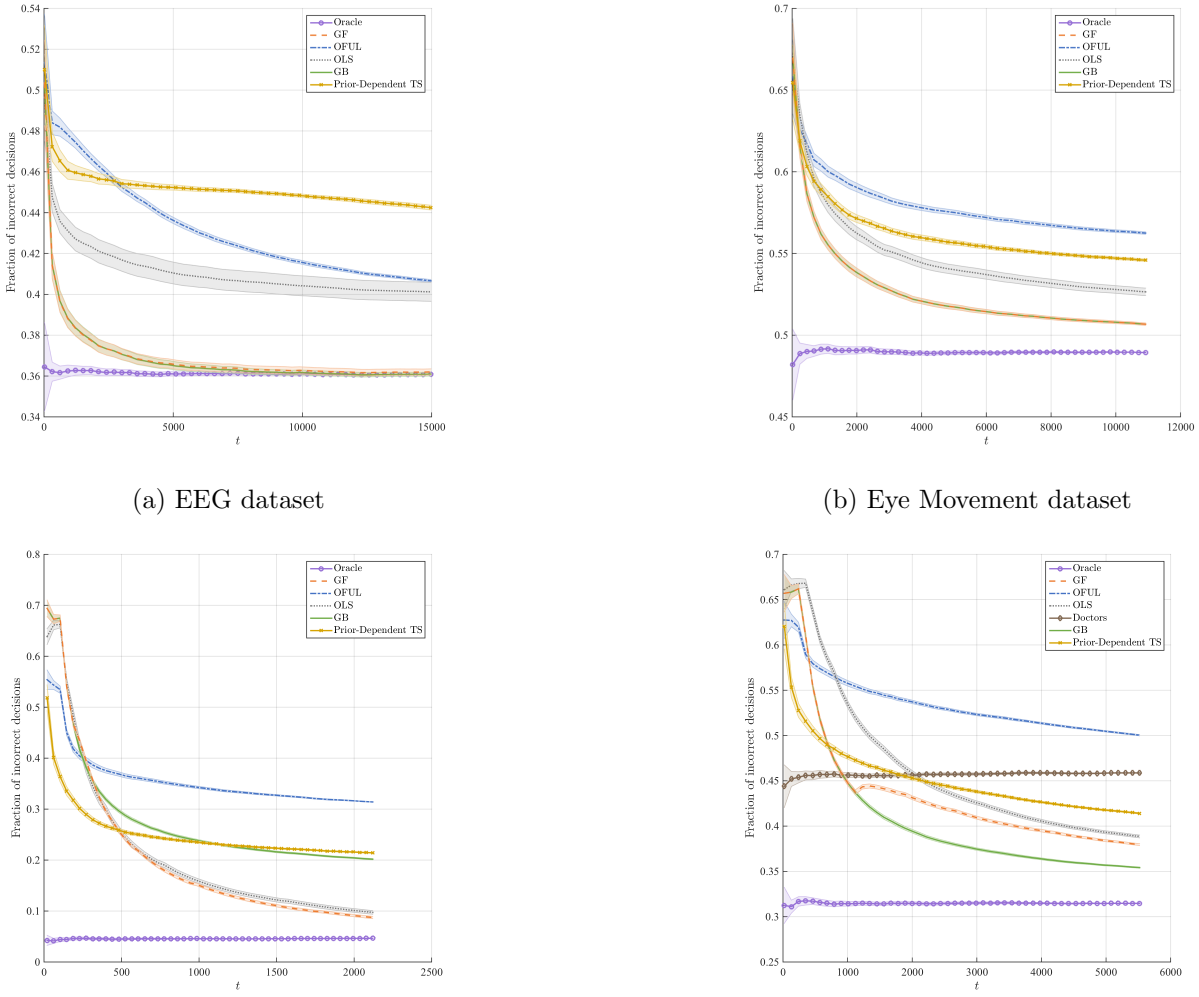
5.2. Simulations on Real Datasets

We now explore the performance of Greedy and Greedy-First with respect to competing algorithms on real datasets. As mentioned earlier, Bietti et al. (2018) performed an extensive empirical study of contextual bandit algorithms on 524 datasets that are publicly available on the OpenML platform, and found that the greedy algorithm outperforms a wide range of bandit algorithms in cumulative regret on more that 400 datasets. We take a closer look at 3 healthcare-focused datasets ((a) EEG, (b) Eye Movement, and (c) Cardiotocography) among these. We also study the (d) warfarin dosing dataset (Consortium 2009), a publicly available patient dataset that was used by Bastani and Bayati (2015) for analyzing contextual bandit algorithms.

Setup: These datasets all involve classification tasks using patient features. Accordingly, we take the number of decisions K to be the number of classes, and consider a binary reward (1 if we output the correct class, and 0 otherwise). The dimension of the features for datasets (a)-(d) is 14, 27, 35 and 93 respectively; similarly, the number of arms is 2, 3, 3, and 3 respectively.

Remark 8. Note that we are now evaluating regret rather than Bayes regret. This is because our arm parameters are given by the true data, and are not simulated from a known prior distribution.

We compare to the same algorithms as in the previous section, i.e., OFUL, prior-dependent TS, prior-free TS, and OLS Bandit. As an additional benchmark, we also include an oracle policy, which uses the best linear model trained on *all the data* in hindsight; thus, one cannot perform better than the oracle policy using linear models on these datasets.



(a) EEG dataset

(b) Eye Movement dataset

(c) Cardiotocography dataset

(d) Warfarin dataset

Figure 4 Expected regret of all algorithms on four real healthcare datasets.

Results: In Figure 4, we plot the regret (averaged over 100 trials with randomly permuted patients) as a function of the number of patients seen so far, along with the 95% confidence intervals. First, in both datasets (a) and (b), we observe that Greedy Bandit and Greedy-First perform the best; Greedy-First recognizes that the greedy algorithm is converging and does not switch to an exploration-based strategy. In dataset (c), the Greedy Bandit gets "stuck" and does not converge to the optimal policy on average. Here, Greedy-First performs the best, followed closely by the OLS Bandit. This result is similar to our results in Fig 2 (c-d), but in this case, exploration appears to be necessary in nearly all instances, explaining the extremely close performance of Greedy-First and OLS Bandit. Finally, in dataset (d), we see that the Greedy Bandit performs the best, followed by Greedy-First. An interesting feature of this dataset is that one arm (high dose) is optimal for a

very small number of patients; thus, dropping this arm entirely leads to better performance over a short horizon than attempting to learn its parameter. In this case, Greedy Bandit is not converging to the optimal policy since it never assigns any patient the high dose. However, Greedy-First recognizes that the high-dose arm is not getting sufficient samples and switches to an exploration-based algorithm. As a result, Greedy-First performs worse than the Greedy Bandit. However, if the horizon were to be extended[2], Greedy-First and the other bandit algorithms would eventually overtake the Greedy Bandit. Alternatively, for non-binary reward functions (e.g., when cost of a mistake for high-dose patients is larger than for other patients) Greedy Bandit would perform poorly.

Looking at these results as a whole, we see that Greedy-First is a robust frontrunner. When exploration is unnecessary, it matches the performance of the Greedy Bandit; when exploration is necessary, it matches or outperforms competing bandit algorithms.

6. Conclusions and Discussions

We prove that a greedy algorithm can be rate optimal in cumulative regret for a two-armed contextual bandit as long as the contexts satisfy *covariate diversity*. Greedy algorithms are significantly preferable when exploration is costly (e.g., result in lost customers for online advertising or A/B testing) or unethical (e.g., personalized medicine or clinical trials). Furthermore, the greedy 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, we provide empirical evidence that the greedy algorithm can outperform standard contextual bandit algorithms when the contexts satisfy covariate diversity. Even when the contexts do not satisfy covariate diversity, we prove that a greedy algorithm is rate optimal with *some probability*, and provide lower bounds on this probability.

However, in many scenarios, the decision-makers may not know whether their problem instance is amenable to a greedy approach, and may still wish to ensure that their algorithm provably converges to the correct policy. In this case, the decision-maker may under-explore by using a greedy algorithm, while a standard bandit algorithm may over-explore (since the greedy algorithm converges to the correct policy with some probability in general). Consequently, we propose the Greedy-First algorithm, which follows a greedy policy in the beginning and only performs exploration when the observed data indicate that exploration is necessary. Greedy-First is rate optimal without the covariate diversity assumption. More importantly, it remains exploration-free when covariate diversity is satisfied, and may provably reduce exploration even when covariate diversity is not satisfied. Our empirical results suggest that Greedy-First outperforms standard bandit algorithms (e.g., UCB, Thompson Sampling, and ϵ-greedy methods) by striking a balance between avoiding exploration and converging to the correct policy.

[2] Our horizon is limited by the number of patients available in the dataset.

References

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

Agrawal, Shipra, Vashist Avadhanula, Vineet Goyal, Assaf Zeevi. 2017. Mnl-bandit: A dynamic learning approach to assortment selection. *arXiv preprint arXiv:1706.03880* .

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

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

Ban, Gah-Yi, N Bora Keskin. 2018. Personalized dynamic pricing with machine learning. *Available at SSRN 2972985* .

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

Bastani, Hamsa, Pavithra Harsha, Georgia Perakis, Divya Singhvi. 2018. Sequential learning of product recommendations with customer disengagement. *Available at SSRN 3240970* .

Bastani, Hamsa, David Simchi-Levi, Ruihao Zhu. 2019. Meta dynamic pricing: Learning across experiments. *Available at SSRN 3334629* .

Bietti, A., A. Agarwal, J. Langford. 2018. A Contextual Bandit Bake-off. *ArXiv e-prints* .

Bird, Sarah, Solon Barocas, Kate Crawford, Fernando Diaz, Hanna Wallach. 2016. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2846909`.

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

Bubeck, Sébastien, Nicolò Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* **5**(1) 1–122.

Chen, Kani, Inchi Hu, Zhiliang Ying. 1999. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics* **27**(4) 1155–1163.

Chick, Stephen E, Noah Gans, Ozge Yapar. 2018. Bayesian sequential learning for clinical trials of multiple correlated medical interventions .

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

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

Consortium, International Warfarin Pharmacogenetics. 2009. Estimation of the warfarin dose with clinical and pharmacogenetic data. *NEJM* **360**(8) 753.

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

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

Filippi, Sarah, Olivier Cappe, Aurélien Garivier, Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*. 586–594.

Gittins, John C. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(2) 148–164.

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

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

Gutin, Eli, Vivek Farias. 2016. Optimistic gittins indices. *Advances in Neural Information Processing Systems*. 3153–3161.

Javanmard, Adel, Hamid Nazerzadeh. 2019. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research* **20**(1) 315–363.

Kallus, Nathan, Madeleine Udell. 2016. Dynamic assortment personalization in high dimensions. *arXiv preprint arXiv:1610.05604* .

Kallus, Nathan, Angela Zhou. 2018. Policy evaluation and optimization with continuous treatments. *arXiv preprint arXiv:1802.06037* .

Kannan, S., J. Morgenstern, A. Roth, B. Waggoner, Z. S. Wu. 2018. A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem. *ArXiv e-prints* .

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

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

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

Kim, Edward S, 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. 2011. The battle trial: personalizing therapy for lung cancer. *Cancer discovery* **1**(1) 44–53.

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

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

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

Lehmann, E.L., G. Casella. 1998. *Theory of Point Estimation*. Springer Verlag.

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

Li, Lihong, Yu Lu, Dengyong Zhou. 2017. Provably optimal algorithms for generalized linear contextual bandits. *arXiv preprint arXiv:1703.00048* .

McCullagh, P., J. A. Nelder. 1989. *Generalized linear models (Second edition)*. London: Chapman & Hall.

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

Mintz, Yonatan, Anil Aswani, Philip Kaminsky, Elena Flowers, Yoshimi Fukuoka. 2017. Non-stationary bandits with habituation and recovery dynamics. *arXiv preprint arXiv:1707.08423* .

Narendra, Kumpati S, Anuradha M Annaswamy. 1987. Persistent excitation in adaptive systems. *International Journal of Control* **45**(1) 127–160.

Nguyen, Nhan T. 2018. *Model-reference adaptive control*. Springer.

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

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

Russo, Daniel. 2019. A note on the equivalence of upper confidence bounds and gittins indices for patient agents. *arXiv preprint arXiv:1904.04732* .

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

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

Tewari, Ambuj, Susan A Murphy. 2017. From ads to interventions: Contextual bandits in mobile health. *Mobile Health*. Springer, 495–517.

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

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

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

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

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

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

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

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

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

Zhou, Zhijin, Yingfei Wang, Hamed Mamani, David G Coffey. 2019. How do tumor cytogenetics inform cancer treatments? dynamic risk stratification and precision medicine using multi-armed bandits. *Dynamic Risk Stratification and Precision Medicine Using Multi-armed Bandits (June 17, 2019)* .

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 $\mathbf{x} \in W$, $a \cdot p_X(-\mathbf{x}) \le b \cdot p_X(\mathbf{x})$.

(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 $\mathbf{x}^\top \mathbf{u} \ge 0$ or $-\mathbf{x}^\top \mathbf{u} \ge 0$ holds, and using conditions (a), (b), and (c) of Lemma 1 we have:

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

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 $a \le b$ which follows from condition (b), and the last inequality uses condition (c). □

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 $i \ne j$ 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,

$$\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left(\int_{B_R^d} x_1^2 \mathrm{d}\mathbf{x}\right) I_d\,. \tag{13}$$

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 (13) 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 7. *The following inequality holds*

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

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

Proof of Lemma 7. We can lower-bound the density $p_{X,\mathrm{trunc}}$ 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

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

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

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

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

Appendix B: Useful Concentration Results

LEMMA 8 **(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*

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

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

LEMMA 9 **(Theorem 3.1 of Tropp (2011))**. *Let $\mathcal{H}_1 \subset \mathcal{H}_2 \cdots$ be a filtration and consider a finite sequence $\{X_k\}$ of positive semi-definite matrices with dimension d adapted to this filtration. Suppose that $\lambda_{\max}(X_k) \leq R$ almost surely. Define the series $Y \equiv \sum_k X_k$ and $W \equiv \sum_k \mathbb{E}[X_k \mid \mathcal{H}_{k-1}]$. Then for all $\mu \geq 0, \gamma \in [0,1)$ we have:*

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

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

$$\mathcal{F}_{i,t}^{\lambda} = \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\right\} \tag{14}$$

$$\mathcal{G}_{i,t}^{\chi} = \left\{\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 < \chi\right\}. \tag{15}$$

LEMMA 6 *The instantaneous expected regret of the Greedy Bandit at time $t \geq 2$ satisfies*

$$r_t(\pi) \leq \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t-1} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right),$$

where $C_3 = \lambda_0^2/(32d\sigma^2x_{\max}^2)$, 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

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

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

$$\mathbb{E}[\text{Regret}_t(\pi) \,|X_t \in \mathcal{R}_l] \leq \mathbb{E}\left[\mathbb{I}(X_t^\top\hat{\beta}(\mathcal{S}_{\pi_t,t-1}) \geq X_t^\top\hat{\beta}(\mathcal{S}_{l,t-1}))X_t^\top(\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l\right]$$

$$
\begin{aligned}
&\leq \sum_{i\neq l} \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1}))X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&= \sum_{i\neq l} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\leq \sum_{i\neq l} \Bigg\{ \mathbb{E}\left[\mathbb{I}(\hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\qquad + \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \overline{\mathcal{F}_{l,t-1}^{\lambda_0/4}})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\qquad + \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \Bigg\} \\
&\leq \sum_{i\neq l} \Bigg\{ \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\qquad + 2b_{\max}x_{\max}\left(\mathbb{P}(\overline{\mathcal{F}_{l,t-1}^{\lambda_0/4}}) + \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}})\right) \Bigg\} \\
&\leq \sum_{i\neq l} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\qquad + 4(K-1)b_{\max}x_{\max}\max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \qquad (16)
\end{aligned}
$$

where in the second line we used a union bound, in the sixth line we used the fact that $\mathcal{F}_{i,t-1}^{\lambda_0/4}$ and $\mathcal{F}_{l,t-1}^{\lambda_0/4}$ 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 $\mathcal{F}_{i,t-1}^{\lambda_0/4}$ and $\mathcal{F}_{l,t-1}^{\lambda_0/4}$, 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 $\delta = 1/\sqrt{(t-1)}$. 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 (16) for i, by conditioning on $X_t \in I^h$ as following

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&= \sum_{h=0}^{h^{\max}} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4})X_t^\top(\beta_l-\beta_i) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t \in I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1)\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t \in I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1)\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t^\top(\beta_l-\beta_i) \in (0, 2\delta x_{\max}(h+1)]] \\
&\leq \sum_{h=0}^{h^{\max}} 4C_0\delta^2 x_{\max}^2(h+1)^2\mathbb{P}\left[X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I_h\right], \qquad (17)
\end{aligned}
$$

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 $\mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}$, 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,

$$\begin{aligned}
&\mathbb{P}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad\leq \mathbb{P}\left[X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad+ \mathbb{P}\left[X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad\leq \mathbb{P}\left[X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\qquad\quad+ \mathbb{P}\left[X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad\leq \mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \geq \delta h, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \geq \delta h, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right], \quad (18)
\end{aligned}$$

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 (15) this can be rewritten as

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{G}_{l,t-1}^{\delta h}}, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] + \mathbb{P}\left[\overline{\mathcal{G}_{i,t-1}^{\delta h}}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&= \mathbb{P}\left[\overline{\mathcal{G}_{l,t-1}^{\delta h}}, \mathcal{F}_{l,t-1}^{\lambda_0/4}\right] + \mathbb{P}\left[\overline{\mathcal{G}_{i,t-1}^{\delta h}}, \mathcal{F}_{i,t-1}^{\lambda_0/4}\right] \\
&\leq 4d\exp\left(-C_3(t-1)(\delta h)^2\right) \\
&= 4d\exp(-h^2),
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}[\text{Regret}_t(\pi)] = \sum_{l=1}^K \mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] \cdot \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\leq \sum_{l=1}^K \left(\sum_{i\neq l} \sum_{h=0}^{h^{\max}} \left[4C_0\delta^2 x_{\max}^2 (h+1)^2 \min\{1, 4d\exp(-h^2)\}\right] + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \right) \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\leq 4(K-1) C_0\delta^2 x_{\max}^2 \left(\sum_{h=0}^{h^{\max}} (h+1)^2 \min\{1, 4d\exp(-h^2)\} \right) + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \\
&\leq 4(K-1) \left(C_0\delta^2 x_{\max}^2 \left(\sum_{h=0}^{h_0} (h+1)^2 + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^2 \exp(-h^2) \right) + b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \right), \quad (19)
\end{aligned}$$

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

$$\sum_{h=h_0+1}^{h^{\max}} (h+1)^2 \exp(-h^2) = \sum_{h=h_0+1}^{h^{\max}} (h^2 + 2h + 1) \exp(-h^2)$$

$$= \sum_{h=h_0+1}^{h^{\max}} h^2 \exp(-h^2) + 2 \sum_{h=h_0+1}^{h^{\max}} h \exp(-h^2) + \sum_{h=h_0+1}^{h^{\max}} \exp(-h^2)$$
$$\leq \int_{h_0}^{\infty} h^2 \exp(-h^2) \mathrm{d}h + \int_{h_0}^{\infty} 2h \exp(-h^2) \mathrm{d}h + \int_{h_0}^{\infty} \exp(-h^2) \mathrm{d}h. \tag{20}$$

Computing the above terms using integration by parts and using the inequality $\int_t^{\infty} \exp(-x^2)\mathrm{d}x \leq \exp(-t^2)/(t+\sqrt{t^2+4/\pi})$ yields

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

where $\bar{C}$ is defined as (4). By replacing this in (19) and substituting $\delta = 1/\sqrt{(t-1)C_3}$ we get

$$r_t(\pi) = \mathbb{E}[\mathrm{Regret}_t(\pi)] \leq \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t-1} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right)$$

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 6 we can write

$$\begin{aligned}
R_T(\pi) &= \sum_{t=1}^{T} r_t(\pi) \\
&\leq 2x_{\max}b_{\max} + \sum_{t=2}^{T}\left[\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t-1} + 4b_{\max}x_{\max}d\exp(-C_1(t-1))\right] \\
&= 2x_{\max}b_{\max} + \sum_{t=1}^{T-1}\left[\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t} + 4b_{\max}x_{\max}d\exp(-C_1t)\right] \\
&\leq 2x_{\max}b_{\max} + \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}(1 + \int_1^T \frac{1}{t}\mathrm{d}t) + 4b_{\max}x_{\max}d\int_1^{\infty}\exp(-C_1t)\mathrm{d}t \\
&= 2x_{\max}b_{\max} + \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}(1+\log T) + \frac{4b_{\max}x_{\max}d}{C_1} \\
&= \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2}\log T + \left(2x_{\max}b_{\max} + \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2} + \frac{160b_{\max}x_{\max}^3 d}{\lambda_0}\right) \\
&= \mathcal{O}(\log T),
\end{aligned}$$

finishing up the proof. □

Appendix D: Additional Details on Greedy-First

We first present the pseudo-code for OLS-Bandit and the heuristic for Greedy-First. 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 OLS Bandit 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):

$$\mathcal{T}_i \equiv \left\{ (2^n - 1) \cdot Kq + j \;\middle|\; n \in \{0, 1, 2, ...\} \text{ and } j \in \{q(i-1)+1, q(i-1)+2, ..., iq\} \right\}. \tag{21}$$

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* $\mathcal{S}_{i,t} = \left\{ t' \;\middle|\; \pi_{t'} = i \text{ and } 1 \leq t' \leq t \right\}$ 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 4 explains the pseudo-code for OLS Bandit.

Algorithm 4 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. (21) 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
    $\hat{\mathcal{K}} = \left\{ i \in K \;\middle|\; X_t^T \hat{\beta}(\mathcal{T}_{i,t-1}) \geq \max_{j \in K} X_t^T \hat{\beta}(\mathcal{T}_{j,t-1}) - h/2 \right\}$
    $\pi_t \leftarrow \arg\max_{i \in \hat{\mathcal{K}}} X_t^T \hat{\beta}(\mathcal{S}_{i,t-1})$
  end if
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
  Play arm π_t, observe $Y_{i,t} = X_t^T \beta_{\pi_t} + \varepsilon_{i,t}$
end for

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

Algorithm 5 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. Proof of Proposition 1

Uniqueness of solution of Equation (6)**.** We first prove that the solution to maximum likelihood equation in Equation (6) is unique whenever the design matrix $\mathbf{X}^\top\mathbf{X}$ is positive definite. The first order optimality condition in Equation (6) implies that

$$\sum_{\ell=1}^{n} X_\ell \left(Y_\ell - A'(X_\ell^\top \hat{\beta})\right) = \sum_{\ell=1}^{n} X_\ell \left(Y_\ell - \mu(X_\ell^\top \hat{\beta})\right) = 0\,. \tag{22}$$

Now suppose that there are two solutions to the above equation, namely $\hat{\beta}_1$ and $\hat{\beta}_2$. Then, we can write

$$\sum_{\ell=1}^{n} X_\ell \left(\mu(X_\ell^\top \hat{\beta}_1) - \mu(X_\ell^\top \hat{\beta}_2)\right) = 0.$$

Using the mean-value theorem, for each $1 \le i \le n$ we have

$$\mu(X_\ell^\top \hat{\beta}_2) - \mu(X_\ell^\top \hat{\beta}_1) = \mu'(X_\ell^\top \tilde{\beta}_\ell)\left(X_\ell^\top(\hat{\beta}_2 - \hat{\beta}_1)\right),$$

where $\tilde{\beta}_\ell$ belongs to the line connecting $\hat{\beta}_1, \hat{\beta}_2$. Replacing this in above equation implies that

$$\sum_{\ell=1}^{n} X_\ell \left(\mu'(X_\ell^\top \tilde{\beta}_\ell)\left(X_\ell^\top(\hat{\beta}_2 - \hat{\beta}_1)\right)\right) = \left(\sum_{\ell=1}^{n} \mu'(X_\ell^\top \tilde{\beta}_\ell) X_\ell X_\ell^\top\right)(\hat{\beta}_2 - \hat{\beta}_1) = 0. \tag{23}$$

Note that μ is strictly increasing meaning that μ' is always positive. Therefore, letting $m = \min_{1\le l\le n}\left\{\mu'(X_\ell^\top \tilde{\beta}_\ell)\right\}$, we have that

$$\sum_{\ell=1}^{n} \mu'(X_\ell^\top \tilde{\beta}_\ell) X_\ell X_\ell^\top \succeq m\mathbf{X}\mathbf{X}^\top.$$

Therefore, if the design matrix $\mathbf{X}\mathbf{X}^\top$ is positive definite, then so is $\sum_{\ell=1}^{n} \mu'(X_\ell^\top \tilde{\beta}_\ell) X_\ell X_\ell^\top$. Hence, Equation (23) implies that $\hat{\beta}_1 = \hat{\beta}_2$.

Proof of Proposition 1. For proving this, we first state and prove a Lemma that will be used later to prove this result.

LEMMA 10. *Consider the generalized linear model with the inverse link function μ. Suppose that we have samples $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, where $Y_i=\mu(X_i^\top\beta_0)+\varepsilon_i$, where $\|X_i\|_2\le x_{\max}$ and $\|\beta_0\|_2\le b_{\max}$. Furthermore, assume that the design matrix $\mathbf{X}^\top\mathbf{X}=\sum_{i=1}^n X_iX_i^\top$ is positive definite. Let $\hat\beta=h_\mu(\mathbf{X},\mathbf{Y})$ be the (unique) solution to the Equation (22) and let θ be an arbitrary positive number. Recall that $m_\theta:=\min\{\mu'(z):z\in[-(\theta+b_{\max})x_{\max},(\theta+b_{\max})x_{\max}]\}$ and suppose $\|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon\|_2\le\theta m_\theta$, then*

$$\|\hat\beta-\beta_0\|_2\le\frac{\|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon\|_2}{m_\theta}.$$

Proving the above Lemma is adapted from Chen et al. (1999). For completeness, we provide a proof here as well. We need the following Lemma which was proved in Chen et al. (1999).

LEMMA 11. *Let H be a smooth injection from $\mathbb{R}^d$ to $\mathbb{R}^d$ with $H(\mathbf{x}_0)=\mathbf{y}_0$. Define $B_\delta(\mathbf{x}_0)=\{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}-\mathbf{x}_0\|\le\delta\}$ and $S_\delta(\mathbf{x}_0)=\partial B_\delta(\mathbf{x}_0)=\{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}-\mathbf{x}_0\|=\delta\}$. Then, $\inf_{\mathbf{x}\in S_\delta(\mathbf{x}_0)}\|H(\mathbf{x})-\mathbf{y}_0\|\ge r$ implies that*

(i) $B_r(\mathbf{y}_0)=\{\mathbf{y}\in\mathbb{R}^d:\|\mathbf{y}-\mathbf{y}_0\|\le r\}\subset H(B_\delta(\mathbf{x}_0))$,

(ii) $H^{-1}(B_r(\mathbf{y}_0))\subset B_\delta(\mathbf{x}_0)$

Proof of Lemma 10. Note that $\hat\beta$ is the solution to the Equation (22) and therefore

$$\sum_{i=1}^n\left(\mu(X_i^\top\hat\beta)-\mu(X_i^\top\beta_0)\right)X_i=\sum_{i=1}^n X_i\varepsilon_i. \tag{24}$$

Using the mean-value theorem for any $\beta\in\mathbb{R}^d$ and $1\le i\le n$ we have

$$\mu(X_i^\top\beta)-\mu(X_i^\top\beta_0)=\mu'(X_i^\top\beta_i')\left(X_i^\top(\beta-\beta_0)\right),$$

where β_i' is a point that lies on the line segment between β and β_0. Define

$$\begin{aligned}G(\beta)&=\left(\sum_{i=1}^n X_iX_i^\top\right)^{-1}\left(\sum_{i=1}^n\left(\mu(X_i^\top\beta)-\mu(X_i^\top\beta_0)\right)X_i\right)\\&=\left(\sum_{i=1}^n X_iX_i^\top\right)^{-1}\left(\sum_{i=1}^n\mu'(X_i^\top\beta_i')\left(X_i^\top(\beta-\beta_0)\right)X_i\right)\\&=\left(\sum_{i=1}^n X_iX_i^\top\right)^{-1}\left(\sum_{i=1}^n\mu'(X_i^\top\beta_i')X_iX_i^\top\right)(\beta-\beta_0)\end{aligned}$$

As $\mu'(\cdot)>0$, $G(\beta)$ is an injection from $\mathbb{R}^d$ to $\mathbb{R}^d$ satisfying $G(\beta_0)=0$. Consider the sets $B_\theta(\beta_0)=\{\beta\in\mathbb{R}^d:\|\beta-\beta_0\|_2\le\theta\}$ and $S_\theta(\beta_0)=\{\beta\in\mathbb{R}^d:\|\beta-\beta_0\|=\theta\}$. If $\beta\in B_\theta(\beta_0)$, for each i,β_i' lies on the line segment between β and β_0 and therefore we have $|X_i^\top\beta_i'|\le\max(X_i^\top\beta_0,X_i^\top\beta)\le x_{\max}(b_{\max}+\theta)$ according to the Cauchy-Schwarz inequality. Then for each $\beta\in B_\theta(\beta_0)$

$$\begin{aligned}\|G(\beta)\|_2^2&=\|G(\beta)-G(\beta_0)\|_2^2\\&=(\beta-\beta_0)^\top\left(\sum_{i=1}^n\mu'(X_i^\top\beta_i')X_iX_i^\top\right)\left(\sum_{i=1}^n X_iX_i^\top\right)^{-2}\left(\sum_{i=1}^n\mu'(X_i^\top\beta_i')X_iX_i^\top\right)(\beta-\beta_0)\\&=m_\theta^2(\beta-\beta_0)^\top\left(\sum_{i=1}^n\frac{\mu'(X_i^\top\beta_i')}{m_\theta}X_iX_i^\top\right)\left(\sum_{i=1}^n X_iX_i^\top\right)^{-2}\left(\sum_{i=1}^n\frac{\mu'(X_i^\top\beta_i')}{m_\theta}X_iX_i^\top\right)(\beta-\beta_0)\\&\ge m_\theta^2(\beta-\beta_0)^\top\left(\sum_{i=1}^n X_iX_i^\top\right)\left(\sum_{i=1}^n X_iX_i^\top\right)^{-2}\left(\sum_{i=1}^n X_iX_i^\top\right)(\beta-\beta_0)\\&=m_\theta^2\|(\beta-\beta_0)\|_2^2,\end{aligned} \tag{25}$$

or in other words $\|G(\beta)\|_2 \geq \|\beta - \beta_0\|_2 m_\theta$. In particular, for any $\beta \in S_\theta(\beta_0)$ we have $G(\beta) \geq \theta m_\theta$. Therefore, letting $\gamma = \theta m_\theta$, Lemma 11 implies that $G^{-1}\left(B_\gamma(0)\right) \subset B_\theta(\beta_0)$. Note that if we let $\mathbf{z} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \varepsilon$, then by the assumption of lemma $\mathbf{z} \in B_\gamma(0)$ and hence there exists $\tilde{\beta}, \|\tilde{\beta} - \beta_0\| \leq \theta$ satisfying $G^{-1}(\mathbf{z}) = \tilde{\beta}$, i.e., $G(\tilde{\beta}) = \mathbf{z}$. Now we claim that $\tilde{\beta} = \hat{\beta}$. The is not very difficult to prove. In particular, according to Equation (24) we know that

$$\sum_{i=1}^{n} \left(\mu(X_i^\top \hat{\beta}) - \mu(X_i^\top \beta_0)\right) X_i = \sum_{i=1}^{n} X_i \varepsilon_i \Longrightarrow G(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^{n} X_i \varepsilon_i\right) = \mathbf{z}.$$

Since the function $G(\cdot)$ is injective, it implies that $\hat{\beta} = \tilde{\beta}$. As a result, $\hat{\beta} \in B_\theta(\beta_0)$ and $G(\hat{\beta}) = \mathbf{z}$. The desired inequality follows according to Equation (25). $\square$

Having this we can prove a Corollary of Lemma 5 for the generalized linear models.

COROLLARY 3. *Consider the generalized linear model with the link function μ. Consider the contextual multi-armed bandit problem, in which upon playing arm i for the context X_t, we observe a reward equal to Y_t satisfying $\mathbb{E}[Y_t] = \mu(X_t^\top \beta_i)$. Furthermore, suppose that the noise terms $\varepsilon_{it} = Y_t - \mu(X_t^\top \beta_i)$ are σ-subgaussian for some $\sigma > 0$. Let $\hat{\beta}(\mathcal{S}_{i,t}) = h_\mu\left(\mathbf{X}(\mathcal{S}_{i,t}), \mathbf{Y}(\mathcal{S}_{i,t})\right)$ be the estimated parameter of arm i. Taking $C_2 = \lambda^2/(2d\sigma^2 x_{\max}^2)$ and $n \geq |S_{i,t}|$, we have for all $\lambda, \chi > 0$,*

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

Proof. Note that if the design matrix $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ is positive definite, then the event $\left\{\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi\right\}$ is the subset of the event

$$\left\{\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1} \mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\| \geq \chi m_\chi\right\}.$$

The reason is very simple. Suppose the contrary, i.e., the possibility of having $\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi$ while $\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1} \mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 < \chi m_\chi$. By using the Lemma 11 for $\theta = \chi$ we achieve that

$$\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \leq \frac{\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1} \mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2}{m_\chi} < \frac{\chi m_\chi}{m_\chi} = \chi,$$

which is a contradiction. Therefore,

$$\begin{aligned}\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \text{ and } \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] &\leq \mathbb{P}\left[\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1} \mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi m_\chi \text{ and } \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\ &\leq 2d \exp\left(-C_2 t^2 (\chi m_\chi)^2 / n\right),\end{aligned}$$

where the last inequality follows from the Lemma 5. $\square$

Now we are ready to prove a Lemma following the same lines of idea as Lemma 6. This lemma can help us to prove the result for the generalized linear models.

LEMMA 12. *Recall that $\mathcal{F}_{i,t}^\lambda = \{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\}$. Suppose that Assumptions 1 and 2 hold. Then, the instantaneous expected regret of the Greedy Bandit for GLMs (Algorithm 2) at time $t \geq 2$ satisfies*

$$r_t(\pi) \leq \frac{4(K-1) L_\mu C_0 \bar{C}_\mu x_{\max}^2}{C_3} \frac{1}{t-1} + 4(K-1) b_{\max} x_{\max} \left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right),$$

where $C_3 = \lambda_0^2/(32d\sigma^2 x_{\max}^2)$, C_0 is defined in Assumption 2, L_μ is the Lipschitz constant of the function $\mu(\cdot)$ on the interval $[-x_{\max} b_{\max}, x_{\max} b_{\max}]$, and $\bar{C}_\mu$ is defined in Proposition 1.

Proof. The proof is very similar to the proof of Lemma 6. 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

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

as μ is L_μ Lipschitz over the interval $[-x_{\max}b_{\max}, x_{\max}b_{\max}]$ and $X_t^\top \beta_j \in [-x_{\max}b_{\max}, x_{\max}b_{\max}]$ for all $j \in [K]$. Now one can follow all the arguments in Lemma 6 up to the point that we use concentration results for $\beta_j - \hat{\beta}_j$. In particular, Equation (18) reads as

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

Using the concentration result on Corollary 3, and noting that X_t is independent of $\hat{\beta}(\mathcal{S}_{j,t-1})$ for all j, the right hand side of above equation turns into

$$\begin{aligned}&\mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \ge \delta h, \mathcal{F}_{l,t-1}^{\lambda_0/4}\right] + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \ge \delta h, \mathcal{F}_{i,t-1}^{\lambda_0/4}\right] \\ &\qquad\le 4d\exp\left(-C_3(t-1)(\delta h)^2 m_{\delta h}^2\right) \\ &\qquad= 4d\exp(-h^2 m_{\delta h}^2).\end{aligned}$$

Now note that δh is at most equal to $b_{\max}$ (since $\mathbf{x}^\top(\beta_i - \beta_l)$ is upper bounded by $2x_{\max}b_{\max}$). As $m_\theta := \min\{\mu'(z) : z \in [-(b_{\max}+\theta)x_{\max}, (b_{\max}+\theta)x_{\max}]\}$, therefore if $\theta_2 > \theta_1$, then $m_{\theta_2} \le m_{\theta_1}$. Hence, for all values of $0 \le h \le h_{\max}$.

$$4d\exp(-h^2 m_{\delta h}^2) \le 4d\exp(-h^2 m_{b_{\max}}^2).$$

We can simply use 1 whenever this number is larger than one as this describes a probability term. Therefore,

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

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

$$\sum_{h=h_0+1}^{h^{\max}} (h+1)^2\exp(-h^2 m_{b_{\max}}^2) \le \int_{h_0}^\infty h^2\exp(-h^2 m_{b_{\max}}^2)\mathrm{d}h + \int_{h_0}^\infty 2h\exp(-h^2 m_{b_{\max}}^2)\mathrm{d}h + \int_{h_0}^\infty \exp(-h^2 m_{b_{\max}}^2)\mathrm{d}h.$$

Using the change of variable $h' = m_{b_{\max}} h$, integration by parts, and the inequality $\int_t^\infty \exp(-x^2)\mathrm{d}x \leq \exp(-t^2)/(t+\sqrt{t^2+4/\pi})$, we obtain that

$$\begin{aligned}
&\sum_{h=0}^{h_0}(h+1)^2 + 4d\sum_{h=h_0+1}^{h^{\max}}(h+1)^2\exp(-h^2)\\
&\quad= \frac{(h_0+1)(h_0+2)(2h_0+3)}{6} + 4d\left(\frac{h_0\frac{m_{b_{\max}}}{2}+\frac{1}{4}}{m_{b_{\max}}^3} + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}}\right)\exp(-h_0^2 m_{b_{\max}}^2)\\
&\quad\leq \frac{1}{3}h_0^3 + \frac{3}{2}h_0^2 + \frac{13}{6}h_0 + 1 + 4d\left(\frac{h_0\frac{m_{b_{\max}}}{2}+\frac{1}{4}}{m_{b_{\max}}^3} + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}}\right)\frac{1}{4d}\\
&\quad\leq \frac{1}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)^3 + \frac{3}{2}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)^2 + \frac{8}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)\\
&\quad+ \frac{1}{m_{b_{\max}}^3}\left(\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)\frac{m_{b_{\max}}}{2}+\frac{1}{4}\right) + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}} = \bar{C}_\mu
\end{aligned}$$

By replacing this in the regret equation above and substituting $\delta = 1/\sqrt{(t-1)C_3}$ we get

$$r_t(\pi) = \mathbb{E}[\mathrm{Regret}_t(\pi)] \leq \frac{4(K-1)L_\mu C_0 \bar{C}_\mu x_{\max}^2}{C_3}\frac{1}{t-1} + 4(K-1)L_\mu b_{\max} x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right)$$

as desired. □

Now we are ready to finish up the proof of Proposition 1. The only other result that we need is an upper bound on the probability terms $\mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]$. The key here is again Lemma 4. Note that in the case of GLMs this lemma again holds. The reason is simply because of the fact that the greedy decision does not change in the presence of the inverse link function μ. In other words, as $\arg\max_{i\in[K]}\mu'(X_t^\top\beta_i) = \arg\max_{i\in[K]} X_t^\top\beta_i$, the minimum eigenvalue of each of the covariance matrices is above $t\lambda_0/4$ with a high probability and that implies what we exactly want.

Remark 9. The result of Lemma 4 remains true for the generalized linear models.

Therefore, we can use this observation to finish the proof of Proposition 1. This consists of summing up the regret terms up to time T.

Proof of Proposition 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 $2L_\mu x_{\max} b_{\max}$ and as $K=2$, by using Lemma 4 and Lemma 12 we can write

$$\begin{aligned}
R_T(\pi) &= \sum_{t=1}^T r_t(\pi)\\
&\leq 2L_\mu x_{\max} b_{\max} + \sum_{t=2}^T L_\mu\left[\frac{4C_0\bar{C}_\mu x_{\max}^2}{C_3}\frac{1}{t-1} + 4b_{\max}x_{\max}d\exp(-C_1(t-1))\right]\\
&= 2L_\mu x_{\max} b_{\max} + \sum_{t=1}^{T-1} L_\mu\left[\frac{4C_0\bar{C}_\mu x_{\max}^2}{C_3}\frac{1}{t} + 4b_{\max}x_{\max}d\exp(-C_1 t)\right]\\
&\leq 2L_\mu x_{\max} b_{\max} + L_\mu\frac{4C_0\bar{C}_\mu x_{\max}^2}{C_3}(1+\int_1^T\frac{1}{t}\mathrm{d}t) + 4L_\mu b_{\max}x_{\max}d\int_1^\infty\exp(-C_1 t)\mathrm{d}t\\
&= 2L_\mu x_{\max} b_{\max} + L_\mu\frac{4C_0\bar{C}_\mu x_{\max}^2}{C_3}(1+\log T) + L_\mu\frac{4b_{\max}x_{\max}d}{C_1}\\
&= L_\mu\left(\frac{128C_0\bar{C}_\mu x_{\max}^4\sigma^2 d}{\lambda_0^2}\log T + \left(2x_{\max}b_{\max} + \frac{128C_0\bar{C}_\mu x_{\max}^4\sigma^2 d}{\lambda_0^2} + \frac{160b_{\max}x_{\max}^3 d}{\lambda_0}\right)\right)\\
&= \mathcal{O}(\log T),
\end{aligned}$$

finishing up the proof. □

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 $\kappa' > 0$:*

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

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

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

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*

$$R_T(\pi) = \begin{cases} \mathcal{O}\left(T^{(1-\alpha)/2}\right) & \text{if } 0 \leq \alpha < 1, \\ \mathcal{O}\left(\log T\right) & \text{if } \alpha = 1, \\ \mathcal{O}(1) & \text{if } \alpha > 1, \end{cases} \tag{26}$$

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 6 and Theorem 1 with a very slight modification. Note that all the arguments in Lemma 6 hold and the only difference is where we want to bound the probability $\mathbb{P}[X_t \in I^h]$ in Equation (17). In this Equation, if we use the α-margin bound as

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

we obtain that

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

which turns the regret bound in Equation (19) into

$$r_t(\pi) \leq (K-1)\Big[C_0' 2^{1+\alpha}\delta^{1+\alpha}x_{\max}^{1+\alpha}\Big(\sum_{h=0}^{h_0}(h+1)^{1+\alpha} + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^{1+\alpha}\exp(-h^2)\Big)\Big] \tag{27}$$
$$+ 4(K-1)b_{\max}x_{\max}\max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0}}),$$

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 $h \geq h_2 = \max(h_0, h_1)$ using $h^2 \geq (3+\alpha)h \geq (3+\alpha)\log h$ we have

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

Furthermore, all the terms corresponding to $h \leq h_2 = \max(h_0, h_1)$ have an upper bound equal to $(h+1)^{1+\alpha}$ (remember that for $h \geq h_0+1$ we have $4d\exp(-h^2) \leq 1$). Therefore, the summation in (27) is bounded above by

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

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 $\delta = 1/\sqrt{(t-1)C_3}$ in the Equation (27) and putting together all the constants we reach to

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

for some function g_1.

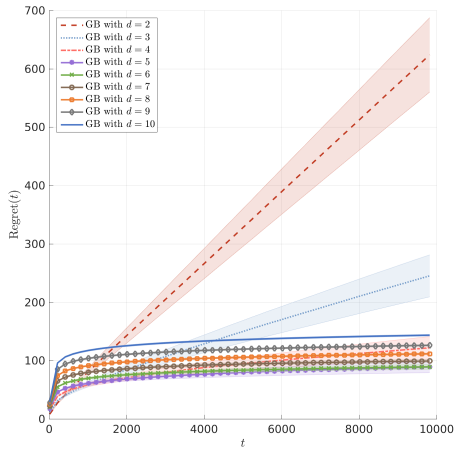
The last part of the proof is summing up the instantaneous regret terms for $t = 1, 2, \ldots, T$. Note that $K = 2$, and using Lemma 4 for $i = 1, 2$, we can bound the probabilities $\mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0}}]$ by $d\exp(-C_1(t-1))$ and therefore

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

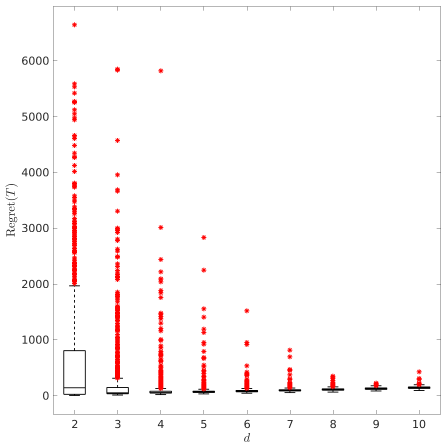
Now note that the integral of $t^{-(1+\alpha)/2}$ over the interval $[1, T]$ satisfies

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

which yields the desired result. □



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



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

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

Appendix F: Additional Simulations

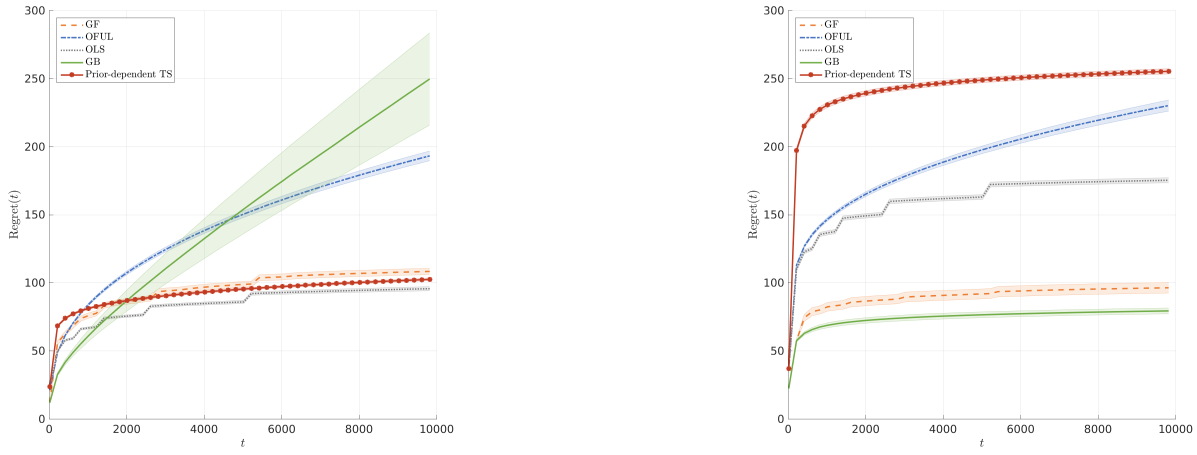
F.1. More than Two Arms ($K > 2$)

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 6, 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.

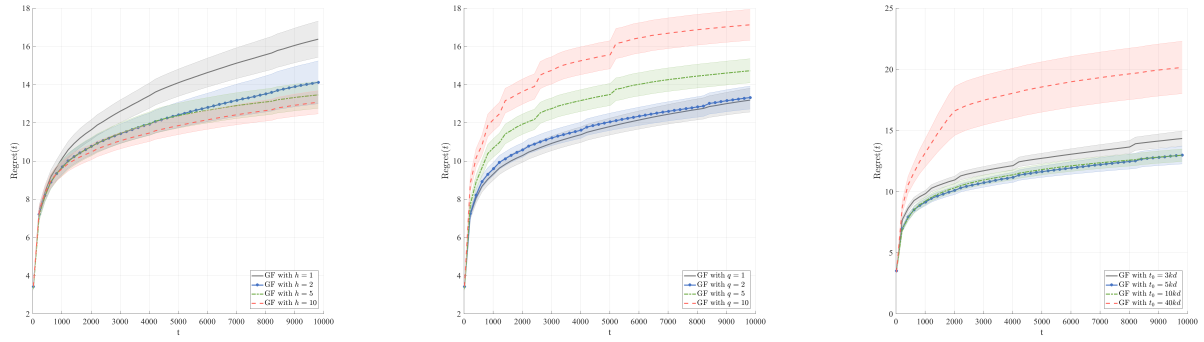
F.2. Sensitivity to parameters

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 7, 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) $K=5, d=3$ (b) $K=5, d=7$

Figure 6 **Simulations for $K>2$ arms.**



(a) Sensitivity with respect to h. (b) Sensitivity with respect to q. (c) Sensitivity with respect to t_0.

Figure 7 **Sensitivity analysis for the expected regret of Greedy-First algorithm with respect to the input parameters h, q, and t_0.**

Appendix G: Missing Proofs of §3.5 and §4.3

Proof of Proposition 2. 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 (8), 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 (8). 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 (8)) 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 (8)) 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 (8) 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 $S^{\text{gb}}(m_1, K_1, \sigma, x_{\max}, \lambda_1)$ and $S^{\text{gb}}(m_2, K_2, \sigma, x_{\max}, \lambda_1)$ 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 (8) 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

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

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

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

By taking limits we reach to the desired conclusion. □

Proof of Proposition 3. We omit proofs regarding to the monotonicity results as they are very similar to those provided in Proposition 2.

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

$$L'(\gamma, \delta, p) \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K + \frac{Kd\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))}. \tag{28}$$

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 $p \leq t_0$ and $\gamma \leq \gamma^*$ we have

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

Substituting this in Equation (28) implies that

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

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

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

and recall that

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

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

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

REMARK 10. 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

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] = \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right) \geq \delta, \|\hat{\beta}(\mathcal{S}_{Km,t}) - \beta_i\|_2 \geq \theta_1\right]. \tag{29}$$

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 (29) turns into

$$\begin{aligned}
\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] &= \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \|M\varepsilon\|_2 \geq \theta_1\right] \\
&= \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \sum_{j=1}^d |m_j^\top \varepsilon| \geq \theta_1\right] \\
&\leq \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \exists j \in [d], |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d}\right] \\
&\leq \sum_{j=1}^d \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d}\right] \\
&= \sum_{j=1}^d \mathbb{P}_{\mathbf{X}} \mathbb{P}_{\varepsilon|\mathbf{X}}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d} \,|\, \mathbf{X} = \mathbf{X}_0\right],
\end{aligned} \tag{30}$$

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

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

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

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

Hence,

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

Putting this back in Equation (30) 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. □

LEMMA 14. *For an arbitrary $Km+1\leq t\leq p-1$ and $i\in[K]$ we have*

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

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

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 15. *For each $t\geq p,i\in\mathcal{K}_{opt}$*

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

Furthermore, for each $t\geq Km+1$ 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 9. Let $i\in\mathcal{K}_{opt}$ and recall that

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

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

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

where X is a random sample from distribution p_X.

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

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

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 $\mathbf{x}^\top\beta_i\ge\max_{l\neq i}\mathbf{x}^\top\beta_l+h$, using two Cauchy-Schwarz inequalities we can write

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

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,

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

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

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

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 $X_kX_k^\top\preceq x_{\max}^2\mathbf{I}$, which is the result of Cauchy-Schwarz inequality, then Lemma 9 implies that

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

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

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

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 16. *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. $\square$

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 17. *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 (14) *and* (15).

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

$$\left\{\left(\cap_{i=1}^K \mathcal{G}_{i,t-1}^{\theta_1}\right) \cap \left(\cup_{i=1}^K \overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\right\} = \left\{\left(\cap_{i=1}^K \mathcal{G}_{i,t-1}^{\theta_1}\right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right\}. \tag{31}$$

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 13 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 14) 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 15). Following this path leads to

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

using Equation (31) 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 15, 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 14 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 17. 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 15. 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 17, 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

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

which by using Lemmas 15 and 16 can be upper bounded by

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

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 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 17 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 15 and 16 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. □

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

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

For deriving a lower bound on the first term let us use the concentration inequality in Lemma 9. Note that here the samples are drawn i.i.d. from the same distribution p_X. Therefore, by applying this Lemma we have

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

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 15. 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

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

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

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

as desired. □

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. □