Exploiting the Natural Exploration In Contextual Bandits

Hamsa Bastani* Mohsen Bayati*† Khashayar Khosravi*

Abstract

The contextual bandit literature has traditionally focused on algorithms that address the exploration-exploitation tradeoff. In particular, greedy policies that exploit current estimates without any exploration may be sub-optimal in general. However, exploration-free greedy policies are desirable in many practical settings where exploration may be prohibitively costly or unethical (e.g. clinical trials). We prove that, for a general class of context distributions, the greedy policy benefits from a natural exploration obtained from the varying contexts and becomes asymptotically rate-optimal for the two-armed contextual bandit. Through simulations, we also demonstrate that these results generalize to more than two arms if the dimension of contexts is large enough. 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 policy or to explore. We prove that this algorithm is asymptotically optimal without any additional assumptions on the distribution of contexts or the number of arms. Extensive simulations demonstrate that Greedy-First successfully reduces experimentation and outperforms existing (exploration-based) contextual bandit algorithms such as Thompson sampling, UCB, or ϵ-greedy.

1 Introduction

The contextual bandit problem [Aue03, LZ08, LCLS10] is a generalization of the traditional multi-armed bandit problem [LR85]. In this setting, the arm rewards depend on an observed context vector X_t at time t. The decision-maker must then learn a mapping between the observed context and the set of available arms. Contextual bandits are a natural model for many applications, including personalized ad or news recommendations [LCLS10] as well as personalized medical care or clinical trials.

A key challenge in such bandit problems is the *exploration-exploitation* tradeoff. The decision-maker needs to explore different arms in order to learn the mapping from contexts to arms, but must also exploit her current estimates in order to minimize expected regret (opportunity cost) at each time step. In particular, simple greedy policies (exploit current estimates without any exploration) may be sub-optimal in general. Therefore, the contextual bandit literature has traditionally focused

*Department of Electrical Engineering, Stanford University
†Graduate School of Business, Stanford University

on regret-minimizing algorithms that include forced-exploration to address this challenge [DHK08, CLRS11, AYPS11, AG13].

However, exploration-free greedy policies are desirable in many practical settings where exploration may be prohibitively costly (e.g., clinical trials) or unethical (e.g., deploying new healthcare quality improvement initiatives). *The first contribution of this paper* is to demonstrate that, under some assumptions, the greedy policy achieves asymptotically optimal regret in a specific but popular setting: the two-armed contextual bandit. In particular, we assume that the contexts are generated i.i.d. from a fixed probability distribution that satisfies a technical condition that we term *covariate diversity*. As discussed in Section 2, this assumption is satisfied by a very general class of natural continuous and discrete probability distributions. Simulations confirm our theoretical results and suggest that the greedy policy outperforms popular contextual bandit algorithms such as UCB, Thompson sampling, or variants of ϵ-greedy [AYPS11, GZ13, AG13, RVR14b] in minimizing cumulative regret. Via simulations we also demonstrate that the superior performance of greedy policy generalizes to more than two arms when the context dimensions are large enough. The key technical challenge of the proof is to show that under the covariate diversity assumption, with a high probability, the minimum eigenvalue for the sample covariance of each arm stays above a fixed positive constant. Combining this fact with a concentration inequality for matrix martingales [Tro11] leads to a sharp convergence proof for greedy estimators. We prove this result for both linear and nonlinear reward functions.

We should note that it is a well known fact that a greedy iterative least squares policy can achieve asymptotically optimal regret if the minimum eigenvalue of the sample covariance stays above a constant [LW82]. Yet, this type of result can only be used after the existence of a lower bound on the minimum eigenvalue is verified. More precisely, in order to invoke [LW82] type results in our setting one needs to show that the smallest eigenvalue of sample covariance for each arm stays larger than a positive constant λ_0 at any step of the algorithm. But this assumption is algorithm dependent since sample covariance of each arm depends on the algorithm's previous actions. We address these issues by establishing a sufficient condition (covariate diversity) that is algorithm independent and is generic enough to hold for many distributions.

Unfortunately, one may not know a priori whether the context distribution satisfies covariate diversity; in such cases, simulation results demonstrate that the greedy algorithm achieves linear regret and fails to converge to the optimal policy. *The second contribution of this paper is to* address this concern. We present Greedy-First, a new algorithm that seeks to reduce exploration when possible by beginning with a greedy policy; however, it uses the observed contexts and rewards to verify (with high probability) if the greedy arm parameter estimates are converging at the asymptotically optimal rate. If not, the algorithm transitions to a standard exploration-based contextual bandit algorithm. We prove that Greedy-First is also asymptotically optimal without our additional covariate diversity assumption or any restriction on the number of arms.

Extensive simulations demonstrate that (i) in the presence of covariate diversity, Greedy-First successfully achieves performance similar to the greedy algorithm (which performs best), and (ii) when the covariate diversity assumption is not met, Greedy-First can still outperform existing contextual bandit algorithms. Empirically, the latter result follows from our observation that the greedy algorithm often converges to the true parameters even without covariate diversity, but achieves linear regret on average since it may fail a small fraction of the time. Greedy-First leverages this observation by following a purely greedy policy until it detects that this approach has failed. Thus, on average, the Greedy-First policy performs better than standard algorithms that include forced-exploration.

1.1 Prior Literature

Several variants of contextual bandit problems have been addressed in the literature (we refer the reader to Chapter 4 of [BCB12] for an informative review). For example, [Sli14, PR+13, RZ10] prove general regret bounds where the arm rewards are given by any smooth, non-parametric functions of the contexts. We focus on the case of linear arm rewards. This setting was first introduced by [Aue03] through the LinRel algorithm and was subsequently improved through the LinUCB algorithm in [CLRS11]. This problem has also been formulated as a linear bandit with changing action space. [DHK08, AYPS11] provide $\mathcal{O}(d\sqrt{T})$ regret bounds when contexts can be generated by an adversary. We note that they also prove "problem-dependent" bounds (where one assumes a constant gap Δ between arm rewards) and regrets are of order $d\log(T)/\Delta$.

As mentioned earlier, this literature typically allows for arbitrary (adversarial) context sequences. We consider the case where contexts are generated i.i.d., which is more suited for certain applications (e.g., clinical trials on treatments for a non-infectious disease). In this setting one can achieve regret of order $\log(T)$ without assuming any gap between arm rewards. To emphasize, our setting is incomparable[1] with the aforementioned problem dependent results; it is more general since arm rewards can be arbitrarily close ($\Delta = 0$) but our contexts are more restrictive (i.i.d. vs adversarial). This setting was first studied by [GZ13] who introduced the OLS Bandit algorithm (a clever extension of ϵ-greedy) and proved an upper bound of $\mathcal{O}(d^3 \log T)$ on its cumulative regret, which was subsequently improved by [BB15] to $\mathcal{O}(d^2 \log d \cdot \log T)$. [GZ13] also provide a lower bound of $\mathcal{O}(\log T)$ regret for this problem (i.e., the contextual bandit with i.i.d. contexts, linear payoffs, and arbitrarily close arm rewards).

However, this substantial literature requires forced-exploration. Greedy policies are desirable in practical settings where exploration may be costly or unethical. Notable exceptions are [Woo79, Sar91], who consider a Bayesian one armed bandit with a single i.i.d. covariate and a parametric reward with a known prior. They show that a greedy policy based on dynamic programming achieves optimal discounted reward as the discount factor converges to 1. [WKP05a] extends this to the single covariate and two arms. [WKP05a, WKP05b] also study situations where the covariate has information about the parameters or when the reward may or may not depend on the covariate. [MRT09] proves that when the greedy policy takes advantage of some known structure between arm rewards, it is asymptotically optimal for the Bayesian setting. Overall, followup work has failed to reproduce this result when the arm parameter is unknown and deterministic or when the regret is minimax and not with respect to a known prior (both of which apply to our setting). For instance, [GZ09] explicitly acknowledges that "we were not able to prove that the myopic policy is rate optimal in our setting." Our work addresses this issue by showing that the greedy (or myopic) policy is optimal under some additional assumptions on the probability distribution over the context space.

[1] For completeness, we will still include the OFUL algorithm of [AYPS11] as one of the benchmarks in our simulation study.

Our approach is also related to recent literature on designing conservative bandit algorithms [WSLS16, KGAYR16] that operate within a safety margin (their loss stays below a certain threshold that is determined by a baseline policy). As a result, similar to us, these papers propose algorithms that control the amount of exploration in order to satisfy the safety constraint. [WSLS16] studies the case of multi-armed bandits where the baseline policy is a fixed arm while [KGAYR16] studies the more general case of contextual linear bandits. Another related paper is [RJZ16] that focuses on learning an underlying linear population model using limited experimentation budget.

Specifically, we consider a decision maker with a limited experimentation budget who must efficiently learn an underlying linear population model. Our main contribution is a novel threshold-based algorithm for selection of most informative observations; we characterize its performance and fundamental lower bounds.

Finally, we note that there are technical parallels between our work and the analysis of the greedy policy and its variants in other literature [LM14, BR12]. For example, in dynamic pricing literature the most commonly-studied dynamic pricing problem (without covariates) is a linear bandit problem without changing action space with a modified reward function [dBZ13], [KZ14b]. Without covariates, there is an uninformative price, which results in incomplete learning and thus greedy policies are provably bad [dBZ13, KZ14b, KZ15]. Thus, variants of the greedy policy with occasional experimentation (similar to GreedyFirst) have been studied [dBZ13, KZ14a, KZ14b, dBZ15]. With covariates, [CLPL16, QB16, JN16, Jav17] there is no such uninformative price because the demand curves are changing. Our setting is different since we need to learn multiple reward functions simultaneously that leads to complex dynamics. For example, in our setting the greedy algorithm could stop learning by dropping an arm (as evidenced by some of our empirical results in Section 5). We prove that this never happens under the covariate diversity assumption which we demonstrate holds for a general class of covariate distributions.

1.2 Organization of the Paper

We start by introducing the model, covariate diversity condition, and discussing examples satisfying the condition in §2. In §3 we present the greedy algorithm, our main results for two-armed bandit, and the proof strategy. We then introduce GreedyFirst and its theoretical guarantees in §4. Finally, we perform empirical simulations using synthetic and semi-synthetic data in §5, and conclude in §6. Technical proofs are provided in the accompanied appendices.

2 Problem Formulation

To simplify the presentation, we will start by the simpler case of linear reward function. The non-linear case is presented in §C.1. Let K denote the number of arms. For any integer n, let $[n]$ denote the set $\{1, ..., n\}$. For each arm, indexed by i where $i \in [K]$, consider an unknown parameter $\beta_i \in \mathbb{R}^d$. At time t, we observe covariates $X_t \in \mathbb{R}^d$ where $\{X_t\}_{t \geq 0}$ is a sequence of i.i.d. samples from an unknown probability distribution P_X. The decision at this point is to pull arm $i \in [K]$, which yields reward $Y_t = X_t^\top \beta_i + \varepsilon_{i,t}$, where $X^\top$ refers to transpose of vector X and $\varepsilon_{i,t}$ are i.i.d. samples from distribution P_ε that is σ-subgaussian. Recall that a random variable

Z is σ-subgaussian if for all $\tau > 0$ we have $\mathbb{E}[e^{\tau Z}] \leq e^{\tau^2\sigma^2/2}$. The expected (pseudo) regret is $r_t \equiv \mathbb{E}\left[\max_{j\in[K]}(X_t^\top \beta_j) - (X_t^\top \beta_i)\right]$, where the expectation is taken with respect to X_t and also the chosen policy, and we seek to minimize the cumulative expected regret $R_T = \sum_{t=1}^{T} r_t$. Our goal is to propose a computationally feasible policy which does not know T and achieves the lowest possible asymptotic expected regret. Next, we will make a set of assumptions that would be required for the theoretical analyses of greedy algorithm. Later in §4 these conditions will be relaxed where we introduce Greedy-First algorithm.

2.1 Assumptions

We first assume that contexts as well as the arm parameters β_i 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, for example see [DHK08].

Assumption 1 (Parameter Set)**.** *There exists a positive constant $x_{\max}$ such that the distribution 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 $i \in [K]$.*

Second, we assume that the distribution of X satisfies a margin condition. This assumption, which is a well-known assumption in the literature [Tsy04], rules out unusual distribution of contexts that put a large mass near the decision boundary (that has zero Lebesgue measure). In particular, we assume that the probability that covariates X lie around hyperplanes $X^\top(\beta_i - \beta_j) = 0$ is bounded ($i \neq j$), but not necessarily on the hyperplane. For example, any distribution that has a bounded probability density function (PDF), with possible point masses on the boundary, satisfies this condition. In Section, C.2 we explain how our regret bounds would change if we assume more general conditions on the decision boundary.

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

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

The last condition is that the context vector comes from a distribution that satisfies a certain property. As we observe later, this property is essential for proving the convergence of our estimated β_i.

Assumption 3 (Covariate Diversity)**.** *Let P_X be the distribution (or law) of context vector $X \in \mathbb{R}^d$. There exists a positive constant λ_0 such that for each vector $\mathbf{u} \in \mathbb{R}^d$ we have*

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

While Assumptions 1-2 seem quite generic, Assumption 3 does not seem straightforward to verify. To help with this, we will state the following lemma that provides a set of sufficient conditions (that are easier to check) that guarantee Assumption 3 and its proof is provided in §A.

Lemma 2.1. *Consider a probability distribution P_X on $\mathbb{R}^d$. Let p_X be the density of X with respect to the Lebesgue measure. 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 $\mathbf{x} \in W$, $(a/b)p_X(-\mathbf{x}) \leq p_X(\mathbf{x})$.

(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 integration is replaced with summation.

In §A.1 we provide a set of generic examples that satisfy Assumptions 1-3. In particular, we show that uniform distribution, truncated gaussian, and a general family of discrete Gibbs distributions satisfy the assumptions. We also discuss that product of any of these distributions that in particular contain situations where discrete and continuous covariates are mixed are also included.

3 Greedy Bandit

We first describe the algorithm and its execution. We state our main result, which bounds the cumulative regret achieved by the Greedy Bandit in §3.2 and provide an overview of the key steps of the proof in §3.3. In §C, we extend our result to the case of nonlinear reward function and more general margin conditions than those described in Assumption 2.

3.1 Algorithm

Notation. Let $\pi_t \in [K]$ denote the arm chosen by our algorithm at time $t \in [T]$. 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. For any subset $\mathcal{S}' \subset [t]$, let $\mathbf{X}(\mathcal{S}')$ be the $|\mathcal{S}'| \times d$ sub-matrix of $\mathbf{X}$ whose rows are X_t for each $t \in \mathcal{S}'$. Similarly, when $\mathcal{S}' \subset \mathcal{S}_{i,t}$, let $Y_i(\mathcal{S}')$ be the length $|\mathcal{S}'|$ vector of corresponding observed rewards $(Y_i)_j$ for each $j \in \mathcal{S}'$. Since $\pi_j = i$ for each $j \in \mathcal{S}'$, the vector $Y_i(\mathcal{S}')$ has no missing entries. 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 created by all the covariates, policy, and observations till time $t-1$ and also covariate X_t. The σ-field $\mathcal{H}_{t-1}$ contains all information that exists at time t for making decision, i.e., π_t is $\mathcal{H}_{t-1}$ measurable. Let $\hat{\Sigma}(\mathbf{X}(\mathcal{S}_{i,t})) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ denote the covariance matrix associated with arm i. For simplifying notations, whenever it is clear what $\mathcal{S}_{i,t}$ is we use $\hat{\Sigma}_{i,t}$ instead of $\hat{\Sigma}(\mathbf{X}(\mathcal{S}_{i,t}))$. For estimating arm parameters, we use the Ordinary Least Squares(OLS) estimator, defined as following. 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 use $\hat{\beta}_{\mathbf{X}(\mathcal{S}_{i,t}), Y(\mathcal{S}_{i,t})}$ as the estimation of β_i at time t. We also denote this estimation by $\hat{\beta}(\mathcal{S}_{i,t})$.

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

Algorithm 1 Greedy Bandit

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

3.2 Main Result: Regret Analysis of Greedy Bandit

Our main result establishes an upper bound of $\mathcal{O}(\log T)$ on the cumulative expected regret of the Greedy Bandit for the two-armed contextual bandit when Assumptions 1-3 are satisfied. The proof is provided in §B.

Theorem 3.1. *If $K = 2$ and Assumptions 1-3 are satisfied, the cumulative expected regret of the Greedy Bandit at time T is at most*

$$R_T(\pi^{gb}) \leq \frac{C_0 \tilde{C} x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} \log T + \tilde{C} \left(\frac{C_0 x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} + \frac{b_{\max} x_{\max}^3 d}{\lambda_0} \right) \tag{3.1}$$

where the constant C_0 is defined in Assumption 2 and the constant $\tilde{C}$ is defined in Eq. (B.4) but it is always between $1/3$ and 51.84 depending on the dimension d.

Lower Bound. [GZ13] established an information-theoretic lower bound of $\mathcal{O}(\log T)$ for any algorithm in a two-armed contextual bandit. We note that they do not make Assumption 3; however, their proof and result still directly apply to our setting since the distribution they use in their proof satisfies Assumption 3 as well. Combined with our upper bound, this result demonstrates that the Greedy Bandit achieves asymptotically optimal cumulative regret in T.

3.3 Proof Strategy

In order to prove that our algorithm achieves a logarithmic regret, we need to prove that the arm estimates $\hat{\beta}$ converge to β with a high probability. However, we cannot use those results on convergence of OLS estimators that require i.i.d. samples from the distribution, as the set $\mathcal{S}_{i,t}$ contains correlated samples. The following proposition, adapted from [BB15] with a few modifications, provides the desired convergence result for our adaptive setting. The proof is provided in §B.

Proposition 3.1 (Tail Inequality for Adaptive Observations, adapted from [BB15])**.** *Consider the K-armed contextual bandit problem with $Y(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \beta_i + \varepsilon(\mathcal{S}_{i,t}),\ i = 1, 2, \ldots, K$. Let π be any policy such that for each t, π_t is measurable with respect to $\mathcal{H}_{t-1}$. For each $i \in [K]$, let E_i denote the event that the minimum eigenvalue of $\hat{\Sigma}_{i,t} = \hat{\Sigma}(\mathbf{X}(\mathcal{S}_{i,t}))$ is not less than λt for some $\lambda > 0$. Then, for all $\chi > 0$,*

$$\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi, E_i\right] \leq 2 \exp\left(\log d - C_2 t \chi^2\right),$$

where $C_2 = \lambda^2/(2d\sigma^2 x_{\max}^2)$.

To show that the minimum eigenvalue of the matrix $\hat{\Sigma}(\mathcal{S}_{i,t})$ is lower bounded by λt for some fixed and positive constant λ with high probability, we use a matrix martingale tail bound due to [Tro11] which is stated in §B for completeness.

Therefore, to establish the regret bound, we take these steps in the reverse order. First, we use Assumption 3 together with Proposition B.1 to prove that with a high probability the sample covariance matrices $\hat{\Sigma}_{i,t}$ have their minimum eigenvalue not less than λt for some constant $\lambda > 0$. This result is proved in Lemma B.1 and is limited to the case that $K = 2$. As we observe later, all other results hold for all values of K. Then, we use Proposition 3.1 to prove that our OLS estimators $\hat{\beta}_i$ converge to β_i with high probability. Finally, we use the standard methodology as in [GZ13] to prove the logarithmic upper bound on regret.

4 Greedy-First Algorithm

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

4.1 Algorithm

Greedy-First algorithm receives as the input λ_0 and t_0, and starts by taking greedy actions according to Greedy Bandit algorithm up to time t_0. After time t_0 in each iteration it checks whether the minimum eigenvalue of sample covariance matrices of all K arms are greater than or equal to $\lambda_0 t/4$. If this condition is satisfied, it means that the covariates for each arm are diverse enough to guarantee the convergence of greedy estimates, with a high probability. On the other hand, if this condition is not met, the algorithm switches to OLS Bandit algorithmwhich is introduced by [GZ13] for two arms and extended to the general setting by [BB15]. For completeness we have presented OLS Bandit in §B.2. Note that, once the algorithm switches from Greedy Bandit to OLS Bandit, it keeps taking actions according to OLS Bandit and it does not switch any longer. In practice,

Algorithm 2 Greedy-First Bandit

Input parameters: λ_0, t_0

Initialize $\hat{\beta}(\mathcal{S}_{i,0})$ at random for $i \in [K]$ and switch to $R = 0$

for $t \in [T]$ **do**

  if $R \neq 0$ **then** break **end (if)**

  Observe $X_t \sim P_X$

  $\pi_t \leftarrow \arg\max_i X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$ (break ties randomly)

  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$

  Play arm π_t, observe $Y_t = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$

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

  Compute covariance matrices $\hat{\Sigma}_{i,t} = \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ for $i \in [K]$

  if $t > t_0$ and $\min_{i \in [K]} \lambda_{\min}\left(\hat{\Sigma}_{i,t}\right) < (\lambda_0 t)/4$ **then** Set $R = t$ **end (if)**

end

Execute OLS Bandit for $t \in [R+1, T]$

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

4.2 Assumptions

We now remove our additional assumptions: namely, that there are only two arms and Assumption 3 on the diversity of covariates. Instead, we will make two weaker assumptions that are typically made in the contextual bandit literature [GZ13]. In particular, Assumption 4 allows for multiple arms (some of which may be uniformly sub-optimal), and Assumption 5 relaxes the assumption on the covariates, e.g., allowing for intercept terms.

Assumption 4 (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 5 (Positive-Definiteness)**.** *Define $\Sigma_i \equiv \mathbb{E}\left[XX^\top \mid X \in U_i\right]$ for all $i \in [K]$. Then, there exists $\lambda_0 > 0$ such that for all $i \in [K]$ the minimum eigenvalue $\lambda_{\min}(\Sigma_i) \geq \lambda_0 > 0$.*

Remark 4.1. *Greedy-First can switch to any contextual bandit algorithm (e.g., OFUL by [AYPS11] or Thompson sampling by [AG13, RVR14a]) instead of the OLS Bandit. Then, the above assumptions would be replaced with analogous assumptions required by that algorithm. Our proof naturally generalizes to adopt the assumptions and regret guarantees of the new algorithm when the Greedy Bandit fails.*

4.3 Regret Analysis of Greedy-First

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

Theorem 4.1. *The cumulative expected regret of the Greedy-First Bandit at time T is at most $C \log T + 2t_0 x_{\max} b_{\max}$, where the constant $C = (K-1)C_{GB} + C_{OB}$. Here C_{GB} is the constant defined in Theorem 3.1 and C_{OB} is the upper bound in the regret decomposition of 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 $\delta = 2d \exp[-t_0 C_1]/C_1$. Here $C_1 = \lambda_0 / 40 x_{\max}^2$, as also defined in Equation (B.1). 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 3.1 only requires Assumption 3 in the proof of Lemma B.1. The remaining steps of the proof hold without the assumption. Thus, if the conclusion of Lemma B.1, $\min_{i \in [K]} \lambda_{\min}(\hat{\Sigma}_{i,t}) \geq \frac{\lambda_0 t}{4}$ holds at every $t \in [t_0 + 1, T]$, then we are guaranteed at most $\mathcal{O}(\log T)$ regret by Theorem 3.1, regardless of whether Assumption 3 holds. Notice that the constant C_{GB} in Theorem 3.1 is replaced by $(K-1)C_{GB}$, as we are considering K arms. This can also be easily implied from Lemma B.3. Note that this condition must hold for Greedy-First to continue pursuing a purely greedy policy, and that happens with high probability when Assumption 3 is true. If this condition fails and Greedy-First switches to a OLS Bandit policy, then we can simply adopt the regret guarantees previously proven for this algorithm.

5 Simulations

In this section we compare the Greedy Bandit and Greedy-First algorithms with the state-of-the-art algorithms on synthetic data and a data with real covariates. The list of other algorithms is: (i) OFUL algorithm of [AYPS11] that is an advanced adaptation of the upper confidence bound (UCB) approach of [LR85]. (ii) Two adaptations of Thompson Sampling [Tho33] by [AG13] and [RVR14b] that will be referred to by *prior-free TS* and *prior-dependent TS* respectively. (iii) OLS Bandit algorithm of [GZ13].

Remark 5.1. *Our naming of Thompson sampling algorithms is based on the fact that the implementation by [RVR14b] assumes the knowledge of prior distribution of parameters β_i while [AG13] does not. But both algorithms assume the knowledge of noise variance. They also require the type of distribution for parameters β_i to be known which would be needed to calculate posterior distribution of the parameters given past observations. In particular, we follow the authors of these papers and implement the version of TS that assumes these variables are gaussian[2]. Finally, we note that the OFUL algorithm also relies on the knowledge of noise variance.*

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

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

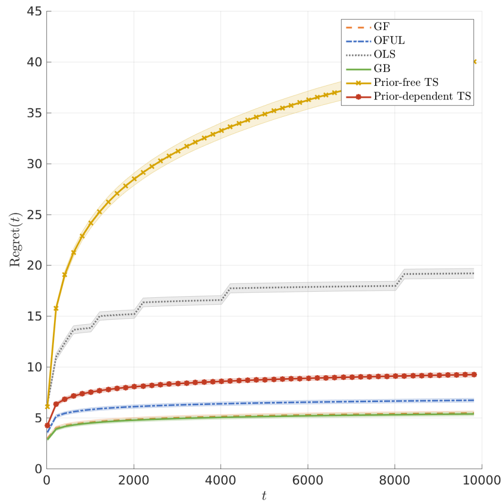
- **Correct prior:** Here we assume that OFUL and both versions of TS know the noise variance. We assume that β_i are sampled independently from $\mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$ Also, prior-dependent TS has access to the true prior distribution of β_i and also the noise variance $\sigma^2 = 0.25$.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

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

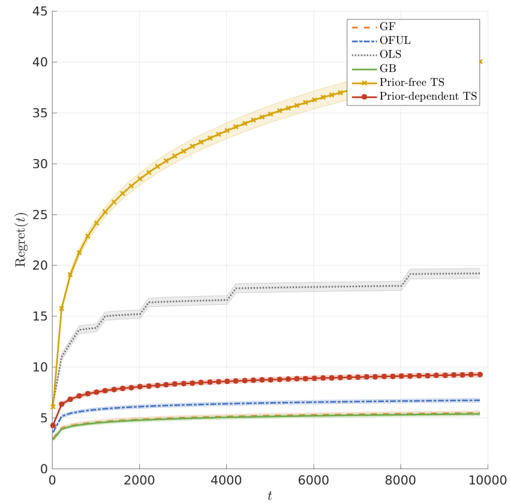
5.1 Additional Simulations

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

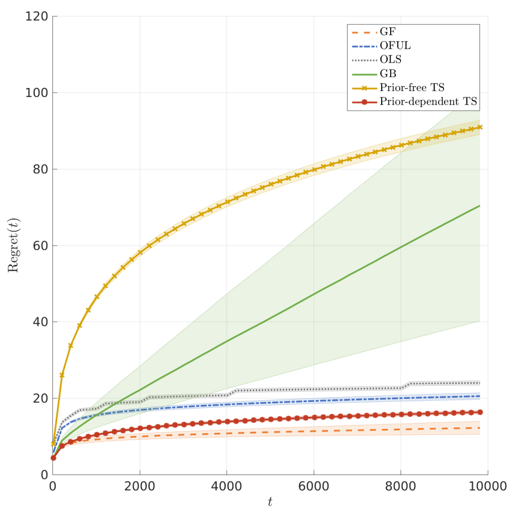
[2]We emphasize that the analyses of [AG13, RVR14b] does not require the parameters to be gaussian. The gaussian assumption is only made for designing the algorithms.



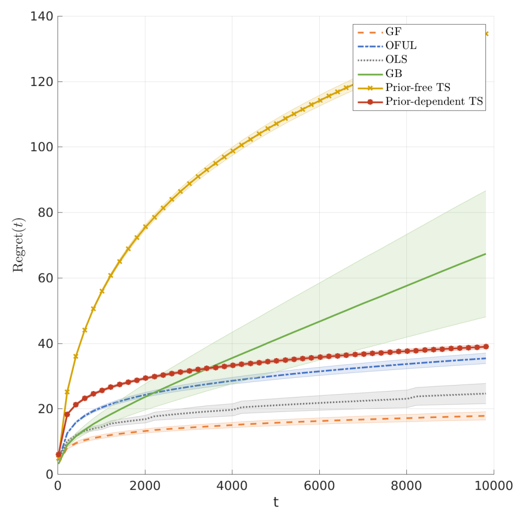
(a) Correct prior and covariate diversity.



(b) Incorrect prior and covariate diversity.



(c) Correct prior and no covariate diversity.



(d) Incorrect prior and no covariate diversity.

Figure 1: Expected regret of all algorithms on synthetic data in four different regimes for the covariate diversity condition and whether OFUL and TS are provided with correct or incorrect information on true prior distribution of the parameters. Out of 1000 runs of each simulation Greedy-First never switched in (a) and (b) and switched only 69 times in (c) and 139 times in (d).

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

- **Real data covariates:** Figure 5 in §D.3 shows results of the situation in Figure 1(a) when covariates are sampled from a real data set. The results have similar qualitative behavior.

- **Sensitivity to parameters** Each of the algorithms requires a set of input parameters. We chose these parameters based on the recommendations in [AY12, RVR14b, BB15]. In particular, OFUL uses $\lambda = 1$ and $\delta = 0.01$. Both versions of Thompson sampling use $\delta = 0.01$, and OLS Bandit uses $h = 5$ and $q = 1$. Greedy Bandit is parameter free and Greedy-First uses the same parameters as OLS Bandit when it switches to that. The only remaining parameter will be t_0 in Greedy-First that will be set at $8Kd$. Figure 4 of §D.1, shows that the choice of parameters h, q, and t_0 will have small impact on performance of Greedy-First.

6 Conclusions and Discussions

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

However, in some scenarios, it is possible to have more than two arms or that the distribution of contexts P_X is not diverse enough. In these cases, the desirable greedy algorithm might sometimes fail. We proposed the Greedy-First algorithm which uses the empirical distributions of contexts to determine whether the Greedy Bandit might successfully proceed or not. This algorithm maintains greedy policy and only switches to an exploratory algorithm, such as OLS Bandit, if the minimum eigenvalue of empirical covariance matrices do not grow fast enough. Simulations suggest that in many problem instances, Greedy-First outperforms OFUL, Thompson Sampling, OLS Bandit, and even Greedy Bandit algorithms. It is also worth noting that, Greedy-First achieves a very similar regret to than of Greedy Bandit for a two-armed contextual bandit problem with diverse covariates, as expected.

References

[AG13] Shipra Agrawal and Navin Goyal, *Thompson sampling for contextual bandits with linear payoffs*, ICML, 2013, pp. 127–135.

[AS+66] Milton Abramowitz, Irene A Stegun, et al., *Handbook of mathematical functions*, Applied mathematics series **55** (1966), no. 62, 39.

[Aue03] Peter Auer, *Using confidence bounds for exploitation-exploration trade-offs*, JMLR **3** (2003), 397–422.

[AY12] Yasin Abbasi-Yadkori, *Online learning for linearly parametrized control problems*, Ph.D. thesis, 2012.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári, *Improved algorithms for linear stochastic bandits*, NIPS, 2011, pp. 2312–2320.

[BB15] Hamsa Bastani and Mohsen Bayati, *Online decision-making with high-dimensional covariates*, 2015, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2661896.

[BCB12] Sbastien Bubeck and Nicol Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Foundations and Trends in Machine Learning **5** (2012), no. 1, 1–122.

[BR12] Josef Broder and Paat Rusmevichientong, *Dynamic pricing under a general parametric choice model*, Oper. Res. **60** (2012), no. 4, 965–980.

[CLPL16] Maxime C Cohen, Ilan Lobel, and Renato Paes Leme, *Feature-based dynamic pricing*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737045.

[CLRS11] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire, *Contextual bandits with linear payoff functions*, AISTATS, 2011, pp. 208–214.

[dBZ13] Arnoud V den Boer and Bert Zwart, *Simultaneously learning and optimizing using controlled variance pricing*, Management Science **60** (2013), no. 3, 770–783.

[dBZ15] Arnoud V. den Boer and Bert Zwart, *Dynamic pricing and learning with finite inventories*, Operations Research **63** (2015), no. 4, 965–978.

[DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade, *Stochastic linear optimization under bandit feedback.*, 2008, pp. 355–366.

[GZ09] Alexander Goldenshluger and Assaf Zeevi, *Woodroofes one-armed bandit problem revisited*, The Annals of Applied Probability **19** (2009), no. 4, 1603–1633.

[GZ13] ———, *A linear response bandit problem*, Stochastic Systems **3** (2013), no. 1, 230–261.

[Jav17] Adel Javanmard, *Perishability of data: Dynamic pricing under varying-coefficient models*, https://arxiv.org/abs/1701.03537.

[JN16] Adel Javanmard and Hamid Nazerzadeh, *Dynamic pricing in high-dimensions*, https://arxiv.org/abs/1609.07574.

[KGAYR16] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy, *Conservative contextual linear bandits*, 2016, https://arxiv.org/abs/1611.06426.

[KZ14a] N. Bora Keskin and Assaf Zeevi, *Chasing demand: Learning and earning in a changing environment*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2389750.

[KZ14b] N Bora Keskin and Assaf Zeevi, *Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies*, Operations Research **62** (2014), no. 5, 1142–1167.

[KZ15] ———, *On incomplete learning and certainty-equivalence control*, preprint (2015).

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire, *A contextual-bandit approach to personalized news article recommendation*, WWW, 2010, pp. 661–670.

[LM14] Tor Lattimore and Remi Munos, *Bounded regret for finite-armed structured bandits*, Advances in Neural Information Processing Systems 27, 2014, pp. 550–558.

[LR85] Tze Leung Lai and Herbert Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in applied mathematics **6** (1985), no. 1, 4–22.

[LW82] Tze Leung Lai and Ching Zong Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, The Annals of Statistics (1982), 154–166.

[LZ08] John Langford and Tong Zhang, *The epoch-greedy algorithm for multi-armed bandits with side information*, NIPS, 2008, pp. 817–824.

[MRT09] Adam J Mersereau, Paat Rusmevichientong, and John N Tsitsiklis, *A structured multi-armed bandit problem and the greedy policy*, IEEE Transactions on Automatic Control **54** (2009), no. 12, 2787–2802.

[PR+13] Vianney Perchet, Philippe Rigollet, et al., *The multi-armed bandit problem with covariates*, The Annals of Statistics **41** (2013), no. 2, 693–721.

[QB16] Sheng Qiang and Mohsen Bayati, *Dynamic pricing with demand covariates*, 2016, https://arxiv.org/abs/1604.07463.

[RJZ16] C. Riquelme, R. Johari, and B. Zhang, *Online Active Linear Regression via Thresholding*, ArXiv e-prints (2016).

[RVR14a] Dan Russo and Benjamin Van Roy, *Learning to optimize via information-directed sampling*, NIPS, 2014, pp. 1583–1591.

[RVR14b] Daniel Russo and Benjamin Van Roy, *Learning to optimize via posterior sampling*, Mathematics of Operations Research **39** (2014), no. 4, 1221–1243.

[RZ10] Philippe Rigollet and Assaf Zeevi, *Nonparametric bandits with covariates*, COLT, 2010, p. 54.

[Sar91] Jyotirmoy Sarkar, *One-armed bandit problems with covariates*, The Annals of Statistics (1991), 1978–2002.

[Sli14] Aleksandrs Slivkins, *Contextual bandits with similarity information*, JMLR **15** (2014), no. 1, 2533–2568.

[Tho33] W. R. Thompson, *On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples*, Biometrika **25** (1933), 285–294.

[Tro11] Joel A Tropp, *User-friendly tail bounds for matrix martingales*, Tech. report, DTIC Document, 2011.

[Tsy04] Alexandre B Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Annals of Statistics (2004), 135–166.

[Wai16] Martin Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Working Publication, 2016.

[WKP05a] Chih-Chun Wang, S. R. Kulkarni, and H. V. Poor, *Bandit problems with side observations*, IEEE Transactions on Automatic Control **50** (2005), no. 3, 338–355.

[WKP05b] Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor, *Arbitrary side observations in bandit problems*, Advances in Applied Mathematics **34** (2005), no. 4, 903 – 938.

[Woo79] Michael Woodroofe, *A one-armed bandit problem with a concomitant variable*, Journal of the American Statistical Association **74** (1979), no. 368, 799–806.

[WPR15] Jonathan Weed, Vianney Perchet, and Philippe Rigollet, *Online learning in repeated auctions*, https://arxiv.org/abs/1511.05720.

[WSLS16] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvari, *Conservative bandits*, Proceedings of The 33rd International Conference on Machine Learning, vol. 48, PMLR, 2016, pp. 1254–1262.

A Properties of Covariate Diversity

Proof of Lemma 2.1. For simplicity, in this proof, we drop the subscript X in p_X. Since for all $\mathbf{u} \in \mathbb{R}^d$ at least one of $\mathbf{x}^\top \mathbf{u} \geq 0$ or $-\mathbf{x}^\top \mathbf{u} \geq 0$ holds, and using conditions (a), (b), and (c) of

Lemma 2.1 we have:

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

Here, the first inequality follows from the fact that $\mathbf{x}\mathbf{x}^\top$ is positive semi-definite, the first equality follows from condition (a) and a change of variable ($\mathbf{x} \to -\mathbf{x}$), the second inequality is by condition (b), the third inequality uses $a \leq b$ which follows from condition (b), and the last inequality uses condition (c). The above argument is implicitly replacing $\int \mathrm{d}P_X$ with an integral with respect to the Lebesgue measure (i.e., $\int p(\mathbf{x})\mathrm{d}\mathbf{x}$). For discrete distributions P_X, the same proof goes through, if we interpret $\int \mathrm{d}P_X$ as $\sum_{\mathbf{x}} p_X(\mathbf{x})$. □ □

A.1 Examples satisfying the assumptions

In this section, using several examples, we demonstrate that the above assumptions contain a wide range of distributions. But first we introduce two new notations. Let B_R^d be the ball of radius R around the origin in $\mathbb{R}^d$ and let the volume of a set $S \subset \mathbb{R}^d$ be defined by $\mathrm{vol}(S) \equiv \int_S \mathrm{d}\mathbf{x}$.

A.1.1 Gibbs Distributions with Positive Covariance

Consider the set $\{\pm 1\}^d \subset \mathbb{R}^d$ equipped with discrete probability distribution P_X that for any $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \{\pm 1\}^d$ satisfies

$$\mathbb{P}(X = \mathbf{x}) \equiv \frac{1}{Z}\, e^{\sum_{1 \leq i,j \leq d} J_{ij} x_i x_j}$$

where $J_{ij} \in \mathbb{R}$ are parameters of the distribution[3] and Z is a normalization term—known as *partition function*. It is easy to see that P_X satisfies Assumptions 1 and 2 since it has a support with finite number of points. Defining $W = \{\pm 1\}^d$, conditions (a) and (b) of Lemma 2.1 hold easily as well. Now, if the covariance of the distribution is positive definite then condition (c) of Lemma 2.1 also holds which means P_X satisfies Assumption 3. Note that this class of distributions includes the well-known Rademacher distribution by taking all J_{ij}'s equal to 0.

[3]In statistical physicis literature J_{ij} maybe random variables themselves, but we assume they are deterministic.

A.1.2 Uniform Distribution

Let P_X be the uniform distribution on the set $V \subset B^d_{x_{\max}}$, where $B^d_R \subset V$ for some $R > 0$. Then Assumptions 1 and 2 as well as conditions (a) and (b) of Lemma 2.1 clearly hold. We would only need to check condition (c). Since P_X is uniform distribution on V this means $p_X(\mathbf{x}) = 1/\mathrm{vol}(V)$, for all $\mathbf{x} \in V$.

Lemma A.1. *For any $R > 0$ we have $\int_{B^d_R} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left[\frac{R^2}{d+2}\mathrm{vol}(B^d_R)\right] I_d$.*

Now, using Lemma A.1, the fact that $V \subset B^d_{x_{\max}}$, and $\mathrm{vol}(B^d_R) = R^d \mathrm{vol}(B^d_{x_{\max}})/x^d_{\max}$ we see that condition (c) of Lemma 2.1 holds with constant $\lambda = R^{d+2}/[(d+2)x^d_{\max}]$.

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

$$\int_{B^d_R} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left(\int_{B^d_R} x_1^2 \mathrm{d}\mathbf{x}\right) I_d \,. \tag{A.1}$$

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 (A.1) with x_i^2 used instead of x_1^2 for all $i \in [d]$, we obtain

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

Comparing this to

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

we obtain that

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

□ □

A.1.3 Truncated Multivariate Gaussian Distribution

Let $p_X(\mathbf{x})$ be a truncated multivariate Gaussian distribution with parameters $\mathrm{N}(\mathbf{0}_d, \Sigma)$ with Σ positive definite. In order to satisfy Assumption 1, we truncate the density to 0 for all $\|\mathbf{x}\|_2 \geq x_{\max}$. The density of truncated Gaussian distribution after renormalization is given by

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

Similar to previous examples, this distribution clearly satisfies Assumptions 1 and 2 as well as conditions (a) and (b) of Lemma 2.1. For proving that the condition (c) of Lemma 2.1 also holds, we have two lemmas that their proofs are provided below.

Lemma A.2. *The following inequality holds*

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

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

Lemma A.3. *Let* $M = \sqrt{\log 2(2d+8)\lambda_{\max}(\Sigma)}$*. If* $x_{\max} \geq M$*, then*

$$\int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,trunc}(\mathbf{x})\mathrm{d}\mathbf{x} \succeq \frac{\Sigma}{2}\,.$$

Lemma A.2 verifies that the condition (c) of Lemma 2.1 holds. However, the bound on density is exponentially decaying so that for large values of $x_{\max}$ it becomes very small. Lemma A.3 solves this issue by proving that the condition (c) in Lemma 2.1 holds for $\lambda = \lambda_{\min}(\Sigma)/2$ whenever $x_{\max} \geq M$. Therefore, this condition holds for $\lambda = \max\{\lambda_{\min}(\Sigma)/2, \lambda_{\mathrm{uni}}\}$.

Remark A.1. *Note that a special case for conditions of Lemma 2.1 to hold would be when W is the entire support of distribution P_X. This is the case in Examples A.1.1 and A.1.3 above. Now, let $X^{(1)}$ be a random vector that satisfies this special case of Lemma 2.1 and has mean 0. Let $X^{(2)}$ be another vector that is independent of $X^{(1)}$ and satisfies the general form of Lemma 2.1. Then it is easy to see that $X = (X^{(1)}, X^{(2)})$ also satisfies conditions of Lemma 2.1. Parts (a) and (b) clearly hold for X. To see why (c) holds, note that the cross diagonal entries in $XX^\top$ are zero since $X^{(1)}$ is mean 0. Therefore, our covariate diversity assumption also holds for general distributions that contain a mixture of discrete and continuous components.*

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

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

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

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

where we used Lemma A.1 in the third line. This concludes the proof. □ □

Proof of Lemma A.3. We want to prove that if $x_{\max}$ is large enough then

$$\int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,\text{trunc}}(\mathbf{x}) \mathrm{d}\mathbf{x} \succeq \frac{\Sigma}{2}. \tag{A.2}$$

For proving this, for any positive definite matrix $C \succ 0$, let us denote the density of $\mathrm{N}(\mathbf{0}_d, C)$ by $p_X(\mathbf{x}; C)$. In other words,

$$p_X(\mathbf{x}; C) = \frac{1}{(2\pi)^{d/2}|C|^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top C^{-1}\mathbf{x}\right)$$

Now using the definition of covariance matrix we obtain that

$$\int_{\mathbb{R}^d} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x}; C) \mathrm{d}\mathbf{x} = C,$$

and as a result we can write that

$$\begin{aligned}
\int_{\mathbb{R}^d} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x}; \Sigma) \mathrm{d}\mathbf{x} &= \Sigma, \\
\int_{\mathbb{R}^d} \mathbf{x}\mathbf{x}^\top \frac{p_X(\mathbf{x}; 2\Sigma)}{4} \mathrm{d}\mathbf{x} &= \Sigma/2.
\end{aligned}$$

Let us compare these two density functions as

$$\begin{aligned}
\frac{p_X(\mathbf{x}; 2\Sigma)/4}{p_X(\mathbf{x}; \Sigma)} &= \frac{\frac{1}{4\sqrt{2\pi}^d |2\Sigma|^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top (2\Sigma)^{-1}\mathbf{x}\right)}{\frac{1}{\sqrt{2\pi}^d |\Sigma|^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x}\right)} \\
&= 2^{-d/2-2} \exp\left(\frac{1}{4}\mathbf{x}^\top \Sigma^{-1}\mathbf{x}\right)
\end{aligned} \tag{A.3}$$

which is greater than or equal to 1 if

$$\frac{1}{4}\mathbf{x}^\top \Sigma^{-1}\mathbf{x} \geq \left(\frac{d}{2}+2\right)\log 2.$$

Now we know that $\mathbf{x}^\top \Sigma^{-1}\mathbf{x} \geq \lambda_{\min}\left(\Sigma^{-1}\right)\|\mathbf{x}\|_2^2$. Therefore, for any $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 \geq M$ where

$$M = \sqrt{(4\log 2)\lambda_{\max}(\Sigma)\left(\frac{d}{2}+2\right)},$$

the ratio in equation (A.3) is greater than or equal to 1. This is true according to $\lambda_{\max}(\Sigma) = 1/\lambda_{\min}(\Sigma^{-1})$. As a result, if $x_{\max} \geq M$ then we have

$$\begin{aligned}
\Sigma &= \int_{\mathbb{R}^d} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} \\
&= \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} + \int_{\|\mathbf{x}\|_2 > x_{\max}} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} \\
&\preceq \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} + \int_{\|\mathbf{x}\|_2 > x_{\max}} \mathbf{x}\mathbf{x}^\top \frac{p_X(\mathbf{x};2\Sigma)}{4}\mathrm{d}\mathbf{x} \\
&\preceq \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d} \mathbf{x}\mathbf{x}^\top \frac{p_X(\mathbf{x};2\Sigma)}{4}\mathrm{d}\mathbf{x} \\
&= \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} + \frac{\Sigma}{2},
\end{aligned}$$

implying

$$\int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x};\Sigma)\mathrm{d}\mathbf{x} \succeq \frac{\Sigma}{2}$$

which implies equation (A.2) as $p_{X,\mathrm{trunc}}(\mathbf{x})$ is obtained after normalizing $p_X(\mathbf{x};\Sigma)$ by a number which is less than or equal to 1. This concludes the proof. □ □

B Proof of Main Theorem

Let us here define the constants that we will use in proofs:

$$C_1 = \frac{0.1\lambda_0}{4x_{\max}^2}, \tag{B.1}$$

$$\lambda = \frac{\lambda_0}{4}, \tag{B.2}$$

$$C_2 = \frac{\lambda^2}{2d\sigma^2 x_{\max}^2}, \tag{B.3}$$

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

We have the following bounds on $\bar{C}$.

Remark B.1. *$\bar{C}$ is a decreasing function of d that is equal to $1/3$ when d goes to $+\infty$. Furthermore, for $d = 2, \bar{C} = 51.84$ which implies that $1/3 \leq \bar{C} \leq 51.84$ always holds.*

B.1 Lemmas for bounding the minimum eigenvalues of empirical covariance matrices

Recall that $\mathbf{X}(\mathcal{S}_{i,t})$ denotes the submatrix of $\mathbf{X}$ whose rows are X_j for $j \in \mathcal{S}_{i,t}$. Moreover, $\hat{\Sigma}(\mathbf{X}(\mathcal{S}_{i,t})) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ denotes the covariance matrix of samples corresponding to the times that the arm i was played. For simplifying notations, we write $\hat{\Sigma}_{i,t}$ instead of $\hat{\Sigma}(\mathbf{X}(\mathcal{S}_{i,t}))$, $i = 1, 2$. The aim is deriving lower bounds on the minimum eigenvalue of the covariance matrices that hold with high probability. These bounds can be obtained using a result on the concentration of random matrices. However, in our setting because of the adaptive nature of decisions, we need a result which does not need an i.i.d. assumption and works in an adaptive setting. Here we use the following result due to [Tro11] to prove the lemma following it.

Proposition B.1 (Theorem 3.1 of [Tro11])**.** *Consider a finite adapted sequence $\{X_k\}$ of positive semi-definite matrices with dimension d, and suppose that $\lambda_{\max}(X_k) \le R$, almost surely. Define the finite series $Y \equiv \sum_k X_k$ and $W \equiv \sum_k \mathbb{E}_{k-1} X_k$, where $\mathbb{E}_m$ denotes the conditional expectation with respect to the filtration of all observed variables till time m. Then for all $\mu \ge 0, \delta \in [0, 1)$ we have:*

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

Lemma B.1. *Let Assumption 1 and 3 hold. When applying Greedy Bandit algorithm for $K = 2$, for each time t and $i = 1, 2$, we have the following bound on the minimum eigenvalues of empirical covariance matrix*

$$\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}_{i,t}\right) \ge \lambda t\right] \ge 1 - \exp(\log d - C_1 t).$$

Proof. We only prove this for the covariance matrix associated with the first arm. The exact same argument shows the similar result for the second arm as well. Note that, at time $k \le t$, we look at $\mathbf{u}_{k-1} = \hat{\beta}(\mathcal{S}_{1,k-1}) - \hat{\beta}(\mathcal{S}_{2,k-1})$ and we pull the first arm if $X_k^\top \mathbf{u}_{k-1} > 0$ and we pull the second arm if $X_k^\top \mathbf{u}_{k-1} < 0$. Note that in the case of tie we draw a random sample W_k from $\mathsf{Binomial}(1, 1/2)$, independent of everything else. We pull the first arm if $W_k = 0$ and we pull the second arm otherwise. Let $\hat{\Sigma}_{1,t}$ denote the covariance matrix of all samples drawn from the first arm then

$$\hat{\Sigma}_{1,t} \equiv \sum_{k=1}^{t} X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0] \right).$$

Since the decision at time k depends on the history of all actions up to time k (and possibly the results of binomial distributions that are independent of everything else), for proving the concentration for $\lambda_{\min}\left(\hat{\Sigma}_{1,t}\right)$, we need to use the adaptive concentration inequality as given in Proposition B.1. First, we define

$$\tilde{\Sigma}_{1,t} = \sum_{k=1}^{t} \mathbb{E}_{k-1} \left[X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0] \right) \right],$$

where in each inner term, the expectation is taken with respect to the filtration of all variables introduced till time $k-1$. As $\mathbf{u}_{k-1}$ belongs to this σ-field, $X \sim P_X$, and W_k is independent of all other variables, when we take expectation with respect to $\mathbb{E}_{k-1}$, each inner term is lower bounded by

$$
\begin{aligned}
&\mathbb{E}_{k-1}\left[X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0)\right)\right] \\
&\quad = \mathbb{E}_{k-1}\left[XX^\top \left(\mathbb{I}[X^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[W_k = 0]\mathbb{I}[X^\top \mathbf{u}_{k-1} = 0)\right)\right] \\
&\quad = \mathbb{E}_X\left[XX^\top \left(\mathbb{I}[X^\top \mathbf{u}_{k-1} > 0] + \frac{1}{2}\mathbb{I}[X^\top \mathbf{u}_{k-1} = 0]\right)\right] \\
&\quad \succeq \frac{1}{2}\mathbb{E}_X\left[XX^\top \mathbb{I}[X^\top \mathbf{u}_{k-1} \geq 0]\right],
\end{aligned}
$$

which by using the Assumption 3 is greater than or equal to $\lambda_0/2$. Therefore, by using the concavity of the function $h(F) = \lambda_{\min}(F)$ over PSD matrices, we can write

$$
\begin{aligned}
\lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right) &= \lambda_{\min}\left(\sum_{k=1}^{t} \mathbb{E}_{k-1}\left[X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0]\right)\right]\right) \\
&\geq \sum_{k=1}^{t} \lambda_{\min}\left(\mathbb{E}_{k-1}\left[X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0]\right)\right]\right) \\
&\geq \frac{\lambda_0}{2} t.
\end{aligned} \tag{B.5}
$$

The last thing we need to show for using the Proposition B.1, is proving an upper bound on $\lambda_{\max}(X_k X_k^\top)$, which can be attained by using a Cauchy-Schwarz inequality as following

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

Therefore, by putting $\delta = 1/2$ in Proposition B.1 we reach to

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

where we used the inequality $-0.5 - 0.5\log(0.5) \leq -0.1$. According to the matrix inequality given in (B.5) the second event inside the probability always happens. Hence,

$$
\mathbb{P}_X\left[\lambda_{\min}\left(\hat{\Sigma}_{1,t}\right) \leq \lambda t\right] \leq \exp\left(\log d - C_1 t\right),
$$

as desired. $\square$ $\square$

From here, we consider the general K-armed contextual bandit problem as we need these results for proving the optimality of Greedy-First algorithm as well. We start by proving the Proposition 3.1. Following the approach in this paper, we first state the Bernstein concentration for martingales.

Lemma B.2 (Bernstein Concentration)**.** *Let $\{D_k, \mathcal{H}_k\}_{k=1}^{\infty}$ be a martingale difference sequence, and let D_k be σ-subgaussian. Then, for all $t > 0$ we have*

$$\mathbb{P}\left[\Big|\sum_{k=1}^{n} D_k\Big| \geq t\right] \leq 2\exp\left\{-t^2/(2n\sigma^2)\right\}.$$

Proof. See Theorem 2.3 of [Wai16] and let $b_k = 0$ and $\nu_k = 0$ for all k. □ □

Now having this in mind, we are ready to prove Proposition 3.1.

Proof of Proposition 3.1. Take an arbitrary i, and note that conditioning on E_i

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

As a result, we can write

$$\begin{aligned}
\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi, E_i\right] &= \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \mid E_i\right]\mathbb{P}\left[E_i\right] \\
&= \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda \mid E_i\right]\mathbb{P}\left[E_i\right] \\
&= \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda, E_i\right] \\
&\leq \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda\right]
\end{aligned}$$

As a result, we need to derive an upper bound on this term. For any matrix Z, Let $Z^{(r)}$ denote the r^{th} column of $\mathbf{Z}$, then we can write

$$\mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{1,t})^\top \varepsilon(\mathcal{S}_{1,t})\|_2 \geq \lambda t \cdot \chi\right] \leq \sum_{r=1}^{d} \mathbb{P}\left[|\varepsilon(\mathcal{S}_{1,t})^T \mathbf{X}(\mathcal{S}_{1,t})^{(r)}| \geq \frac{\lambda t \cdot \chi}{\sqrt{d}}\right].$$

Let us expand $\varepsilon(\mathcal{S}_{1,t})^T \mathbf{X}(\mathcal{S}_{1,t})^{(r)} = \sum_{j=1}^{t} \varepsilon_j X_{j,r} \mathbb{I}\left[j \in \mathcal{S}_{1,t}\right] = \sum_{j=1}^{t} \varepsilon_j X_{j,r}\left[j \in \mathcal{S}_{1,j}\right]$, where we used $\mathbb{I}\left[j \in \mathcal{S}_{1,t}\right] = \mathbb{I}\left[j \in \mathcal{S}_{1,j}\right]$. As $|X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{1,j}\right]| \leq x_{\max}$ holds and as ε_j is a σ-subgaussian random variable, therefore $\varepsilon_j X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{1,j}\right]$ is a $(x_{\max}\sigma)$-subgaussian random variable. Recall that $\mathcal{H}_{j-1}$ is the σ-field created by random vectors $X_1, X_2, \ldots, X_j$, policy $\pi_1, \pi_2, \ldots, \pi_{j-1}$, and the random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{j-1}$. Note that the variable $X_{j,r}$ is $\mathcal{H}_{j-1}$-measurable. Moreover, $\mathbb{I}\left[j \in \mathcal{S}_{1,j}\right]$ is also $\mathcal{H}_{j-1}$ measurable, as it only depends on $X_1, X_2, \ldots, X_j$, $\pi_1, \pi_2, \ldots, \pi_{j-1}$, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{j-1}$. Taking expectation gives $\mathbb{E}[\varepsilon_j X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{1,j}\right] \mid \mathcal{H}_{j-1}] = X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{1,j}\right]\mathbb{E}[\varepsilon_j \mid \mathcal{H}_{j-1}] = 0$. This implies that the sequence $D_j = \varepsilon_j X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{1,j}\right]$, $j = 1, 2, \ldots, t$ is a martingale difference sequence adapted to the filtration $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_t$. Therefore, the Bernstein inequality implies that

$$\mathbb{P}\left[\Big|\sum_{j=1}^{t} \varepsilon_j X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{1,t}\right]\Big| \geq \frac{\lambda t \cdot \chi}{\sqrt{d}}\right] \leq 2\exp\left(-\frac{t\lambda^2\chi^2}{2d\sigma^2 x_{\max}^2}\right),$$

which turns into

$$\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{1,t})-\beta_1\|_2 \geq \chi, E_i\right] \leq 2d\exp\left(-\frac{t\lambda^2\chi^2}{2d\sigma^2 x_{\max}^2}\right) = 2\exp\left(\log d - C_2 t\chi^2\right),$$

yielding the desired inequality. □ □

B.2 Regret analysis of Greedy Bandit and Greedy-First

Before we start with the proofs, we present the pseudo-code for OLS-Bandit and also the the heuristic for Greedy-First that were not presented in §4 due to space limitations. The OLS bandit algorithm is introduced by [GZ13] and generalized by [BB15]. 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. The algorithm requires definition of *forced-sample sets*. In particular, let us prescribe a set of times when we forced-sample arm i (regardless of the observed covariates X_t):

$$\mathcal{T}_i \equiv \left\{(2^n-1)\cdot Kq + j \;\middle|\; n \in \{0,1,2,...\} \text{ and } j \in \{q(i-1)+1, q(i-1)+2, ..., iq\}\right\}. \quad \text{(B.6)}$$

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 $\mathcal{T}_{i,t} \subset \mathcal{S}_{i,t}$.

Algorithm 3 OLS Bandit

Input parameters: q, h
Initialize $\hat{\beta}(\mathcal{T}_{i,0})$ and $\hat{\beta}(\mathcal{S}_{i,0})$ by 0 for all i in $[K]$
Use q to construct force-sample sets $\mathcal{T}_i$ using Eq. (B.6) 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**
    $\pi_t \leftarrow i$
  end
  else
    $\hat{\mathcal{K}} = \left\{i \in K \;\middle|\; X_t^\top \hat{\beta}(\mathcal{T}_{i,t-1}) \geq \max_{j\in K} X_t^\top \hat{\beta}(\mathcal{T}_{j,t-1}) - h/2\right\}$
  end
  $\pi_t \leftarrow \arg\max_{i\in\hat{\mathcal{K}}} X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
  Play arm π_t, observe $y_t = X_t^\top \beta_{\pi_t} + \varepsilon_{i,t}$
end

Before starting with the proof we also note that our proof is adapted from [GZ13, BB15] with a few modifications.

Algorithm 4 Heuristic Greedy-First Bandit

Input parameters: t_0

Execute Greedy Bandit for $t \in [t_0]$, set $\lambda' = \frac{1}{2}\min_{i\in[K]} \lambda_{\min}\left(\hat{\Sigma}_{i,t_0}\right)$

if $\lambda' \neq 0$ **then**

| Execute Greedy-First Bandit for $t \in [t_0+1, T]$ with $\lambda_0 = \lambda'$

end

else

| Execute OLS Bandit for $t \in [t_0+1, T]$

end

Lemma B.3. *Considering Assumptions 1 and 2 hold, we have the following upper bound on the expected regret of Greedy Bandit at time t:*

$$r_t(\pi^{gb}) \leq \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}\frac{1}{t} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[E_i{}^c]\right).$$

In above $E_i = \{\lambda_{\min}(\hat{\Sigma}_{i,t}) \geq \lambda t\}$, C_0 is the constant which comes from the Assumption 2, C_2 is defined in (B.3), and $\bar{C}$ is given in (B.4).

Proof. Let us divide the space of contexts into K different parts $\{A_l\}_{l=1}^K$ such that A_l is the subspace in which the arm l is optimal. In other words,

$$A_l = \{X \mid X^\top \beta_l \geq \max_{j\neq l} X^\top \beta_j\}.$$

In above, we can assign the ties arbitrarily. Now we compute the regret that is incurred in each of these sets. More precisely,

$$r_t(\pi^{\text{gb}}) = \sum_{l=1}^K r_t(\pi^{\text{gb}} \mid A_l)\mathbb{P}(A_l).$$

Let us consider the lth term in the above summation. Note that $\pi^{\text{gb}}(t) = \arg\max_{i=1}^K X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$, meaning that conditioning on A_l, we incur a regret with expectation equal to $X_t^\top(\beta_l - \beta_{\pi^{\text{gb}}(t)})$. So that conditioning on A_l, we can write the expected regret at time t as

$$r_t(\pi^{\text{gb}} \mid A_l) = \mathbb{E}\left[X_t^\top(\beta_l - \beta_{\pi^{\text{gb}}(t)}) \mid A_l\right],$$

where the expectation is taken with respect to filtration $\mathcal{H}_{t-1}$. Let us define

$$B_{il} = \left\{X_t \mid X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})\right\}, \text{ for all } i, l \in [K], i \neq l.$$

Note that we may incur a nonzero regret if $X_t^\top \hat{\beta}(\mathcal{S}_{\pi^{\text{gb}}(t),t-1}) > X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ or if $X_t^\top \hat{\beta}(\mathcal{S}_{\pi^{\text{gb}}(t),t-1}) = X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ 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 $X_t^\top \hat{\beta}(\mathcal{S}_{\pi^{\text{gb}}(t),t-1}) = X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ we do not incur any regret if W_t indicates arm l as the action to be taken. Nevertheless, as regret is a non-negative quantity, we can write

$$
\begin{aligned}
r_t(\pi^{\mathrm{gb}} \mid A_l) &\leq \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{\pi^{\mathrm{gb}}(t),t-1}) \geq X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})) X_t^\top(\beta_l - \beta_{\pi^{\mathrm{gb}}(t)}) \mid A_l\right] \\
&\leq \sum_{i \neq l} \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})) X_t^\top(\beta_l - \beta_i) \mid A_l\right] \\
&= \sum_{i \neq l} \mathbb{E}\left[\mathbb{I}(B_{il}) X_t^\top(\beta_l - \beta_i) \mid A_l\right] \\
&\leq \sum_{i \neq l} \left\{\mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) X_t^\top(\beta_l - \beta_i) \mid A_l\right]\right. \\
&\quad \left. + \mathbb{E}\left[\mathbb{I}(B_{il}, E_1{}^{\mathsf{c}}) X_t^\top(\beta_l - \beta_i) \mid A_l\right] + \mathbb{E}\left[\mathbb{I}(B_{il}, E_i{}^{\mathsf{c}}) X_t^\top(\beta_l - \beta_i) \mid A_l\right]\right\} \\
&\leq \sum_{i \neq l} \left\{\mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) X_t^\top(\beta_l - \beta_i) \mid A_l\right] + 2 b_{\max} x_{\max} \left(\mathbb{P}(E_l{}^{\mathsf{c}}) + \mathbb{P}(E_i{}^{\mathsf{c}})\right)\right\} \\
&\leq \sum_{i \neq l} \mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) X_t^\top(\beta_l - \beta_i) \mid A_l\right] + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(E_i{}^{\mathsf{c}}) \qquad \text{(B.7)}
\end{aligned}
$$

where in the second line we used a union bound, in the sixth line we used the fact that E_i and E_l are independent of A_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 E_i and E_l, we can use Proposition 3.1 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 event

$$
I^h = \{X_t^\top(\beta_l - \beta_i) \in (2\delta x_{\max} h, 2\delta x_{\max}(h+1)]\},
$$

where $\delta = 1/\sqrt{tC_2}$. Note that since, $X_t^\top(\beta_l - \beta_i)$ is bounded above by $2b_{\max}x_{\max}$, 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 (B.7) for i, by conditioning on I^h as following

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) X_t^\top(\beta_l - \beta_i) \mid A_l\right] \\
&= \sum_{h=0}^{h^{\max}} \mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) X_t^\top(\beta_l - \beta_i) \mid A_l, I^h\right] \mathbb{P}[I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1) \mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) \mid A_l, I^h\right] \mathbb{P}[I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1) \mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) \mid A_l, I^h\right] \mathbb{P}[X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]] \\
&\leq \sum_{h=0}^{h^{\max}} 4 C_0 \delta^2 x_{\max}^2 (h+1)^2 \mathbb{P}\left[B_{il}, E_l, E_i \mid A_l, I^h\right], \qquad \text{(B.8)}
\end{aligned}
$$

where in the first inequality we used the fact that, conditioning on 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 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 E_l, E_i, and I^h we want to bound the probability that we pull a wrong arm. Note that conditioning on I^h, the event $X_t^\top\left(\hat{\beta}(\mathcal{S}_{i,t-1}) - \hat{\beta}(\mathcal{S}_{l,t-1})\right) \geq 0$ can be rewritten as

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

This implies that 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. Therefore, we can write that

$$
\begin{aligned}
&\mathbb{P}\left[\mathbb{I}(B_{il}, E_l, E_i) \mid A_l, I^h\right] \\
&\quad \leq \mathbb{P}\left[X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, E_l, E_i \mid A_l, I^h\right] + \mathbb{P}\left[X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, E_l, E_i \mid A_l, I^h\right] \\
&\quad \leq \mathbb{P}\left[X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, E_l \mid A_l, I^h\right] + \mathbb{P}\left[X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, E_i \mid A_l, I^h\right] \\
&\quad \leq \mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \geq \delta h, E_l \mid A_l, I^h\right] + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \geq \delta h, E_i \mid A_l, I^h\right] \\
&\quad = \mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \geq \delta h, E_l\right] + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \geq \delta h, E_i\right] \\
&\quad \leq 4d \exp\left(-C_2 t (\delta h)^2\right) \\
&\quad = 4d \exp(-h^2),
\end{aligned}
$$

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, in the fifth line we used the fact that both A_l and I^h only depend on X_t which is independent of $\hat{\beta}(\mathcal{S}_{j,t-1})$ for all j, and in the sixth line we used Proposition 3.1. We can also bound this probability by 1, which is better than $4d\exp(-h^2)$ for small values of h. Hence, using $\sum_{l=1}^K \mathbb{P}[A_l] = 1$ we can write the regret as

$$
\begin{aligned}
r_t(\pi^{\text{gb}}) &= \sum_{l=1}^K r_t(\pi^{\text{gb}} \mid A_l)\mathbb{P}(A_l) \\
&\leq \sum_{l=1}^K \left(\sum_{i \neq l} \sum_{h=0}^{h^{\max}} \left[4C_0 \delta^2 x_{\max}^2 (h+1)^2 \min\{1, 4d\exp(-h^2)\}\right] + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(E_i{}^{\mathsf{c}}) \right) \mathbb{P}(A_l) \\
&\leq 4(K-1) C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h^{\max}} (h+1)^2 \min\{1, 4d\exp(-h^2)\} \right) + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(E_i{}^{\mathsf{c}}) \\
&\leq 4(K-1) \left(C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h_0} (h+1)^2 + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^2 \exp(-h^2) \right) + b_{\max} x_{\max} \max_i \mathbb{P}(E_i{}^{\mathsf{c}}) \right),
\end{aligned}
\tag{B.9}
$$

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

are both decreasing for $x \geq 1$ and therefore

$$\begin{aligned}
\sum_{h=h_0+1}^{h^{\max}} (h+1)^2 \exp(-h^2) &= \sum_{h=h_0+1}^{h^{\max}} (h^2+2h+1) \exp(-h^2) \\
&= \sum_{h=h_0+1}^{h^{\max}} h^2 \exp(-h^2) + 2 \sum_{h=h_0+1}^{h^{\max}} h \exp(-h^2) + \sum_{h=h_0+1}^{h^{\max}} \exp(-h^2) \\
&\leq \int_{h_0}^{\infty} h^2 \exp(-h^2) \mathrm{d}h + \int_{h_0}^{\infty} 2h \exp(-h^2) \mathrm{d}h + \int_{h_0}^{\infty} \exp(-h^2) \mathrm{d}h. \quad \text{(B.10)}
\end{aligned}$$

Now let us compute an upper bound on each of these three terms separately:

- For the first term we apply the integration by parts to obtain

$$\begin{aligned}
\int_{h_0}^{\infty} h^2 \exp(-h^2) \mathrm{d}h &= -\frac{h}{2} \exp(-h^2)\Big|_{h=h_0}^{h=+\infty} + \frac{1}{2} \int_{h_0}^{\infty} \exp(-h^2) \mathrm{d}h \\
&\leq \frac{h_0}{2} \exp(-h_0^2) + \frac{1}{2} \frac{\exp(-h_0^2)}{h_0 + \sqrt{h_0^2 + 4/\pi}} \\
&\leq \exp(-h_0^2) \left(\frac{h_0}{2} + \frac{1}{4} \right),
\end{aligned}$$

  where we used $h_0 \geq 1$ together with the well-known inequality

$$\int_{t}^{\infty} \exp(-x^2) \mathrm{d}x \leq \frac{\exp(-t^2)}{t + \sqrt{t^2 + 4/\pi}}, \tag{B.11}$$

  which is carried from Handbook of Mathematical Functions by [AS+66].

- For the second term we write

$$\int_{h_0}^{\infty} 2h \exp(-h^2) \mathrm{d}h = \exp(-h^2)\Big|_{h=h_0}^{h=+\infty} = \exp(-h_0^2).$$

- For the third term we have

$$\int_{h_0}^{\infty} \exp(-h^2) \mathrm{d}h \leq \frac{\exp(-h_0^2)}{h_0 + \sqrt{h_0^2 + 4/\pi}} \leq \frac{\exp(-h_0^2)}{2},$$

  where we again used equation (B.11).

Now if we replace these bounds on (B.10) we obtain

$$\sum_{h=h_0+1}^{h^{max}} (h+1)^2 \exp(-h^2) \leq \frac{h_0}{2} \exp(-h_0^2) + \frac{7}{4} \exp(-h_0^2).$$

Therefore, the summation in (B.9) is equal to

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

where $\bar{C}$ is defined as (B.4). By replacing this in (B.9) and $\delta = 1/\sqrt{tC_2}$ we get

$$r_t(\pi^{\mathrm{gb}}) \le \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}\frac{1}{t} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[E_i{}^{\mathsf{c}}]\right)$$

as desired. □ □

Now we are ready to prove the final theorem. This is basically summing up the regret terms up to time T.

Proof of Theorem 3.1. Note that in this theorem $K = 2$ and Assumption 3 holds. The expected cumulative regret is the sum of expected regret for times up to time T. Using Lemma B.1 and Lemma B.3 we can write

$$\begin{aligned}
R_T(\pi^{\mathrm{gb}}) &= \sum_{t=1}^{T} r_t(\pi^{\mathrm{gb}})\\
&\le \sum_{t=1}^{T}\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}\frac{1}{t} + 4b_{\max}x_{\max}d\exp(-C_1 t)\\
&\le \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}(1+\int_1^T\frac{1}{t}\mathrm{d}t) + 4b_{\max}x_{\max}d\int_0^\infty\exp(-C_1 t)\mathrm{d}t\\
&= \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_2}(1+\log T) + \frac{4b_{\max}x_{\max}d}{C_1}\\
&= \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2}\log T + \left(\frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2} + \frac{160b_{\max}x_{\max}^3 d}{\lambda_0}\right)\\
&= \mathcal{O}(\log T),
\end{aligned}$$

finishing up the proof. □ □

Proof of Theorem 4.1. First, we will show that Greedy-First achieves asymptotically optimal regret. Note that the expected regret during the first t_0 rounds is upper bounded by $2x_{\max}b_{\max}t_0$. For the period $[t_0+1, T]$ we consider two cases: (i) the algorithm pursues a purely greedy strategy, i.e., $R = 0$, or (ii) the algorithm switches to the OLS Bandit algorithm, i.e., $R \in [t_0+1, T]$.

Case 1: By construction, we know that

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

for all $t > t_0$. This is true according to the fact that Greedy-First only switches when the minimum of minimum eigenvalues of the covariance matrices is less than $\lambda_0 t/4$. Therefore, if the algorithm does not switch, it basically implies that the minimum eigenvalue of all covariance matrices are greater that or equal to $\lambda_0 t/4$ for all values of $t > t_0$. Then, the conclusion of Lemma B.1 holds in this time range (E_i holds for all $i \in [K]$). Consequently, even if Assumption 3 does not hold and $K \neq 2$, Lemma B.3 holds and proves an upper bound on the expected regret r_t. This implies that the regret bound of Theorem 3.1, after multiplying by $(K-1)$, holds for the Greedy-First. Therefore, Greedy-First is guaranteed to achieve $(K-1)C_{GB}\log(T-t_0)$ regret in the period $[t_0+1, T]$, for some constant C_{GB} that depends only on p_X, b and σ. Hence, the regret in this case is upper bounded by $2x_{\max}b_{\max}t_0 + (K-1)C_{GB}\log T$.

Case 2: Once again, by construction, we know that $\min_{i\in[K]} \lambda_{\min}\left(\hat{\Sigma}_{i,t}\right) \geq \lambda_0 t/4$ for all $t \in [t_0+1, R]$. Then, using the same argument as above, Theorem 3.1 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 [BB15] guarantees that the OLS Bandit will achieve at most $C_{OB}\log(T-R)$ regret under Assumptions 1, 2, 4, 5 and when we apply this algorithm to $t \in [R+1, T]$. Thus, the total regret is at most $2x_{\max}b_{\max}t_0 + ((K-1)C_{GB} + C_{OB})\log T$.

Thus, the Greedy-First algorithm always achieves $\mathcal{O}(\log T)$ cumulative expected regret. Next, we prove that when Assumption 3 holds and $K = 2$, the Greedy-First algorithm maintains a purely greedy policy with high probability. In particular, Lemma B.1 states that if the specified λ_0 satisfies

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

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

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

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

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

This conclues the proof. □ □

C Extensions to nonlinear rewards and α-margin boundary conditions

C.1 Extension to nonlinear reward functions

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

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

Note that although many interesting functions such as exponential reward are generally not Lipschitz over $I = \mathbb{R}$, they become Lipschitz once the interval $I = (a, b)$ is bounded. In our case, $X_t^\top \beta_i$ always belongs to $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$, according to Assumption 1 and a Cauchy-Schwarz inequality, the function ψ needs to be Lipschitz on the interval $I = [-b_{\max}x_{\max}, b_{\max}x_{\max}]$. In particular, any differentiable function $\psi(z)$ with a derivative that is bounded above by $D \geq 0$ over the interval $[-b_{\max}x_{\max}, b_{\max}x_{\max}]$ is Lipschitz with the same constant D. This can be implied using the Mean-Value Theorem as

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

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

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

$$Y_t = \psi\left(X_t^\top \beta_i + \varepsilon_{i,t}\right). \tag{C.1}$$

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

$$\mathbb{E}\left[\psi\left(X_t^\top \beta_i\right) - \psi\left(X_t^\top \beta_j\right)\right], \tag{C.2}$$

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

Algorithm 5 Modified Greedy Bandit

Given: Function ψ

Initialize $\hat{\beta}(\mathcal{S}_{i,0})$ at random for $i \in [K]$

for $t \in [T]$ **do**

  Observe $X_t \sim P_X$

  $\pi_t \leftarrow \arg\max_i \psi\left(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})\right)$ (break ties randomly)

  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$

  Play arm π_t, observe $Y_t = \psi\left(X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}\right)$

  Let $Z_t = \psi^{-1}(Y_t)$

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

end

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

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

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

where $R_T(\pi^{gb})$ given in (3.1), is the regret of Greedy Bandit when the reward is linear, and $R_T^{\psi}(\pi^{gb})$ is the regret of Modified Greedy Bandit when the reward comes from the model in Equation (C.1). According to Theorem 3.1, this further implies that

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

with C_{GB} defined in Theorem 3.1.

Proof. The regret proof of the Greedy Bandit algorithm (designed for the linear reward) can be readily extended to the Modified Greedy Bandit algorithm with two slight modifications that we need to: 1) verify that the concentration results for $\hat{\beta}_i$ holds, similar to what we had in Proposition 3.1, and 2) derive an upper bound on the expected regret that is incurred in the presence of ψ, based on the expected regret of the linear case.

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

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

the concentration result in Proposition 3.1 holds.

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

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

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

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

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

$$R_T^{\psi}(\pi^{\text{gb}}) \le MR_T(\pi^{\text{gb}}),$$

which implies the first result. The second result is a straightforward application of Theorem 3.1.

□ □

C.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 [WPR15], we assume that the distribution of contexts P_X satisfies a more general α-margin condition as following.

Assumption 6 (α-Margin Condition)**.** *For $\alpha \ge 0$, we say that the distribution P_X satisfies the α-margin condition, if there exists a constant $C_0 > 0$ such that for each $\kappa > 0$:*

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

Although it is straightforward to verify that any distribution P_X satisfies the 0-margin condition, for each $\alpha > 0$, it is easy to construct a distribution violating the α-margin condition. 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

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

the distribution P_X satisfies the α-margin condition for all $\alpha \ge 0$.

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 C.2. *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*

$$R_T(\pi^{gb}) = \begin{cases} \mathcal{O}\left(T^{(1-\alpha)/2}\right) & \text{if } 0 \leq \alpha < 1, \\ \mathcal{O}\left(\log T\right) & \text{if } \alpha = 1, \\ \mathcal{O}(1) & \text{if } \alpha > 1, \end{cases} \tag{C.3}$$

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 B.3 and Theorem 3.1 with a very slight modification. Note that all the arguments in Lemma B.3 hold and the only difference is where we want to bound the probability $\mathbb{P}[I^h]$ in Equation (B.8). 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

$$\mathbb{E}\left[\mathbb{I}(B_{il}, E_l, E_i) X_t^\top(\beta_l - \beta_i) \mid A_l\right] \leq \sum_{h=0}^{h^{\max}} 2^{1+\alpha} C_0 \delta^{1+\alpha} x_{\max}^{1+\alpha}(h+1)^{1+\alpha}\, \mathbb{P}\left[B_{il}, E_l, E_i \mid A_l, I^h\right],$$

which turns the regret bound in Equation (B.9) into

$$\begin{aligned} r_t(\pi^{\text{gb}}) \leq (K-1)&\Big[C_0 2^{1+\alpha}\delta^{1+\alpha}x_{\max}^{1+\alpha}\Big(\sum_{h=0}^{h_0}(h+1)^{1+\alpha} + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^{1+\alpha}\exp(-h^2)\Big)\Big] \\ &+ 4(K-1)b_{\max}x_{\max}\max_i \mathbb{P}({E_i}^{\mathsf{c}}), \end{aligned} \tag{C.4}$$

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 to t comes only 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 (C.4) 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 $\delta = 1/\sqrt{tC_2}$ in the Equation (C.4) and putting together all the constants we reach to

$$r_t(\pi^{\text{gb}}) = (K-1)g_1(d, \alpha, C_0, x_{\max}, \sigma, \lambda)t^{-(1+\alpha)/2} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[E_i{}^{\mathsf{c}}]\right)$$

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 B.1 for $i = 1, 2$, we can bound the probabilities $\mathbb{P}[E_i{}^{\mathsf{c}}]$ by $d\exp(-C_1 t)$ and therefore

$$\begin{aligned}
R_T(\pi^{\text{gb}}) &= \sum_{t=1}^{T} g_1(d, \alpha, C_0, x_{\max}, \sigma, \lambda)t^{-(1+\alpha)/2} + 4b_{\max}x_{\max}d\exp(-C_1 t) \\
&\le g_1(d, \alpha, C_0, x_{\max}, \sigma, \lambda)\left[1 + \left(\int_{t=1}^{T} t^{-(1+\alpha)/2}\mathrm{d}t\right)\right] + 4db_{\max}x_{\max}\int_0^\infty \exp(-C_1 t)\mathrm{d}t \\
&= g_1(d, \alpha, C_0, x_{\max}, \sigma, \lambda)\left[1 + \left(\int_{t=1}^{T} t^{-(1+\alpha)/2}\mathrm{d}t\right)\right] + \frac{4b_{\max}x_{\max}d}{C_1}.
\end{aligned}$$

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

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

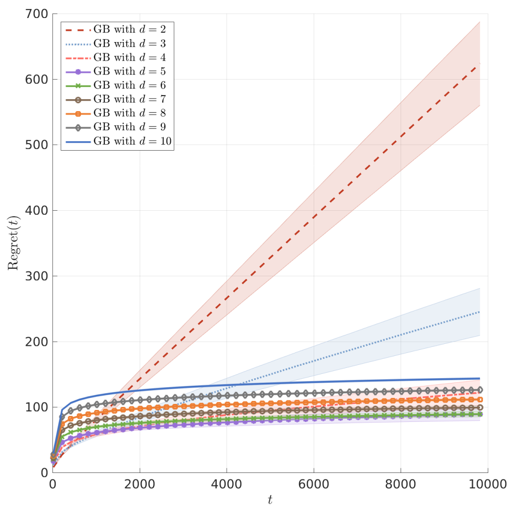
which yields the desired result. □ □

D Additional Simulations

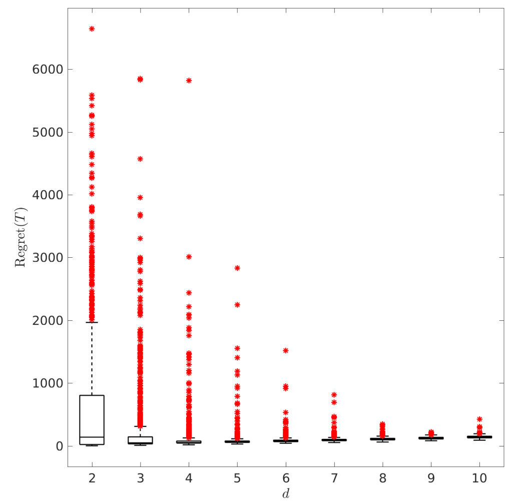
D.1 More than Two Arms ($K > 2$)

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 Section 5 for $K = 5$ and $d \in \{3, 7\}$ while keeping the other parameters as in Section 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



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



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

Figure 2: These figures show a sharp change in the performance of Greedy Bandit for $K = 5$ arms as d increases.

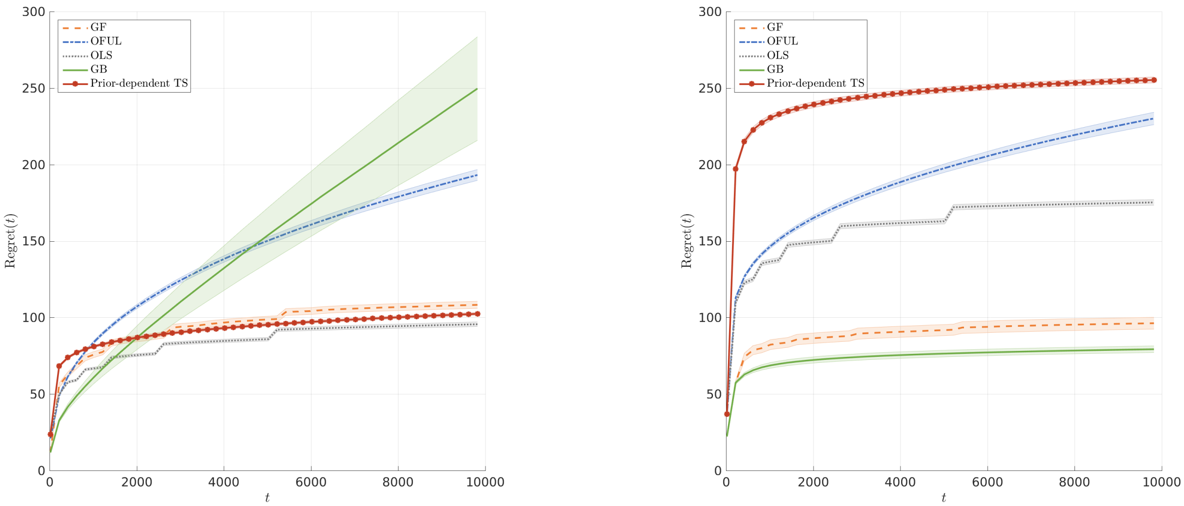
the correct prior regime for OFUL and TS. The results, as shown in Figure 3, demonstrate that Greedy-First nearly ties with Greedy Bandit as the winner when $d = 7$. but for $d = 3$ that Greedy Bandit performs poorly, Greedy-First performs very close to the best algorithms.

D.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 1 for the case that covariate diversity does not hold. As it can be observed from Figure 4, 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.

D.3 Real Data Covariates

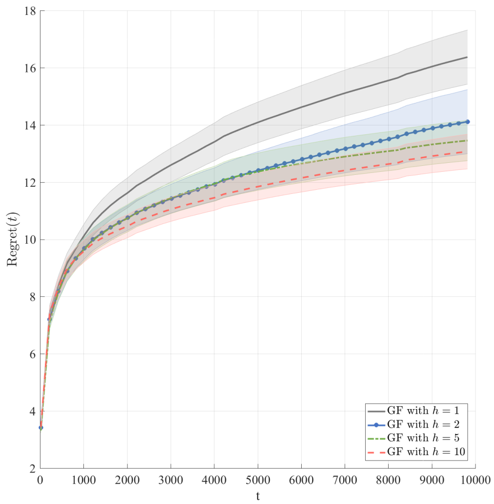
To test the robustness of our results, we repeat simulations of Section 5 with a single change that the covariates are not generated by us but instead, they are sampled from a large real data set. In particular, the covariates are obtained from a large data set of hotel listings on Expedia that is publicly available at https://www.kaggle.com/c/expedia-personalized-sort. The covariates are Star Rating, Average Reviews, Brand, Position, Weekend Indicator, Location Popularity, and an



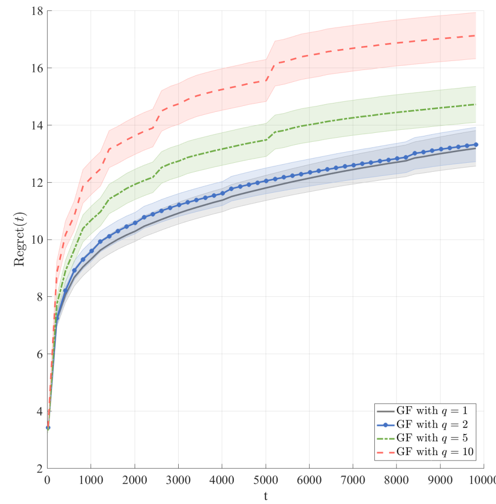
(a) $K = 5, d = 3$ (b) $K = 5, d = 7$

Figure 3: Simulations for $K > 2$ arms.

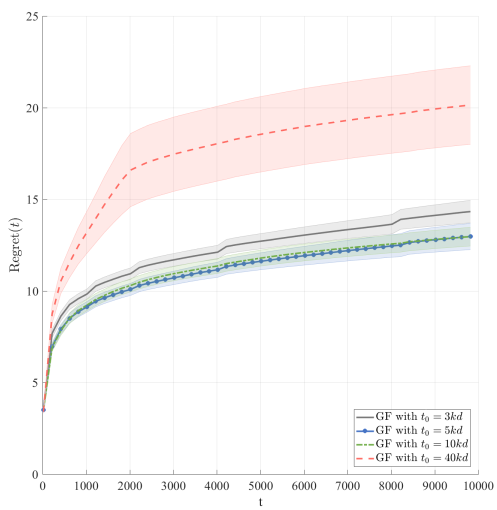
indicator variable for Summer for 267832 user bookings. For any one of 1000 simulations of the bandit algorithms, we sample $T = 10000$ covariate vectors from this dataset at random. We use the correct prior regime of Section 5 and run the simulations for $K = 2$ and $K = 3$. Figure 5 shows the results which has the same qualitative behavior as in Figure 1(a) that the covariate diversity holds.



(a) Sensitivity with respect to h.

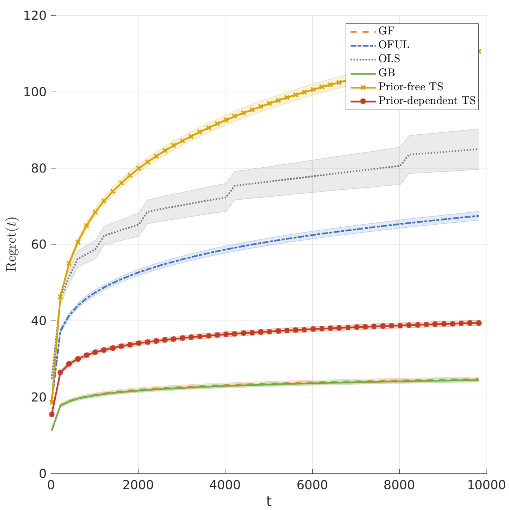


(b) Sensitivity with respect to q.

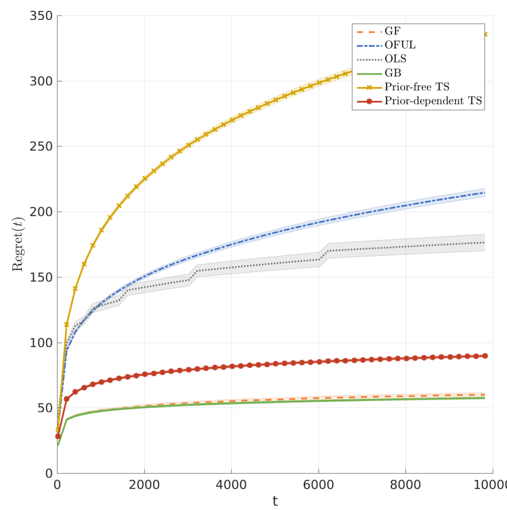


(c) Sensitivity with respect to t_0.

Figure 4: Sensitivity analysis for the expected regret of Greedy-First algorithm with respect to the input parameters h, q, and t_0.



(a) $K = 2$



(b) $K = 3$

Figure 5: Performance of various algorithms on the Expedia data.