Online Decision-Making with High-Dimensional Covariates

Hamsa Bastani
Wharton School, Operations Information and Decisions, hamsab@wharton.upenn.edu

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

Big data has enabled decision-makers to tailor decisions at the individual-level in a variety of domains such as personalized medicine and online advertising. This involves learning a model of decision rewards conditional on individual-specific covariates. In many practical settings, these covariates are *high-dimensional*; however, typically only a small subset of the observed features are predictive of a decision's success. We formulate this problem as a K-armed contextual bandit with high-dimensional covariates, and present a new efficient bandit algorithm based on the LASSO estimator. We prove that our algorithm's cumulative expected regret scales at most poly-logarithmically in the covariate dimension d; to the best of our knowledge, this is the first such bound for a contextual bandit. The key step in our analysis is proving a new tail inequality that guarantees the convergence of the LASSO estimator despite the non-i.i.d. data induced by the bandit policy. Furthermore, we illustrate the practical relevance of our algorithm by evaluating it on a simplified version of a medication dosing problem. A patient's optimal medication dosage depends on the patient's genetic profile and medical records; incorrect initial dosage may result in adverse consequences such as stroke or bleeding. We show that our algorithm outperforms existing bandit methods as well as physicians to correctly dose a majority of patients.

Key words: contextual bandits, adaptive treatment allocation, online learning, high-dimensional statistics, LASSO, personalized decision-making

1. Introduction

The growing availability of user-specific data provides a unique opportunity for decision-makers to *personalize* service decisions for individuals. In healthcare, doctors can personalize treatment choices based on patient biomarkers and clinical history. For example, the BATTLE trial demonstrated that the effectiveness of different chemotherapeutic agents on a cancer patient depends on the molecular biomarkers found in the patient's tumor biopsy; thus, personalizing the chemotherapy regimen led to increased treatment success rates (Kim et al. 2011). Similarly, in marketing, companies may achieve greater conversion rates by targeting advertisements or promotions based on user demographics and search keywords. Personalization is typically achieved by (i) learning a model that predicts a user's outcome for each available decision as a function of the user's observed covariates, and (ii) using this model to inform the chosen decision for subsequent new users (see, e.g., He et al. 2012, Ban and Rudin 2014, Bertsimas and Kallus 2014, Chen et al. 2015).

However, the increased variety of potentially relevant user data poses *greater* challenges for learning such predictive models because user covariates may be *high-dimensional*. For instance, medical decision-making may involve extracting patient covariates from electronic health records (containing information on lab tests, diagnoses, procedures, and medications) or genetic or molecular biomarker profiles. The resulting number of covariates in medical decision-making problems can be as many as a few thousand (in Bayati et al. 2014) or tens of thousands (in Razavian et al. 2015). Similarly, user covariates in web marketing are often high-dimensional since they include relevant but fine-grained data on past clicks and purchases (Naik et al. 2008). Learning accurate predictive models from high-dimensional data statistically requires many user samples. These samples are often obtained through randomized trials on initial users, but this may be prohibitively costly in the high-dimensional setting.

Predictive algorithms such as the LASSO (Chen et al. 1998, Tibshirani 1996) help alleviate this issue by producing good estimates using far fewer user samples than traditional statistical models (Candes and Tao 2007, Bickel et al. 2009, Bühlmann and Van De Geer 2011). In particular, the LASSO identifies a *sparse* subset of predictive covariates, which is an effective approach for treatment effect estimation in practice (Belloni et al. 2014, Athey et al. 2016). For example, the BATTLE cancer trial found that only a few of many available patient biomarkers were predictive of the success of any given treatment (Kim et al. 2011). Similarly, variable selection is often used to predict Internet users' click-through rates in online advertising (see e.g., Yan et al. 2014).

However, we must be careful not to sacrifice asymptotic performance when using such techniques. They create substantial bias in our estimates to increase predictive accuracy for small sample sizes. Thus, it is valuable to incorporate new observations and carefully tune the bias-variance tradeoff over time to ensure good performance for both initial users (data-poor regime) and later users (data-rich regime). This can be done *online*: after making a decision, we learn from the resulting reward, e.g., how well a treatment performed on a patient, or the profit from an advertisement. This process suffers from *bandit feedback*, i.e., we only obtain feedback for the chosen decision and we do not observe (counterfactual) rewards for alternate actions. For example, we may incorrectly conclude that a particular action is low-reward early on and discard it based on (uncertain) estimates; then, we may never identify our mistake and perform poorly in the long-term since we will not observe the counterfactual reward for this action without choosing it. Therefore, while we seek to leverage our current estimates to optimize decisions (*exploitation*), we must also occasionally experiment with each available action to improve our estimates (*exploration*).

This exploration-exploitation tradeoff has been studied in the framework of contextual bandits (Auer 2003, Langford and Zhang 2008). Although many algorithms have been proposed and analyzed in the literature, they typically optimize asymptotic performance (when the number of users T grows large) and may not perform well in the data-poor regime. In particular, the performance of all existing algorithms scales polynomially in the number of covariates d, and provide no theoretical guarantees when the number of users T is of order d (see, e.g., Goldenshluger and Zeevi 2013), even when the underlying model is known to be sparse (Abbasi-Yadkori et al. 2012). Thus, such algorithms may essentially randomize on the initial $\mathcal{O}(d)$ individuals, which as discussed earlier, may be prohibitively costly in high-dimensional settings.

In this paper, we propose a new algorithm (the LASSO Bandit) that addresses these shortcomings. In particular, we adapt the LASSO estimator to the bandit setting and tune the resulting bias-variance tradeoff over time to gracefully transition from the data-poor to data-rich regime. We prove theoretical guarantees that our algorithm achieves good performance as soon as the number of users T is poly-logarithmic in d, which is an *exponential* improvement over existing theory. Simulations confirm our theoretical results. Finally, we empirically demonstrate the potential benefit of our algorithm in a medical decision-making context by evaluating it on the clinical task of warfarin dosing with real patient data. In general, evaluating a bandit algorithm retrospectively on data is challenging because we require access to counterfactuals; we choose warfarin dosing as our case study since this unique dataset gives us access to such counterfactuals under some simplifying assumptions. We find that our algorithm significantly outperforms other bandit methods, and outperforms the benchmark policy used in practice by physicians after observing 200 patients. In particular, the LASSO Bandit successfully leverages limited available data to make better decisions for initial patients, while continuing to perform well in the data-rich regime.

1.1. Main Contributions

We introduce the LASSO Bandit, a new statistical decision-making algorithm that efficiently leverages high-dimensional user covariates in the bandit setting by learning LASSO estimates of decision rewards. Below we highlight our contributions in three categories.

Algorithm. Our algorithm builds on an existing algorithm in the low-dimensional bandit setting by Goldenshluger and Zeevi (2013) that uses ordinary least squares estimation. We use LASSO estimation in the high-dimensional setting, which introduces the key additional step of selecting a *regularization path*. We specify such a path to optimally control the convergence of our LASSO estimators by trading off bias and variance over time.

Theory. We measure performance using the standard notion of *expected cumulative regret*, which is the total expected deficit in reward achieved by our algorithm compared to an oracle that knows all the problem parameters. Our main result establishes that the LASSO Bandit asymptotically achieves expected cumulative regret that scales poly-logarithmically with the dimension of the covariates. The technical challenge is that the bandit policy induces non-i.i.d. samples from each arm during the exploitation phase. In particular, even though the sequence of all covariates are i.i.d. samples from a fixed distribution, the subset of covariates for which the outcome of a fixed arm is observed may not be i.i.d. In low-dimensional settings, this is typically addressed using martingale matrix Chernoff inequalities (Tropp 2015). We prove analogous results in the high-dimensional setting for the convergence of the LASSO estimator using matrix perturbation theory and martingale concentration results. In particular, we prove a new tail inequality for the LASSO (that may be of independent interest) which holds with high probability even when an unknown portion of the samples are generated by a non-i.i.d. process.

We further derive an optimal specification for the LASSO regularization parameters, and prove that the resulting cumulative regret of the LASSO Bandit over T users is at most $\mathcal{O}\big(s_0^2\,[\log T + \log d]^2\big)$, where $s_0 \ll d$ is the number of relevant covariates. To the best of our knowledge, the LASSO Bandit achieves the first regret bound that scales poly-logarithmically in both d and T, making it suitable for leveraging high-dimensional data without experimenting on a large number of users. As a secondary contribution, our techniques can also be used to improve existing regret bounds in the low-dimensional setting by a factor of d for the OLS Bandit (a variant of the algorithm by Goldenshluger and Zeevi (2013)) under the same problem setting and weaker assumptions.

Empirics. We compare the performance of the LASSO Bandit against existing algorithms in the bandit literature. Simulations on synthetic data demonstrate that the LASSO Bandit significantly outperforms these alternatives in cumulative regret. Surprisingly, we find that our algorithm can significantly improve upon these baselines even in "low-dimensional" settings.

More importantly, we evaluate the potential value of our algorithm in a medical decision-making context using a real patient dataset on warfarin (a widely-prescribed anticoagulant). Here, we apply the LASSO Bandit to learn an optimal dosing strategy using patients' clinical and genetic factors. We show that our algorithm significantly outperforms existing bandit algorithms to correctly dose a majority of patients. Furthermore, our algorithm outperforms the current benchmark policy used in practice by physicians after observing 200 patients. Finally, we evaluate the trade-off between increased patient risk and improved dosing, and find that our algorithm increases the risk of incorrect dosing for a small number of patients in return for a large improvement in average dosing accuracy. We note that we do not take advantage of certain information structures that are specific to the warfarin dosing problem (see §5 for details); exploiting this structure could potentially result in even better algorithms tailored specifically for warfarin dosing, but developing such an algorithm is beyond the scope of our paper.

1.2. Related Literature

As discussed earlier, there is a significant OR/MS literature on learning predictive models from historical data, and using such models to inform context-specific decision-making (e.g., Ban and Rudin 2014, Bertsimas and Kallus 2014). In contrast, our work addresses the problem of learning these predictive models online under bandit feedback (i.e., we only observe feedback for the chosen decision, as is often the case in practice), which results in an exploration-exploitation trade-off.

There is a rich literature on the exploration-exploitation tradeoff in the contextual bandit framework (also known as contextual bandits or linear bandits with changing action space) from OR/MS, computer science, and statistics. One approach is to make no parametric assumptions on arm rewards. For example, Slivkins (2014), Perchet and Rigollet (2013) and Rigollet and Zeevi (2010) analyze settings where the arm rewards are given by any smooth, non-parametric function of the observed covariates. However, these algorithms perform very poorly in high dimension as the cumulative regret depends exponentially on the covariate dimension d.

Thus, much of the bandit literature (including the present paper) has focused on the case where the arm rewards are linear functions of the covariates; this setting was first introduced by Auer (2003) and was subsequently improved by UCB-type algorithms by Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), Chu et al. (2011), Abbasi-Yadkori et al. (2011) and Deshpande and Montanari (2012). (Note that some of these papers study the linear bandit, which is different from a contextual bandit; however, the theoretical guarantees of a linear bandit can be mapped to theoretical guarantees for a contextual bandit *if* the feasible action set for the linear bandit is allowed to change exogenously over time (Abbasi-Yadkori 2012).) These algorithms use the idea of optimism-in-the-face-of-uncertainty (OFU), which elegantly solves the exploration-exploitation tradeoff by maintaining confidence sets for arm parameter estimates and choosing arms optimistically from within these confidence sets. Follow-up work demonstrated that similar guarantees can be achieved using a posterior sampling algorithm (Agrawal and Goyal 2013, Russo and Van Roy 2014b). We also note that Carpentier and Munos (2012) tackle a linear bandit in the high-dimensional sparse setting but they use a non-standard definition of regret and also do not consider the relevant case where the action set changes over time.

However, this literature typically does not make any assumptions on how the user covariates X_t are generated. In particular, they allow for arbitrarily constructed covariate sequences that may be generated by an adversary to make learning difficult; Chapter 3 of Bubeck and Cesa-Bianchi (2012) provides a detailed survey of "adversarial bandits". For example, if X_t is equal to a fixed vector X that does not change over time, it is impossible to learn more than one parameter per arm. This may explain why the current-best cumulative regret bounds are given by: $\mathcal{O}(d\sqrt{T})$

in the low-dimensional setting (Dani et al. 2008, Abbasi-Yadkori et al. 2011) and $\mathcal{O}(\sqrt{ds_0T})$ in the high-dimensional sparse setting (Abbasi-Yadkori et al. 2012). Note that such algorithms still achieve regret that is polynomial in d and T, implying slow rates of convergence. In particular, when $T=\mathcal{O}(d)$ (the regime of of interest here), these regret bounds are no longer sublinear in T.

Remark 1. Several of the above-mentioned papers also have "problem-dependent" bounds that scale as $\mathcal{O}(\log T)$ for the linear bandit (see, e.g., Abbasi-Yadkori et al. 2011). These bounds only apply when there is a fixed constant gap between the mean rewards of any pair of arms; they do *not* apply to a contextual bandit since there is no such constant gap. In our setting, the mean rewards of arm i and j can be arbitrarily close as a function of the observed covariates X_t at time t. We remark further on this point in §2.1.

Yet, assuming covariate sequences can be selected completely arbitrarily constitutes a pessimistic environment that is unlikely to occur in practical settings. For example, in healthcare, the treatment choices made for one patient do not directly affect the health status of the next patient, suggesting that covariates are roughly i.i.d. Thus, we focus on the case where covariates are generated i.i.d. from an unknown fixed distribution, where we can achieve exponentially better regret bounds. This insight was first noted by Goldenshluger and Zeevi (2013), who presented a novel algorithm that carefully trades off between a biased and an unbiased arm parameter estimate; as a result, they prove a corresponding upper bound of $\mathcal{O}(d^3\log T)$ on cumulative regret, which significantly improves the $\mathcal{O}(d\sqrt{T})$ bound for arbitrary covariate sequences as T grows large. We adapt this idea to the high-dimensional setting using LASSO estimators. However, we require a much tighter regret analysis as well as new convergence results on LASSO estimators, which we use to prove a regret bound of $\mathcal{O}(s_0^2[\log T+\log d]^2)$. Note that we relax the polynomial dependence on d to a poly-logarithmic factor by leveraging sparsity. As a consequence of our new proof technique, we also improve the regret bound in the low-dimensional setting from $\mathcal{O}(d^3\log T)$ (Goldenshluger and Zeevi 2013) to $\mathcal{O}\left(d^2\log^{\frac{3}{2}} d\cdot\log T\right)$. These results hold while allowing for some arms to be uniformly sub-optimal; in contrast, the formulation in Goldenshluger and Zeevi (2013) requires the assumption that every arm is optimal for some subset of users.

Remark 2. It is worth comparing both bounds in the low-dimensional setting where all covariates are relevant, i.e., $s_0=d$. In this setting, we show that the OLS Bandit achieves $\mathcal{O}\left(d^2\log^{\frac{3}{2}} d\cdot\log T\right)$ regret, while the LASSO Bandit achieves a slightly worse upper bound of $\mathcal{O}(d^2[\log T+\log d]^2)$ regret. This difference arises from the weaker convergence results established for the LASSO as opposed to the least squares estimator (see §4). However, when $s_0\ll d$ (as is often the case in practical high-dimensional settings), the LASSO Bandit can achieve exponentially better regret (in the ambient dimension d) by leveraging sparse structure.

Past theoretical analysis of high-dimensional bandits has not used LASSO techniques. In particular, Carpentier and Munos (2012) use random projections, Deshpande and Montanari (2012) use ℓ_2-regularized regression, and Abbasi-Yadkori et al. (2012) use SeqSEW. Our proofs rely on existing literature on oracle inequalities that guarantee convergence of LASSO estimators (Candes and Tao 2007, Bickel et al. 2009, Negahban et al. 2012, Bühlmann and Van De Geer 2011); a technical contribution of our work is proving a new LASSO tail inequality that can be used on non-i.i.d. data induced by the bandit policy, which may be of independent interest.

There has also been interest in posterior sampling and information-directed sampling methods (Russo and Van Roy 2014a,b), which show evidence of improved empirical performance on standard bandit problems. These algorithms do not yet have theoretical guarantees for our setting that are competitive with existing bounds described above. Developing algorithms of this flavor and corresponding regret bounds for our setting may be a promising avenue for future work.

Finally, our paper is also related to recent papers in the operations management literature at the intersection of machine learning and multi-armed bandits. Kallus and Udell (2016) use low-rank matrix completion for dynamic assortment optimization with a large number of customers, and Elmachtoub et al. (2017) introduce a novel bootstrap-inspired method for performing Thompson sampling using decision trees. In contrast, our work focuses on developing provable guarantees for bandits with covariates under the LASSO estimator; to that end, we introduce new theoretical results for the LASSO with adapted sequences of (possibly non-i.i.d) observations.

The remainder of the paper is organized as follows. We describe the problem formulation and assumptions in §2. We present the LASSO Bandit algorithm and our main result on the algorithm's performance in §3; the key steps of the proof are outlined in §4. Finally, empirical results on simulated data as well as our evaluation on real patient data for the task of warfarin dosing are presented in §5. All proofs, robustness checks, and our secondary result in the low-dimensional setting are relegated to the appendices.

2. Problem Formulation

We now describe the standard problem formulation for a bandit with covariates and linear arm rewards (as introduced by Auer (2003) and others). We start by introducing some notation that will be used throughout the paper.

Notation. For any integer n, we will let $[n]$ denote the set $\{1,...,n\}$. For any index set $I \subset [d]$, and a vector $\beta \in \mathbb{R}^d$, let $\beta_I \in \mathbb{R}^d$ be the vector obtained by setting the elements of β that are not in I to zero. For a vector $v \in \mathbb{R}^m$, let the support of v (denoted $\text{supp}(v)$) to be the set of indices corresponding to nonzero entries of v. For any vector X or matrix $\mathbf{X}$, the infinity norm (i.e., $\|\cdot\|_\infty$)

is the maximum absolute value of its entries. We also use $\mathbb{R}^+$ and $\mathbb{Z}^+$ to refer to positive reals and integers respectively, and use $\mathbb{R}^{d\times d}_{\succeq 0}$ for the set of d by d positive semidefinite matrices.

Let T be the number of (unknown) time steps; at each time step, a new user arrives and we observe her individual covariates X_t. The observed sequence of covariates $\{X_t\}_{t\geq 1}$ are random vectors that are drawn i.i.d. from a distribution $\mathcal{P}_X$ over a deterministic set $\mathcal{X}\subset\mathbb{R}^d$ (see Remark 3 for a precise definition). The decision-maker has access to K arms (decisions) and each arm yields an uncertain user-specific reward (e.g., patient outcome or profit from a user conversion). Each arm i has an unknown parameter $\beta_i\in\mathbb{R}^d$. At time t, if we pull arm $i\in[K]$, we yield reward

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

where the $\varepsilon_{i,t}$ are independent σ-subgaussian random variables (see Definition 1 below) that are also independent of the sequence $\{X_{t'}\}_{t'\geq 1}$. In §EC.6.3, we numerically show how our approach can be used even when the reward is a nonlinear function of the covariates by using basis expansion methods from statistical learning to approximate nonlinear functions.

Definition 1. A real-valued random variable z is σ-subgaussian if $\mathbb{E}[e^{tz}]\leq e^{\sigma^2t^2/2}$ for every $t\in\mathbb{R}$.

This definition implies $\mathbb{E}[z]=0$ and $\mathrm{Var}[z]\leq\sigma^2$. Many classical distributions are subgaussian; typical examples include any bounded, centered distribution, or the normal distribution. Note that the errors need not be identically distributed.

Remark 3. The reward function contains two stochastic sources: the covariate vector X_t and the noise. Therefore, we define the precise notion of the probability space. Each X_t is a $\mathcal{H}$-measurable vector-valued function on probability space $(\Omega_X,\mathcal{H}_X,\Pr_X)$. We also refer to the distribution that X_t induces on $\mathbb{R}^d$ by $\mathcal{P}_X$, i.e., for any Borel set A of $\mathbb{R}^d$ we have $\Pr_X(X_t\in A)=\mathcal{P}_X(A)$. Similarly, each noise $\varepsilon_{i,t}$ is a real-valued random variable with probability space $(\Omega_\varepsilon,\mathcal{H}_\varepsilon,\Pr_\varepsilon)$. Throughout the paper all probabilities and expectations are with respect to the product measure $\Pr_X\times\Pr_\varepsilon$. To simplify notation, we will use $\mathbb{E}$ and $\Pr$ to refer to "expectation" and "probability" with respect to this product measure, unless the probability measure is specified as a subindex.

Our goal is to design a sequential decision-making policy π that learns the arm parameters $\{\beta_i\}$ over time in order to maximize expected reward for each individual. Let $\pi_t\in[K]$ denote the arm chosen by policy π at time $t\in[T]$. We compare ourselves to an *oracle* policy π^* that already knows the $\{\beta_i\}$ (but not the noise ε) and thus always chooses the best expected arm $\pi_t^*=\max_j(X_t^\top\beta_j)$. Thus, if we choose arm $\pi_t=i$ at time t, we incur expected regret

$$r_t\equiv\mathbb{E}\left[\max_j(X_t^\top\beta_j)-X_t^\top\beta_i\right]\,,$$

which is simply the difference in expected reward between π_t^* and π_t. We seek a policy π that minimizes the cumulative expected regret $R_T \equiv \sum_{t=1}^{T} r_t$. In particular, if R_T is small for policy π, then the performance of π is similar to that of the oracle.

We additionally introduce the *sparsity parameter* $s_0 \in [d]$, which is the smallest integer such that for all $i \in [K]$, we have $\|\beta_i\|_0 \leq s_0$. (Note that this is trivially satisfied for $s_0 = d$.) Our algorithm has strong performance guarantees when $s_0 \ll d$, i.e., when the arm rewards are determined by only a small subset (of size s_0) of the d observed user-specific covariates in X.

2.1. Assumptions

We now describe the assumptions we require on the problem parameters for our regret analysis. These assumptions are adapted from the bandit literature and will be attributed in the text below. For simplicity, we introduce a specific example and show how each assumption translates to the example. Later, we describe more generic examples that are encompassed by our formulation.

Simple Example: Let the induced probability distribution of covariates, $\mathcal{P}_X$, be the uniform distribution over the d-dimensional unit cube $[0,1]^d$. Consider three arms whose corresponding arm parameters are given by $\beta_1 = (1, 0, ..., 0)$, $\beta_2 = (0, 1, 0, ..., 0)$, and $\beta_3 = (1/4, 1/4, 0, ..., 0)$.

Assumption 1 **(Parameter set)**. *There exist positive constants $x_{\max}$ and b such that $\|x\|_\infty \leq x_{\max}$ for all $x \in \mathcal{X}$ and $\|\beta_i\|_1 \leq b$ for all $i \in [K]$. The former implies that any realization of the random variable X_t satisfies $\|X_t\|_\infty \leq x_{\max}$ for all t.*

Our first assumption is that the observed covariate vector X_t as well as the arm parameters β_i are bounded. This is a standard assumption made in the bandit literature (see, e.g., Rusmevichientong and Tsitsiklis 2010), ensuring that the maximum regret at any time step is bounded, i.e., all realizations of X_t satisfy $|X_t^\top \beta_i| \leq b\, x_{\max}$ by Cauchy-Schwarz for dual norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$ on $\mathbb{R}^d$. This is likely satisfied since user covariates and outcomes are bounded in practice. Our example clearly satisfies this assumption with $x_{\max} = 1$ and $b = 1$.

Assumption 2 **(Margin condition)**. *There exists a constant $C_0 \in \mathbb{R}^+$ such that for all i and j in $[K]$ where $i \neq j$, $\Pr\left[0 < |X^\top(\beta_i - \beta_j)| \leq \kappa\right] \leq C_0 \kappa$ for all $\kappa \in \mathbb{R}^+$.*

Our second assumption is a margin condition that ensures that the density of the covariate distribution $\mathcal{P}_X$ should be bounded near a decision boundary, i.e., the intersection of the hyperplane given by $\left\{x^\top \beta_i = x^\top \beta_j\right\}$ and $\mathcal{X}$ for any $i \neq j \in [K]$. (Note that the distribution of $\mathcal{P}_X$ can be such that point masses on the decision boundary are allowed.) This assumption was introduced in the classification literature by Tsybakov (2004) and highlighted in a bandit setting by Goldenshluger and Zeevi (2013). Intuitively, even small errors in our parameter estimates can cause us to choose the wrong action (between arms i and j) for a realization of the covariate vector X_t close to the decision boundary since the rewards for both arms are nearly equal. Thus, we can perform poorly if a disproportionate fraction of observed covariate vectors are drawn near these hyperplanes. Since the uniform distribution has a bounded density everywhere in the simple example above, this assumption is satisfied; a simple geometric argument yields $C_0 = 2\sqrt{2}$.

ASSUMPTION 3 **(Arm optimality)**. *Let $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ be mutually exclusive sets that include all K arms. Then there exists some $h > 0$ such that: (a) sub-optimal arms $i \in \mathcal{K}_{sub}$ satisfy $x^\top \beta_i < \max_{j\neq i} x^\top \beta_j - h$ for every $x \in \mathcal{X}$; and (b) for a constant $p_* > 0$, each optimal arm $i \in \mathcal{K}_{opt}$ has a corresponding set*

$$U_i \equiv \left\{ x \in \mathcal{X} \;\middle|\; x^\top \beta_i > \max_{j\neq i} x^\top \beta_j + h \right\},$$

such that $\min_{i\in\mathcal{K}_{opt}} \Pr\left[X \in U_i\right] \geq p_*$.

Our third assumption is a less restrictive version of an assumption introduced in Goldenshluger and Zeevi (2013). In particular, we assume that our K arms can be split into two sets:

a. Sub-optimal arms $\mathcal{K}_{sub}$ that are *strictly* sub-optimal for all covariate vectors in $\mathcal{X}$, i.e., there exists a constant $h_{sub} > 0$ such that for each $i \in \mathcal{K}_{sub}$, $x^\top \beta_i < \max_{j\neq i} x^\top \beta_j - h_{sub}$ for every $x \in \mathcal{X}$.
b. A non-empty set of optimal arms $\mathcal{K}_{opt}$ that are *strictly* optimal with positive probability for some covariate vectors $x \in \mathcal{X}$, i.e., there exists a constant $h_{opt} > 0$ and some region $U_i \subset \mathcal{X}$ (with $\Pr[X \in U_i] = p_i > 0$) for each $i \in \mathcal{K}_{opt}$ such that $x^\top \beta_i > \max_{j\neq i} x^\top \beta_j + h_{opt}$ for all covariate vectors x in U_i.

In other words, we assume that every arm is either optimal (by a margin h_{opt}) for *some* users (Assumption 3b), or sub-optimal (by a margin h_{sub}) for *all* users (Assumption 3a). For simplicity, in Assumption 3, we define the *localization parameter* $h = \min\{h_{opt}, h_{sub}\}$ and $p_* = \min_{i\in\mathcal{K}_{opt}} p_i$. By construction, the regions U_i are separated from all decision boundaries (by at least h in reward space); thus, intuitively, small errors in our parameter estimates are unlikely to make us choose the wrong arm under the event $X \in U_i$ for some $i \in \mathcal{K}_{opt}$. Thus, we will play each optimal arm $i \in \mathcal{K}_{opt}$ at least $p_* T$ times in expectation with high probability (i.e., whenever the event $X \in U_i$ occurs). This ensures that we can quickly learn accurate parameter estimates for all optimal arms over time. We will discuss the choice of h later (see Remark 8 and Appendix EC.6.2).

In our simple example, one can easily verify that $\mathcal{K}_{opt} = \{\beta_1, \beta_2\}$ and $\mathcal{K}_{sub} = \{\beta_3\}$. We can choose any value $h \in (0, 1/2]$ with corresponding $p_* = (1 - h\sqrt{2})^2$ for this setting.

REMARK 4. We emphasize that Assumption 3 differs from the "gap" assumption made in problem-dependent bounds in the bandit literature (see, e.g., Abbasi-Yadkori et al. 2011), which assumes that there exists some gap $\Delta > 0$ between the rewards of the optimal arm i^* and the next best arm, i.e., $\Delta \le \min_{j,\, x\in\mathcal{X}} x^\top(\beta_{i^*} - \beta_j)$. In a general contextual bandit, no $\Delta > 0$ satisfies the gap assumption since the user covariate vector X can be drawn arbitrarily close to the decision boundary for some β_k (i.e., arbitrarily close to the set $\{x \in \mathcal{X} \mid x^\top\beta_{i^*} = x^\top\beta_k\}$). Rather, Assumption 3 posits that such a gap exists ($\Delta = h$) only with some probability $p_* > 0$. While the "gap" assumption does not hold for most covariate distributions (e.g., uniform), our assumption holds for a very wide class of continuous and discrete covariate distributions (as we will discuss below).

We state a definition for our final assumption, which is drawn from the high-dimensional statistics literature (Bühlmann and Van De Geer 2011).

Definition 2 (Compatibility condition). For any set of indices $I \subseteq [d]$ and a positive and deterministic constant ϕ, define the set of matrices

$$\mathcal{C}(I,\phi) \equiv \left\{ M \in \mathbb{R}^{d\times d}_{\succeq 0} \mid \forall v \in \mathbb{R}^d \text{ s.t. } \|v_{I^c}\|_1 \le 3\|v_I\|_1, \text{ we have } \|v_I\|_1^2 \le |I|\,(v^\top M v)/\phi^2 \right\} .$$

Assumption 4 **(Compatibility condition)**. *There exists a constant $\phi_0 > 0$ such that for each $i \in \mathcal{K}_{opt}$, $\Sigma_i \in \mathcal{C}(supp(\beta_i), \phi_0)$, where we define $\Sigma_i \equiv \mathbb{E}\,[XX^\top \mid X \in U_i]$.*

Our fourth and final assumption concerns the covariance matrix[1] of samples restricted to the regions U_i for each $i \in \mathcal{K}_{opt}$. In particular, we require that $\Sigma_i \equiv \mathbb{E}_{X\sim\mathcal{P}_X}\,[XX^\top \mid X \in U_i]$ belongs to the set $\mathcal{C}(\text{supp}(\beta_i), \phi_0)$ with some constant $\phi_0 > 0$ (Definition 2). This assumption is required for the identifiability of LASSO estimates trained on samples $X \in U_i$ (Candes and Tao 2007, Bickel et al. 2009, Negahban et al. 2012, Bühlmann and Van De Geer 2011). As we discussed earlier in Assumption 3, for each $i \in \mathcal{K}_{opt}$, we expect to play arm i at least $p_*T = \mathcal{O}(T)$ times based on samples $X \in U_i$. The compatibility condition ensures that a LASSO estimator trained on these samples will converge to the true parameter vector β_i with high probability as the number of samples grows to infinity. We will discuss the LASSO estimator and its convergence properties in detail in §3.1.

Note that a standard assumption in ordinary least squares (OLS) estimation is that the matrix Σ_i be *positive-definite*, i.e., $\lambda_{\min}(\Sigma_i) > 0$. It can be easily verified that if Σ_i is positive-definite, then it belongs to $\mathcal{C}(I, \sqrt{\lambda_{\min}(\Sigma_i)})$ for *any* set $I \subseteq [d]$. Thus, the compatibility condition is weaker than the requirement that Σ_i be positive-definite.

In our example, the events $X \in U_i$ (defined by any allowable choice of $h \in (0, 1/2]$) for each $i \in \mathcal{K}_{opt}$ have positive probability, and the matrices Σ_i are positive definite. Note that smaller choices of h (which can generally be chosen arbitrarily close to zero) result in larger sets U_i by definition, and therefore yield larger values of $\lambda_{\min}(\Sigma_i)$. For example, $h = 0.1$ corresponds to $\lambda_{\min}(\Sigma_i) \approx 0.01$. Thus, the covariance matrices Σ_i also satisfy the compatibility condition.

[1] Throughout the paper, the "covariance matrix" of X refers to the matrix $\mathbb{E}[XX^\top]$, even when $\mathbb{E}[X] \neq 0$.

REMARK 5. Throughout the proof, we will study events of type $\{M \notin \mathcal{C}(\text{supp}(\beta), \phi)\}$ for appropriate β, ϕ, random (sample-covariance) matrices M, and find upper bounds for their probabilities. These events are clearly measurable since they can be written as intersections of countably many measurable sets. Specifically, for any vector $v \in \mathbb{R}^d$ that satisfies $\|v_{I^c}\|_1 \le 3\|v_I\|_1$, the function G_v that sends a random matrix M to $|I|v^\top Mv/\phi^2 - \|v_I\|_1^2$ is measurable; consequently $G_v^{-1}([0,\infty))$ is also measurable. Since $|I|v^\top Mv/\phi^2 - \|v_I\|_1^2$ and $\|v_{I^c}\|_1 - 3\|v_I\|_1$ are both continuous in v, and using the fact that any vector v can be approximated with arbitrary accuracy with a rational vector in $\mathbb{R}^d$, the event $\{M \notin \mathcal{C}(\text{supp}(\beta), \phi)\}$ can be written as a countable intersection of measurable sets of the form $G_u^{-1}([0,\infty))$ for all rational $u \in \mathbb{R}^d$ satisfying $\|u_{I^c}\|_1 \le 3\|u_I\|_1$.

Finally, we give a few more examples of settings that satisfy all four of our assumptions.

Discrete Covariates: In many applications, the covariate vector may have discrete rather than continuous coordinates. It is easy to verify that our assumptions are satisfied for any discrete distribution with finite support, as long as its support does not lie in a hyperplane. For instance, we can take the probability distribution $\mathcal{P}_X$ over covariate vectors to be any discrete distribution over the vertices of the d-dimensional unit cube $\{0,1\}^d$. Note that Assumption 2 is still satisfied because all the vertices lie on the decision boundary (where $x^\top\beta_1 = x^\top\beta_2$) or are separated from this boundary by at least a constant distance. In fact, any discrete distribution over a finite number of points satisfies Assumption 2.

Generic Example: We now describe a generic example that satisfies all the above assumptions. Consider a bounded set $\mathcal{X}$ in $\mathbb{R}^d$ (Assumption 1). We call some coordinates "continuous" (all possible realizations $x \in \mathcal{X}$ take on continous values along these coordinates) and some "discrete" (all possible realizations $x \in \mathcal{X}$ take on a finite number of values along these coordinates). Assume further that Assumption 2 holds (e.g., if $\mathcal{P}_X$ is the product measure for a distribution of continuous and discrete coordinates, then the distribution of continuous coordinates has a bounded density and the probability of each value for the discrete coordinates is positive). These conditions are met by most distributions in practice. Next, we impose that no arm lies on the edge of the convex hull of all K arms (Assumption 3), i.e., every arm is either a vertex (optimal locally) or is contained inside the convex hull (sub-optimal everywhere). (Note that if the arm parameters are randomly selected from a uniform distribution on $\{\beta \in \mathbb{R}^d \mid \|\beta\|_\infty \le b\}$, this condition would hold with probability one.) Finally, we assume that with large enough probability, the covariates are linearly independent on each U_i so that the covariance matrix Σ_i is positive-definite (Assumption 4).

3. LASSO Bandit Algorithm

We begin by providing some brief intuition about the LASSO Bandit algorithm. Our policy produces LASSO estimates $\hat{\beta}_i$ for the parameter of each arm $i \in [K]$ based on past samples X_t where arm i was played. A typical approach for addressing the exploration-exploitation tradeoff is to *forced-sample* each arm at prescribed times; this produces i.i.d. data for unbiased estimation of the arm parameters, which can then be used to play myopically at all other times (i.e., choose the best arm based on current estimates). However, such an algorithm will provably incur at least $\Omega(\sqrt{T})$ regret since we will require many forced-samples for the estimates to converge fast enough.

Instead, our estimates may converge faster if we use *all* past samples (including non-i.i.d. samples from myopic play) from arm i to estimate β_i. However, since these samples are not i.i.d., standard convergence guarantees for LASSO estimators do not apply and we cannot ensure that the estimated parameters $\hat{\beta}_i$ converge to the true parameters β_i. We tackle this by adapting an idea from the low-dimensional bandit algorithm by Goldenshluger and Zeevi (2013), i.e., maintaining two sets of estimators for each arm: (i) *forced-sampling estimates* trained only on forced-samples, and (ii) *all-sample estimates* trained on all past samples when arm i was played. The former estimator is trained on i.i.d. samples (and therefore has convergence guarantees) while the latter estimator has the advantage of being trained on a much larger sample size (but naively, has no convergence guarantees). The LASSO Bandit uses the forced-sampling estimator in a pre-processing step to select a subset of arms; it then uses the all-sample estimator to choose the estimated best arm from this subset. We prove that using the forced-sampling estimator for the pre-processing step guarantees convergence of the all-sample estimator. A key novel ingredient of our algorithm is specifying the regularization paths to control the convergence of our LASSO estimators by carefully trading off bias and variance over time. Intuitively, we build low-dimensional linear models in the data-poor regime by limiting the number of allowed covariates; this allows us to make reasonably good decisions even with limited data. As we collect more data, we allow for increasingly complex models (consisting of more covariates), eventually recovering the standard OLS model.

Additional notation. Let the *design matrix* $\mathbf{X}$ be the $T \times d$ matrix whose rows are X_t. Similarly, let Y_i be the length T vector of observations $X_t^\top \beta_i + \varepsilon_{i,t}$. Since we only obtain feedback when arm i is played, entries of Y_i may be missing. We define the *all-sample set* $\mathcal{S}_i = \{t \mid \pi_t = i\} \subset [T]$ for arm i as the set of times when arm i was played. 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 $\mathcal{S}' \subset \mathcal{S}_i$, let $Y_i(\mathcal{S}')$ be the length $|\mathcal{S}'|$ vector of corresponding observed rewards $Y_i(t)$ for each $t \in \mathcal{S}'$. Since $\pi_t = i$ for each $t \in \mathcal{S}'$, $Y_i(\mathcal{S}')$ has no missing entries. Lastly, for any matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$, let $\hat{\Sigma}(\mathbf{Z}) = \mathbf{Z}^\top \mathbf{Z}/n$ be its sample covariance matrix. For any subset $\mathcal{A} \subset [n]$, we use the short notation $\hat{\Sigma}(\mathcal{A})$ to refer to $\hat{\Sigma}(\mathbf{Z}(\mathcal{A}))$.

3.1. LASSO Estimation

Consider a linear model $Y = \mathbf{X}\beta + \varepsilon$, with design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, response vector $Y \in \mathbb{R}^n$, and noise vector $\varepsilon \in \mathbb{R}^n$ whose entries are independent σ-subgaussian random variables. We define the LASSO estimator for estimating the parameter β (with $\|\beta\|_0 = s_0$):

Definition 3 (LASSO). Given a *regularization parameter* $\lambda \geq 0$, the LASSO estimator is

$$\hat{\beta}_{\mathbf{X},Y}(\lambda) \equiv \arg\min_{\beta'} \left\{ \frac{\|Y - \mathbf{X}\beta'\|_2^2}{n} + \lambda\|\beta'\|_1 \right\} . \tag{1}$$

The LASSO estimator satisfies the following *tail inequality*.

Proposition 1 **(LASSO Tail Inequality for Adapted Observations)**. *Let X_t denote the t^{th} row of $\mathbf{X}$ and $Y(t)$ denote the t^{th} entry of Y. The sequence $\{X_t : t = 1, ..., n\}$ forms an adapted sequence of observations, i.e., X_t may depend on past regressors and their resulting observations $\{X_{t'}, Y(t')\}_{t'=1}^{t-1}$. Also, assume that all realizations of random vectors X_t satisfy $\|X_t\|_\infty \leq x_{\max}$. Then for any $\phi > 0$ and $\chi > 0$, if $\lambda = \lambda(\chi, \phi) \equiv \chi\phi^2/(4s_0)$, we have*

$$\Pr\left[\|\hat{\beta}_{\mathbf{X},Y}(\lambda) - \beta\|_1 > \chi\right] \leq 2\exp[-C_1(\phi)n\chi^2 + \log d] + \Pr\left[\hat{\Sigma}(\mathbf{X}) \notin \mathcal{C}(supp(\beta), \phi)\right] ,$$

where $C_1(\phi) \equiv \phi^4/(512 s_0^2 \sigma^2 x_{\max}^2)$.

Remark 6. Note that the convergence rate χ and compatibility condition parameter ϕ determine the regularization parameter $\lambda(\chi, \phi)$; this will be reflected in the choice of regularization parameters in our algorithm, and is further discussed in Remark 7. Therefore, when we say "choosing regularization parameter λ", it is implicitly assumed that the parameter χ is selected appropriately.

Proposition 1 is a more general version of the standard LASSO oracle inequality (e.g., see Theorem 6.1 in Bühlmann and Van De Geer (2011)). Our version allows for adapted sequences of observations and errors that are σ-subgaussian conditional on all past observations. The result follows from modifying the proof of the standard LASSO oracle inequality[2] using martingale theory and is provided in Appendix EC.1.

LASSO for the bandit setting. Returning to our original problem, we consider the task of estimating the parameter β_i for each arm $i \in [K]$. Using any subset of past samples $\mathcal{S}' \subset \mathcal{S}_i$ where arm i was played and any choice of parameter λ, we can use the corresponding LASSO estimator $\hat{\beta}_{\mathbf{X}(\mathcal{S}'),Y(\mathcal{S}'),\lambda}$, which we denote by the simpler notation $\hat{\beta}(\mathcal{S}', \lambda)$, to estimate β_i. In order to prove regret bounds, we need to establish convergence guarantees for such estimates. From Proposition 1, in order to bound the error $\|\hat{\beta}(\mathcal{S}', \lambda) - \beta_i\|_1$ for each arm $i \in [K]$, we need to (i) ensure with high probability $\hat{\Sigma}(\mathcal{S}') \in \mathcal{C}(\text{supp}(\beta_i), \phi)$ for some constant ϕ and (ii) appropriately choose parameters λ over time to control the rate of convergence. Thus, the main challenge in the algorithm and analysis is constructing and maintaining sets $\mathcal{S}'$ such that with high probability $\hat{\Sigma}(\mathcal{S}') \in \mathcal{C}(\text{supp}(\beta_i), \phi)$ (although the rows of $\mathbf{X}(\mathcal{S}')$ are not i.i.d.) with sufficiently fast convergence rates.

[2] "Oracle inequality" refers to the fact that the LASSO achieves the same accuracy $\|\hat{\beta}_{\mathbf{X},Y}(\lambda) - \beta\|_1$ (up to logarithmic factors) compared to an oracle that knows supp(β) in advance (see Chapter 6 of Bühlmann and Van De Geer (2011)).

3.2. Description of Algorithm

The LASSO Bandit takes as input the *forced sampling parameter* $q \in \mathbb{Z}^+$ (which is used to construct the forced-sample sets), a *localization parameter* $h > 0$ (defined in Assumption 3)[3], as well as initial regularization parameters $\lambda_1, \lambda_{2,0}$. These parameters will be specified in Theorem 1.

Forced-Sample Sets: We prescribe a set of times when we forced-sample arm i (regardless of the observed covariates X_t):

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

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

All-Sample Sets: As before, let $\mathcal{S}_{i,t} = \left\{ t' \;\middle|\; \pi_{t'} = i \text{ and } 1 \le t' \le t \right\}$ denote the set of times we play arm i up to time t. Note that by definition $\mathcal{T}_{i,t} \subset \mathcal{S}_{i,t}$.

At any time t, the LASSO Bandit maintains two sets of parameter estimates for each β_i:

1. the forced-sample estimate $\hat{\beta}(\mathcal{T}_{i,t-1}, \lambda_1)$ based only on forced samples observed from arm i,
2. the all-sample estimate $\hat{\beta}(\mathcal{S}_{i,t-1}, \lambda_{2,t})$ based on all samples observed from arm i.

Execution: If the current time t is in $\mathcal{T}_i$ for some arm i, then arm i is played. Otherwise, two actions are possible. First, we use the forced-sample estimates to find the highest estimated reward achievable across all K arms. We then select the subset of arms $\hat{\mathcal{K}} \subset [K]$ whose estimated rewards are within $h/2$ of the maximum achievable. After this pre-processing step, we use the all-sample estimates to choose the arm with the highest estimated reward within the set $\hat{\mathcal{K}}$.

Algorithm LASSO Bandit

Input parameters: $q, h, \lambda_1, \lambda_{2,0}$
Initialize $\mathcal{T}_{i,0}$ and $\mathcal{S}_{i,0}$ by the empty set, and $\hat{\beta}(\mathcal{T}_{i,0}, \lambda_1)$ and $\hat{\beta}(\mathcal{S}_{i,0}, \lambda_{2,0})$ by 0 in $\mathbb{R}^d$ for all i in $[K]$
Use q to construct force-sample sets $\mathcal{T}_i$ using Eq. (2) for all i in $[K]$
for $t \in [T]$ **do**
  Observe user covariates $X_t \sim \mathcal{P}_X$
  if $t \in \mathcal{T}_i$ for any i **then**
    $\pi_t \leftarrow i$ (forced-sampling)
  else
    $\hat{\mathcal{K}} = \left\{ i \in [K] \;\middle|\; X_t^\top \hat{\beta}(\mathcal{T}_{i,t-1}, \lambda_1) \ge \max_{j \in [K]} X_t^\top \hat{\beta}(\mathcal{T}_{j,t-1}, \lambda_1) - h/2 \right\}$ is the set of near-optimal arms according to the forced-sample estimators
    $\pi_t \leftarrow \arg\max_{i \in \hat{\mathcal{K}}} X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}, \lambda_{2,t-1})$ is the best arm within $\hat{\mathcal{K}}$ according to the all-sample estimators
  end if
  Update all-sample sets $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$ and regularization $\lambda_{2,t} \leftarrow \lambda_{2,0} \sqrt{\frac{\log t + \log d}{t}}$
  Play arm π_t, observe $Y(t) = X_t^\top \beta_{\pi_t} + \varepsilon_{i,t}$
end for

---

[3] Note that if some $\bar{h}$ satisfies Assumption 3, then any $h \in (0, \bar{h}]$ also satisfies the assumption. Therefore, a conservatively small value can be chosen in practice, but this will be reflected in the constant in the regret bound.

REMARK 7. The choices of regularization parameters λ_1 and $\lambda_{2,t}$ are motivated by the following rough intuition. In Proposition 1, the regularization parameter impacts two quantities: the size of the error χ, and the probability of error $\exp[-C_1 n\chi^2 + \log d]$. (Note that it does not affect the term $\Pr\left[\hat{\Sigma}(\mathbf{X}) \notin \mathcal{C}(\text{supp}(\beta), \phi)\right]$.) For our regret analysis of the forced sample estimator, it suffices to keep the estimation error χ under $h/(4x_{\max})$ with as high a probability as possible; this can be achieved by taking λ_1 to be a constant. In contrast, for the all-sample estimator we wish to maintain both small estimation error χ, as well as a small probability of error; the above recipe for $\lambda_{2,t}$ trades these two terms nearly equally by guaranteeing the probability of error to be of order $1/\sqrt{t}$ and estimation error χ to be of order $\sqrt{\log(t)/t}$.

3.3. Main Result: Regret Analysis of LASSO Bandit

Our main result establishes that the LASSO Bandit asymptotically achieves expected cumulative regret that scales logarithmically with the dimension of covariates:

THEOREM 1. *When $q \geq 4\lceil q_0 \rceil$, $K \geq 2$, $d > 2$, $t \geq C_5$, and we take $\lambda_1 = (\phi_0^2 p_* h)/(64 s_0 x_{\max})$ and $\lambda_{2,0} = [\phi_0^2/(2s_0)]\sqrt{1/(p_* C_1)}$, we have the following (non-asymptotic) upper bound on the expected cumulative regret of the LASSO Bandit at time T by:*

$$
\begin{aligned}
R_T &\leq C_3 (\log T)^2 + [2Kbx_{\max}(6q+4) + C_3 \log d] \log T + (2bx_{\max}C_5 + 2Kbx_{\max} + C_4) \\
&= \mathcal{O}\left(s_0^2 [\log T + \log d]^2\right),
\end{aligned}
$$

where the constants $C_1(\phi_0)$, $C_2(\phi_0)$, $C_3(\phi_0, p_*)$, $C_4(\phi_0, p_*)$, and C_5 are given by

$$
C_1(\phi_0) \equiv \frac{\phi_0^4}{512 s_0^2 \sigma^2 x_{\max}^2}, \qquad C_2(\phi_0) \equiv \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 x_{\max}^2}\right), \qquad C_3(\phi_0, p_*) \equiv \frac{1024 K C_0 x_{\max}^2}{p_*^3 C_1},
$$

$$
C_4(\phi_0, p_*) \equiv \frac{8Kbx_{\max}}{1 - \exp\left[-\frac{p_*^2 C_2^2}{32}\right]}, \qquad C_5 \equiv \min\left\{t \in \mathbb{Z}^+ \mid t \geq 24Kq \log t + 4(Kq)^2\right\},
$$

and we take $q_0 \equiv \max\left\{\frac{20}{p_*}, \frac{4}{p_* C_2^2}, \frac{12 \log d}{p_* C_2^2}, \frac{1024 x_{\max}^2 \log d}{h^2 p_*^2 C_1}\right\} = \mathcal{O}\left(s_0^2 \log d\right)$.

Lower Bound. Goldenshluger and Zeevi (2013) prove an information-theoretic lower bound on the expected cumulative regret of $\mathcal{O}(\log T)$ for a (low-dimensional) contextual bandit. Since our formulation encompasses their setting, the same lower bound also applies to our setting. In particular, they consider (i) low-dimension $s_0 = d$, and (ii) two arms $K = 2$, (iii) both of which are assumed to be optimal arms $\mathcal{K}_{opt} = \{1, 2\}$. Thus, our upper bound of $\mathcal{O}\left([\log T]^2\right)$ for the expected cumulative regret may be up to a $\log T$ factor away from being optimal in T. It remains an open question whether tighter convergence guarantees can be developed for the LASSO estimator so that our analysis of the LASSO Bandit can be improved to meet the current lower bound.

In the interest of space, we do not provide a rigorous proof of the lower bound; however, we describe a road map of the proof. First, a lower bound of $\mathcal{O}(d \log T)$ in the low-dimensional setting follows by extending the proof of Goldenshluger and Zeevi (2013) using the multi-dimensional (rather than the scalar) Van Trees inequality. In high-dimensional settings, this naturally gives rise to a $\mathcal{O}(s_0 \log T)$ lower bound. To see this, consider the case where the support of the arm parameters is known; then, the decision-maker can discard irrelevant covariates, and the problem reduces to the low-dimensional setting with a new covariate dimension of s_0 (rather than d).

Remark 8. The localization parameter h (specified in Assumption 3) can be thought of as a tolerance parameter. In practice, decision-makers may choose h to be a threshold value such that arms are considered sub-optimal if they are not optimal for some users by at least h. For example, in healthcare, we may not wish to prescribe a treatment that does not improve patient outcomes above existing treatments by at least some threshold value. However, if no such value is known, one can consider supplying an initial value of h_0 and tuning this value down over time. In particular, our algorithm provides similar regret guarantees (with some minor updates to the proof) if we choose $h = h_0/\sqrt{\log t}$ for any initial choice $h_0 > 0$. Thus, once t is large enough such that $h < \bar{h}$ (where $\bar{h}$ is an unknown value that satisfies Assumption 3), we recover the desired statistical properties of our algorithm even if the initial parameter h_0 is incorrectly specified to be too large; however, the regret during the initial time periods may suffer as a result. We exclude the proof for brevity.

4. Key Steps of the Analysis of LASSO Bandit

In this section, we outline the proof strategy for Theorem 1. First, we need to obtain convergence guarantees for the forced-sample and all-sample estimators to compute the expected regret incurred while using such estimators. As discussed earlier, this is challenging because the all-sample estimator is trained on non-i.i.d. data, and thus standard LASSO convergence results do not apply. We prove a new general LASSO tail inequality that holds even when the rows of the design matrix are not i.i.d. (§4.1). We then use this result to obtain convergence guarantees for the forced-sample (§4.2) and all-sample estimators (§4.3) under a fixed regularization path. Finally, we sum up the expected regret from the errors in the estimators (§4.4).

4.1. A LASSO Tail Inequality for non-i.i.d. Data

We now prove a general result for the LASSO estimator. In particular, consider a linear model

$$W = \mathbf{Z}\beta + \varepsilon\,,$$

where $\mathbf{Z}_{n\times d}$ is the design matrix, $W_{n\times 1}$ is the response vector and $\varepsilon_{n\times 1}$ is the vector of errors whose entries are independent σ-subgaussians. The rows Z_t of $\mathbf{Z}$ are random vectors such that all their realizations are bounded, i.e., $\|Z_t\|_\infty \le x_{\max}$ for all $t \in [n]$. We also assume $\|\beta\|_0 = s_0$. Following the notation introduced earlier in §3.1, for any subset $\mathcal{A} \subset [n]$ we define the analogous quantities $\mathbf{Z}(\mathcal{A})$, $W(\mathcal{A})$, and $\hat{\Sigma}(\mathcal{A})$. Then, for any $\lambda \ge 0$ we have a LASSO estimator trained on samples in $\mathcal{A}$:

$$\hat{\beta}(\mathcal{A}, \lambda) \equiv \arg\min_{\beta'} \left\{ \frac{\|W(\mathcal{A}) - \mathbf{Z}(\mathcal{A})\beta'\|_2^2}{|\mathcal{A}|} + \lambda\|\beta'\|_1 \right\} .$$

Note that we have not made any distributional (or i.i.d.) assumptions on the samples in $\mathcal{A}$. We now consider that some unknown subset $\mathcal{A}' \subset \mathcal{A}$ comprises of i.i.d. samples from a distribution $\mathcal{P}_Z$, i.e., $\{Z_t \mid t \in \mathcal{A}'\} \sim \mathcal{P}_Z \times \cdots \times \mathcal{P}_Z$. Letting $\Sigma \equiv \mathbb{E}_{Z\sim\mathcal{P}_Z}[ZZ^\top]$, we further assume that $\Sigma \in \mathcal{C}(\mathrm{supp}(\beta), \phi_1)$ for a constant $\phi_1 \in \mathbb{R}^+$. We will show that if the number $|\mathcal{A}'|$ of i.i.d. samples is sufficiently large, then we can prove a convergence guarantee for the LASSO estimator $\hat{\beta}(\mathcal{A}, \lambda)$ trained on samples in $\mathcal{A}$, which includes non-i.i.d. samples. (Note that $\mathcal{A}'$ is unknown; if not, we can simply use the estimator $\hat{\beta}(\mathcal{A}', \lambda)$ trained only on the i.i.d. samples in $\mathcal{A}'$.) In particular, suppose that at least some constant fraction of the samples in $\mathcal{A}$ belong in $\mathcal{A}'$, i.e., $|\mathcal{A}'|/|\mathcal{A}| \ge p/2$ for a positive constant p. We then have the following result.

Lemma 1. *For any $\chi > 0$, if $d > 1$, $|\mathcal{A}'|/|\mathcal{A}| \ge p/2$, $|\mathcal{A}| \ge 6\log d/(pC_2(\phi_1)^2)$, and $\lambda = \lambda(\chi, \phi_1\sqrt{p}/2) = \chi\phi_1^2 p/(16 s_0)$, then the following tail inequality holds:*

$$\Pr\left[\|\hat{\beta}(\mathcal{A}, \lambda) - \beta\|_1 > \chi\right] \le 2\exp\left[-C_1\left(\frac{\phi_1\sqrt{p}}{2}\right)|\mathcal{A}|\chi^2 + \log d\right] + \exp\left[-pC_2(\phi_1)^2\,|\mathcal{A}|/2\right] .$$

Recall that the constants C_1 and C_2 are defined in §3.3. The full proof is given in Appendix EC.2, but we describe the main steps here. We first show that $\hat{\Sigma}(\mathcal{A}') \in \mathcal{C}(\mathrm{supp}(\beta), \phi_1/\sqrt{2})$ with high probability. This involves showing that $\|\hat{\Sigma}(\mathcal{A}') - \Sigma\|_\infty$ is small with high probability using random matrix theory. Next, we use this fact along with the Azuma-Hoeffding inequality to show that $\hat{\Sigma}(\mathcal{A}) \in \mathcal{C}(\mathrm{supp}(\beta), \phi_1\sqrt{p}/2)$ with high probability. Applying Proposition 1 to $\hat{\beta}(\mathcal{A}, \lambda)$ will give the desired tail inequality even though part of the data is not generated i.i.d. from $\mathcal{P}_Z$.

4.2. LASSO Tail Inequality for the Forced-Sample Estimator

We now obtain a tail inequality for the forced-sample estimator $\hat{\beta}(\mathcal{T}_{i,t}, \lambda_1)$ of each arm $i \in [K]$.

Proposition 2. *For all $i \in [K]$, the forced sample estimator $\hat{\beta}(\mathcal{T}_{i,t}, \lambda_1)$ satisfies*

$$\Pr\left[\|\hat{\beta}\left(\mathcal{T}_{i,t}, \lambda_1\right) - \beta_i\|_1 > \frac{h}{4x_{\max}}\right] \le \frac{5}{t^4} ,$$

when $\lambda_1 = \phi_0^2 p_* h/(64 s_0 x_{\max})$, $t \ge (Kq)^2$, $q \ge 4\lceil q_0 \rceil$, and q_0 satisfies the definition in §3.3.

Note that the matrix $\hat{\Sigma}(\mathcal{T}_{i,t})$ concentrates around $\mathbb{E}_{X\sim\mathcal{P}_X}[XX^\top]$. Thus, although this estimator is trained on i.i.d. samples from $\mathcal{P}_X$, the above tail inequality does not follow directly from Proposition 1 since we have only assumed that the compatibility condition holds for $\Sigma_i = \mathbb{E}_{X\sim\mathcal{P}_X}[XX^\top | X \in U_i]$ rather than $\mathbb{E}_{X\sim\mathcal{P}_X}[XX^\top]$ (Assumption 4). This is easily resolved by showing $\mathcal{T}'_{i,t} \equiv \{t' \in \mathcal{T}_{i,t} \mid X_{t'} \in U_i\}$ is a set of i.i.d. samples from $\mathcal{P}_{X|X\in U_i}$, and then applying Lemma 1 with $\mathcal{A} = \mathcal{T}_{i,t}$, $\mathcal{A}' = \mathcal{T}'_{i,t}$, and $\mathcal{P}_Z = \mathcal{P}_{X|X\in U_i}$. The full proof is given in Appendix EC.3.

4.3. LASSO Tail Inequality for the All-Sample Estimator

We now provide a tail inequality for the all-sample estimator of optimal arms $\mathcal{K}_{opt}$. The challenge is that the all-sample sets $\mathcal{S}_{i,t}$ depend on choices made online by the algorithm. More precisely, the algorithm selects arm i at time t based both on X_t and on previous observations $\{X_{t'}\}_{1\le t'<t}$ (which are used to estimate β_i). As a consequence, the variables $\{X_t \mid t \in \mathcal{S}_{i,t}\}$ may be correlated.

Moreover, unlike the forced-sample estimator, we do not have a guarantee that a constant fraction of the all-sample sets $\mathcal{S}_{i,t}$ are i.i.d. In particular, only the $|\mathcal{T}_{i,t}| = \mathcal{O}(\log T)$ forced samples are guaranteed to be i.i.d., but we will prove that $|\mathcal{S}_{i,t}| = \mathcal{O}(T)$ for optimal arms $i \in \mathcal{K}_{opt}$ with high probability. Thus, we cannot apply Lemma 1 directly with $\mathcal{A} = \mathcal{S}_{i,t}$ and $\mathcal{A}' = \mathcal{T}'_{i,t}$ as before. We resolve this by showing that (i) our algorithm uses the forced-sample estimator $\mathcal{O}(T)$ times with high probability, and (ii) a constant fraction of the samples where we use the forced-sample estimator are i.i.d. from the regions U_i. We then invoke Lemma 1 with a modified $\mathcal{A}'$ such that $|\mathcal{A}'| = \mathcal{O}(T)$. In particular, we define the event

$$A_t \equiv \left\{ \|\hat{\beta}(\mathcal{T}_{i,t}, \lambda_1) - \beta_i\|_1 \le \frac{h}{4x_{\max}} \ , \quad \forall i \in [K] \right\} . \tag{3}$$

Since the event A_t only depends on forced-samples, the random variables $\{X_t \mid A_{t-1} \text{ holds}\}$ are i.i.d. (with distribution $\mathcal{P}_X$). Furthermore, if we let

$$\mathcal{S}'_{i,t} \equiv \left\{ t' \in [t] \mid A_{t'-1} \text{ holds}, \ X_{t'} \in U_i \ , \text{ and } t' \notin \cup_{j\in[K]} \mathcal{T}_{j,t} \right\} .$$

then the random variables $\{X_{t'} \mid t' \in \mathcal{S}'_{i,t}\}$ are i.i.d. (with distribution $\mathcal{P}_{X|X\in U_i}$). Finally, we will show that for $i \in \mathcal{K}_{opt}$, the event $A_{t'-1}$ ensures that LASSO Bandit chooses arm i at time t' when $X_{t'} \in U_i$, so $\mathcal{S}'_{i,t} \subset \mathcal{S}_{i,t}$. Finally, we will use Proposition 2 to show that events $A_{t'-1}$ occur frequently enough so that $|\mathcal{S}'_{i,t}|$ is sufficiently large. Then, we can use Lemma 1 with $\mathcal{A} = \mathcal{S}_{i,t}$ and $\mathcal{A}' = \mathcal{S}'_{i,t}$ to prove Proposition 3. (Note that we will not need to prove convergence of the all-sample estimator for sub-optimal arms $\mathcal{K}_{sub}$.)

PROPOSITION 3. *The all-sample estimator $\hat{\beta}(\mathcal{S}_{i,t}, \lambda_{2,t})$ for $i \in \mathcal{K}_{opt}$ satisfies the tail inequality*

$$\Pr\left[\|\hat{\beta}\left(\mathcal{S}_{i,t}, \lambda_{2,t}\right) - \beta_i\|_1 > 16\sqrt{\frac{\log t + \log d}{p_*^3 C_1(\phi_0) t}} \right] < \frac{2}{t} + 2\exp\left[-\frac{p_*^2 C_2(\phi_0)^2}{32} \cdot t \right] , \tag{4}$$

when $\lambda_{2,t} = [\phi_0^2/(2s_0)]\sqrt{(\log t + \log d)/(p_* C_1(\phi_0) t)}$ and $t \ge C_5$.

In particular, Proposition 3 guarantees $\|\hat{\beta}\left(\mathcal{S}_{i,t}, \lambda_{2,t}\right) - \beta_i\|_1 = \mathcal{O}(\sqrt{\log t/t})$ with high probability while Proposition 2 only guarantees $\|\hat{\beta}\left(\mathcal{T}_{i,t}, \lambda_1\right) - \beta_i\|_1 = \mathcal{O}(1)$ with high probability. However, note that the all-sample estimator tail inequality only holds for optimal arms $\mathcal{K}_{opt}$ while the forced-sample estimator tail inequality holds for all arms $[K]$. Thus, the LASSO Bandit uses the all-sample estimator to choose the best estimated arm because of its significantly faster convergence. However, the algorithm requires a pre-processing step using the forced-sample estimator to (i) ensure that we obtain $\mathcal{O}(T)$ i.i.d. samples for each $i \in \mathcal{K}_{opt}$ (required for the proof of Proposition 3), and (ii) to prune out sub-optimal arms $\mathcal{K}_{sub}$ with high probability (as we will show in the next subsection) for which Proposition 3 does not hold. The full proof is given in Appendix EC.4.

4.4. Bounding the Cumulative Expected Regret

We now use our convergence results to compute the cumulative regret of LASSO Bandit. We divide our time periods $[T]$ into three groups:

(a) Initialization ($t \le C_5$), or forced sampling ($t \in \mathcal{T}_{i,T}$ for some $i \in [K]$).

(b) Times $t > C_5$ when the event A_{t-1} does not hold.

(c) Times $t > C_5$ when the event A_{t-1} holds and we do not perform forced sampling, i.e., the LASSO Bandit plays the estimated best arm from $\hat{\mathcal{K}}$ (chosen by the forced-sampling estimator) using the all-sample estimator.

Note that these groups may not be disjoint but their union contains $[T]$. We bound the regret from each period separately and sum the results to obtain an upper bound on the cumulative regret. Our division of groups (b) and (c) is motivated by the fact that when A_{t-1} holds, the forced-sample estimator (i) includes the correct arm as part of the chosen subset of arms $\hat{\mathcal{K}}$ and (ii) does not include any sub-optimal arms from $\mathcal{K}_{sub}$ in $\hat{\mathcal{K}}$. Thus, when A_{t-1} holds, we can apply the convergence properties of the all-sample estimator (which only hold for optimal arms) to $\hat{\mathcal{K}}$ without the concerns that $\hat{\mathcal{K}}$ may not include the true optimal arm or that it may include sub-optimal arms.

The cumulative expected regret from time periods in group (a) at time T is bounded by at most $2bx_{\max}(6qK\log T + C_5)$ (Lemma EC.15). This follows from the fact that the worst-case regret at any time step is at most $2bx_{\max}$ (Assumption 1), while there are only C_5 initialization samples and at most $6Kq\log T$ forced samples up to time T (Lemma EC.8).

Next, the cumulative expected regret from time periods in group (b) at time T is bounded by at most $2Kbx_{\max}$ (Lemma EC.17). This follows from the tail inequality for the forced-sample estimator (Proposition 2), which bounds the probability that event A_t does not hold at time t by at most $5K/t^4$. The result follows from summing this quantity over time periods $C_5 < t \le T$.

Finally, the cumulative expected regret from time periods (c) at time T is bounded by at most $(8Kbx_{\max} + C_3\log d)\log T + C_3(\log T)^2 + C_4$ (Lemma EC.20). To show this, we first observe that if event A_t holds, then the set $\hat{\mathcal{K}}$ (chosen by the forced-sample estimator) contains the optimal arm $i^* = \arg\max_{i\in[K]} X_t^\top \beta_i$ and no sub-optimal arms from the set $\mathcal{K}_{sub}$ (Lemma EC.18). Then, we sum the expected regret using Proposition 3 for all optimal arms. Our all-sample estimators for each optimal arm satisfy $\|\hat{\beta}(\mathcal{S}_{i,t}, \lambda_{2,t}) - \beta_i\|_1 = \mathcal{O}(\sqrt{\log t/t})$ with high probability; thus, as shown in Lemma EC.19, we only incur regret if the observed covariate vectors are within a $\mathcal{O}(\sqrt{\log t/t})$ distance from a decision boundary (which occurs with small probability based on Assumption 2). Finally, if the error of some optimal arm's parameter estimate $\|\hat{\beta}(\mathcal{S}_{i,t}, \lambda_{2,t}) - \beta_i\|_1$ is much larger than $\mathcal{O}(\sqrt{\log t/t})$, we incur worst-case regret, but this occurs with exponentially small probability.

4.5. Proof of the Main Result

Summing up the regret contributions from the previous subsection gives us our main result.

Proof of Theorem 1 The total expected cumulative regret of the LASSO Bandit up to time T is upper-bounded by summing all the terms from Lemmas EC.15, EC.17, and EC.20:

$$\begin{aligned}
R_T &\le \overbrace{2bx_{\max}(6qK\log T + C_5)}^{\text{Regret from}(a)} + \overbrace{2Kbx_{\max}}^{\text{Regret from }(b)} + \overbrace{(8Kbx_{\max} + C_3\log d)\log T + C_3(\log T)^2 + C_4}^{\text{Regret from}(c)} \\
&= C_3(\log T)^2 + [2Kbx_{\max}(6q+4) + C_3\log d]\log T + (2bx_{\max}C_5 + 2Kbx_{\max} + C_4) \\
&= \log T\,[C_3\log T + 2Kbx_{\max}(6q+4) + C_3\log d] + (2bx_{\max}C_5 + 2Kbx_{\max} + C_4)\,.
\end{aligned}$$

Now, using $q = \mathcal{O}(s_0^2 \log d)$, and the fact that $C_0, \ldots, C_5$, b, $x_{\max}$, and ϕ_0 are constants,

$$R_T = \mathcal{O}\left(\log T\left[\log T + s_0^2\log d\right]\right) = \mathcal{O}\left(s_0^2\left[\log T + \log d\right]^2\right). \quad \square$$

5. Empirical Results

The objective of this section is to compare the performance of LASSO Bandit with existing algorithms that have theoretical guarantees in our setting. We present two sets of empirical results evaluating our algorithm on both sparse synthetic data (§5.1), and a simplified version of the warfarin dosing problem using a real patient dataset (§5.2).

5.1. Synthetic Data

We evaluate the LASSO Bandit on a synthetically-generated dataset to address two questions: (1) How does the LASSO Bandit's performance compare against existing algorithms empirically?; (2) Is the LASSO Bandit robust to the choice of input parameters?

We compare the LASSO Bandit against (i) the UCB-based algorithm OFUL-LS (Abbasi-Yadkori et al. 2011), which is an improved version of the algorithm suggested in Dani et al. (2008), (ii) a sparse variant OFUL-EG for high-dimensional settings (Abbasi-Yadkori et al. 2012, Abbasi-Yadkori 2012), and (iii) the OLS Bandit by Goldenshluger and Zeevi (2013). Our results demonstrate that the LASSO Bandit significantly outperforms these benchmarks. Separately, we find that the LASSO Bandit is robust to changes in input parameters by even an order of magnitude.

REMARK 9. We only consider algorithms that have theoretical guarantees for our problem. In particular, recall that linear bandit algorithms can only be translated to the contextual bandit if they consider a changing action space (see Abbasi-Yadkori (2012) for details on the connection between variations of the linear bandit and contextual bandit). Two notable linear bandit algorithms that do not meet this criteria are Carpentier and Munos (2012) and Agrawal and Goyal (2013). We also do not include the Thompson sampling algorithm of Russo and Van Roy (2014a) since they use a different performance metric of Bayes risk, which is the expected value of the standard notion of regret (that we use) with respect to a Bayesian prior over the unknown arm parameters. In practice, the decision-maker may not have access to the true prior.

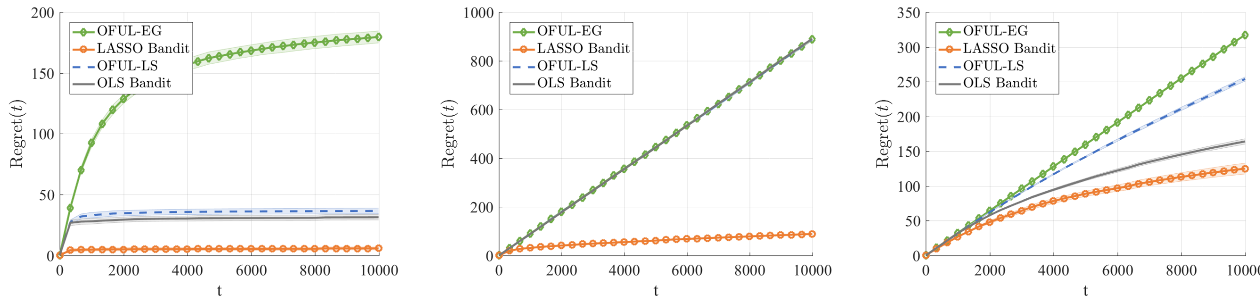
Synthetic Data Generation. We consider three scenarios for K, d, and s_0: a) $K=2$, $d=100$, $s_0=5$; (b) $K=10$, $d=1000$, $s_0=2$; and (c) $K=50$, $d=20$, $s_0=2$. In each case, we consider K arms (treatments) and d user covariates, where only a randomly chosen subset of s_0 covariates are predictive of the reward for each treatment, i.e., for each $i \in [K]$, the arm parameters β_i are set to zero except for s_0 randomly selected components that are drawn from a uniform distribution on $[0,1]$. We note that the OFUL-EG algorithm requires an additional technical assumption that $\sum_{i=1}^{K} \|\beta_i\|_1 = 1$. We scale our β_i's accordingly so that this assumption is met.

Next, at each time t, user covariates X_t are independently sampled from a Gaussian distribution $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and truncated so that $\|X_t\|_\infty = 1$. Finally, we set the noise variance to be $\sigma^2 = 0.05^2$.

Algorithm Inputs. Bandit algorithms require the decision-maker to specify a variety of input parameters that are often unknown in practice. For instance, Theorem 1 suggests specific input parameters for the LASSO Bandit (e.g., σ, ϕ_0) and similarly, the benchmark OFUL and OLS Bandit algorithms require analogous specifications. Therefore, in order to simulate a realistic environment where no past (properly randomized) data is available to tune these parameters, we make ad-hoc choices for the input parameters of the LASSO and OLS Bandit algorithms, and use parameters suggested in computational experiments by the authors of the OFUL-LS and OFUL-EG algorithms (Abbasi-Yadkori 2012). Note that these parameters cannot be estimated from historical data since we suffer from bandit feedback and estimating some parameters requires knowledge of every arm's reward at a given time step. As a robustness check, we later vary the input parameters of the LASSO Bandit to better understand the sensitivity of its performance to these heuristic choices.

For the LASSO and OLS Bandit algorithms, we choose the forced-sampling parameter $q=1$ and the localization parameter $h=5$. For the LASSO Bandit, we further set the initial regularization parameters to $c=\lambda_1=\lambda_{2,0}=0.05$. For the OFUL algorithms, as suggested by Abbasi-Yadkori (2012), we set $\lambda=1$ and $\delta=10^{-4}$, and furthermore, we set $\eta=1$ for OFUL-EG.

Results. Figure 1 compares the cumulative regret (averaged over 5 trials) of the LASSO Bandit against other bandit algorithms on the aforementioned synthetic data for $T=10,000$ steps. The



(a) $K = 2,\ d = 100,\ s_0 = 5$ (b) $K = 10,\ d = 1000,\ s_0 = 2$ (c) $K = 50,\ d = 20,\ s_0 = 2$

Figure 1 Comparison of the cumulative regret of the LASSO Bandit against other bandit algorithms on synthetic data for different values of K, d, and s_0.

shaded region around each curve is the 95% confidence interval across the 5 trials. We see that the LASSO Bandit significantly outperforms benchmarks in cumulative regret.

Figure 1(a) shows that the LASSO Bandit continues to achieve significantly less per-time-step regret than the alternative algorithms even for large t. For example, when $t \approx 1,000$, we have that $d \ll t$ and thus we are in a *low-dimensional* regime. However, the slope of the cumulative regret curve of the LASSO Bandit is visibly smaller than that of the alternative algorithms at $t \approx 1,000$. This shows that the LASSO Bandit may be useful even in low-dimensional regimes since other algorithms continue to overfit the arm parameters.

Figure 1(b) considers a larger number of covariates d. As expected, we see that the performance gap between the LASSO Bandit and the other algorithms increases significantly; this is because the benchmark algorithms do not take advantage of sparsity and perform exploration for at least $\mathcal{O}(Kd)$ samples in order to define linear regression estimates for each arm. Figure 1(c) considers a larger number of arms and fewer covariates. Here, we see that the performance gap between the LASSO Bandit and alternative methods decreases; this is because the LASSO Bandit does not provide any improvement over existing algorithms in K, and provides limited improvement when the number of covariates is very small.

Additional Simulations. To study the robustness of the above simulations, we provide a comprehensive set of simulations in Appendix EC.6 to test the performance of the LASSO Bandit as the parameters or modeling assumptions (required for the theory) are varied. First, we study how the regret of the LASSO Bandit scales with respect to each of the parameters K, d, and s_0 separately (see §EC.6.1); we find that the regret appears to grow logarithmically with d, and linearly with K and s_0. Next, we perform sensitivity analysis to the input parameters h, q, and c (see §EC.6.2). We find that the cumulative regret performance is not substantially impacted despite experimenting with the parameters by up to an order of magnitude; this suggests that the LASSO Bandit is robust, which is important since input parameters are likely to be misspecified in practice.

Another interesting direction is considering nonlinear reward functions. The LASSO Bandit can be used in conjunction with basis expansion methods from statistical learning to approximate any nonlinear function (Hastie et al. 2001). In §EC.6.3, we numerically demonstrate that such a version of our method can perform very well with nonlinear rewards.

Finally, in §EC.6.4, we consider settings where the covariate distribution $\mathcal{P}_X$ does not satisfy the margin condition (Assumption 2) or the arm optimality condition (Assumption 3).

5.2. Case Study: Warfarin Dosing

Preliminaries. A finite-armed adaptive clinical trial with patient covariates is an ideal application for our problem formulation and algorithm. For instance, in the aforementioned BATTLE clinical trial (Kim et al. 2011), the arms would be the four chemotherapeutic agents, the patient covariates would be the biomarkers from the patient's tumor biopsy, and the reward would be the patient's expected length of cancer remission. Our algorithm (and other algorithms for the contextual bandit) would seek to learn a mapping between patient biomarkers and the optimal chemotherapeutic assignment to maximize overall patient remission rates. (Even in such a setting, we have made a number of simplifications, e.g., the ability to observe instantaneous rather than delayed feedback. Modeling the full complexity of the problem is beyond the scope of our paper.)

Therefore, we would ideally evaluate our algorithm on a real patient dataset from such an application. However, performing such an evaluation retrospectively on observational data is challenging because we require access to counterfactuals. In particular, our algorithm may choose a different action than the one taken in the data; thus, we need an unbiased estimate of the resulting reward to evaluate the algorithm's performance. Estimating such counterfactuals is known to be very difficult in healthcare since there are many unobserved confounders that can significantly bias our results.

As a consequence, we choose a unique application (warfarin dosing), where we do have access to counterfactuals. However, in order to simulate bandit feedback, we will suppress this counterfactual information to the bandit algorithms, thereby handicapping ourselves relative to an optimal algorithm. This lets us benchmark the performance of our algorithm against existing bandit methods in an unbiased manner on a real patient dataset (where our technical assumptions may not hold).

Warfarin Problem. Warfarin is the most widely used oral anticoagulant agent in the world (Wysowski et al. 2007). Correctly dosing warfarin remains a significant challenge as the appropriate dosage is highly variable among individuals (by a factor of up to 10) due to patient clinical, demographic and genetic factors.

Physicians currently follow a fixed-dose strategy: they start patients on 5mg/day (the appropriate dose for the majority of patients) and slowly adjust the dose over the course of a few weeks by tracking the patient's anticoagulation levels. However, an incorrect initial dosage can result in highly adverse consequences such as stroke (if the initial dose is too low) or internal bleeding (if the initial dose is too high). Every year, nearly 43,000 emergency department visits in the United States are due to adverse events associated with inappropriate warfarin dosing (Budnitz et al. 2006). Thus, we tackle the problem of learning and assigning an appropriate *initial dosage* to patients by leveraging patient-specific factors.

Dataset. We use a publicly available patient dataset that was collected by staff at the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) for 5700 patients who were treated with warfarin from 21 research groups spanning 9 countries and 4 continents. Importantly, this data contains the true patient-specific optimal warfarin doses (which are initially unknown but are eventually found through the physician-guided dose adjustment process over the course of a few weeks) for 5528 patients. It also includes patient-level covariates such as clinical factors, demographic variables, and genetic information that have been found to be predictive of the optimal warfarin dosage (Consortium 2009). These covariates include:

- *Demographics:* gender, race, ethnicity, age, height, weight
- *Diagnosis:* reason for treatment (e.g. deep vein thrombosis, pulmonary embolism, etc.)
- *Pre-existing diagnoses:* indicators for diabetes, congestive heart failure or cardiomyopathy, valve replacement, smoker status
- *Medications:* indicators for potentially interacting drugs (e.g. aspirin, Tylenol, Zocor, etc.)
- *Genetics:* presence of genotype variants of CYP2C9 and VKORC1

Details on the dataset can be found in Supplementary Appendix 1 of Consortium (2009). These covariates were hand-selected by professionals as being relevant to the task of warfarin dosing based on medical literature; there are no extraneously added variables.

Finally, we note that the authors of Consortium (2009) report that an ordinary least-squares linear model fits the data best (i.e. achieves the best cross-validation accuracy) compared to alternative models (such as support vector regression, regression trees, model trees, multivariate adaptive regression splines, least-angle regression, LASSO, etc.) for the objective of predicting the correct warfarin dosage as a function of the given patient-level variables. The results of Consortium (2009) suggest that there is no underlying sparsity in this data. Thus, one might expect low-dimensional algorithms like the OLS Bandit or OFUL-LS to perform no worse than the LASSO Bandit; surprisingly, we find that this is not the case in the online setting.

Bandit Formulation. We formulate the problem as a 3-armed bandit with covariates.

Arms: We bucket the optimal dosages using the "clinically relevant" dosage differences suggested in (Consortium 2009): (1) Low: under 3mg/day (33% of cases), (2) Medium: 3-7mg/day (54% of cases), and (3) High: over 7mg/day (13% of cases). In particular, patients who require a low (high) dose would be at risk for excessive (inadequate) anti-coagulation under the physician's medium starting dose. We estimate the true arm parameters β_i using linear regressions on the entire dataset.

Covariates: We construct 93 patient-specific covariates, including indicators for missing values.

Reward: For each patient, we set the reward to 0 if the dosing algorithm chooses the arm corresponding to the patient's true optimal dose. Otherwise, the reward is set to -1. We choose this simple reward function so that the regret directly measures the number of incorrect dosing decisions. Other objectives (e.g., the cost of treating adverse outcomes for under- vs. over-dosing) can be easily considered by adjusting the definition of the reward function accordingly.

As an aside, note that we have chosen a binary reward for simplicity although we are modeling the reward as a linear function. Yet, the LASSO Bandit performs well in this setting, suggesting that it can also be valuable for discrete outcomes.

Evaluation and Results. We consider 10 random permutations[4] of patients and simulate the following policies:

1. **LASSO Bandit**, described in §3 of this paper,
2. **OLS Bandit**, described in Goldenshluger and Zeevi (2013),
3. **OFUL-LS**, described in Abbasi-Yadkori et al. (2011),
4. **OFUL-EG**, described in Abbasi-Yadkori et al. (2012)[5], and
5. **Doctors**, who currently always assign an initial medium dose (Consortium 2009),
6. **Oracle**, which assigns the optimal estimated dose given the true arm parameters β_i.

Note that a true oracle policy cannot be implemented since arm parameters β_i are not available. Instead, we consider an "approximate" version of the oracle that estimates the arm parameters β_i upfront using all the patient outcomes (that have not yet been observed by the other algorithms). This oracle may still make incorrect decisions since it only has access to estimated arm parameters. We consider two versions of the oracle policy: **Linear Oracle** that estimates β_i via linear regression, and **Logit Oracle** that estimates β_i via logistic regression (since the outcomes are binary).

[4] We also repeated the analysis using bootstrap samples (random subsets with replacement) and the results were similar. We present the results for permuted samples because the confidence intervals produced by the bootstrapped samples may be optimistic (since they may overfit to samples drawn multiple times from the original data with replacement). In the offline setting, Efron and Tibshirani (1993) provide methods for correcting such bias; such methods may be extendable to our online setting, but this is beyond the scope of this paper.

[5] The original OFUL-EG requires the assumption that $\sum_{i=1}^{k} \|\beta_i\|_1 = 1$ (Abbasi-Yadkori 2012); however, there is no way to guarantee that this holds on a real dataset where we do not know the $\{\beta_i\}$. Thus, we modify the confidence sets using the EG($\pm$) algorithm (Kivinen and Warmuth 1997), which does not require this assumption.

We sequentially draw random permutations of patients and simulate the actions and feedback of each of the six policies. Note that the data contains each patient's true optimal dosage, but we suppress this information from the learning algorithms; we use the true dosage as counterfactuals to evaluate the reward of each algorithm after it chooses an action. Figure 2 compares the average fraction of incorrect dosing decisions under each policy as a function of the number of patients seen in the data; the shaded error bars represent statistical fluctuations of the rewards over the 10 permutations.

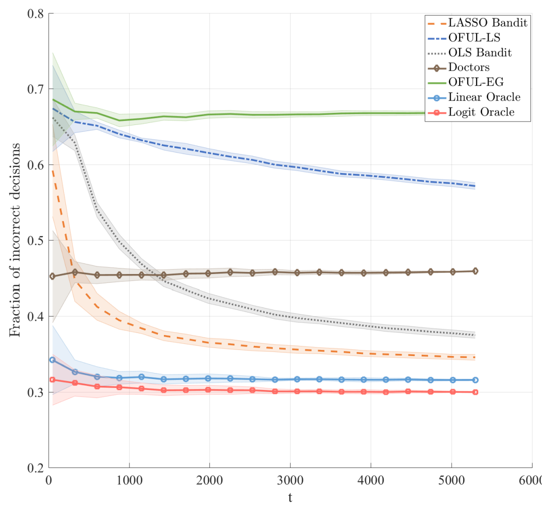


Figure 2 Comparison of the fraction of incorrectly dosed patients under the oracle, LASSO Bandit, OLS Bandit, OFUL-LS, OFUL-EG, and doctor policies as a function of number of patients in the warfarin data.

We first note that the LASSO Bandit outperforms the three other bandit algorithms for any number of patients across all permutations. The results show three regimes:

Small Data. When there are very few samples (<200 patients), the doctor's policy of assigning the medium dose (which is optimal for the majority of patients) performs best on average.

Moderate Data. When there are a moderate number of samples (200 - 1500 patients), the LASSO Bandit effectively learns the arm parameters and outperforms the doctor's policy; however, the remaining bandit algorithms still perform worse than physicians.

Big Data. When there are a large number of samples (1500 - 5000 patients), both the LASSO and OLS bandit policies outperform the physician's policy and begin to look comparable. However, the OFUL-LS and OFUL-EG algorithms still perform worse than doctors.

Note that all three existing bandit algorithms required more than 1500 patient samples before outperforming the doctor's static policy; this may be prohibitively costly in a healthcare setting and may hinder adoption of learning strategies in practice. In contrast, we see that the LASSO

		LASSO Bandit Policy Assigned Dosage			Physician Policy Assigned Dosage			% of Patients
		Low	*Medium*	*High*	*Low*	*Medium*	*High*	
True Dosage	*Low*	**57%**	42%	**1%**	**0%**	100%	**0%**	**33%**
	Medium	14%	**83%**	3%	0%	**100%**	0%	**54%**
	High	**3%**	90%	**7%**	**0%**	100%	**0%**	**13%**

Table 1 Fraction of patients (stratified by their true dose) who were assigned each dose (low/medium/high) under the LASSO Bandit and physician policies. Blue numbers indicate the fraction of patients who were dosed correctly; red numbers indicate the fraction of patients who were dosed incorrectly by two buckets.

Bandit starts outperforming the doctor's policy after only 200 patients, resulting in a significant improvement of outcomes for initial patients. Thus, although an OLS linear model fits the entire dataset better than a LASSO model, it may be more effective to use the LASSO Bandit in an online setting in order to more efficiently use information as it is collected. In particular, the LASSO Bandit uses regularization to first build simple predictive models (with few covariates), and gradually builds more complex predictive models (by including more covariates over time); this helps us make reasonable decisions in the small-data regime without sacrificing performance in the big-data regime.

Risk Implications. One concern that arose in conversations with clinicians is that although the LASSO Bandit policy achieves a higher dosing accuracy overall (compared to doctors), it may assign a "significantly worse" dose to some patients. In particular, the bandit algorithm may assign a low dose to a patient whose true dose is high (or vice-versa); on the other hand, the doctor always hedges her bet by assigning the medium dose. To better illustrate the risk consequences, we tabulate the assigned vs. true dosages for the LASSO Bandit and doctor's policies after 5,000 patients (see Table 1). The red numbers indicate the fraction of patients assigned a significantly worse dose and the blue numbers indicate the fraction of patients assigned the correct dose. We find that there is only a 0.7% weighted probability that a patient receives a significantly worse dose under the LASSO Bandit policy. On the other hand, the LASSO Bandit correctly doses 57% of the patients for whom low dosage is optimal; in contrast, the physician policy does not dose any of these patients correctly (thereby subjecting them to excessive anti-coagulation) although they account for a third of the patient population. This trade-off can be explored further by adjusting the reward function; in particular, we have used a binary loss for mis-dosing, but the loss can be a function of the magnitude of mis-dosing.

Remark 10. Several simplifying assumptions were made in this case study. For example, warfarin dosing is not a truly bandit problem, since we always observe the optimal arm (patient's true dose) even if we play the wrong arm (assign the wrong dose initially) as the doctor tunes the dosage over time. Yet, we use this setting as a case study to evaluate bandit policies because the data contains the true counterfactual outcomes without performing an experiment. For problems with true bandit feedback, we do not observe counterfactual rewards for actions that were not chosen in the data, so we cannot evaluate the counterfactual performance of the LASSO Bandit. In practice, the LASSO Bandit would be most useful for bandit settings where the patient can only receive one treatment and the counterfactual outcomes under other treatments cannot be observed, e.g., the problem of choosing chemotherapy agents as described in the introduction (Kim et al. 2011).

6. Conclusions

We present the LASSO Bandit algorithm for contextual bandit problems with high-dimensional covariates, and we prove the first regret bound that grows only poly-logarithmically in both the number of covariates and the number of patients. We empirically find that the LASSO Bandit is more versatile than existing methods: although it is designed for high-dimensional sparse settings, it outperforms the OLS Bandit even in *low-dimensional* and *non-sparse* problems. We illustrate the LASSO Bandit's practical relevance by evaluating it on a medical decision-making problem of warfarin dosing; we find that it surpasses existing bandit methods as well as physicians to correctly dose a majority of patients and thereby improve overall patient outcomes. We note that several simplifying assumptions were made in this evaluation, and thus, modeling the full complexity of the problem would be a valuable direction to pursue in future work.

Limitations and future directions. We conclude by discussing a number of limitations of the LASSO Bandit algorithm. First, it is not suitable in applications with a large number of arms, since our regret bounds scale superlinearly with K. This is because our model treats each arm as an independent decision, and so the outcome of each arm provides no information on other arms. However, in certain applications (e.g., combination chemotherapy where each arm is a combination of several base drugs, or assortment optimization where each assortment is a combination of several products), one can improve performance by taking advantage of the correlation between arms. Second, our algorithm relies on a prescribed schedule for exploration. Such pure exploration phases may be prohibitively costly or unethical in settings such as medical decision-making. In such situations, methods such as UCB that only explore within a certain confidence set may be more desirable. One could even consider algorithms that avoid exploration. Finally, our algorithm, similar to UCB or OLS Bandit, requires a number of input parameters which should ideally be optimized for the desired application. An interesting research question would be how to optimize these parameters in a data-driven fashion.

Acknowledgments

The authors gratefully acknowledge the National Science Foundation (Graduate Research Fellowship Grant No. DGE-114747, NSF EAGER award CMMI:1451037, NSF CAREER award CMMI: 1554140, and NSF grant CCF:1216011) and the Stanford Cyber Security Initiative for financial support.

This paper has also benefitted from valuable feedback from Stephen Chick, Nima Hamidi, Hamid Nazerzadeh, Stefanos Zenios, anonymous referees, and various seminar participants. They have been instrumental in guiding us to improve the paper.

References

Abbasi-Yadkori, Yasin. 2012. Online learning for linearly parametrized control problems. Ph.D. thesis, Edmonton, Alta., Canada. URL `https://yasinov.github.io/Yasin-PhD-Thesis.pdf`. AAINR89307.

Abbasi-Yadkori, Yasin, Dávid Pál, Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2312–2320. URL `http://papers.nips.cc/paper/4417-improved-algorithms-for-linear-stochastic-bandits.pdf`.

Abbasi-Yadkori, Yasin, David Pal, Csaba Szepesvari. 2012. Online-to-confidence-set conversions and application to sparse stochastic bandits. Neil D. Lawrence, Mark Girolami, eds., *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 22. PMLR, La Palma, Canary Islands, 1–9. URL `http://proceedings.mlr.press/v22/abbasi-yadkori12.html`.

Agrawal, Shipra, Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. Sanjoy Dasgupta, David McAllester, eds., *Proceedings of the 30th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 28. PMLR, Atlanta, Georgia, USA, 127–135. URL `http://proceedings.mlr.press/v28/agrawal13.html`.

Alon, N, J Spencer. 1992. The probabilistic method. *Wiley, New York* .

Athey, S., G. W. Imbens, S. Wager. 2016. Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. *ArXiv e-prints* URL `https://arxiv.org/abs/1604.07125`.

Auer, Peter. 2003. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3** 397–422.

Ban, Gah-Yi, Cynthia Rudin. 2014. The big data newsvendor: Practical insights from machine learning URL `https://ssrn.com/abstract=2559116`.

Bayati, Mohsen, Mark Braverman, Michael Gillam, Karen Mack, George Ruiz, Mark Smith, Eric Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS ONE* **9**(10).

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81**(2) 608–650.

Bertsimas, Dimitris, Nathan Kallus. 2014. From predictive to prescriptive analytics. *Working Paper* URL `https://arxiv.org/abs/1402.5481`.

Bickel, Peter, Ya'acov Ritov, Alexandre Tsybakov. 2009. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 1705–1732.

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

Budnitz, DS, DA Pollock, KN Weidenbach, AB Mendelson, TJ Schroeder, JL Annest. 2006. National surveillance of emergency department visits for outpatient adverse drug events. *Journal of the American Medical Association* **296** 1858–1866.

Bühlmann, Peter, Sara Van De Geer. 2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candes, Emmanuel, Terence Tao. 2007. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* 2313–2351.

Carpentier, Alexandra, Remi Munos. 2012. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. Neil D. Lawrence, Mark Girolami, eds., *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, vol. 22. 190–198.

Chen, Scott Shaobing, David L Donoho, Michael A Saunders. 1998. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20**(1) 33–61.

Chen, Xi, Zachary Owen, Clark Pixton, David Simchi-Levi. 2015. A statistical learning approach to personalization in revenue management URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2579462`.

Chu, Wei, Lihong Li, Lev Reyzin, Robert Schapire. 2011. Contextual bandits with linear payoff functions. Geoffrey Gordon, David Dunson, Miroslav Dudík, eds., *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 15. PMLR, Fort Lauderdale, FL, USA, 208–214. URL `http://proceedings.mlr.press/v15/chu11a.html`.

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

Dani, Varsha, Thomas P. Hayes, Sham M. Kakade. 2008. Stochastic Linear Optimization under Bandit Feedback. Rocco A. Servedio, Tong Zhang, Rocco A. Servedio, Tong Zhang, eds., *Conference On Learning Theory*. Omnipress, 355–366. URL `http://dblp.uni-trier.de/rec/bibtex/conf/colt/DaniHK08`.

Deshpande, Yash, Andrea Montanari. 2012. Linear bandits in high dimension and recommendation systems. URL `https://arxiv.org/abs/1301.1722`.

Efron, B, RJ Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman Hall.

Elmachtoub, Adam N, Ryan McNellis, Sechan Oh, Marek Petrik. 2017. A practical method for solving contextual bandit problems using decision trees. *Proceedings of the Thirty-third Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.

Goldenshluger, Alexander, Assaf Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* **19**(4) 1603–1633.

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

Hastie, Trevor, Robert Tibshirani, Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer New York Inc.

He, Biyu, Franklin Dexter, Alex Macario, Stefanos Zenios. 2012. The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manufacturing & Service Operations Management* **14**(1) 99–114.

Kallus, N., M. Udell. 2016. Dynamic Assortment Personalization in High Dimensions URL `https://arxiv.org/abs/1610.05604`.

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

Kivinen, Jyrki, Manfred K. Warmuth. 1997. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation* **132**(1).

Langford, John, Tong Zhang. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. J. C. Platt, D. Koller, Y. Singer, S. T. Roweis, eds., *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., 817–824. URL `http://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandits-with-side-information.pdf`.

Naik, Prasad, Michel Wedel, Lynd Bacon, Anand Bodapati, Eric Bradlow, Wagner Kamakura, Jeffrey Kreulen, Peter Lenk, David M Madigan, Alan Montgomery. 2008. Challenges and opportunities in high-dimensional choice data analyses. *Marketing Letters* **19**(3-4) 201–213.

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, Bin Yu. 2012. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.* **27**(4) 538–557. doi:10.1214/12-STS400. URL `https://doi.org/10.1214/12-STS400`.

Perchet, Vianney, Philippe Rigollet. 2013. The multi-armed bandit problem with covariates. *The Annals of Statistics* **41**(2) 693–721.

Razavian, Narges, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, David Sontag. 2015. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* **3**(4) 277–287.

Rigollet, Philippe, Assaf Zeevi. 2010. Nonparametric bandits with covariates. Adam Tauman Kalai, Mehryar Mohri, eds., *Conference On Learning Theory*. Omnipress, 54–66. URL `http://dblp.uni-trier.de/db/conf/colt/colt2010.html#RigolletZ10`.

Rusmevichientong, Paat, John N Tsitsiklis. 2010. Linearly parameterized bandits. *Mathematics of Operations Research* **35**(2) 395–411.

Russo, Daniel, Benjamin Van Roy. 2014a. Learning to optimize via information-directed sampling. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 1583–1591. URL `http://papers.nips.cc/paper/5463-learning-to-optimize-via-information-directed-sampling.pdf`.

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

Slivkins, Aleksandrs. 2014. Contextual bandits with similarity information. *Journal of Machine Learning Research* **15**(1) 2533–2568.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tropp, Joel. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning* **8**(1-2) 1–230.

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

Wainwright, Martin. 2016. *High-dimensional statistics: A non-asymptotic viewpoint*. Book Draft (Working Publication). URL `https://www.stat.berkeley.edu/~wainwrig/nachdiplom/Chap2_Sep10_2015.pdf`.

Wysowski, Diane K, Parivash Nourjah, Lynette Swartz. 2007. Bleeding complications with warfarin use: a prevalent adverse effect resulting in regulatory action. *Internal Medicine* **167**(13) 1414–1419.

Yan, Ling, Wu-Jun Li, Gui-Rong Xue, Dingyi Han. 2014. Coupled group lasso for web-scale ctr prediction in display advertising. Eric P. Xing, Tony Jebara, eds., *Proceedings of the 31st International Conference on Machine Learning*, vol. 32. PMLR, Bejing, China, 802–810. URL `http://proceedings.mlr.press/v32/yan14.html`.

Online Appendix

This appendix contains proofs of the theoretical results, additional simulations testing the robustness of LASSO Bandit, and new results for the OLS Bandit.

EC.1. Proof of LASSO Tail Inequality for Adapted Observations

Recall the setup of §3.1. Let X_t denote the t^{th} row of $\mathbf{X}$ and $Y(t)$ denote the t^{th} entry of Y. The sequence $\{X_t : t = 1, ..., n\}$ forms an adapted sequence of observations, i.e., X_t may depend on past regressors and their resulting observations $\{X_{t'}, Y(t')\}_{t'=1}^{t-1}$. We also assume that all realizations of the random variable X_t, $t \in [n]$, satisfy $\|X_t\|_\infty \le x_{\max}$.

Before proving Proposition 1, we state and prove some lemmas, starting with the Bernstein concentration inequality for adapted sequences.

LEMMA EC.1 **(Bernstein Concentration)**. *Let $\{D_k, \mathfrak{S}_k\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that D_k is σ-subgaussian in an adapted sense, i.e., for all $\alpha \in \mathbb{R}$, $\mathbb{E}[e^{\alpha D_k} | \mathfrak{S}_{k-1}] \le e^{\alpha^2\sigma^2/2}$ almost surely. Then, for all $t \ge 0$, $\Pr[\,|\sum_{k=1}^n D_k| \ge t] \le 2\exp[-t^2/(2n\sigma^2)]$.*

Lemma EC.1 is from Theorem 2.3 of Wainwright (2016) when $\alpha_* = a_k = 0$ and $\nu_k = \sigma$ for all k.

LEMMA EC.2. *Define the event*

$$\mathcal{F}(\lambda_0(\gamma)) \equiv \left\{ \max_{r \in [d]} \left(2|\varepsilon^\top X^{(r)}|/n \right) \le \lambda_0(\gamma) \right\},$$

where $X^{(r)}$ is the r^{th} column of $\mathbf{X}$ and $\lambda_0(\gamma) \equiv 2\sigma x_{\max}\sqrt{(\gamma^2 + 2\log d)/n}$. Then, we have $\Pr[\mathcal{F}(\lambda_0(\gamma))] \ge 1 - 2\exp[-\gamma^2/2]$.

Proof of Lemma EC.2 Let $\mathfrak{S}_t$ be the sigma algebra generated by random variables $X_1, \ldots, X_{t-1}$, and $Y(1), \ldots, Y(t-1)$. First, using a union bound, we can write

$$\Pr[\mathcal{F}(\lambda_0(\gamma))] \ge 1 - \sum_{r=1}^{d} \Pr\left[|\varepsilon^\top X^{(r)}| > n\lambda_0(\gamma)/2 \right].$$

Now, for each $r \in [d]$, let $D_{t,r} = \varepsilon_t X_{t,r}$ and note that $D_{1,r}, ...D_{n,r}$ is a martingale difference sequence adapted to the filtration $\mathfrak{S}_1 \subset ... \subset \mathfrak{S}_n$ since $\mathbb{E}[\varepsilon_t X_{t,r} | \mathfrak{S}_t] = 0$. On the other hand, each $D_{t,r}$ is a $(x_{\max}\sigma)$-subgaussian random variable adapted to $\{\mathfrak{S}_t\}_{t=1}^n$, since

$$\mathbb{E}[e^{\alpha D_{t,r}} | \mathfrak{S}_{t-1}] \;\le\; \mathbb{E}_{X_t}[e^{\alpha^2 X_{t,r}^2 \sigma^2/2} | \mathfrak{S}_{t-1}] \;\le\; e^{\alpha^2 (x_{max}\sigma)^2/2}.$$

Then, using Lemma EC.1, $\Pr[\mathcal{F}(\lambda_0(\gamma))] \ge 1 - 2d\exp[-(\gamma^2 + 2\log d)/2] = 1 - 2\exp[-\gamma^2/2]$. $\square$

LEMMA EC.3 **(From page 105 of (Bühlmann and Van De Geer 2011))**. *For anly $\lambda_0 \in \mathbb{R}^+$, when $\lambda \ge 2\lambda_0$, on event $\mathcal{F}(\lambda_0)$, we have*

$$2\|\mathbf{X}(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta}_{supp(\beta)^c}\|_1 \le 3\lambda\|\hat{\beta}_{supp(\beta)} - \beta_{supp(\beta)}\|_1.$$

Now we are ready to prove Proposition 1.

Proof of Proposition 1 Let $\lambda_0(\gamma) = 2\sigma x_{\max}\sqrt{(\gamma^2 + 2\log d)/n}$ and let λ be an arbitrary constant such that $\lambda \geq 2\lambda_0(\gamma)$. If both events $\mathcal{F}(\lambda_0(\gamma))$ and $\{\hat{\Sigma}(\mathbf{X}) \in \mathcal{C}(\text{supp}(\beta), \phi)\}$ hold, then

$$
\begin{aligned}
2\|\mathbf{X}(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta} - \beta\|_1 &= 2\|\mathbf{X}(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1 + \lambda\|\hat{\beta}_{\text{supp}(\beta)^c}\|_1 \\
&\leq 4\lambda\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1 \\
&\leq 4\lambda\sqrt{s_0}\|\mathbf{X}(\hat{\beta} - \beta)\|_2/\sqrt{n\phi^2} \\
&\leq \|\mathbf{X}(\hat{\beta} - \beta)\|_2^2/n + 4\lambda^2 s_0/\phi^2\,.
\end{aligned}
$$

Here the three inequalities use Lemma EC.3, the definition of $\mathcal{C}(\text{supp}(\beta), \phi)$ (Definition 2), and the inequality $4uv \leq u^2 + 4v^2$, respectively. Thus, for $\lambda \geq 2\lambda_0(\gamma)$,

$$
\begin{aligned}
\Pr\left\{\|\hat{\beta} - \beta\|_1 \leq \frac{4\lambda s_0}{\phi^2}\right\} &\geq \Pr\left[\mathcal{F}(\lambda_0(\gamma)) \cap \{\hat{\Sigma}(\mathbf{X}) \in \mathcal{C}(\text{supp}(\beta), \phi)\}\right] \\
&\geq \Pr\left[\mathcal{F}(\lambda_0(\gamma))\right] - \Pr\left[\hat{\Sigma}(\mathbf{X}) \notin \mathcal{C}(\text{supp}(\beta), \phi)\right] \\
&\geq 1 - 2\exp\left[-\gamma^2/2\right] - \Pr\left[\hat{\Sigma}(\mathbf{X}) \notin \mathcal{C}(\text{supp}(\beta), \phi)\right]\,.
\end{aligned}
$$

Summarizing, we have shown that,

$$
\lambda \geq 2\lambda_0(\gamma) \Longrightarrow \Pr\left\{\|\hat{\beta} - \beta\|_1 > \frac{4\lambda s_0}{\phi^2}\right\} \leq 2\exp\left[-\gamma^2/2\right] + \Pr\left[\hat{\Sigma}(\mathbf{X}) \notin \mathcal{C}(\text{supp}(\beta), \phi)\right]\,. \tag{EC.1}
$$

Now we choose $\gamma = \gamma(\chi) \equiv \sqrt{2nC_*\chi^2 - 2\log d}$ for a suitable constant C_*, to be determined. Then, the exponent of the error probability becomes $-\gamma(\chi)^2/2 = -nC_*\chi^2 + \log d$. We will now show that $C_* = C_1(\phi)$ will guarantee the condition $\lambda(\chi, \phi) \geq 2\lambda_0(\gamma(\chi))$. In particular,

$$
2\lambda_0(\gamma(\chi)) \;=\; 4\sigma x_{\max}\sqrt{\left[\gamma(\chi)^2 + 2\log d\right]/n} \;=\; 4\sigma x_{\max}\chi\sqrt{2C_*} \;=\; \frac{16\sigma x_{\max} s_0}{\phi^2}\,\underbrace{\frac{\chi\phi^2}{4s_0}}_{\lambda(\chi,\phi)}\,\sqrt{2C_*}\,.
$$

Therefore, for the inequality $\lambda(\chi, \phi) \geq 2\lambda_0(\gamma(\chi))$ to hold, we need $\phi^4 \geq C_*(512 s_0^2\sigma^2 x_{\max}^2)$, which is satisfied by $C_* = C_1(\phi)$. Now, we can invoke (EC.1) for $\lambda = \lambda(\chi, \phi)$, and use the inverse relation $\chi = 4\lambda(\chi, \phi)s_0/\phi^2$ to finish the proof. $\square$

EC.2. Proof of LASSO Tail Inequality for Non-i.i.d. Data

Recall the setup from §4.1 as well as assumptions of Lemma 1. The proof involves showing that $\|\hat{\Sigma}(\mathcal{A}') - \Sigma\|_\infty$ is small with high probability using random matrix theory. Next, we use the Azuma-Hoeffding inequality to show that $\hat{\Sigma}(\mathcal{A}) \in \mathcal{C}(\text{supp}(\beta), \phi_1\sqrt{p}/2)$ with high probability. This result provides a tail inequality for LASSO estimates $\hat{\beta}(\mathcal{A}, \lambda)$, even when part of the data is not generated i.i.d. from $\mathcal{P}_Z$.

EC.2.1. Empirical covariance matrix via random matrix theory

LEMMA EC.4. *Given i.i.d. observations $Z_1,...,Z_n \in \mathbb{R}^d$ from the distribution $\mathcal{P}_Z$ such that all realizations of Z_t satisfy $\|Z_t\|_\infty \le x_{\max}$ for all $t \in [n]$, then for all $w > 0$,*

$$\Pr\left[\|\hat{\Sigma} - \Sigma\|_\infty \ge 2x_{\max}^2 \left(w + \sqrt{2w} + \sqrt{\frac{2\log(d^2+d)}{n}} + \frac{\log(d^2+d)}{n}\right)\right] \le e^{-nw},$$

where $\Sigma \equiv \mathbb{E}_{\mathcal{P}_Z}[ZZ^\top]$ and $\hat{\Sigma} \equiv \sum_{t=1}^n Z_t Z_t^\top / n$.

Proof of Lemma EC.4 First, define the family $\{\gamma_{jk}\}_{1\le j\le k\le d}$ of real-valued functions that take as input random variables $Z \sim \mathcal{P}_Z$. Precisely, for all $1 \le j \le k \le d$,

$$\gamma_{jk}(Z) \equiv \frac{Z(j)Z(k) - \mathbb{E}[Z(j)Z(k)]}{2x_{\max}^2},$$

where $Z(j)$ refers to j^{th} coordinate of vector Z. It is easy to see that each such γ_{jk} satisfies $\mathbb{E}[\gamma_{jk}(Z)] = 0$ and

$$\frac{1}{n}\sum_{t=1}^n \mathbb{E}|\gamma_{jk}(Z_t)|^m \le 1 \ , \quad m = 2, 3, \dots ,$$

implying that the family $\{\gamma_{jk}\}_{1\le j\le k\le d}$ satisfies condition (14.5) in page 489 of Bühlmann and Van De Geer (2011) with $K = 1$. Therefore, we can apply Lemma 14.13 from page 490 of Bühlmann and Van De Geer (2011) and obtain, for all $w > 0$,

$$\Pr\left[\max_{1\le j\le k\le d}\left|\frac{1}{n}\sum_{t=1}^n \gamma_{jk}(Z_t)\right| \ge w + \sqrt{2w} + \sqrt{\frac{2\log(d^2+d)}{n}} + \frac{\log(d^2+d)}{n}\right] \le e^{-nw}.$$

Now, the result follows from the fact that $\|\hat{\Sigma} - \Sigma\|_\infty/(2x_{\max}^2) = \max_{1\le j\le k\le d}|\sum_{t=1}^n \gamma_{jk}(Z_t)|/n$. $\square$

EC.2.2. Compatibility condition for non-i.i.d. samples

Recall $\mathcal{A}$, $\mathcal{A}'$, Σ, β, $\mathbf{Z}$, W, and the corresponding assumptions on them from §4.1. We will first show that $\hat{\Sigma}(\mathcal{A}')$ satisfies the compatibility condition (with respect to supp(β) and an appropriate constant) with high probability.

LEMMA EC.5. *If $\Sigma \in \mathcal{C}(supp(\beta), \phi_1)$ for constant $\phi_1 > 0$ and $\|\Sigma - \Sigma'\|_\infty \le \phi_1^2/(32 s_0)$ holds, then $\Sigma' \in \mathcal{C}(supp(\beta), \phi_1/\sqrt{2})$.*

Proof of Lemma EC.5 The proof follows directly from Corollary 6.8 in page 152 of Bühlmann and Van De Geer (2011). $\square$

LEMMA EC.6. *If the assumptions of Lemma 1 hold, then*

$$\Pr\left[\hat{\Sigma}(\mathcal{A}') \in \mathcal{C}(supp(\beta), \frac{\phi_1}{\sqrt{2}})\right] \ge 1 - e^{-C_2(\phi_1)^2|\mathcal{A}'|}.$$

Proof of Lemma EC.6 Given the assumptions of Lemma 1, we have $|\mathcal{A}'| \geq 3\log(d)/C_2(\phi_1)^2$. Together with $d > 1$, this implies that $\log(d^2+d)/|\mathcal{A}'| \leq C_2^2(\phi_1)$. Therefore, for $w = C_2^2(\phi_1)$, we have,

$$
\begin{aligned}
2x_{\max}^2\left(w+\sqrt{2w}+\sqrt{\frac{2\log(d^2+d)}{|\mathcal{A}'|}}+\frac{\log(d^2+d)}{|\mathcal{A}'|}\right) &\leq 4x_{\max}^2\left[C_2(\phi_1)^2+\sqrt{2}C_2(\phi_1)\right] \\
&\leq 8x_{\max}^2 C_2(\phi_1) \\
&\leq \frac{\phi_1^2}{32s_0},
\end{aligned}
$$

where the last two inequalities follow from the definition of $C_2(\phi_1) = \min[1/2, \phi_1^2/(256s_0x_{\max}^2)]$. Thus, it follows from Lemma EC.4 that

$$
\Pr\left[\|\Sigma - \hat{\Sigma}(\mathcal{A}')\|_\infty \geq \frac{\phi_1^2}{32s_0}\right] \leq e^{-C_2(\phi_1)^2|\mathcal{A}'|}
$$

The result then follows directly from Lemma EC.5. □

LEMMA EC.7. *Given the assumptions of Lemma 1, if we have $\hat{\Sigma}(\mathcal{A}') \in \mathcal{C}(supp(\beta), \phi_1')$ for some constant $\phi_1' > 0$, then $\hat{\Sigma}(\mathcal{A}) \in \mathcal{C}(supp(\beta), \phi_1'\sqrt{|\mathcal{A}'|/|\mathcal{A}|})$.*

Proof of Lemma EC.7 By definition, we can write

$$
\hat{\Sigma}(\mathcal{A}) = \frac{|\mathcal{A}'|}{|\mathcal{A}|}\hat{\Sigma}(\mathcal{A}') + \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}\setminus\mathcal{A}'} Z_t Z_t^\top = \frac{|\mathcal{A}'|}{|\mathcal{A}|}\hat{\Sigma}(\mathcal{A}') + \frac{|\mathcal{A}\setminus\mathcal{A}'|}{|\mathcal{A}|}\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}').
$$

Then, for all v satisfying $\|v_{\mathrm{supp}(\beta)^c}\|_1 \leq 3\|v_{\mathrm{supp}(\beta)}\|_1$,

$$
v^\top\hat{\Sigma}(\mathcal{A})v = \frac{|\mathcal{A}'|}{|\mathcal{A}|}v^\top\hat{\Sigma}(\mathcal{A}')v + \frac{|\mathcal{A}\setminus\mathcal{A}'|}{|\mathcal{A}|}v^\top\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')v \geq \frac{|\mathcal{A}'|}{|\mathcal{A}|}\frac{\phi_1'^2\,\|v_{\mathrm{supp}(\beta)}\|_1^2}{s_0},
$$

using the fact that $\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')$ is positive semi-definite. □

Now we have all the ingredients to complete the proof of Lemma 1.

Proof of Lemma 1: Applying Lemmas EC.6 and EC.7 implies that

$$
\hat{\Sigma}(\mathcal{A}) \in \mathcal{C}\left(\mathrm{supp}(\beta), \frac{\phi_1}{\sqrt{2}}\sqrt{\frac{|\mathcal{A}'|}{|\mathcal{A}|}}\right),
$$

with probability at least $1-\exp[-C_2(\phi_1)^2|\mathcal{A}'|]$. This implies that $\hat{\Sigma}(\mathcal{A}) \in \mathcal{C}(\mathrm{supp}(\beta), \phi_1\sqrt{p}/2)$ with probability at least $1-\exp[-p\,C_2(\phi_1)^2|\mathcal{A}|/2]$. Applying Proposition 1 with compatibility constant $\phi = \phi_1\sqrt{p}/2$ yields the result. □

EC.3. Proof of LASSO Tail Inequality for Forced-Sample Estimator

In this section, we prove a tail inequality for the forced sample estimator $\hat{\beta}(\mathcal{T}_{i,t}, \lambda_1)$ by applying Lemma 1. Recall that at each $t \in \mathcal{T}_{i,t}$, we draw a random covariate vector X_t, sampled i.i.d. from $\mathcal{P}_X$, and play arm i. Moreover, we assumed that $\Sigma_i \in \mathcal{C}(\text{supp}(\beta_i), \phi_0)$ where $\Sigma_i = \mathbb{E}_{X\sim\mathcal{P}_{X|X\in U_i}}[XX^\top]$ and also that $\Pr[X_t \in U_i] \geq p_*$.

Lemma EC.8. *If $t \geq (Kq)^2$, then $(1/2)q\log t \leq |\mathcal{T}_{i,t}| \leq 6q\log t$.*

Proof of Lemma EC.8 Define the n^{th} round of forced sampling of all the arms

$$L_n \equiv \{(2^n-1)Kq+1, \ldots, (2^n)Kq\},$$

for $n \geq 0$. By construction, arm i is sampled $|\mathcal{T}_i \cap L_n| = q$ times during L_n, so

$$\left|\mathcal{T}_i \cap \left(\bigcup_{r=0}^{n-1} L_r\right)\right| = nq\,.$$

Then for each $t \in L_n$, $nq \leq |\mathcal{T}_{i,t}| \leq (n+1)\,q$. To show the lower bound, note that for $t \in L_n$, we have $t \leq (2^n)Kq$, i.e., $\log_2[t/(Kq)] \leq n$. Therefore, using $t \geq (Kq)^2$,

$$|\mathcal{T}_{i,t}| \geq nq \geq q\,\log_2\frac{t}{Kq} \geq q\,(\log t - \log Kq) \geq (1/2)q\log t\,.$$

To show the upper bound, note that for $t \in L_n$, $t \geq (2^n-1)\,Kq$, i.e., $n \leq \log_2[1+t/(Kq)]$, so

$$|\mathcal{T}_{i,t}| \leq (n+1)q \leq \left[\log_2\left(\frac{t}{Kq}+1\right)+1\right]q \leq \frac{\log(2t+2\sqrt{t})}{\log 2}q \leq 6q\log t\,. \quad \square$$

Lemma EC.9. *Let $\mathcal{T}'_{i,t} \subset \mathcal{T}_{i,t}$ be the set of all $r \in \mathcal{T}_{i,t}$ such that $X_r \in U_i$. Then for each $r \in \mathcal{T}_{i,t}$ we have $r \in \mathcal{T}'_{i,t}$ independently with probability at least p_*. In addition, $\{X_r\}_{r\in\mathcal{T}'_{i,t}}$ are i.i.d. from $\mathcal{P}_{X|X\in U_i}$.*

Proof of Lemma EC.9 By construction, for each $r \in \mathcal{T}_{i,t}$, X_r is drawn i.i.d. from $\mathcal{P}_X$ and therefore with probability at least p_*, $X_r \in U_i$, i.e., $s \in \mathcal{T}'_{i,t}$. Also, note that the events $X_r \in U_i$ are independent for different values of r since the original sequence $\{X_r\}_{s\in\mathcal{T}_{i,t}}$ is i.i.d., implying that each $r \in \mathcal{T}'_{i,t}$, X_r is an i.i.d. sample of $\mathcal{P}_{X|X\in U_i}$. $\square$

Using Lemma EC.9 we see that the inclusion of each member of $\mathcal{T}_{i,t}$ in $\mathcal{T}'_{i,t}$ is a Bernoulli i.i.d. random variable with mean at least p_*. Therefore, we get the following result using Chernoff bound.

Lemma EC.10. *If $t \geq (Kq)^2$, for $\mathcal{T}_{i,t}$ and $\mathcal{T}'_{i,t}$ defined as in Lemma EC.9 the following holds*

$$\Pr\left[\frac{|\mathcal{T}'_{i,t}|}{|\mathcal{T}_{i,t}|} \geq \frac{p_*}{2}\right] \geq 1 - \frac{2}{t^4}\,.$$

Proof of Lemma EC.10 We use the following version of the Chernoff inequality (Corollary A.1.14 in page 268 of Alon and Spencer 1992, for $\varepsilon = 1/2$ and $c_\varepsilon \approx 0.1082$): Letting y be the sum of mutually independent indicator random variables with $\mu = \mathbb{E}[y]$, we have that

$$\Pr\left[\,|y-\mu| > \mu/2\,\right] < 2\exp\left[-0.1\mu\right].$$

Applying this to indicator random variables $\mathbb{I}(r \in \mathcal{T}'_{i,i})$ for all $r \in \mathcal{T}_{i,t}$ and using

$$\mu = \mathbb{E}\left[\sum_{r\in\mathcal{T}_{i,t}} \mathbb{I}(r \in \mathcal{T}'_{i,t})\right] \geq p_* |\mathcal{T}_{i,t}|\,,$$

we can write that

$$\Pr\left[\,|\mathcal{T}'_{i,t}| < (p_*/2)|\mathcal{T}_{i,t}|\,\right] < 2e^{-\frac{p_*}{10}|\mathcal{T}_{i,t}|}.$$

Next, using Lemma EC.8, $t \geq (Kq)^2$, $q \geq 4q_0$, and the definition of q_0 from §3.3 we have

$$\Pr\left[\,|\mathcal{T}'_{i,t}| < (p_*/2)|\mathcal{T}_{i,t}|\,\right] < 2e^{-(p_*/5)q_0\log t} \leq \frac{2}{t^4}. \quad \square$$

Now we are ready to prove Proposition 2.

Proof of Proposition 2 By construction, and the definition of q_0 from §3.3,

$$|\mathcal{T}_{i,t}| \geq (1/2)q\log t \geq 2q_0\log t \geq \frac{6\log(d)}{p_*\,C_2(\phi_0)^2}.$$

Then combining Lemma EC.9-EC.10 and Lemma 1, with $\mathcal{P}_Z = \mathcal{P}_{X|X\in U_i}$, $\chi = h/(4x_{\max})$, $p = p_*$, and $\lambda_1 = \lambda(h/(4x_{\max}, \phi_0\sqrt{p_*}/2)$ we obtain

$$\begin{aligned}
\Pr\left[\|\hat\beta\left(\mathcal{T}_{i,t},\lambda_1\right) - \beta_i\|_1 > \frac{h}{4x_{\max}}\right] &\leq 2e^{-C_1\left(\frac{\phi_0\sqrt{p_*}}{2}\right)2q_0\log t\frac{h^2}{16x_{\max}^2}+\log d} + e^{-p_* C_2(\phi_0)^2 q_0\log t} + \frac{2}{t^4}\\
&\leq 2e^{-p_*^2 C_1(\phi_0)q_0\log t\frac{h^2}{128x_{\max}^2}+\log d} + \frac{1}{t^4} + \frac{2}{t^4}\\
&\leq \frac{5}{t^4}.
\end{aligned}$$

The last two inequalities use the definition of q_0 and the fact that $t \geq (Kq)^2$ to show that the exponent of each term on the right hand side is at most $-4\log t$. $\square$

EC.4. Proof of LASSO Tail Inequality for All-Sample Estimator

In this section, we prove the tail inequality for the all-sample estimator $\hat\beta(\mathcal{S}_{i,t}, \lambda_{2,t})$ for arms in $\mathcal{K}_{opt}$. The approach mirrors the steps taken in Appendix EC.3. However, there is an additional complication due to the correlation between rows of $\mathbf{X}(\mathcal{S}_{i,t})$ that was discussed in §4.3. Recall the events A_t defined in Eq. (3).

LEMMA EC.11. *For each $i \in [K]$, if the events $X_t \in U_i$ and A_{t-1} hold, and $t \notin \cup_{j\in[K]}\mathcal{T}_{j,t}$, the LASSO Bandit uses the forced-sample estimator $\hat{\beta}(\mathcal{T}_{i,t-1}, \lambda_1)$ to arrive at $\hat{\mathcal{K}} = \{i\}$, implying that it plays the optimal arm at time t.*

Proof of Lemma EC.11 Since $X_t \in U_i$, we know

$$X_t^\top \beta_i \geq h + \max_{j\neq i} X_t^\top \beta_j \,.$$

Then, for any $j \in [K] \setminus \{i\}$, since A_{t-1} holds,

$$\begin{aligned} X_t^\top \left[\hat{\beta}(\mathcal{T}_{i,t-1}) - \hat{\beta}(\mathcal{T}_{j,t-1})\right] &= X_t^\top \left[\hat{\beta}(\mathcal{T}_{i,t-1}) - \beta_i\right] - X_t^\top \left[\hat{\beta}(\mathcal{T}_{j,t-1}) - \beta_j\right] + X_t^\top (\beta_i - \beta_j) \\ &\geq -x_{\max}\frac{h}{4x_{\max}} - x_{\max}\frac{h}{4x_{\max}} + h \\ &\geq h/2 \,. \end{aligned}$$

Thus, at time t, $\hat{\mathcal{K}} = \{i\}$, i.e., the LASSO Bandit will play arm i. □

LEMMA EC.12. *For all t with $t \geq (Kq)^2$ the event A_t occurs with probability at least $1 - 5K/t^4$.*

Proof of Lemma EC.12 For each $i \in [K]$ and all $t \geq (Kq)^2$, we have from Proposition 2,

$$\Pr\left[\|\hat{\beta}(\mathcal{T}_{i,t}, \lambda_1) - \beta_i\|_1 > \frac{h}{4x_{\max}}\right] \leq \frac{5}{t^4} \,.$$

Taking a union bound over all K arms gives us the result. □

LEMMA EC.13. *Let $i \in [K]$. Recall from §4.3 that $\mathcal{S}'_{i,t} \subset [t]$ is the set of all time periods r such the events $X_r \in U_i$ and A_{r-1} hold and we are not forced-sampling any arm $j \in [K]$. Then the following properties are satisfied.*

(1) The set of random variables $\{X_r \mid r \in \mathcal{S}'_{i,t}\}$ are i.i.d. from distribution $\mathcal{P}_{X|X\in U_i}$.

(2) For each $r \in [t] \setminus \cup_{j\in[K]}\mathcal{T}_{j,t}$, we have $r \in \mathcal{S}'_{i,t}$ with probability at least $p_*/2$ when $t \geq (Kq)^2$.

(3) $\mathcal{S}'_{i,t} \subset \mathcal{S}_{i,t}$.

Proof of Lemma EC.13 For (1), since A_{r-1} is only a function of samples in $\mathcal{T}_{i,r-1}$, A_{r-1} is independent of X_r. Therefore, random variables $\{X_r | A_{r-1} \text{ holds}\}$ are i.i.d. samples from $\mathcal{P}_X$. Now, the presence of each X_r in U_i is simply rejection sampling; thus, using the fact that $r \notin \cup_{j\neq i}\mathcal{T}_{j,t}$ is deterministic, for each $r \in \mathcal{S}'_{i,t}$, X_r is distributed i.i.d. from $\mathcal{P}_{X|X\in U_i}$. For (2), we know that $X \in U_i$ with probability at least p_* and Lemma EC.12 implies that A_{r-1} holds with probability at least $1 - 5K(r-1)^{-4}$ when $(r-1) \geq (Kq)^2$. Note that $(r-1) \geq (Kq)^2 \geq 16K^2$ (since $q \geq 4\lceil q_0 \rceil \geq 4$), which implies that A_{r-1} holds with probability at least $1 - 5K(r-1)^{-4} \geq 1/2$. Then, $r \in \mathcal{S}'_{i,t}$ with probability at least $p_*/2$. Finally, for (3), from Lemma EC.11, we know that for $X_r \sim \mathcal{P}_X$, if events $X_r \in U_i$ and A_{r-1} holds and $r \notin \cup_{j\in[K]}\mathcal{T}_{j,t}$, then $r \in \mathcal{S}_{i,t}$, so $\mathcal{S}'_{i,t} \subset \mathcal{S}_{i,t}$. □

LEMMA EC.14. *If $t \geq C_5$, for $\mathcal{S}'_{i,t}$ defined as in §4.3 the following holds*

$$\Pr\left[\,|\mathcal{S}'_{i,t}| \geq t\,p_*/4\,\right] \geq 1 - e^{tp_*^2/36}\,.$$

Proof of Lemma EC.14 Note that we need to take a more refined approach than in Lemma EC.10 since the events $r \in \mathcal{S}'_{i,t}$, $r \in [t]$, are not independent. By definition of $\mathcal{S}'_{i,t}$ we have for all $r \in [t]/\mathcal{T}_{i,t}$,

$$\mathbb{I}(r \in \mathcal{S}'_{i,t}) = \mathbb{I}(A_{r-1}) \cdot \mathbb{I}(X_r \in U_i) \cdot \mathbb{I}(r \notin \cup_{j\in[K]} \mathcal{T}_{j,t})\,.$$

Note that by construction of the forced-sampling sets, for each round of forced-sampling $n \in \{0,1,2,...\}$, the time periods $t \in [2^n Kq+1, (2^{n+1}-1)Kq]$ are played contiguously without any forced-sampling (or updates to the forced-sampling estimators). Then, for any time $t \notin \{\mathcal{T}_{j,t'}\}_{j\in[K],t'\in[T]}$ where we do not perform forced sampling, let n_t be the largest integer satisfying $t > 2^{n_t+1}Kq$. Given arm i, we can define

$$M_{i,t} \equiv \sum_{r=2^{n_t}Kq+1}^{(2^{n_t+1}-1)Kq} \mathbb{I}(r \in \mathcal{S}'_{i,t}) + \sum_{r=2^{n_t+1}Kq+1}^{t} \mathbb{I}(r \in \mathcal{S}'_{i,t}) \;<\; \sum_{r=1}^{t} \mathbb{I}(r \in \mathcal{S}'_{i,t}) = |\mathcal{S}'_{i,t}|\,.$$

In other words, $M_{i,t}$ is strictly smaller than $|S_{i,t}|$, because it only considers the time interval

$$V_t = [2^{n_t}Kq+1, (2^{n_t+1}-1)Kq] \cup [2^{n_t+1}Kq+1, t] \;\subset\; [1,t]\,.$$

By construction, the time interval $[2^{n_t}Kq+1, (2^{n_t+1}-1)Kq]$ contains no forced-sampling, and thus the forced-sample estimator is never updated in this period; the same holds for $[2^{n_t+1}Kq+1, t]$. As a result, note that the event $A_r = A_{2^{n_t}Kq}$ for all times r in the first interval, and $A_r = A_{2^{n_t+1}Kq}$ for all times r in the second interval. Thus, we can expand

$$\begin{aligned}
M_{i,t} = &\sum_{r=2^{n_t}Kq+1}^{(2^{n_t+1}-1)Kq} \mathbb{I}(A_{2^{n_t}Kq}) \cdot \mathbb{I}(X_r \in U_i) \cdot \mathbb{I}(r \notin \cup_{j\in[K]} \mathcal{T}_{j,t}) \\
&+ \sum_{r=2^{n_t+1}Kq+1}^{t} \mathbb{I}(A_{2^{n_t+1}Kq}) \cdot \mathbb{I}(X_r \in U_i) \cdot \mathbb{I}(r \notin \cup_{j\in[K]} \mathcal{T}_{j,t}) \\
\geq\, &\mathbb{I}(A_{2^{n_t}Kq}) \cdot \mathbb{I}(A_{2^{n_t+1}Kq}) \cdot \sum_{r\in V_t} \mathbb{I}(X_r \in U_i)\,,
\end{aligned}$$

We are then left with a sum over independent random variables drawn from $\mathcal{P}_X$, each occurring with probability p_* by Assumption 3. The definition of n_t as the largest integer such that $t > 2^{n_t+1}Kq$ implies that

$$\begin{aligned}
|V_t| &= (t - 2^{n_t+1}Kq) + (2^{n_t+1}Kq - Kq - 2^{n_t}Kq) \;=\; t - 2^{n_t}Kq - Kq \\
&> t - t/2 - Kq \;=\; t/2 - Kq \\
&> 3t/8\,,
\end{aligned}$$

where we note that $t > 8Kq$ for $t \geq C_5$. Next, note that since $t \geq C_5 > 4(Kq)^2$ and $t < 2^{n_t+2}Kq$ by definition of n_t, we can write

$$2^{n_t}Kq > t/4 \geq (Kq)^2\,,$$

allowing us to apply Lemma EC.12 for $r = 2^{n_t}Kq$ and $r = 2^{n_t+1}Kq$. A simple union bound yields

$$\begin{aligned}\Pr[A_{2^{n_t}Kq} \text{ and } A_{2^{n_t+1}Kq}] &\geq 1 - \frac{5K}{(t/4)^2} - \frac{5K}{(t/2)^2} \\ &\geq 1 - \frac{100}{K^3q^4} > 8/9\,,\end{aligned}$$

by the definition of q. Then,

$$\begin{aligned}\mathbb{E}[M_{i,t}] &\geq \Pr[A_{2^{n_t}Kq} \text{ and } A_{2^{n_t+1}Kq}] \cdot p_* |V_t| \\ &\geq \frac{tp_*}{3}\,,\end{aligned}$$

Applying the Hoeffding inequality, we have that

$$\begin{aligned}\Pr\left[M_{i,t} < \mathbb{E}[M_{i,t}] - \eta\right] &\leq e^{-2\eta^2/|V_t|} \\ &\leq e^{-4\eta^2/t}\,,\end{aligned}$$

Plugging in $\eta = tp_*/12$, we have

$$\Pr\left[M_{i,t} < tp_*/4\right] \leq e^{-tp_*^2/36}\,.$$

Applying the fact that $M_{i,t} \leq |S'_{i,t}|$, we have that

$$\Pr\left[|\mathcal{S}'_{i,t}| < \frac{t\,p_*}{4}\right] \leq e^{-tp_*^2/36}\,.$$

□

Now, we are ready to prove Proposition 3.

Proof of Proposition 3: From Lemma EC.14, for $t \geq C_5$ we have $|\mathcal{S}'_{i,t}| \geq p_* t/4$ with probability $1 - \exp[-t\,p_*^2/128]$. Therefore, using a union bound, we can apply Lemma 1 with $p = p_*/2$, $\mathcal{A} = \mathcal{S}_{i,t}$, $\mathcal{A}' = \mathcal{S}'_{i,t}$, and $\lambda = \chi\phi_0^2 p_*/(32 s_0)$ to arrive at,

$$\begin{aligned}&\Pr\left[\|\hat{\beta}\left(\mathcal{S}_{i,t},\lambda\right) - \beta_i\|_1 > \chi\right] \\ &\qquad \leq 2\exp\left[-C_1\left(\frac{\phi_0\sqrt{p_*/2}}{2}\right)\frac{tp_*}{4}\chi^2 + \log d\right] + \exp\left[-\frac{tp_*^2 C_2(\phi_0)^2}{16}\right] + \exp\left[-\frac{tp_*^2}{36}\right] \\ &\qquad\qquad \leq 2\exp\left[-\frac{tp_*^3 C_1(\phi_0)}{256}\chi^2 + \log d\right] + 2\exp\left[-\frac{tp_*^2 C_2(\phi_0)^2}{32}\right]\,,\end{aligned}$$

where the last inequality uses $C_2(\phi_0) \le 1/2$. Note that the condition $|\mathcal{A}'|/|\mathcal{A}| \ge p_*/4$ holds when $|\mathcal{S}'_{i,t}| \ge p_* t/4$ (since $|\mathcal{A}| \le t$). Also, the condition $|\mathcal{S}_{i,t}| \ge 6\log(d)/(p_* C_2(\phi_0)^2)$ is satisfied, using $|\mathcal{S}_{i,t}| \ge |\mathcal{S}'_{i,t}| \ge p_* t/4$, $t \ge C_5$, $q \ge 4q_0$, and the definition of q_0. Taking

$$\chi = 16\sqrt{\frac{\log t + \log d}{t p_*^3 C_1(\phi_0)}},$$

gives us the desired result. Note that this choice of χ implies $\lambda = \lambda_{2,t}$. $\square$

EC.5. Bounding the Regret in the High-Dimensional Setting

Recall from our proof strategy in §4.4, that we divide our time steps $[T]$ into three groups:

(a) Initialization ($t \le C_5$) or forced sampling ($t \in \mathcal{T}_{i,T}$ for some $i \in [K]$).

(b) Times $t > C_5$ when the event A_{t-1} does not hold.

(c) Times $t > C_5$ when the event A_{t-1} holds and we do not perform forced sampling.

These groups may not be disjoint but their union contains $[T]$.

We now compute an upper bound on the regret for time periods in each group (a)-(c) and sum the results. First, the following lemma gives the worst-case regret for time periods in (a).

Lemma EC.15. *The cumulative expected regret of the LASSO Bandit from initialization ($t < C_5$) and forced sampling ($t \in \mathcal{T}_{i,t}$ for some $i \in [K]$) up to time T is at most*

$$2Kqbx_{\max}(6\log T + C_5).$$

Proof of Lemma EC.15: From Lemma EC.8, at most $6Kq\log T$ forced samples occur up to time T. We also have C_5 initialization samples. Using Cauchy-Schwarz, we can bound the worst-case regret in each time period by $\max_{i,j} X^\top(\beta_i - \beta_j) \le 2bx_{\max}$. The result follows directly. $\square$

Before moving to time periods in (b)-(c), we state the following helpful lemma:

Lemma EC.16. *If f is a monotone decreasing and integrable function on the range $[r-1, s]$, then*

$$\sum_{t=r}^{s} f(t) \le \int_{r-1}^{s} f(t)dt.$$

Proof of Lemma EC.16:

$$\sum_{t=r}^{s} f(t) \le \sum_{t=r}^{s} \int_{t-1}^{t} f(\tilde{t})d\tilde{t} = \int_{r-1}^{s} f(t)dt. \quad \square$$

Next, we find the worst-case regret from time periods in (b) at time T.

Lemma EC.17. *The cumulative expected regret of LASSO Bandit from time periods $C_5 < t \le T$ where A_{t-1} does not hold is at most $2Kbx_{\max}$.*

Proof of Lemma EC.17 From Lemma EC.12, the probability that A_{t-1} does not hold is at most $5Kt^{-4} \leq Kt^{-3}$ since $t \geq C_5 > 5$. Now we can sum this quantity for $t \in [C_5, T-1]$. Using Lemma EC.16,

$$\sum_{t=C_5}^{T-1} \frac{K}{t^3} \leq K \int_1^T \frac{1}{t^3}\, dt \leq \frac{K}{2}(1 - \frac{1}{T^2}) \leq K\,.$$

As before, the worst-case regret at time t is $2b\,x_{\max}$, and the result follows. □

Before analyzing the regret from group (c), we show that if the event A_{t-1} holds, then the set $\hat{\mathcal{K}}$ chosen by the forced-sample estimator has two desirable properties: (i) it contains the true optimal arm, and (ii) it does not contain any sub-optimal arms. Thus, we can apply the convergence properties of the all-sample estimator (which only hold among optimal arms) to analyze the regret from choosing an arm within $\hat{\mathcal{K}}$.

Lemma EC.18. *If A_{t-1} holds, then the set $\hat{\mathcal{K}}$ contains the optimal arm $i^* = \arg\max_{i\in[K]} X_t^\top \beta_i$ and no sub-optimal arms from the set $\mathcal{K}_{sub}$.*

Proof of Lemma EC.18 To simplify notation, we call our forced-sample arm estimators $\hat{\beta}(\mathcal{T}_{i,t-1}, \lambda_1)$ at time t as $\hat{\beta}_i$. Since A_{t-1} holds, we have that for any pair of arms $i, j \in [K]$,

$$\begin{aligned} X_t^\top \hat{\beta}_i - X_t^\top \hat{\beta}_j &= X_t^\top(\hat{\beta}_i - \beta_i) + X_t^\top(\beta_j - \hat{\beta}_j) + X^\top(\beta_i - \beta_j) \\ &\leq h/2 + X_t^\top(\beta_i - \beta_j)\,. \end{aligned}$$

Thus, if we let $i = \arg\max_{\ell\in[K]} X_t^\top \hat{\beta}_\ell$ and $j = i^*$, we see that $X_t^\top(\hat{\beta}_i - \hat{\beta}_{i^*}) \leq h/2$ since $X_t^\top(\beta_i - \beta_{i^*}) < 0$ (by definition of i^*). Then, the optimal arm $i^* \in \hat{\mathcal{K}}$.

On the other hand, consider $i = \arg\max_{\ell\in[K]} X_t^\top \hat{\beta}_\ell$ and any sub-optimal arm $j \in \mathcal{K}_{sub}$. Then, $X_t^\top \hat{\beta}_i - X_t^\top \hat{\beta}_j \geq X_t^\top \hat{\beta}_{i^*} - X_t^\top \hat{\beta}_j$, and furthermore, since A_{t-1} holds:

$$\begin{aligned} X_t^\top \hat{\beta}_{i^*} - X_t^\top \hat{\beta}_j &= X_t^\top(\hat{\beta}_{i^*} - \beta_{i^*}) + X_t^\top(\beta_j - \hat{\beta}_j) + X^\top(\beta_{i^*} - \beta_j) \\ &\geq -h/2 + X_t^\top(\beta_{i^*} - \beta_j)\,. \end{aligned}$$

Recall that for every sub-optimal arm $j \in \mathcal{K}_{sub}$, we have $X_t^\top \beta_j < X_t^\top \beta_{i^*} - h$. Then, we can write

$$\begin{aligned} X_t^\top(\hat{\beta}_i - \hat{\beta}_j) &\geq X_t^\top \hat{\beta}_{i^*} - X_t^\top \hat{\beta}_j \\ &> -h/2 + h \;=\; h/2\,. \end{aligned}$$

Thus, $j \notin \hat{\mathcal{K}}$ for every sub-optimal arm $j \in \mathcal{K}_{sub}$. □

Finally, the next two lemmas bound the regret from time periods in (c) by separately summing over expected regret when the all-sample tail inequality does and does not hold. We simplify our notation by calling our all-sample estimators $\hat{\beta}(\mathcal{S}_{i,t-1}, \lambda_{2,t-1})$ at time t as $\hat{\beta}_i$, where we recall $\lambda_{2,t} = [\phi_0^2/(2s_0)]\sqrt{(\log t + \log d)/(p_* C_1 t)}$.

LEMMA EC.19. *If $t > C_5$, A_t holds, and we do not perform forced sampling, then the expected regret at time $t+1$ is bounded by*

$$(8Kbx_{\max})/t + 8Kbx_{\max}\exp\left[-(p_*^2C_2(\phi_0)^2t)/32\right] + C_3(\phi_0,p_*)\cdot(\log t+\log d)/t\,,$$

where $C_3(\phi_0,p_*) = 1024KC_0x_{\max}^2/(p_*^3C_1(\phi_0))$.

Proof of Lemma EC.19 Without loss of generality, assume that arm 1 is optimal: $\arg\max_{i\in[K]} X_{t+1}^\top\beta_i = 1$. Then, the expected regret at time $t+1$ is given by

$$r_{t+1} = \mathbb{E}\left(\sum_{i\in\hat{\mathcal{K}}} X_{t+1}^\top(\beta_1-\beta_i)\mathbb{I}[\text{choose arm } i]\right) \le \mathbb{E}\left(\sum_{i\in\hat{\mathcal{K}}} X_{t+1}^\top(\beta_1-\beta_i)\mathbb{I}\left[X_{t+1}^\top\hat\beta_i \ge X_{t+1}^\top\hat\beta_1\right]\right)\,,$$

where the last inequality uses the fact that event $\{i = \arg\max_{j\in[K]} X_{t+1}^\top\hat\beta_j\}$ is a subset of the event $\{X_{t+1}^\top\hat\beta_i \ge X_{t+1}^\top\hat\beta_1\}$, and that $X_{t+1}^\top(\beta_1-\beta_i)\ge 0$ (since we have assumed that arm 1 is optimal). Thus, we can bound r_{t+1} through the regret incurred by each arm in $\hat{\mathcal{K}}$ with respect to the optimal arm independently of the other arms. We now define the event $B_i \equiv \left\{X_{t+1}^\top(\beta_1-\beta_i) > 2\delta x_{\max}\right\}$, where we take $\delta \equiv 16\sqrt{(\log t+\log d)/(p_*^3C_1t)}$. Then, we can write

$$\begin{aligned} r_{t+1} \le\ & \mathbb{E}\left(\sum_{i\in\hat{\mathcal{K}}} X_{t+1}^\top(\beta_1-\beta_i)\mathbb{I}\left[(X_{t+1}^\top\hat\beta_i \ge X_{t+1}^\top\hat\beta_1)\cap B_i\right]\right) \\ & + \mathbb{E}\left(\sum_{i\in\hat{\mathcal{K}}} X_{t+1}^\top(\beta_1-\beta_i)\mathbb{I}\left[(X_{t+1}^\top\hat\beta_i \ge X_{t+1}^\top\hat\beta_1))\cap B_i^c\right]\right)\,, \end{aligned}$$

which by definition of B_i and using $X_{t+1}^\top(\beta_1-\beta_i)\le 2b\,x_{\max}$ gives

$$r_{t+1} \le 2b\,x_{\max}\mathbb{E}\left[\sum_{i\in\hat{\mathcal{K}}}\mathbb{I}[(X_{t+1}^\top\hat\beta_i \ge X_{t+1}^\top\hat\beta_1)\cap B_i]\right] + 2\delta\,x_{\max}\mathbb{E}\left[\sum_{i\in\hat{\mathcal{K}}}\mathbb{I}(B_i^c)\right]\,, \tag{EC.2}$$

Note that the intersection of event B_i and the event of choosing arm $i\ne 1$ implies that

$$0 \ge X_{t+1}^\top\hat\beta_1 - X_{t+1}^\top\hat\beta_i > X_{t+1}^\top\left(\hat\beta_1-\beta_1\right) + X_{t+1}^\top\left(\beta_i-\hat\beta_i\right) + 2\delta\,x_{\max}\,.$$

Thus, it must be that either $X_{t+1}^\top\left(\hat\beta_1-\beta_1\right) < -\delta\,x_{\max}$ or $X_{t+1}^\top\left(\beta_i-\hat\beta_i\right) < -\delta\,x_{\max}$. Therefore,

$$\begin{aligned} \Pr\left[(X_{t+1}^\top\hat\beta_i \ge X_{t+1}^\top\hat\beta_1)\cap B_i\right] &\le \Pr\left[\|\beta_1-\hat\beta_1\|_1 > \delta\right] + \Pr\left[\|\hat\beta_i-\beta_i\|_1 > \delta\right] \\ &\le \frac{4}{t} + 4\exp\left[-\frac{p_*^2C_2^2}{32}t\right]\,, \end{aligned} \tag{EC.3}$$

using a union bound and the tail inequality for the all sample estimator.

We can also bound $\Pr[B_i^c]$ using Assumption 2 on the margin condition: $\Pr[B_i^c] = \Pr\left[X_{t+1}^\top(\beta_1-\beta_i) \le 2\delta\, x_{\max}\right] \le 2C_0\delta\, x_{\max}$. Using this and Eq. (EC.3) in Eq. (EC.2) we obtain

$$\begin{aligned} r_{t+1} &\le K\left\{\frac{8b\,x_{\max}}{t} + 8bx_{\max}\exp\left[-\frac{p_*^2C_2^2}{32}t\right] + 4C_0\delta^2x_{\max}^2\right\} \\ &\le \frac{8Kbx_{\max}}{t} + 8Kbx_{\max}\exp\left[-\frac{p_*^2C_2^2}{32}t\right] + C_3\cdot\frac{\log t + \log d}{t}. \quad \square \end{aligned}$$

Lemma EC.20. *The cumulative expected regret from using the all-sample estimator up to time T is bounded by*

$$[8Kbx_{\max} + C_3(\phi_0,p_*)\log d]\log T + C_3(\phi_0,p_*)\,(\log T)^2 + C_4(\phi_0,p_*)\,,$$

where $C_4(\phi_0,p_*) = (8Kbx_{\max})/(1-\exp\left[-\frac{p_*^2C_2(\phi_0)^2}{32}\right])$.

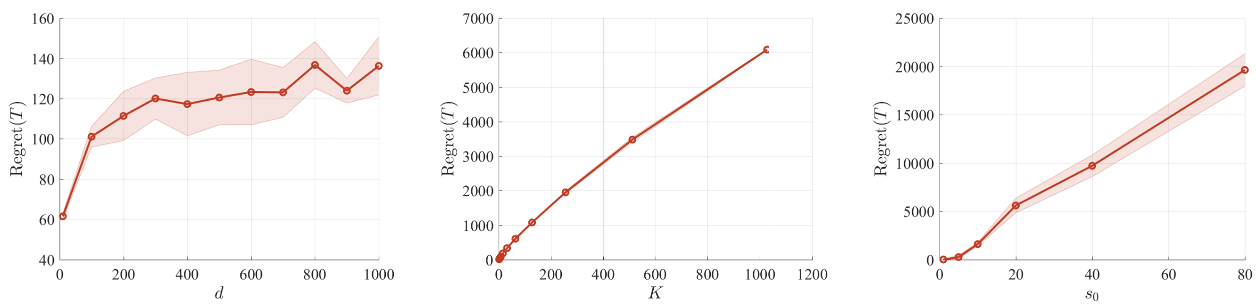
Proof of Lemma EC.20 We sum regret from Lemma EC.19

$$\begin{aligned} &\sum_{t=C_5}^{T-1}\left[(8Kbx_{\max})/t + 8Kbx_{\max}\exp\left[-\frac{p_*^2C_2(\phi_0)^2}{32}t\right] + C_3(\phi_0,p_*)(\log t + \log d)/t\right] \\ &\qquad \le [8Kbx_{\max} + C_3(\phi_0,p_*)\log d]\log T + C_3(\phi_0,p_*)\,(\log T)^2 + C_4(\phi_0,p_*)\,. \quad \square \end{aligned}$$

EC.6. Additional Simulations

EC.6.1. Dependence on K, d, and s_0

First, we study how the LASSO Bandit's cumulative expected regret scales as a function of each parameter K, d, and s_0. The results (see Figure EC.1) show that the regret appears to grow logarithmically with d, but almost linearly with K and s_0.



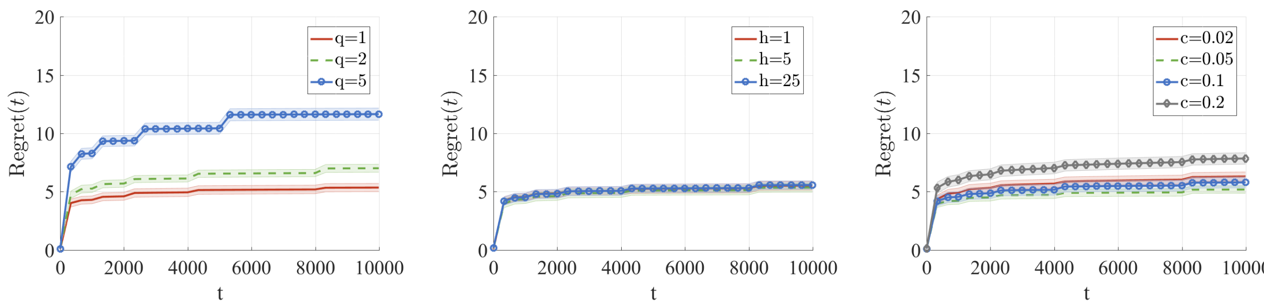
(a) $K=10$, $s_0=2$ (b) $d=20$, $s_0=2$ (c) $d=500$, $K=2$

Figure EC.1 These plots show how the regret of the LASSO Bandit scales as any single parameter d, K, s_0 is varied while the others are fixed.

EC.6.2. Robustness to Algorithm Inputs

We now study the cumulative expected regret of the LASSO Bandit while varying any one of: (i) the forced sampling parameter $q \in \{1, 2, 5\}$, (ii) the localization parameter $h \in \{1, 5, 25\}$, and (iii) the regularization coefficient $c \in \{0.02, 0.05, 0.1, 0.2\}$. We only focus on scenario (a) from above.

The results are computed over $T = 10,000$ time steps and averaged over 30 trials (see Figure EC.2). We find that the cumulative regret performance is not substantially impacted despite experimenting with the parameters by up to an order of magnitude; this suggests that the LASSO Bandit is robust, which is important since input parameters are likely to be misspecified in practice.



(a) (b) (c)

Figure EC.2 Cumulative regret for the LASSO Bandit for varying values of (a) the forced sampling parameter q, (b) the localization parameter h, and (c) the coefficient c of the regularization parameters.

EC.6.3. Nonlinear Reward Function

Another interesting direction is considering nonlinear reward functions. The LASSO Bandit can be used in conjunction with basis expansion methods from statistical learning to approximate any nonlinear function (Hastie et al. 2001). Specifically, given a covariate vector $X = (x_1, \ldots, x_d)$, we can consider a large vector with length $O(d^n)$ consisting of all *distinct* monomials of maximum degree at most n, denoted by $X \otimes_n$. Then we can use a linear model with covariate vector $X \otimes_n$ to approximate a reward function that is up to a n^{th}-degree polynomial in X. Assuming that the true model is a sparse function of these monomials (i.e., the reward only depends on s_0 entries of $X \otimes_n$), the LASSO Bandit algorithm could be employed. From a theoretical perspective, one has to study the behavior of the constant ϕ_0 for the compatibility condition of the covariance matrix of $X \otimes_n$ in order to prove theoretical guarantees for this approach; however, such an analysis is beyond the scope of this paper and we simply empirically test the approach. We repeat scenario (a) from above ($K = 2$, $d = 100$, $s_0 = 5$) where the true reward function of each arm is a (distinct) polynomial of degree $n = 3$. We compare two versions of the LASSO Bandit: (1) naïve-LASSO Bandit that uses only the raw covariates X and does not expand them to $X\otimes_3$, and (2) NL-LASSO Bandit that uses $X\otimes_3$. For comparison, we also include a nonlinear version of the other bandit algorithms that, similar to NL-LASSO Bandit, use the expanded covariate vector and refer to them by NL-OLS-Bandit, NL-OFUL-LS, and NL-OFUL-EG respectively. Figure EC.3 shows the results. NL-LASSO Bandit outperforms all other methods. It is interesting to see that the naïve-LASSO Bandit is competitive for small t since it avoids overfitting more effectively with a smaller covariate space; however, the regret is linear since its model is misspecified, and it loses to the other approaches as T grows.

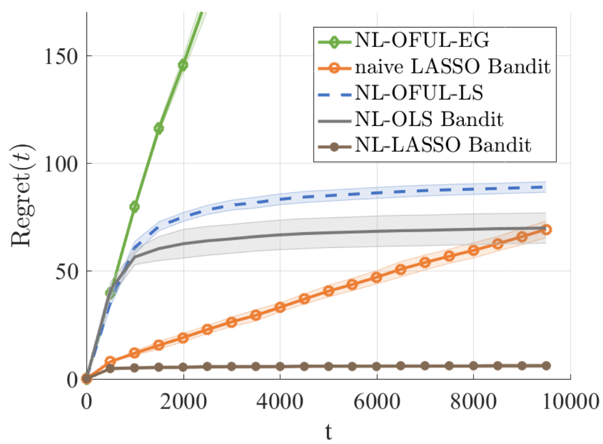


Figure EC.3 Comparison of all methods with parameters as in Figure 1(a), but the reward functions are polynomials of degree 3. The suffix NL means that the algorithm uses the expanded version of the covariate vector that contains all monomials of degree at most 3.

EC.6.4. When Assumptions 2 and 3 Fail

We now study two settings where some of the assumptions required by our theory fail. First, we look at Assumption 3 (arm optimality). We consider $K = 3$, $d = 100$, $s_0 = 2$, $\beta_1 = [1, 1, 0, \dots, 0]$, $\beta_2 = [0, 0, 1, 1, 0, \dots, 0]$, and $\beta_3 = r(\beta_1 + \beta_2)/2$ for $r \in \{0.9, .99, 1, 1.01, 1.1\}$. In this situation, when r is close or equal to 1, the arm optimality condition fails for arm 3. Figure EC.4 shows that performance of the LASSO Bandit is robust as r varies around 1. However, there is a small but noticeable loss when $r = 1.1$, which is due to the failure of the arm optimality condition. In particular, there are cases where arm 3 is the optimal arm and we incur regret if it is not played; however, the magnitude of this regret (relative to pulling the second best arm) is small, making the overall loss from the assumption's failure small. This simulation suggests that Assumption 3 could possibly be relaxed at the expense of a more cumbersome regret analysis.

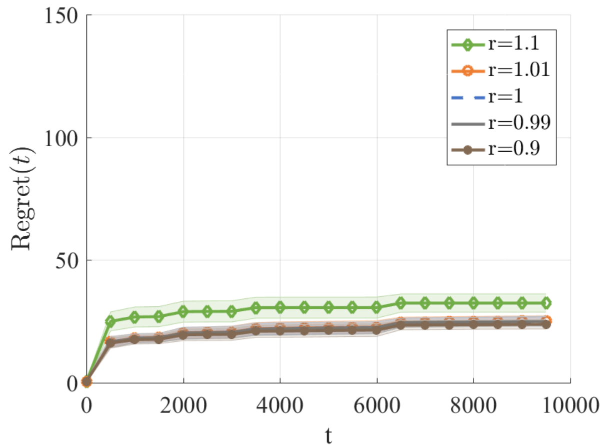


Figure EC.4 Performance of LASSO Bandit when arm optimality condition fails (when an arm is r times a convex combination of the other two arms).

Next, we study Assumption 2 (the margin condition). We consider $K=2$, $d=10$, $s_0=3$, $\beta_1=[1,0,1,0,\ldots,0]$ and $\beta_2=[1,1,1,0,\ldots,0]$. The covariates are generated according to the following procedure. First, a $(d-2)$-dimensional vector $\tilde{X}$ is sampled from the truncated normal (as above). Then, we add two coordinates at the beginning based on sampling a random variable U from the uniform distribution on $[-1,1]$, independent of $\tilde{X}$. Then, our d-dimensional covariate vector X is given by

$$X(r)=\begin{cases}1 & \text{if } r=1\\ \operatorname{sign}(U)|U|^{1+\epsilon} & \text{if } r=2\\ \tilde{X}(r) & \text{if } r>2\,.\end{cases}$$

Now, note that

$$\begin{aligned}\Pr\left[0<|X^\top(\beta_1-\beta_2)|<\kappa\right] &= \Pr\left[0<|X(2)|<\kappa\right]\\ &= \Pr\left[0<|U|^{1+\epsilon}<\kappa\right]=2(\kappa)^{\frac{1}{1+\epsilon}}>\kappa^{\frac{1}{1+\epsilon}}\,,\end{aligned}$$

which implies that the margin condition fails for any $\epsilon>0$. We simulate the LASSO Bandit for $\epsilon\in\{0,1\}$ and the results are shown in Figure EC.5. When $\epsilon=0$ (margin condition holds) the regret grows at a slower rate than when $\epsilon=1$ (margin condition fails). In fact, when $d=1$, Goldenshluger and Zeevi (2009) prove a lower bound that the regret scales as $\mathcal{O}\left(T^{\frac{\epsilon}{2(1+\epsilon)}}\right)$ for $\epsilon>0$. Generalizing a variant of their result to $d>1$ is an open direction, but matches our simulation results.

EC.7. OLS Bandit Algorithm and Analysis

In this section, we propose the OLS Bandit, which is a variant of the algorithm by Goldenshluger and Zeevi (2013) for the low-dimensional setting. We then apply the analytical tools we developed in the proof of the LASSO Bandit to prove an upper bound of $\mathcal{O}\left(d^2\log^{\frac{3}{2}}d\cdot\log T\right)$ on the cumulative expected regret of the OLS Bandit; this is an improvement over the existing $\mathcal{O}(d^3\log T)$ bound.

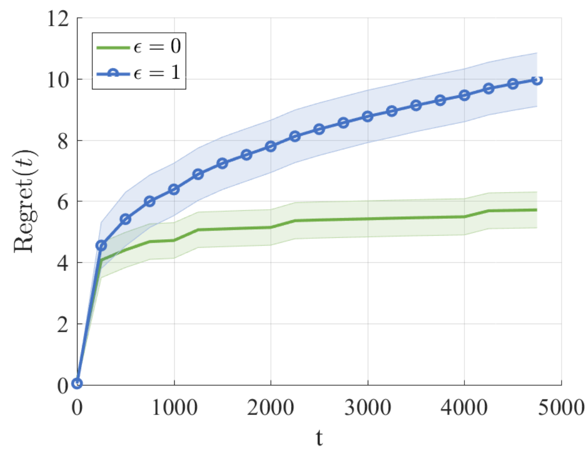


Figure EC.5 Performance of the LASSO Bandit when the margin condition holds ($\epsilon = 0$) versus when it fails ($\epsilon = 1$). In the latter case, the probability of observing covariate vectors that lie within a margin κ of the decision boundary is of order $\sqrt{\kappa}$.

Remark EC.1. Our analysis yields a better bound because we employ matrix martingale concentration results (Tropp 2015) to bound the difference of the true and sample covariance matrices, i.e., $\|\hat{\Sigma} - \Sigma\|_\infty$; in contrast, Goldenshluger and Zeevi (2013) rely on applying the union bound, which contributes an extra factor of d.

Assumptions. We make similar but weaker assumptions on the problem formulation as Goldenshluger and Zeevi (2013). In particular, prior work only allowed for two arms and required each arm to be optimal for some subset of users; in contrast, our formulation tackles the K-armed bandit and further allows for some arms $\mathcal{K}_{sub}$ to be uniformly sub-optimal.

Consequently, we make the same assumptions as that of the LASSO Bandit (including Assumptions 1-3 in §2.1) but we replace Assumption 4 on the LASSO compatibility condition with the following stronger requirement of positive-definiteness:

Assumption EC.1 **(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 a deterministic constant $\phi_0 \in \mathbb{R}^+$ such that for all $i \in [K]$ the minimum eigenvalue $\lambda_{\min}(\Sigma_i) \geq \phi_0^2 > 0$.*

OLS Estimation. Recall the notation we established in §3.1. Consider a linear model $Y = \mathbf{X}\beta + \varepsilon$, with design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, response vector $Y \in \mathbb{R}^n$, and noise vector $\varepsilon \in \mathbb{R}^n$ whose entries are independent σ-subgaussian random variables.

Definition EC.1 (OLS). If $\hat{\Sigma}(\mathbf{X}) = \mathbf{X}^\top\mathbf{X}/n$ is positive definite, the OLS estimator for the parameter β is defined by:

$$\hat{\beta}_{\mathbf{X},Y} \equiv \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y\,. \tag{EC.4}$$

The OLS estimator converges with high probability according to the following tail inequality.

PROPOSITION EC.1 **(OLS Tail Inequality)**. *Let X_t denote the t^{th} row of $\mathbf{X}$ and $Y(t)$ denote the t^{th} entry of Y. Also, assume that $\{X_t : t = 1, ..., n\}$ forms an adapted sequence of observations, i.e., X_t may depend on past regressors and their resulting observations $\{X_{t'}, Y(t')\}_{t'=1}^{t-1}$. If all realizations of random variables X_t satisfy $\|X_t\|_\infty \leq x_{\max}$, the following tail inequality holds for all $\chi > 0$ and all constants $\phi > 0$:*

$$\Pr\left[\|\hat{\beta} - \beta\|_1 \leq \chi\right] \geq 1 - \exp\left[-\tilde{C}_1(\phi)n\chi^2 + \log 2d\right] - \Pr\left[\lambda_{min}\left(\hat{\Sigma}(\mathbf{X})\right) \leq \phi^2\right],$$

where we define $\tilde{C}_1(\phi) \equiv \phi^4 / (2d^2 x_{\max}^2 \sigma^2)$.

Algorithm. We introduce the OLS Bandit algorithm below (Algorithm 2), which proceeds analogously to the LASSO Bandit (Algorithm 1). In particular, we define and use the forced-sample sets $\mathcal{T}_{i,t}$ and all-sample sets $\mathcal{S}_{i,t}$ in the same way. The key difference is that we now use OLS instead of LASSO estimation (note that we no longer require a path of regularization parameters).

Algorithm OLS Bandit

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

EC.7.1. New Upper Bound on Regret of OLS Bandit

THEOREM EC.1. *When $q \geq 4\lceil \tilde{q}_0 \rceil$, $K \geq 2$, $d > 2$, and $T \geq (Kq)^2$, we have an upper bound on the expected cumulative regret at time T:*

$$\begin{aligned} R_T &\leq 2qKbx_{\max}(6\log T + Kq) + 2Kbx_{\max} + \frac{8K\max(C_0, 1)x_{\max}^2[\log(12d)]^{3/2}}{\tilde{C}_3}\log T + \tilde{C}_4 Kbx_{\max} \\ &= \mathcal{O}\left(d^2[\log d]^{3/2} \cdot \log T\right), \end{aligned}$$

where constants $\tilde{C}_1 = \tilde{C}_1(\phi_0)$, $\tilde{C}_2 = \tilde{C}_2(\phi_0)$, $\tilde{C}_3 = \tilde{C}_3(\phi_0, p_*)$, and $\tilde{C}_4 = \tilde{C}_4(\phi_0, p_*)$ are defined by

$$\tilde{C}_1 \equiv \frac{\phi_0^4}{2d^2 x_{\max}^2 \sigma^2},\ \tilde{C}_2 \equiv \min\left(\frac{1}{2}, \frac{\phi_0^2}{8x_{\max}^2}\right),\ \tilde{C}_3 \equiv \frac{p_*^3 \tilde{C}_1}{256},\ \textit{and}\ \tilde{C}_4 \equiv \frac{8}{1 - \exp\left[-\tilde{C}_2 \frac{p_*^2}{64}\right]},$$

C_0 is defined in Assumption 2, and we take

$$\tilde{q}_0 \equiv \max\left\{\frac{20}{p_*}, \frac{8\log d}{p_*\tilde{C}_2}, \frac{1024x_{\max}^2 \log 2d}{h^2 p_*^2 \tilde{C}_1}\right\} = \mathcal{O}\left(d^2 \log d\right) .$$

Key Steps. The proof strategy is similar to that of the LASSO Bandit. First, we prove a technical lemma (analogous to Lemma 1) that shows a tail inequality holds for the OLS estimator if only a constant (but unknown) fraction of the rows of the design matrix are independent (Lemma EC.21). We use this lemma to prove analogous tail inequalities for the forced-sample estimator (Proposition EC.2) and the all-sample estimator (Proposition EC.3) in §EC.7.4. Finally, we use these tail inequalities to sum up the expected regret contributions from the three groups of time periods:

(a) Initialization ($t \le (Kq)^2$) and forced sampling ($t \in \mathcal{T}_{i,T}$ for some $i \in [K]$).
(b) Times $t > (Kq)^2$ when the event A_{t-1} does not hold.
(c) Times $t > (Kq)^2$ when the event A_{t-1} holds and we do not perform forced sampling, i.e., the OLS Bandit plays the estimated best arm from $\hat{\mathcal{K}}$ using the all-sample estimator.

Summing the results concludes the proof of Theorem EC.1. The proof is given in §EC.7.5.

EC.7.2. OLS Tail Inequality for Adapted Observations

Proof of Proposition EC.1 For simplicity, we start with the ℓ_2 norm. Note that, if the event $\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) > \phi^2$ holds,

$$\begin{aligned}
\|\hat{\beta} - \beta\|_2 &= \|(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \varepsilon\|_2 \\
&\le \|(\mathbf{X}^\top \mathbf{X})^{-1}\|_2 \cdot \|\mathbf{X}^\top \varepsilon\|_2 \\
&= \frac{1}{n\phi^2}\|\mathbf{X}^\top \varepsilon\|_2 .
\end{aligned}$$

Then, for any $\tilde{\chi} > 0$, we can write

$$\begin{aligned}
\Pr\left[\|\hat{\beta} - \beta\|_2 \le \tilde{\chi}\right] &\ge \Pr\left[\left(\|\mathbf{X}^\top \varepsilon\|_2 \le n\tilde{\chi}\phi^2\right) \cap \left(\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) > \phi^2\right)\right] \\
&\ge 1 - \sum_{r=1}^{d} \Pr\left[|\varepsilon^\top X^{(r)}| > \frac{n\tilde{\chi}\phi^2}{\sqrt{d}}\right] - \Pr\left[\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) \le \phi^2\right] ,
\end{aligned}$$

where we have let $X^{(r)}$ denote the r^{th} column of $\mathbf{X}$. We can expand $\varepsilon^\top X^{(r)} = \sum_{t\in[n]} \varepsilon(t) X_t(r)$, where we note that $D_{t,r} \equiv \varepsilon(j) X_j(r)$ is a $(x_{\max}\sigma)$-subgaussian random variable by Definition 1, conditioned on the sigma algebra $\mathfrak{S}_{t-1}$ that is generated by random variables $X_1, \ldots, X_{t-1}, Y(1), \ldots, Y(t-1)$. Defining $D_{0,r} = 0$, the sequence $D_{0,r}, D_{1,r}, \ldots, D_{n,r}$ is a martingale difference sequence adapted to the filtration $\mathfrak{S}_1 \subset \ldots \subset \mathfrak{S}_n$ since $\mathbb{E}[\varepsilon(t) X_t(r) | \mathfrak{S}_{t-1}] = 0$. Using Lemma EC.1,

$$\begin{aligned}
\Pr\left[\|\hat{\beta} - \beta\|_2 \le \tilde{\chi}\right] &\ge 1 - \sum_{r=1}^{d} \Pr\left[|\varepsilon^\top X^{(r)}| > \frac{n\tilde{\chi}\phi^2}{\sqrt{d}}\right] - \Pr\left[\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) \le \phi^2\right] \\
&\ge 1 - 2d \exp\left[-\frac{n\tilde{\chi}^2\phi^4}{2dx_{\max}^2\sigma^2}\right] - \Pr\left[\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) \le \phi^2\right] .
\end{aligned}$$

Now, to bound the ℓ_1 norm, we can use Cauchy-Schwarz to write (for $\tilde{\chi} = \chi/\sqrt{d}$)

$$\begin{aligned}
\Pr\left[\|\hat{\beta} - \beta\|_1 \le \chi\right] &\ge \Pr\left[\|\hat{\beta} - \beta\|_2 \le \tilde{\chi}\right] \\
&\ge 1 - 2d\exp\left[-\frac{n\tilde{\chi}^2\phi^4}{2dx_{\max}^2\sigma^2}\right] - \Pr\left[\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) \le \phi^2\right] \\
&= 1 - \exp\left[-\underbrace{\frac{\phi^4}{2d^2x_{\max}^2\sigma^2}}_{\tilde{C}_1(\phi)} n\chi^2 + \log(2d)\right] - \Pr\left[\lambda_{\min}\left(\hat{\Sigma}(\mathbf{X})\right) \le \phi^2\right]. \quad \square
\end{aligned}$$

EC.7.3. Positive-Definiteness for non-i.i.d. samples

In this section we prove a tail inequality for OLS with non-i.i.d. data, analogous to the result of §4.1. In particular, consider a linear model $Y = \mathbf{Z}\beta + \varepsilon$, with random design matrix $\mathbf{Z} \in \mathbb{R}^{n\times d}$ such that all realizations of $\mathbf{Z}$ satisfy $\|\mathbf{Z}\|_\infty \le x_{\max}$, response vector $Y \in \mathbb{R}^n$, and noise vector $\varepsilon \in \mathbb{R}^n$ whose entries are independent σ-subgaussian random variables. Consider a fixed subset $\mathcal{A}$ of $[n]$, and if $\mathcal{A}' \subset \mathcal{A}$ is such that $\{Z_t \mid t \in \mathcal{A}'\}$ is an i.i.d. subset of random variables with distribution $\mathcal{P}_Z$ with $\lambda_{\min}(\mathbb{E}[ZZ^T]) = \phi_1^2$ and $|\mathcal{A}'|/|\mathcal{A}| \ge p/2$ for positive constants ϕ_1 and p. Similar to §3.1, we use the short notation $\hat{\beta}(\mathcal{A})$ and $\hat{\Sigma}(\mathcal{A})$ to refer to the OLS estimator and sample covariance matrix on the set $\mathcal{A}$. In this section, we will show that $\hat{\Sigma}(\mathcal{A})$ is positive-definite with minimum eigenvalue bounded below by $\phi_1^2 p/4 = \left(\phi_1\sqrt{p}/2\right)^2$ with high probability, and then apply Proposition EC.1 to obtain the following result.

Lemma EC.21. *Under the assumptions above, the following tail inequality holds for all $\chi > 0$:*

$$\Pr\left[\|\hat{\beta}(\mathcal{A}) - \beta\|_1 \ge \chi\right] \le \exp\left[-\tilde{C}_1\left(\frac{\phi_1\sqrt{p}}{2}\right)|\mathcal{A}|\chi^2 + \log 2d\right] + \exp\left[-p\tilde{C}_2(\phi_1)|\mathcal{A}|/2 + \log d\right],$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are defined in §EC.7.1.

Before formally proving Lemma EC.21, we state and prove some results.

First, we will show that $\hat{\Sigma}(\mathcal{A}')$ has minimum eigenvalue bounded below with high probability.

Lemma EC.22. *The minimum eigenvalue of $\hat{\Sigma}(\mathcal{A}')$ is bounded below by $\phi_1^2/2$ with probability* $1 - \exp\left[-\tilde{C}_2(\phi_1)|\mathcal{A}'| + \log d\right]$.

Proof of Lemma EC.22 First, note that

$$\begin{aligned}
\lambda_{\max}\left(\hat{\Sigma}(\mathcal{A}')\right) &= \max_{\|u\|=1} u^\top \hat{\Sigma}(\mathcal{A}')u \\
&= \max_{\|u\|=1} \frac{1}{|\mathcal{A}'|}\sum_{t\in\mathcal{A}'}(Z_t^\top u)^2 \ \le \ x_{\max}^2
\end{aligned}$$

Then, it follows from the matrix Chernoff inequality, Corollary 5.2 in Tropp (2015), that

$$\Pr\left[\lambda_{\min}(\hat{\Sigma}(\mathcal{A}')) > \frac{\phi_1^2}{2}\right] \geq 1 - d \cdot \exp\left[-\frac{|\mathcal{A}'|\phi_1^2}{8x_{\max}^2}\right] \geq 1 - \exp\left[-\tilde{C}_2(\phi_1)|\mathcal{A}'| + \log d\right]$$

, if we take $\delta = 1/2$ and $R = x_{\max}^2$. □

LEMMA EC.23. *If the minimum eigenvalue of $\hat{\Sigma}(\mathcal{A}')$ is bounded below by $\phi_1'^2$, then the minimum eigenvalue of $\hat{\Sigma}(\mathcal{A})$ is bounded below by $\phi_1'^2|\mathcal{A}'|/|\mathcal{A}|$.*

Proof of Lemma EC.23 From our definition, we can write

$$\begin{aligned}\hat{\Sigma}(\mathcal{A}) &= \frac{|\mathcal{A}'|}{|\mathcal{A}|}\hat{\Sigma}(\mathcal{A}') + \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}/\mathcal{A}'} Z_t Z_t^\top \\ &= \frac{|\mathcal{A}'|}{|\mathcal{A}|}\hat{\Sigma}(\mathcal{A}') + \frac{|\mathcal{A}\setminus\mathcal{A}'|}{|\mathcal{A}|}\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')\,.\end{aligned}$$

Now, using the fact that the minimum eigenvalue is a concave function, it immediately follows that

$$\begin{aligned}\lambda_{\min}\left(\hat{\Sigma}(\mathcal{A})\right) &\geq \frac{|\mathcal{A}'|}{|\mathcal{A}|}\lambda_{\min}\left(\hat{\Sigma}(\mathcal{A}')\right) + \frac{|\mathcal{A}\setminus\mathcal{A}'|}{|\mathcal{A}|}\lambda_{\min}\left(\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')\right) \\ &\geq \frac{|\mathcal{A}'|}{|\mathcal{A}|}\phi_1'^2\,,\end{aligned}$$

where the last inequality relies on the fact that $\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')$ is always positive semi-definite. □

Now, we are ready to prove the main result of this section.

Proof of Lemma EC.21 Combining Lemmas EC.22 and EC.23, and using $|\mathcal{A}'| \geq p|\mathcal{A}|/2$ implies that

$$\Pr\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{A})\right) \leq \frac{\phi_1^2 p}{4}\right] \leq \exp\left[-p\tilde{C}_2(\phi_1)|\mathcal{A}|/2 + \log d\right]\,.$$

Now, we can apply Proposition EC.1 with $\phi = \phi_1\sqrt{p}/2$ to obtain the result. □

EC.7.4. Proof of Tail Inequalities for OLS Force-Sample and All-Sample Estimators

PROPOSITION EC.2. *When $t \geq (Kq)^2$, the forced sample estimator $\hat{\beta}(\mathcal{T}_{i,t})$ satisfies the tail inequality*

$$\Pr\left[\|\hat{\beta}\left(\mathcal{T}_{i,t}\right) - \beta_i\|_1 > \frac{h}{4x_{\max}}\right] \leq \frac{4}{t^4}\,.$$

Proof of Proposition EC.2 Since forced-sampling schedule is the same as LASSO Bandit, using Lemma EC.8, $|\mathcal{T}_{i,t}| \geq (q/2)\log t \geq 2\tilde{q}_0$. Also, by Assumption EC.1, Σ_i has minimum eigenvalue bounded below by ϕ_0^2. If $|\mathcal{T}'_{i,t}|/|\mathcal{T}_{i,t}| \geq p_*/2$, Lemma EC.9 allows us to apply Lemma EC.21, with $\chi = h/(4x_{\max})$, to show that

$$\begin{aligned}&\Pr\left[\|\hat{\beta}\left(\mathcal{T}_{i,t}\right) - \beta_i\|_1 > \frac{h}{4x_{\max}}\right] \\ &\quad\leq \exp\left[-\tilde{C}_1\left(\frac{\phi_0\sqrt{p_*}}{2}\right)|\mathcal{T}_{i,t}|\frac{h^2}{16x_{\max}^2} + \log 2d\right] + \exp\left[-p_*\tilde{C}_2(\phi_0)|\mathcal{T}_{i,t}|/2 + \log d\right] \\ &\quad\quad\leq \exp\left[-\tilde{q}_0\log t \cdot \frac{p_*^2\tilde{C}_1(\phi_0)h^2}{128x_{\max}^2} + \log 2d\right] + \exp\left[-p_*\tilde{C}_2(\phi_0)\tilde{q}_0\log t + \log d\right]\,.\end{aligned}$$

Combining this with the probability that $|\mathcal{T}'_{i,t}|/|\mathcal{T}_{i,t}| \geq p_*/2$ (Lemma EC.10), and using a union bound gives

$$\Pr\left[\|\hat{\beta}\left(\mathcal{T}_{i,t}\right)-\beta_i\|_1 > \frac{h}{4x_{\max}}\right] \leq \exp\left[-\tilde{q}_0 \log t \cdot \frac{p_*^2 \tilde{C}_1(\phi_0) h^2}{128 x_{\max}^2} + \log 2d\right] + \exp\left[-p_* \tilde{C}_2(\phi_0)\tilde{q}_0 \log t + \log d\right] + 2/t^4 \,.$$

Now, using definition of $\tilde{q}_0$, in particular

$$\tilde{q}_0 \geq \frac{8 \log d}{p_* \tilde{C}_2} \quad \text{and} \quad \tilde{q}_0 \geq \frac{1024 x_{\max}^2 \log 2d}{h^2 p_*^2 \tilde{C}_1}\,,$$

and the fact that $d > 2$ and $t > (Kq)^2$, the result follows. $\square$

We again define the event A_t in the same way as (3) in order to prove the tail inequality for the all-sample OLS estimator.

Proposition EC.3. *When $t \geq (Kq)^2$, for $i \in \mathcal{K}_{opt}$, the all-sample estimator $\hat{\beta}(\mathcal{S}_{i,t})$ satisfies the tail inequality*

$$\Pr\left[\|\hat{\beta}\left(\mathcal{S}_{i,t}\right)-\beta_i\|_1 \leq \chi\right] \geq 1 - \exp\left[-t\chi^2 \frac{p_*^3 \tilde{C}_1(\phi_0)}{256} + \log 2d\right] - 2\exp\left[-\tilde{C}_2(\phi_0)\frac{p_*^2}{64}t + \log d\right] .$$

Proof of Proposition EC.3 First, we note that Lemma EC.12 holds for the OLS estimator as well since the forced-sample tail inequality for the OLS estimator (Proposition EC.2) is slightly stronger than the forced-sample tail inequality for the LASSO estimator (Proposition 2), $5/t^4$ versus $4/t^4$ error bound.

From Lemma EC.13, we have that at time $t \geq (Kq)^2$, each of $\{1,...,t\} \setminus \cup_{j=1}^{K} \mathcal{T}_{j,t}$ belongs to $\mathcal{S}'_{i,t}$ with probability at least $p_*/2$. Applying Lemma EC.14 and Lemma EC.21 with $p = p_*/2$ and $|\mathcal{A}| \geq p_* t/4$, we get, using a union bound,

$$\begin{aligned}
\Pr\left[\|\hat{\beta}\left(\mathcal{S}_{i,t}\right)-\beta_i\|_1 > \chi\right] &\leq \exp\left[-\tilde{C}_1\left(\frac{\phi_0\sqrt{p_*/2}}{2}\right)\frac{tp_*}{4}\chi^2 + \log 2d\right] + \exp\left[-\frac{\tilde{C}_2(\phi_0)p_* t}{16} + \log d\right] + \exp\left[-\frac{p_*^2}{128}t\right] \\
&\leq \exp\left[-t\chi^2 \frac{p_*^3 \tilde{C}_1(\phi_0)}{256} + \log 2d\right] + 2\exp\left[-\tilde{C}_2(\phi_0)\frac{p_*^2}{64}t + \log d\right],
\end{aligned}$$

where we have used $\tilde{C}_2(\phi_0) \leq 1/2$ in the last step. $\square$

EC.7.5. Bounding the Regret in the Low-Dimensional Setting

We can now use the above tail inequalities to sum up the expected regret contributions from the three groups of time periods:

(a) Initialization ($t \leq (Kq)^2$) or forced sampling ($t \in \mathcal{T}_{i,T}$ for some $i \in [K]$).

(b) Times $t > (Kq)^2$ when the event A_{t-1} does not hold.

(c) Times $t > (Kq)^2$ when the event A_{t-1} holds and we do not perform forced sampling, i.e., the OLS Bandit plays the estimated best arm from $\hat{\mathcal{K}}$ using the all-sample estimator.

We first note that the regret bounds of §EC.5 for groups (a) times where $t \leq (Kq)^2$ or we are force-sampling, and (b) time periods where A_{t-1} does not hold can be re-used. This is because the forced-sampling schedule is the same and the tail inequality we prove for the OLS forced-sample estimator is strictly stronger than the tail inequality for the LASSO forced-sample estimator. We now focus on bounding the regret from time periods (c) when $t > (Kq)^2$, we are not force-sampling, and A_{t-1} holds.

In this section, we simplify our notation by letting $\hat{\beta}_i = \hat{\beta}\left(\mathcal{S}_{i,t}\right)$ for all $i \in [K]$. We also define the constant $\tilde{C}_3(\phi_0, p_*) = p_*^3 \tilde{C}_1(\phi_0)/256$, but to simplify the notation, drop the references to ϕ_0 and p_* in all constants $\tilde{C}_1, \tilde{C}_2, \ldots$ since the values for ϕ_0 and p_* will be fixed in the remaining.

Lemma EC.24. *If Algorithm 2 does not use the forced-sample estimator and A_{t-1} holds, then the expected regret at time t is bounded by*

$$\frac{8K\max(C_0,1)x_{\max}^2[\log(12d)]^{3/2}}{t\tilde{C}_3} + 8Kbx_{\max}e^{-\tilde{C}_2\frac{p_*^2}{64}t}.$$

Proof of Lemma EC.24 Recall from Lemma EC.18 that since A_{t-1} holds, the set $\hat{\mathcal{K}}$ contains the optimal arm $i^* = \arg\max_{i\in[K]} X_t^\top \beta_i$ and no sub-optimal arms from the set $\mathcal{K}_{sub}$. Without loss of generality, assume that arm 1 is optimal, i.e., $1 = \arg\max_{i\in\{1,\ldots,K\}} X_t^\top \beta_i$. Then, the expected regret at time t is given by

$$\begin{aligned}\mathbb{E}[r_t] &= \mathbb{E}\left(\sum_{i\in\hat{\mathcal{K}}, i\neq 1} X_t^\top\left(\beta_1 - \beta_i\right)\cdot \mathbb{I}\left[i = \arg\max_{j\in\{1,\ldots,K\}} X_t^\top \hat{\beta}_j\right]\right)\\ &\leq \mathbb{E}\left(\sum_{i\in\hat{\mathcal{K}}, i\neq 1} X_t^\top\left(\beta_1 - \beta_i\right)\mathbb{I}\left[X_t^\top\hat{\beta}_i > X_t^\top\hat{\beta}_1\right]\right)\end{aligned}$$

where the inequality follows from the fact that the event where $i = \arg\max_{j\in\{1,\ldots,K\}} X_t^\top \hat{\beta}_j$ is a subset of the event $X_t^\top\hat{\beta}_i > X_t^\top\hat{\beta}_1$, and that $\mathbb{E}\left[X_t^\top(\beta_1-\beta_i)\right] \geq 0$ (since we have assumed that arm 1 is optimal). Thus, we can bound r_t through the regret incurred by each arm with respect to the optimal arm independently of the other arms. We now define, for each $r = 0,1,2,3,\ldots$ the event

$$B_r^i = \left\{2x_{\max}r\delta \leq X_t^\top(\beta_1-\beta_i) < 2x_{\max}(r+1)\delta\right\}$$

where δ is a parameter we will choose later to minimize regret. Note that, since $X_t^\top(\beta_1-\beta_i) < 2x_{\max}b$, B_r^i is empty for $r+1 > b/\delta$. Then, we can write

$$\mathbb{E}[r_t] < 2x_{\max}\delta\,\mathbb{E}\left(\sum_{r=0}^{\lfloor b/\delta\rfloor-1}(r+1)\sum_{i\in\hat{\mathcal{K}}, i\neq 1}\mathbb{I}\left[(X_t^\top\hat{\beta}_i > X_t^\top\hat{\beta}_1)\cap B_r^i\right]\right) \tag{EC.5}$$

by the definition of B_r^i.

Note that the event $(X_t^\top \hat{\beta}_i > X_t^\top \hat{\beta}_1) \cap B_r^i$ for $i \neq 1$ implies that

$$
\begin{aligned}
0 &> X_t^\top \hat{\beta}_1 - X_t^\top \hat{\beta}_i \\
&= \left[X^\top \hat{\beta}_1 - X_t^\top \beta_1\right] + \left[X_t^\top \beta_i - X_t^\top \hat{\beta}_i\right] + \left[X_t^\top \beta_1 - X_t^\top \beta_i\right] \\
&\geq \left[X_t^\top \left(\hat{\beta}_1 - \beta_1\right)\right] + \left[X_t^\top \left(\beta_i - \hat{\beta}_i\right)\right] + 2x_{\max} r\delta\,.
\end{aligned}
$$

Thus, it must be that either $X_t^\top \left(\hat{\beta}_1 - \beta_1\right) > x_{\max} r\delta$ or $X_t^\top \left(\beta_i - \hat{\beta}_i\right) > x_{\max} r\delta$ which means, using a union bound,

$$
\begin{aligned}
\Pr\left[(X_t^\top \hat{\beta}_i > X_t^\top \hat{\beta}_1) \cap B_r^i\right] &\leq \Pr\left[(X_t^\top \left(\hat{\beta}_1 - \beta_1\right) > x_{\max} r\delta) \cap B_r^i\right] + \Pr\left[(X_t^\top \left(\beta_i - \hat{\beta}_i\right) > x_{\max} r\delta) \cap B_r^i\right] \\
&\leq \Pr\left[(\|\beta_1 - \hat{\beta}_1\|_1 > r\delta) \cap B_r^i\right] + \Pr\left[(\|\hat{\beta}_i - \beta_i\|_1 > r\delta) \cap B_r^i\right] \\
&= \Pr\left[\|\beta_1 - \hat{\beta}_1\|_1 > r\delta\right] \Pr\left[B_r^i\right] + \Pr\left[\|\hat{\beta}_i - \beta_i\|_1 > r\delta\right] \Pr\left[B_r^i\right]\,.
\end{aligned}
$$

Note that the last equality uses the fact that event B_r^i depends on the randomness of X_t that is completely independent of past samples that impact the randomness of $\hat{\beta}_1$ and $\hat{\beta}_i$.

Next, recall that the tail inequality (Proposition EC.3) implies that for all $j \in \hat{\mathcal{K}}$, and all $r, \delta \geq 0$,

$$
\Pr\left[\|\beta_j - \hat{\beta}_j\|_1 > r\delta\right] \leq \min\left\{1,\ \exp\left[-\tilde{C}_3 r^2 \delta^2 t + \log 2d\right] + 2\exp\left[-\tilde{C}_2 \frac{p_*^2}{64} t\right]\right\}\,.
$$

Combining this with the fact that, via Assumption 2 on margin condition,

$$
\Pr[B_r^i] \leq \Pr\left[X_t^\top (\beta_1 - \beta_i) \leq 2x_{\max}(r+1)\delta\right] \leq 2C_0 x_{\max}(r+1)\delta\,,
$$

we get,

$$
\begin{aligned}
\Pr\left[(X_t^\top \hat{\beta}_i > X_t^\top \hat{\beta}_1) \cap B_r^i\right] &\leq 2\Pr[B_r^i] \min\left\{1, e^{-\tilde{C}_3 r^2 \delta^2 t + \log 2d} + 2e^{-\tilde{C}_2 \frac{p_*^2}{64} t}\right\} \\
&\leq 2\Pr[B_r^i] \min\left\{1, e^{-\tilde{C}_3 r^2 \delta^2 t + \log 2d}\right\} + 4\Pr[B_r^i] e^{-\tilde{C}_2 \frac{p_*^2}{64} t} \\
&\leq 4C_0 x_{\max}(r+1)\delta \min\left\{1, e^{-\tilde{C}_3 r^2 \delta^2 t + \log 2d}\right\} + 4\Pr[B_r^i] e^{-\tilde{C}_2 \frac{p_*^2}{64} t} \\
&\leq 4C_0 x_{\max}(r+1)\delta \min\left\{1, e^{-\tilde{C}_3 r^2 \delta^2 t + \log 2d}\right\} + 4e^{-\tilde{C}_2 \frac{p_*^2}{64} t}\,. \qquad \text{(EC.6)}
\end{aligned}
$$

Note that, for a large r the term $e^{-\tilde{C}_3 t r^2 \delta^2 + \log 2d}$ will be small. Therefore, we will use the term 1 for small r and the second term for large r. Combining (EC.5) and (EC.6), setting $\delta = 1/\sqrt{\tilde{C}_3 t}$, and defining

$$
R \equiv R(d, t, \delta) = \left\lfloor \sqrt{\log(12d)} \right\rfloor\,,
$$

we have

$$\mathbb{E}[r_t] \leq \frac{8KC_0x_{\max}^2}{t\tilde{C}_3}\left[\sum_{r=0}^{R}(r+1)^2 + 2d\sum_{r=R+1}^{\lfloor b\sqrt{\tilde{C}_3 t}\rfloor -1}(r+1)^2 e^{-r^2}\right] + \left[\frac{8Kbx_{\max}}{\sqrt{\tilde{C}_3 t}}\sum_{r=R+1}^{\lfloor b\sqrt{\tilde{C}_3 t}\rfloor -1}(r+1)e^{-\tilde{C}_2\frac{p_*^2}{64}t}\right]$$
$$\leq \frac{8KC_0x_{\max}^2[\log(12d)]^{3/2}}{t\tilde{C}_3} + \frac{16Kdx_{\max}^2}{t\tilde{C}_3}\sum_{r=R+1}^{\lfloor b\sqrt{\tilde{C}_3 t}\rfloor -1}(r+1)^2e^{-r^2} + 8Kbx_{\max}e^{-\tilde{C}_2\frac{p_*^2}{64}t}.$$

Now, note that

$$\begin{aligned}\sum_{r=R+1}^{\infty}(r+1)^2e^{-r^2} &\leq 4\sum_{r=R+1}^{\infty}r^2e^{-r^2}\\ &\leq 4\int_R^\infty u^2e^{-u^2}du\\ &= 2Re^{-R^2} + 2\int_R^\infty e^{-u^2}du\end{aligned}$$

where the second inequality follows from Lemma EC.16 and the equality is via integration by parts. Therefore,

$$\begin{aligned}\sum_{r=R+1}^{\infty}(r+1)^2e^{-r^2} &\leq 2Re^{-R^2} + 2\int_R^\infty \left(\frac{u}{R}\right)e^{-u^2}du\\ &= 2Re^{-R^2} + \frac{e^{-R^2}}{R}\\ &\leq \frac{\sqrt{\log(12d)}}{3d}.\end{aligned}$$

Summarizing,

$$\begin{aligned}\mathbb{E}[r_t] &\leq \frac{8KC_0x_{\max}^2[\log(12d)]^{3/2}}{t\tilde{C}_3} + \frac{6Kx_{\max}^2[\log(12d)]^{1/2}}{t\tilde{C}_3} + 8Kbx_{\max}e^{-\tilde{C}_2\frac{p_*^2}{64}t}\\ &\leq \frac{8K\max(C_0,1)x_{\max}^2[\log(12d)]^{3/2}}{t\tilde{C}_3} + 8Kbx_{\max}e^{-\tilde{C}_2\frac{p_*^2}{64}t}. \quad \square\end{aligned}$$

Lemma EC.25. *The cumulative expected regret from the time periods in group (c), times $t \in [T] \setminus [(Kq)^2]$ when the event A_{t-1} holds and we do not perform forced sampling, is bounded by*

$$\frac{8K\max(C_0,1)x_{\max}^2[\log(12d)]^{3/2}\log T}{\tilde{C}_3} + \tilde{C}_4Kbx_{\max},$$

where

$$\tilde{C}_4 = \frac{8}{1-\exp\left[-\tilde{C}_2\frac{p_*^2}{64}\right]}.$$

Proof of Lemma EC.25 Using Lemma EC.24,

$$\sum_{t=(Kq)^2+1}^{T}\mathbb{E}[r_t] \leq \frac{8K\max(C_0,1)x_{\max}^2[\log(12d)]^{3/2}\log T}{\tilde{C}_3} + \frac{8Kbx_{\max}}{1-\exp\left[-\tilde{C}_2\frac{p_*^2}{64}\right]}. \quad \square$$

Summing up the regret contributions from the previous subsection gives us our main result.

Proof of Theorem EC.1 The total expected cumulative regret of the OLS Bandit up to time T is upper-bounded by summing all the terms from Lemmas EC.15, EC.17, and EC.25):

$$R_T \leq 2qKbx_{\max}(6\log T + Kq) + 2Kbx_{\max} + \frac{8K\max(C_0,1)x_{\max}^2[\log(12d)]^{3/2}\log T}{\tilde{C}_3} + \tilde{C}_4 Kbx_{\max}. \quad \square$$