Online Decision-Making with High-Dimensional Covariates

Hamsa Bastani
Stanford University, Electrical Engineering

Mohsen Bayati
Stanford University, Graduate School of Business

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 multi-armed bandit with high-dimensional covariates, and present a new efficient bandit algorithm based on the LASSO estimator. Our regret analysis establishes that our algorithm achieves near-optimal performance in comparison to an oracle that knows all the problem parameters. The key step in our analysis is proving a new oracle 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 real-world clinical problem of warfarin dosing. A patient’s optimal warfarin 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: multi-armed bandits with covariates, adaptive treatment allocation, online learning, high-dimensional statistics, LASSO, statistical decision-making, personalized medicine

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, Bertsimas and Kallus 2014, Rudin and Vahn 2015, 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 10,000 (in Wiens et al. 2012) or 40,000 (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. 1995, 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 since they substantially increase bias in our estimates to increase predictive accuracy. 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 multi-armed bandits with covariates (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 $O(d)$ 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 LASSO methods 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 demonstrate the practical value of our algorithm in a medical decision-making context by evaluating it on the clinical task of warfarin dosing. 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. 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. We demonstrate through both theoretical guarantees and empirical evaluations that the LASSO Bandit significantly improves performance in the data-poor regime compared to existing algorithms without sacrificing performance in the asymptotic data-rich regime.

Algorithm. Our algorithm adapts ideas from existing work in the low-dimensional bandit setting (Goldenshluger and Zeevi 2013). The key step is selecting the regularization path, which controls the convergence of our LASSO estimators by trading off bias and variance over time. Apart from the high-dimensional setting, we make several extensions that improve the applicability of such bandit algorithms. For example, Goldenshluger and Zeevi (2013) only allow two choices and require that each choice is optimal for some subset of users; such assumptions are often not met in practice. In contrast, we allow for multiple choices, some of which may be uniformly sub-optimal.

Theory. We measure performance using cumulative regret, which is the total 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 near-optimal cumulative regret. The technical challenge is that the bandit policy induces non-i.i.d. samples from each arm during the exploitation phase; thus, existing results for the convergence of LASSO estimates do not hold. We tackle this problem by proving a new tail inequality for the LASSO which holds with high probability even when an unknown portion of the samples are generated by a non-i.i.d. process.

Yet, even with our new tail inequality, existing proof techniques do not establish regret bounds that are not polynomial in $d$. We develop new analytical tools based on matrix perturbation theory and prove that the cumulative regret of the LASSO Bandit over $T$ users is at most $O(s_0^2 \log T + \log d^2)$, 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 literature, including (i) UCB-based algorithm OFUL (Dani et al. 2008, Abbasi-Yadkori et al. 2011), (ii) a sparse variant of OFUL for high-dimensional settings (Abbasi-Yadkori et al. 2012), and (iii) the OLS Bandit by Goldenshluger and Zeevi (2013). Evaluations 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 settings that might be considered low-dimensional.

More importantly, we demonstrate the practical value of our algorithm by evaluating it on the medical problem of dosing warfarin (a widely prescribed anticoagulant). Assigning the appropriate initial dosage is crucial in order to avoid stroke or internal bleeding; however, the correct dose is highly variable due to patient-specific factors (Wysowski et al. 2007). We apply the LASSO Bandit to a real patient dataset (Consortium 2009) and 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. Our results suggest that learning algorithms like the LASSO Bandit may be valuable in practice to improve overall patient outcomes.
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., Bertsimas and Kallus 2014, Rudin and Vahn 2015). 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 multi-armed bandits with covariates 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, we focus 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 encompasses our setting if the feasible action set is allowed to change exogenously over time.) 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. However, this literature typically does not make any assumptions on how the user covariates $X_t$ are generated. In particular, they allow for adversarially constructed covariate sequences that make learning difficult. This may explain why the current-best cumulative regret bounds are given by: $O(d\sqrt{T})$ in the low-dimensional setting (Dani et al. 2008, Abbasi-Yadkori et al. 2011) and $O(\sqrt{d s_0 T})$ 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 $T = O(d)$ (the regime of of interest in our work), these regret bounds are no longer sublinear in $T$. We note that these papers also have “problem-dependent” bounds that scale as $O(\log T)$ for the linear bandit; these bounds apply when there is a fixed constant gap between the mean rewards of any pair of arms. Thus, these bounds do not apply to our setting since there is no such constant gap. In particular, the mean rewards of arm $i$ and $j$ can be arbitrarily close depending on the observed covariates $X_t$ at time $t$. 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 rely on a fixed action space (thus, their results cannot be mapped to a contextual bandit).
But adversarial covariate sequences constitute 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 $O(d^3 \log T)$ on cumulative regret, which significantly improves the $O(d \sqrt{T})$ bound for possibly adversarial covariate sequences as $T$ grows large. We adapt this idea to the high-dimensional setting. 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 $O(s_0^2 [\log T + \log d] \log T)$. Note that we relax the polynomial dependence on $d$ to a poly-logarithmic factor. As a consequence of our new proof technique, we also improve the regret bound in the low-dimensional setting from $O(d^3 \log T)$ (Goldenshluger and Zeevi 2013) to $O\left(d^2 \log^3 d \cdot \log T\right)$. These results hold while allowing for some arms to be uniformly sub-optimal; in contrast, the formulation in Goldenshluger and Zeevi (2013) requires a strong assumption that every arm is optimal for some subset of users.

It is also worth noting that 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. 2009, 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.

Finally, there has been recent 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.

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. Empirical results on simulated data as well as our evaluation on real patient data for the task of warfarin dosing are presented in §5. Finally, we state our secondary result in the low-dimensional setting in §6 and conclude in §7. Proofs and technical details are relegated to the appendices.
2. Problem Formulation

For any integer \( n \), we will let \([n]\) denote the set \( \{1, \ldots, n\} \). Let \( T \) be the number of (unknown) time steps; at each time step, a new user arrives and we observe her individual covariates \( X_t \in \mathcal{X} \subset \mathbb{R}^d \).

The observed sequence of covariates \( \{X_t\}_{t \geq 0} \) are drawn i.i.d. from a fixed (unknown) distribution \( P_X \). The decision-maker has access to \( K \) arms (decisions), which each yield uncertain user-specific rewards (e.g., patient outcome or profit from a user conversion). Each arm \( i \) is associated with an unknown parameter \( \beta_i \in \mathbb{R}^d \). At time \( t \), if we pull arm \( i \in [K] \), we yield reward

\[ X_t^T \beta_i + \varepsilon_{i,t}, \]

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

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

This definition implies \( E[z] = 0 \) and \( \text{Var}[z] \leq \sigma^2 \). 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.

Thus, the goal is to design a sequential decision-making policy \( \pi \) 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 \( \pi \) at time \( t \in [T] \). We compare ourselves to an oracle policy \( \pi^* \) that already knows the \( \{\beta_i\} \) (but not the noise \( \varepsilon \)) and thus always chooses the best expected arm \( \pi^*_t = \max_j (X_t^T \beta_j) \). Thus, if we choose arm \( \pi_t = i \) at time \( t \), we incur expected regret

\[ r_t \equiv E \left[ \max_j (X_t^T \beta_j) - X_t^T \beta_i \right], \]

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

We also define 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. 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 probability distribution \( P_X \) over patient covariates be the uniform distribution over the \( d \)-dimensional unit cube \([0,1]^d\). Consider three arms whose corresponding arm parameters are given by \( \beta_1 = (1,0,\ldots,0) \), \( \beta_2 = (0,1,0,\ldots,0) \), and \( \beta_3 = (1/4,1/4,0,\ldots,0) \).
Assumption 1 (Parameter set). There exist positive constants $x_{\text{max}}$ and $b$ such that $\|X_t\|_\infty \leq x_{\text{max}}$ for all $t$ and $\|\beta_i\|_1 \leq b$ for all $i \in [K]$.

Our first assumption is that the observed covariates $X_t$ as well as the arm parameters $\beta_i$ 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., $|X_t^T \beta_i| \leq bx_{\text{max}}$ by Cauchy-Schwarz. This is likely satisfied since user covariates and outcomes are bounded in practice. Our example clearly satisfies this assumption with $x_{\text{max}} = 1$ 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 $i \neq j$, $\Pr[0 < |X^T (\beta_i - \beta_j)| \leq \kappa] \leq C_0 \kappa$ for all $\kappa \in \mathbb{R}^+$.

Our second assumption is a margin condition that ensures that the covariate distribution $P_X$ cannot diverge locally near a decision boundary, i.e., the hyperplane given by $\{X^T \beta_i = X^T \beta_j\}$ in $\mathcal{X}$ for any $i \neq j \in [K]$. (Note that point masses on the decision boundary are allowed.) Intuitively, even small errors in our parameter estimates can cause us to choose the wrong action (between arms $i$ and $j$) for covariates $X_t$ 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 covariates are drawn near these hyperplanes. Since $P_X(x) < \infty$ 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 $\mathcal{K}_{\text{opt}}$ and $\mathcal{K}_{\text{sub}}$ be mutually exclusive sets that include all $K$ arms. Sub-optimal arms $i \in \mathcal{K}_{\text{sub}}$ satisfy $X^T \beta_i < \max_{j \neq i} X^T \beta_j - h$ for some $h > 0$ and every $X \in \mathcal{X}$. On the other hand, each optimal arm $i \in \mathcal{K}_{\text{opt}}$, has a corresponding set

$$U_i \equiv \left\{ X \mid X^T \beta_i > \max_{j \neq i} X^T \beta_j + h \right\}.$$

We assume there exists $p_* > 0$ such that $\min_{i \in \mathcal{K}_{\text{opt}}} \Pr[U_i] \geq p^*$.

Our third assumption is that our $K$ arms can be split into two sets:

- Optimal arms $\mathcal{K}_{\text{opt}}$ that are strictly optimal for at least some subset of covariates. In other words, there exists a constant $h_{\text{opt}} > 0$ and some region $U_i \subset \mathcal{X}$ (that has positive probability $p_i > 0$) for each $i \in \mathcal{K}_{\text{opt}}$ such that $X^T \beta_i > \max_{j \neq i} X^T \beta_j + h_{\text{opt}}$ for all covariates $X \in U_i$.

- Sub-optimal arms $\mathcal{K}_{\text{sub}}$ that are strictly sub-optimal for all covariates, i.e., there exists a constant $h_{\text{sub}} > 0$ such that for each $i \in \mathcal{K}_{\text{sub}}$, $X^T \beta_i < \max_{j \neq i} X^T \beta_j - h_{\text{sub}}$ for every $X \in \mathcal{X}$.

In particular, we assume that every arm is either optimal (by a margin $h_{\text{opt}}$) for some patients or sub-optimal (by a margin $h_{\text{sub}}$) for all patients. For simplicity, we define the localization parameter $h = \min \{h_{\text{opt}}, h_{\text{sub}}\}$ and $p_i = \min_{i \in \mathcal{K}_{\text{opt}}} p_i$. 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 when \( X \in U_i \) for some \( i \in K_{opt} \). Thus, we will play each optimal arm \( i \in K_{opt} \) at least \( p_* T \) times in expectation with high probability (i.e., whenever \( X_t \in U_i \)). This ensures that we can quickly learn accurate parameter estimates for all optimal arms over time.

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

We briefly introduce some notation and state a definition for our final assumption.

**Notation.** For any index set \( I \subset [d] \), let \( \beta_I \in \mathbb{R}^d \) be the vector obtained by setting the elements of \( \beta \) that are not in \( I \) to zero. We also define, for a vector \( v \in \mathbb{R}^m \), the support of \( v \) (denoted \( \text{supp}(v) \)) to be the set of indices corresponding to nonzero entries of \( v \).

**Definition 2 (Compatibility condition).** For any matrix \( M \in \mathbb{R}^{d \times d} \) and a set of indices \( I \subseteq [d] \), we say the pair \((M, I)\) satisfies the compatibility condition with positive constant \( \phi_0 \) if for all \( v \in \mathbb{R}^d \) satisfying \( \|v_I\|_1 \leq 3\|v\|_1 \), it holds that \( \|v_I\|_2 \leq \frac{\|v\|_2}{\phi_0} (v^T M v) \).

**Assumption 4 (Compatibility condition).** For each \( i \in K_{opt} \), the pair \((\Sigma_i, \text{supp}(\beta_i))\) satisfies the compatibility condition with constant \( \phi_0 > 0 \), where we define \( \Sigma_i \equiv \mathbb{E}[X X^T | X \in U_i] \).

Our fourth and final assumption concerns the covariance matrix of samples restricted to the regions \( U_i \) for each \( i \in K_{opt} \). In particular, we require that \( \Sigma_i \equiv \mathbb{E}_{X \sim \mathcal{P}_X}[X X^T | X \in U_i] \) satisfies a compatibility condition 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. 2009, Bühlmann and Van De Geer 2011). As we discussed earlier in Assumption 3, for each \( i \in K_{opt} \), we expect to play arm \( i \) at least \( p_* T = O(T) \) times based on samples \( X \in U_i \). The compatibility condition ensures that a LASSO estimator trained on these samples will converge. 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 covariance matrix be positive-definite, i.e., \( \lambda_{\min}(\Sigma_i) = \lambda_0 > 0 \). It can be easily verified that if \( \Sigma_i \) is positive-definite, then it satisfies the compatibility condition for any \( I \subseteq [d] \) with constant \( \phi_0 = \sqrt{\lambda_0} \). Thus, the compatibility condition is strictly weaker than the requirement that \( \Sigma_i \) be positive-definite; for example, the compatibility condition allows for some limited collinearity in the covariates, which can occur often in high-dimensional settings (see Chapter 6 in Bühlmann and Van De Geer (2011)).

In general, non-degenerate regions of \( \mathcal{X} \) that have positive support under \( \mathcal{P}_X \) will have a positive-definite covariance matrix. This requirement is clearly met in our example for the regions \( U_i \) (defined by any allowable choice of \( h \in (0, 1/2] \)) for each \( i \in K_{opt} \). 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 \( \lambda_0 \). For example, \( h = 0.1 \) corresponds to \( \lambda_{\min}(\Sigma_i) = \lambda_0 \approx 0.01 \). Thus, the covariance matrices \( \Sigma_i \) also satisfy the compatibility condition.

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 entries. 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 \( P_X \) over patient covariates to be any discrete distribution over the vertices of the \( d \)-dimensional unit cube \( \{0, 1\}^d \). Note that \( P_X \) now diverges at each vertex, but Assumption 2 is still satisfied because all the vertices lie on the decision boundary (where \( X^T \beta_1 = X^T \beta_2 \)) 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 more generic example that satisfies all the above assumptions. Let the probability distribution \( P_X \) have support on a bounded context set \( \mathcal{X} \). Consider the case where \( P_X \) does not diverge anywhere on \( \mathcal{X} \) and is bounded below from zero on an open set around the origin. Examples of such probability distributions include uniform, truncated Gaussian, etc. We further assume that no arm lies on the edge of the convex hull of all \( K \) arms, i.e., every arm is either a vertex (optimal locally) or is contained inside the convex hull (sub-optimal everywhere). When arms are generated uniformly randomly, this condition holds with probability one.

As mentioned earlier, some components of \( P_X \) may be discrete over a finite support.

### 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 \( \beta_i \). 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 \( \beta_i \). We tackle this by 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. Furthermore, we specify 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.

3.1. LASSO Estimation

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

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

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

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

The LASSO estimator converges with high probability according to the following oracle inequality (Proposition 1) if $X$ satisfies the compatibility condition (Definition 2).

**Proposition 1 (LASSO Oracle Inequality for Adapted Observations).** Let $X_i$ denote the $t^{th}$ row of $X$ and $y_t$ denote the $t^{th}$ entry of $Y$. The sequence $\{X_t : t = 1, \ldots, n\}$ 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 $(\hat{\Sigma}(X), \text{supp}(\beta))$ satisfies the compatibility condition with constant $\phi_0$ and $\|X_t\|_\infty \leq x_{\text{max}}$, the following oracle inequality holds for $\chi \equiv 4\|\beta\|_0 \lambda / \phi_0^2$:

$$\Pr \left( \|\hat{\beta}_{X,Y}(\lambda) - \beta\|_1 > \chi \right) \leq \exp \left( -C_1 n \chi^2 + \log d \right),$$

where we define $C_1 = \phi_0^2 / (512 \|\beta\|_0^2 \sigma^2 x_{\text{max}}^2)$. 

Note that the regularization parameter $\lambda$ determines the error size and convergence rate $\chi$.

**Remark 1.** 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 with independent $\sigma$-subgaussian errors. The result follows from modifying the proof of the standard LASSO oracle inequality using martingale theory (see Appendix A).

Now, we consider the task of estimating the parameter $\beta_i$ for each arm $i \in [K]$. Using any subset of past samples $S' \subset S_i$ where arm $i$ was played and any choice of regularization parameter $\lambda \geq 0$, we can use the corresponding LASSO estimator $\hat{\beta}_{X(S'),Y(S'),\lambda}$ (which we denote by the simpler notation $\hat{\beta}(S',\lambda)$) to estimate $\beta_i$. In order to prove regret bounds, we need to establish convergence guarantees for such estimates. From Proposition 1, in order to bound the error $\|\hat{\beta}(S',\lambda) - \beta_i\|_1$ for each arm $i \in [K]$, we need to (i) ensure $\hat{\Sigma}(S')$ satisfies the compatibility condition for some constant and (ii) appropriately choose the regularization parameter $\lambda$ over time to control the rate of convergence. Thus, the main challenge in the algorithm and analysis is constructing and maintaining sets $S'$ such that the compatibility condition is satisfied (although the rows of $X(S')$ are not i.i.d.) with sufficiently fast convergence rates.

### 3.2. Description of Algorithm

The inputs are 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)$^1$, 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 $X_t$):

\[
T_i = \left\{(2^n - 1) \cdot Kq + j \mid n \in \{0, 1, 2, \ldots\} \text{ and } j \in \{q(i-1) + 1, q(i-1) + 2, \ldots, iq\}\right\}. \tag{2}
\]

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

**All-Sample Sets.** As before, let $S_{i,t} = \{t' \mid \pi_{t'} = i \text{ and } 1 \leq t' \leq t\}$ denote the set of times we play arm $i$ up to time $t$. Note that by definition $T_{i,t} \subset S_{i,t}$

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

1. the forced-sample estimate $\hat{\beta}(T_{i,t-1}, \lambda_1)$ based only on forced samples observed from arm $i$,
2. the all-sample estimate $\hat{\beta}(S_{i,t-1}, \lambda_2)$ based on all samples observed from arm $i$.

\[^1\text{Note that if some } \bar{h} \text{ satisfies Assumption 3, then any } h \in (0, \bar{h}] \text{ also satisfies the assumption. Therefore, a conservatively small value can be chosen in practice, but this will be reflected in the constant in the regret bound.}\]
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 $\hat{K} \subset [K]$ 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 $\hat{K}$.

**Algorithm LASSO Bandit**

**Input parameters:** $q, h, \lambda_1, \lambda_2, 0$

Initialize $\hat{\beta}(\mathcal{T}_{t,0}, \lambda_1)$ and $\hat{\beta}(\mathcal{S}_{i,0}, \lambda_2, 0)$ by 0 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 $X_i \in \mathcal{P}_X$.
  if $t \in \mathcal{T}_i$ for any $i$ then
    $\pi_t \leftarrow t$
  else
    $\hat{K} = \{ i \in K \mid X_i^T \hat{\beta}(\mathcal{T}_{i,t-1}, \lambda_1) \geq \max_{j \in K} X_j^T \hat{\beta}(\mathcal{T}_{j,t-1}, \lambda_1) - h/2 \}$
    $\pi_t \leftarrow \arg \max_{i \in \hat{K}} X_i^T \hat{\beta}(\mathcal{S}_{i,t-1}, \lambda_{2,t-1})$
  end if
  $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{ t \}$, $\lambda_{2,t} \leftarrow \lambda_{2,0} \sqrt{\frac{\log t + \log d}{t}}$
  Play arm $\pi_t$, observe $y_t = X_i^T \beta_{\pi_t} + \varepsilon_{i,t}$
end for

3.3. Main Result: Regret Analysis of LASSO Bandit

**Theorem 1.** When $q \geq 4 \lfloor q_0 \rfloor$, $K \geq 2$, $\log d > 1$, $t \geq (Kq)^2$, and we take $\lambda_1 = \frac{\phi_0^2 p^2 h}{64 s_0 x_{\max}}$ and $\lambda_2, 0 = \frac{\phi_0^2 s_{\max}}{2 s_0 \sqrt{p^* C_1}}$, we have the following (non-asymptotic) upper bound on the expected at cumulative regret of the LASSO Bandit at time $T$ by:

$$R_T \leq C_3 \left( \log T \right)^2 + \left( 2K bx_{\max} (6q + 5) + C_3 \log d \right) \log T + \left( 2q^2 K^2 bx_{\max} + 6K bx_{\max} + C_4 \right)$$

where the constants are given by

$$C_1 \equiv \frac{\phi_0^4}{512 s_0^2 \sigma^2 x_{\max}^2}, \quad C_2 \equiv \frac{\phi_0^2}{384 s_0 x_{\max}^2}, \quad C_3 \equiv \frac{1024 K C_0 x_{\max}^4}{p^* C_1}, \quad \text{and} \quad C_4 \equiv 2 \left( 1 - \exp \left[ - p^2 (C_2 \wedge [1/2]) \right] \right)^{-1}$$

and we take $q_0 \equiv \max \left\{ \frac{20}{p^* C_2}, \frac{4}{p^* C_2}, \frac{4x_{\max} \log d}{p^* C_2}, \frac{256 x_{\max} \log d}{p^* C_2} \right\} = O(s_0^2 \sigma^2 \log d)$.

**Lower Bound.** Goldenshluger and Zeevi (2013) prove an information-theoretic lower bound on the expected cumulative regret of $O(\log T)$ for a (low-dimensional) multi-armed bandit with covariates. Since our formulation encompasses their setting, the same lower bound applies to our setting as well. 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 $\mathcal{K}_{\text{opt}} = \{1, 2\}$. Thus, we can conclude that our upper
bound of $O\left(\log T\right)^2$ for the expected cumulative regret of the LASSO Bandit is near-optimal in $T$ since it matches the lower bound up to a factor of $\log T$.

**Remark 2.** 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 oracle 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. An Oracle Inequality for non-i.i.d. Data

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

$$W = Z\beta + \varepsilon$$

where $Z_{n \times d}$ 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 $\sigma$-subgaussians. Let the rows $Z_t$ of $Z$ be bounded, i.e., $\|Z_t\|_{\infty} \leq z_{\text{max}}$ for all $t \in [n]$. Following the notation introduced earlier in §3.1, for any subset $A \subset [n]$ we define the analogous quantities $Z(A), W(A), \hat{\Sigma}(A)$. Then, for any $\lambda \geq 0$ we have a LASSO estimator trained on samples in $A$:

$$\hat{\beta}(A, \lambda) \equiv \arg\min_{\beta} \left\{ \left\| W(A) - Z(A)\beta \right\|^2_{2} / |A| + \lambda \|\beta\|_1 \right\} .$$

Note that we have not made any distributional (or i.i.d.) assumptions on the samples in $A$. We now consider that some unknown subset $A' \subset A$ comprises of i.i.d. samples with distribution
\( P \), i.e., \( \{ Z_t \mid t \in A' \} \sim P \times \cdots P \). Letting \( \Sigma = E_{Z \sim P} [ZZ^T] \), we further assume that \( P \) is such that \((\Sigma, \text{supp}(\beta))\) satisfies the compatibility condition with constant \( \phi_0 \). We will show that if the number \( |A'| \) of i.i.d. samples is sufficiently large, then we can prove a convergence guarantee for a LASSO estimator \( \hat{\beta}(A, \lambda) \) trained on samples in \( A \), which includes non-i.i.d. samples. (Note that \( A' \) is unknown; if not, we can simply use the estimator \( \hat{\beta}(A', \lambda) \) trained only on the i.i.d. samples in \( A' \).) In particular, suppose that at least some constant fraction of the samples in \( A \) belong in \( A' \), i.e., \( |A'|/|A| \geq p/2 \) for a positive constant \( p \). We then have the following convergence guarantee:

**Lemma 1.** If \( |A'|/|A| \geq p/2, |A| \geq \frac{4z_{\text{max}} \log d}{pC_2} \) and \( \lambda = \frac{\phi_0 p}{16 \| \beta \|_0} \), the following oracle inequality holds \( \forall \chi > 0 \):

\[
\Pr \left[ \| \hat{\beta}(A, \lambda) - \beta \|_1 > \chi \right] \leq \exp \left[ -\frac{|A| C_1 p^2 \chi^2}{16 z_{\text{max}}^2} + \log d \right],
\]

with probability at least \( 1 - \exp \left[ -pC_2 |A|/2 \right] \).

Note that we the constants \( C_1, C_2 \) were defined in \( \S 3.3 \) but we have replaced \( x_{\text{max}} \) and \( s_0 \) with the analogous quantities \( z_{\text{max}} \) and \( \| \beta \|_0 \) respectively.

**Remark 3.** Note that in Lemma 1 we can use the union bound and state a simpler oracle inequality: given all the assumptions of Lemma 1, for all \( \chi > 0 \),

\[
\Pr \left[ \| \hat{\beta}(A, \lambda) - \beta \|_1 > \chi \right] \leq \exp \left[ -\frac{|A| C_1 p^2 \chi^2}{16 z_{\text{max}}^2} + \log d \right] + \exp \left[ -pC_2 |A|/2 \right].
\]

However, stating a bound for the two events separately simplifies our analysis of the regret for using the all-sample estimator (Lemmas 21 and 22 in Appendix E).

The proof proceeds as follows. We first show that \((\hat{\Sigma}(A'), \text{supp}(\beta))\) satisfies the compatibility condition with constant \( \phi_0/\sqrt{2} \) with high probability. This involves showing that \( \| \hat{\Sigma}(A') - \Sigma \|_\infty \) is small with high probability using results on matrix perturbations. Next, we use this fact along with the Azuma-Hoeffding inequality to show that \((\hat{\Sigma}(A), \text{supp}(\beta))\) satisfies the compatibility condition with constant \( \phi_0 \sqrt{2}/2 \) with high probability. Applying Proposition 1 to this result implies that an oracle inequality holds with high probability for LASSO estimates \( \hat{\beta}(A, \lambda) \) although part of the data is not generated i.i.d. from \( P \). The full proof is given in Appendix B.

### 4.2. Oracle Inequality for the Forced-Sample Estimator

We now obtain an oracle inequality for the forced-sample estimator \( \hat{\beta}(T_{i,t}, \lambda_1) \) of each arm \( i \in [K] \).

**Proposition 2.** The forced sample estimator \( \hat{\beta}(T_{i,t}, \lambda) \) satisfies the oracle inequality

\[
\Pr \left[ \| \hat{\beta}(T_{i,t}, \lambda_1) - \beta_i \|_1 > \frac{h}{4x_{\text{max}}} \right] \leq \exp \left[ -q_0 \log t \cdot \frac{p^2 h^2 C_1}{256 x_{\text{max}}^4} + \log d \right] + \frac{2}{t},
\]

when \( \lambda_1 = \frac{\phi_0 p h}{64q_0 x_{\text{max}}} \) and \( t \geq (Kq)^2 \).
Note that the sample covariance matrix from the forced-samples \( \hat{\Sigma}(T_{i,t}) \) concentrates around \( \mathbb{E}_{X \sim \mathcal{P}_X}[XX^T] \). Thus, although this estimator is trained on i.i.d. samples from \( \mathcal{P}_X \), the above oracle inequality does not follow directly from Proposition 1 since we have only assumed that the compatibility condition holds for \( \Sigma = \mathbb{E}_{X \sim \mathcal{P}_X}[XX^T] \) rather than \( \mathbb{E}_{X \sim \mathcal{P}_X}[XX^T] \) \( (\text{Assumption 4}) \).

This is easily resolved by showing \( \mathcal{T}'_{i,t} \equiv \{ t' \in T_{i,t} \mid X_{t'} \in U_i \} \) is a set of i.i.d. samples from \( \mathcal{P}_{X|X \in U_i} \), and then applying Lemma 1 with \( A = T_{i,t} \) and \( A' = T_{i,t}' \). The full proof is given in Appendix C.

4.3. Oracle Inequality for the All-Sample Estimator

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

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

\[
A_t \equiv \left\{ \| \hat{\beta}(T_{i,t}, \lambda_1) - \beta_i \|_1 \leq \frac{h}{4x_{\max}}, \quad \forall i \in [K] \right\}.
\]  

Since the event \( A_t \) only depends on forced-samples, the random variables \( \{X_t \mid A_{t-1} \text{ holds and } t \notin T_{i,t}\} \) are i.i.d. (with distribution \( \mathcal{P}_X \)). Furthermore, if we let \( S'_{i,t} \equiv \{ t' \in [t] \mid A_{t'-1} \text{ holds and } t \notin T_{i,t} \text{ and } X_{t'} \in U_i \} \), then the random variables \( \{X_{t'} \mid t' \in S'_{i,t}\} \) 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 \( S'_{i,t} \subseteq S_{i,t} \). Finally, we will use Proposition 2 to show that \( |S'_{i,t}| \) is sufficiently large, so we can use Lemma 1 with \( A = S_{i,t} \) and \( A' = S'_{i,t} \) 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 \( \hat{\beta}(S_{i,t}, \lambda_{2,t}) \) for \( i \in \mathcal{K}_{opt} \) satisfies the oracle inequality

\[
\Pr \left[ \| \hat{\beta}(S_{i,t}, \lambda_{2,t}) - \beta_i \|_1 > 16x_{\max} \sqrt{\frac{\log t + \log d}{p^3 C_1^1 t}} \right] < \frac{1}{t},
\]

with probability at least \( 1 - 2 \exp \left[ -\frac{p^2 (C_2 \sqrt{t}/2)}{16} \right] t \) when \( \lambda_{2,t} = \frac{\phi_d x_{\max}}{2x_0} \sqrt{\frac{\log t + \log d}{p^3 C_1^1 t}} \) and \( t \geq (Kq)^2 \).
In particular, Proposition 3 guarantees \( \|\hat{\beta}(S_{i,t}, \lambda_2) - \beta_i\|_1 = O(\sqrt{\log t/t}) \) with high probability while Proposition 2 only guarantees \( \|\hat{\beta}(T_{i,t}, \lambda_1) - \beta_i\|_1 = O(1) \) with high probability. However, note that the all-sample estimator oracle inequality only holds for optimal arms \( K_{opt} \) while the forced-sample estimator oracle 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 \( O(T) \) i.i.d. samples for each \( i \in K_{opt} \) (required for the proof of Proposition 3), and (ii) to prune out sub-optimal arms \( K_{sub} \) 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 D.

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 (\( t \leq (Kq)^2 \)) and forced sampling (\( t \in T_{i,T} \) 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 LASSO Bandit plays the estimated best arm from \( \hat{K} \) (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 \( \hat{K} \) and (ii) does not include any sub-optimal arms from \( K_{sub} \) in \( \hat{K} \). Thus, when \( A_{t-1} \) holds, we can apply the convergence properties of the all-sample estimator (which only hold for optimal arms) to \( \hat{K} \) without the concerns that \( \hat{K} \) 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 \( 2qKbx_{\max}(6\log T + Kq) \) (Lemma 17). This follows from the fact that the worst-case regret at any time step is at most \( bx_{\max} \) (Assumption 1), while there are only \( (Kq)^2 \) initialization samples and at most \( 6Kq \log T \) forced samples up to time \( T \) (Lemma 10).

Next, the cumulative expected regret from time periods in group (b) at time \( T \) is bounded by at most \( 2Kbx_{\max}(2\log T + 3) \) (Lemma 19). This follows from the oracle 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 \( K \exp[-\log t \log d + \log d + 2K/t] \). The result follows from summing this quantity over time periods from \((Kq)^2 < t \leq T\).

Finally, the cumulative expected regret from time periods (c) at time \( T \) is bounded by at most \( (4Kbx_{\max} + C_3 \log d) \log T + C_3 (\log T)^2 + C_4 \) (Lemma 22). To show this, we first observe that
if event $A_t$ holds, then the set $\hat{K}$ (chosen by the forced-sample estimator) contains the optimal arm $i^* = \arg\max_{i \in [K]} X^T_t \beta_i$ and no sub-optimal arms from the set $K_{sub}$ (Lemma 20). Then, we separately sum our expected regret when the all-sample oracle inequalities (Proposition 3) does and does not hold for all optimal arms. In the former case, our all-sample estimators for each optimal arm satisfy $\|\hat{\beta}(S_{i,t}, \lambda_{2,t}) - \beta_i\|_1 = \mathcal{O}(\sqrt{\log t/t})$ with high probability; thus, as shown in Lemma 21, we only incur regret if the observed covariates are within a $\mathcal{O}(\sqrt{\log t/t})$ distance from a decision boundary (which occurs with small probability as guaranteed by Assumption 2). In the latter case, we incur worst-case regret, but this occurs with exponentially vanishing 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 17, 19, and 22:

$$R_T \leq \sum_{a} \text{Regret from (a)} + \sum_{b} \text{Regret from (b)} + \sum_{c} \text{Regret from (c)}$$

\[ R_T \leq 2qKbx_{\max}(6 \log T + Kq) + 2Kbx_{\max}(\log T + 3) + (4Kbx_{\max} + C_3 \log d) \log T + C_3 (\log T)^2 + C_4 \]

5. Empirical Results

We present two sets of empirical results. First, we evaluate the LASSO Bandit on sparse synthetic data and demonstrate its superior performance compared to existing bandit algorithms in the literature. Next, we evaluate our algorithm on the task of warfarin dosing using a real patient dataset where our technical assumptions may not hold. We find that the LASSO bandit still performs favorably, suggesting that the LASSO Bandit can provide significant value in practical decision-making contexts even if they do not match the ideal theoretical setting described in §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, surprisingly, even in low-dimensional regimes. Separately, we find that the LASSO Bandit is robust to changes in input parameters by even an order of magnitude.

Synthetic Data Generation. We consider two arms (treatments) and $d = 100$ user covariates, where only a randomly chosen subset of $s_0 = 5$ covariates are predictive of the reward for each
treatment. In particular, for each $i = 1, 2$, the arm parameters $\beta_i$ are set to zero except for $s_0$ randomly selected components that are drawn from a uniform distribution on $[0, 1]$.

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

**Algorithm Inputs.** Bandit algorithms require the decision-maker to specify a variety of problem constants that are often unknown in practice. For instance, Theorem 1 suggests specific input parameters for the LASSO Bandit (e.g., $\sigma, \phi_0$) and similarly, the benchmark OFUL and OLS Bandit algorithms require analogous specifications. Therefore, in order to simulate a realistic environment, we make ad-hoc choices for 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). As a robustness check, we 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 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.

Furthermore, we note that the LASSO Bandit continues to achieve significantly less per-time-step regret than the alternative algorithms even for large $t$. For examples, 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.

**Robustness to Algorithm Inputs.** We now compare the cumulative 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 $c \in \{0.02, 0.05, 0.1, 0.2\}$. The results are computed over $T = 10,000$ time steps and averaged over 10 trials (see Figure 2). We find that the cumulative regret performance is not hugely impacted despite experimenting with the parameters by up to an order of magnitude. This suggests that the LASSO Bandit is robust, which is important in practice since the input parameters are likely to be specified incorrectly.

---

2 The OFUL-EG algorithm requires an additional technical assumption that $\|\beta_1\|_1 + \|\beta_2\|_1 = 1$. We scale our $\beta_i$’s accordingly so that this assumption is met.

3 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.
Figure 1  Comparison of the cumulative regret of the LASSO Bandit against other bandit algorithms on synthetic data. Figure (a) is a zoomed-in version of figure (b).
Figure 2  Cumulative regret for LASSO Bandit on synthetic data for varying values of inputs, i.e. (a) the forced sampling parameter $q$, (b) the localization parameter $h$, and (c) the coefficient $c$ of the regularization parameters.

5.2. Case Study: Warfarin Dosing

We now demonstrate the practical value of the LASSO Bandit in a medical decision-making context by evaluating it on the clinical task of warfarin dosing. In general, evaluating a bandit algorithm retrospectively on data is challenging because we require access to counterfactuals. In particular, the bandit algorithm may choose a different action than the one taken in the data and thus, we need an unbiased estimate of the resulting reward. We choose warfarin dosing as our case study since this unique dataset gives us access to such counterfactuals.
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 LASSO, support vector regression, regression trees, etc.) for the objective of predicting the correct warfarin dosage as a function of the given patient-level variables.

**Remark 4.** The above result suggests 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 bandit 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 $\beta_i$ 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$^4$ 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.

Evaluation and Results. We consider 10 bootstrap samples of patients and simulate the following policies:

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

We sequentially draw random bootstrapped samples of patients and simulate the actions and feedback of each of the six policies. (Note that we can evaluate the reward of any action since we know each patient’s true optimal dosage.) Figure 3 compares the 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 95% confidence intervals.

We first note that the LASSO Bandit outperforms the three other bandit algorithms for any number of patients with statistical significance. 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 - 1000 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.

$^4$ We have chosen a 0-1 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 is robust to such model mis-specifications in practice.

$^5$ 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(±) algorithm (Kivinen and Warmuth 1997), which does not require this assumption.
Figure 3  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 seen in the warfarin dosing data.

**Big Data.** When there are a large number of samples (1000 - 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\(^6\) algorithms still perform worse than doctors.

Note that all three existing bandit algorithms required more than a 1000 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 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 build increasingly complex predictive models (by allowing more covariates in the model over time), which helps it make reasonable decisions in the small-data regime without sacrificing performance in the big-data regime.

\(^6\) The theory behind the OFUL-EG algorithm is very dependent on sparsity in the data. This may explain why it performs so poorly in this experiment since all covariates seem to be predictive in the warfarin dataset.
**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.

<table>
<thead>
<tr>
<th>True Dosage</th>
<th>LASSO Bandit Policy Assigned Dosage</th>
<th>Physician Policy Assigned Dosage</th>
<th>% of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Low</td>
<td>57%</td>
<td>42%</td>
<td>1%</td>
</tr>
<tr>
<td>Medium</td>
<td>14%</td>
<td>83%</td>
<td>3%</td>
</tr>
<tr>
<td>High</td>
<td>3%</td>
<td>90%</td>
<td>7%</td>
</tr>
</tbody>
</table>

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.

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.

**Remark 5.** Finally, we note that the warfarin dosing task is not a truly bandit problem; in particular, we always observe the optimal arm (patient’s true dose) even if we play the wrong arm (assign the wrong dose initially) because the doctor tunes the dosage over time. Yet, we use this setting as a case study to evaluate bandit policies since we have access to 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. However, 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. 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 $O(d^2 \log \frac{d}{2} \log T)$ on the cumulative expected regret of the OLS Bandit; this is an improvement over the existing $O(d^3 \log T)$ bound.

**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 $K_{sub}$ 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 5 (Positive-Definiteness).** Define $\Sigma_i \equiv \mathbb{E} [XX^T | X \in U_i]$ for all $i \in [K]$. Then, there exists $\phi_0 > 0$ such that for all $i \in [K]$ the minimum eigenvalue $\lambda_{\min} (\Sigma_i) \geq \phi_0 > 0$.

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

**Definition 4 (OLS).** We define the OLS estimator for estimating the parameter $\beta$:

$$\hat{\beta}_{X,Y} \equiv (X^T X)^{-1} X^T Y. \quad (5)$$

The OLS estimator converges with high probability according to the following tail inequality (see Proposition 4) if the covariance matrix $\hat{\Sigma}(X)$ is positive definite.

**Proposition 4 (OLS Tail Inequality).** Let $X_t$ denote the $t^{th}$ row of $X$ and $y_t$ denote the $t^{th}$ entry of $Y$. The sequence $\{X_t : t = 1, \ldots, n\}$ 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 the minimum eigenvalue of $\hat{\Sigma}(X)$ is greater than $\phi_0$ and $\|X_t\|_\infty \leq x_{\max}$, the following oracle inequality holds for all $\chi > 0$:

$$\Pr \left[ \|\hat{\beta} - \beta\|_1 \leq \chi \right] \geq 1 - \exp \left[ -D_1 n \chi^2 + \log d \right],$$

where we define $D_1 := \phi_0^2 / (2d^2 x_{\max}^2 \sigma^2)$.

**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 $T_{i,t}$ and all-sample sets $S_{i,t}$ 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).
6.1. New Upper Bound on Regret

**Theorem 2.** When \( q \geq 4[q_0], \ K \geq 2, \ \log d > 1, \) and \( T \geq (Kq)^2, \) we have an upper bound on the expected cumulative regret at time \( T: \)

\[
R_T \leq 2qKbx_{\text{max}}(6\log T + Kq) + 2Kbx_{\text{max}}(2\log T + 3) + \frac{212Kd_{\text{max}}\log d}{D_3} \log T + D_4
\]

where we define the constants

\[ D_1 := \frac{\phi_0^2}{2d^2x_{\text{max}}^2\sigma^2}, \quad D_2 := \frac{\phi_0}{8x_{\text{max}}}, \quad D_3 := \frac{p^2D_1}{32}, \quad \text{and} \quad D_4 := \frac{3}{1 - \exp\left[-\frac{p^2(D_2\wedge p_*)/2}{64}\right]}, \]

and we take

\[ q_0 = \max\left\{\frac{8\log 2d}{p_*(D_2 \wedge p_*)}, \frac{256x_{\text{max}}^2\log 2d}{h^2p_*^2D_1}\right\} = O\left(d^2\log d\right). \]

**Algorithm OLS Bandit**

**Input parameters:** \( q, h \)

Initialize \( \hat{\beta}(\mathcal{T}_{i,0}) \) and \( \hat{\beta}(\mathcal{S}_{i,0}) \) by 0 for all \( i \) in \([K]\)

Use \( q \) to construct force-sample sets \( \mathcal{T}_i \) using Eq. (2) for all \( i \) in \([K]\)

for \( t \in [T] \) do

Observe \( X_t \in \mathcal{P}_X \)

if \( t \in \mathcal{T}_i \) for any \( i \) then

\( \pi_t \leftarrow i \)

else

\( \hat{\mathcal{K}} = \{i \in K \mid X^T_i \hat{\beta}(\mathcal{T}_{i,t-1}) \geq \max_{j \in K} X^T_i \hat{\beta}(\mathcal{T}_{j,t-1}) - h/2\} \)

\( \hat{\pi}_t \leftarrow \arg\max_{i \in \hat{\mathcal{K}}} X^T_i \hat{\beta}(\mathcal{S}_{i,t-1}) \)

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

Play arm \( \pi_t \), observe \( y_t = X^T_{\pi_t} \beta_{\pi_t} + \varepsilon_{\pi,t} \)

end if

end for

**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 even if only a constant fraction of the rows of the design matrix are independent. In particular, for a fixed subset \( \mathcal{A} \) of \([n]\), 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 \( \mathcal{P}_Z \) and \( \mathcal{A}'/|\mathcal{A}| \geq p/2 \) for a positive constant \( p \), then \( \hat{\Sigma} (\mathcal{A}) \) is positive-definite with minimum eigenvalue bounded below by \( \phi_0\sqrt{p}/2 \) with high probability. (See Appendix F.)

**Lemma 2.** Under the assumptions above, the following tail inequality holds \( \forall \chi > 0: \)

\[
\Pr\left[\|\hat{\beta}(\mathcal{A}) - \beta\|_1 \leq \chi\right] \geq 1 - \exp\left[-|\mathcal{A}|\chi^2 \cdot \frac{pD_1}{4} + \log 2d\right],
\]

with probability at least \( 1 - \exp\left[-pD_2|\mathcal{A}|/2 + \log d\right]. \)
We use this lemma to prove analogous tail inequalities for the forced-sample estimator (Proposition 5) and the all-sample estimator (Proposition 6) in Appendix G. Finally, we use these tail inequalities to sum up the expected regret contributions from the three groups of time periods:

(a) Initialization \( (t \leq (Kq)^2) \) and forced sampling \( (t \in T_{i,T} \text{ for some } i \in [K]) \).
(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 \( \hat{K} \) using the all-sample estimator.

Summing the results concludes the proof of Theorem 2. The proof is given in Appendix H.

**Remark 6.** Our upper bound on the cumulative expected regret of the OLS Bandit scales as \( O(\log T) \) while that of the LASSO Bandit scales as \( O([\log T]^2) \). This is because the tail inequality for the OLS estimator (Proposition 4) holds for any \( \chi > 0 \), while the oracle inequality for the LASSO estimator (Proposition 1) only holds for a single value of \( \chi = \frac{16\|\beta\|_0}{\phi_p} \lambda \), which is fixed by the choice of regularization parameter \( \lambda \). It remains an open question whether our analysis of the LASSO Bandit can be improved to meet the current lower bound of \( O(\log T) \).

### 7. Conclusions

We present the LASSO Bandit algorithm for multi-armed 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 the 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.

There are several directions for future work. First, our results can be extended from the linear setting to more general function classes such as generalized linear models, which have proven to be useful in several applications (Li et al. 2012). Second, our algorithm relies on a prescribed schedule for exploration. One could explore UCB or Thompson sampling variants of our algorithm. Such methods have been found to improve empirical performance in other bandit settings. Finally, our algorithm relies on forced exploration, which may be prohibitively costly or unethical in settings such as medical decision-making. One could also explore the performance of greedy algorithms, which may be more appropriate for some applications.

### Acknowledgments

The authors gratefully acknowledge the National Science Foundation for financial support through the following grants: Graduate Research Fellowship Grant No. DGE-114747, NSF EAGER award CMMI:1451037, NSF CAREER award CMMI: 1554140, and NSF grant CCF:1216011.
This paper has also benefitted from valuable feedback from Steve Chick, Hamid Nazerzadeh, and various seminar participants. They have been instrumental in guiding us to improve the paper.

References


Appendix

A. Proof of LASSO Oracle Inequality for Adapted Observations

Recall that we are considering the linear model $Y = X\beta + \varepsilon$. The design matrix is $X \in \mathbb{R}^{n \times d}$, the response vector is $Y \in \mathbb{R}^n$ and the entries of the noise vector $\varepsilon \in \mathbb{R}^n$ are independent draws from a $\sigma$-subgaussian distribution. Let $X_t$ denote the $t^{th}$ row of $X$ and $y_t$ denote the $t^{th}$ entry of $Y$. The sequence $\{X_t: t = 1, \ldots, n\}$ forms an adapted sequence of observations, i.e., $X_t$ may depend on past regressors and their resulting observations $\{X_{t'}, y_{t'}\}_{t'=1}^{t-1}$. Finally, we assume $\|X_t\|_\infty \leq x_{\text{max}}$ with probability 1.

We now prove a LASSO oracle inequality (Proposition 1) for adapted sequences of observations with independent $\sigma$-subgaussian errors. The result follows from modifying the proof of the standard LASSO oracle inequality (e.g., see Theorem 6.1 in Bühlmann and Van De Geer (2011)) using martingale theory.

**Lemma 3 (Bernstein Concentration).** Let $\{D_k, \mathcal{G}_k\}_{k=1}^\infty$ be a martingale difference sequence, and let $D_k$ be $\sigma$-subgaussian. Then, for all $t \geq 0$,

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

**Proof of Lemma 3** See Theorem 2.3 of Wainwright (2016) and take $b_k = 0$ and $\nu_k = \sigma$ for all $k$. □
Lemma 4. Define the following event

\[ \mathcal{F} \equiv \left\{ \max_{r \in [d]} \left( \frac{1}{n} | 2\varepsilon^T X^{(r)} | \right) \leq \lambda_0 \right\}, \]

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

Proof of Lemma 4. Let \( D_{t,r} = \varepsilon_r X_{t,r} \), and note that \( \varepsilon^T X^{(r)} = \sum_{t=1}^{n} D_{t,r} \). Since

\[ \mathbb{E}_{\varepsilon,X}[e^{\varepsilon^T D}] \leq \mathbb{E}_X[e^{\varepsilon^2 X^{2}\sigma^2/2}] \leq e^{\varepsilon^2 x_{\text{max}}^2 \sigma^2/2}, \]

\( D_{t,r} \) is a \((x_{\text{max}} \sigma)\)-subgaussian random variable by Definition 1. Next, we note \( D_0, D_1, \ldots, D_n \) is a martingale adapted to the filtration \( \mathcal{G}_1 \subset \ldots \subset \mathcal{G}_n \) since \( \mathbb{E}[\varepsilon_r X_{t,r} | X_1, \ldots, X_{t-1}, y_1, \ldots, y_{t-1}] = 0 \). Then,

\[ \text{Pr}[\mathcal{F}] \geq 1 - d \text{Pr} \left[ \max_{k=1}^{n} D_k > \sigma x_{\text{max}} \sqrt{n} \sqrt{t^2 + 2 \log d} \right] \]
\[ \geq 1 - 2d \exp\left[-(t^2 + 2 \log d)/2\right] \]
\[ = 1 - 2 \exp[-t^2/2]. \]

The first inequality comes from a union bound over \( r \in [d] \) and the second inequality comes from applying Bernstein’s concentration bound for martingale difference sequences (Lemma 3).

Lemma 5 (From page 105 of (Bühlmann and Van De Geer 2011)). When \( \lambda \geq 2\lambda_0 \) and \( \mathcal{F} \) holds,

\[ 2\|X(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1 \leq 3\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1. \]

Now we are ready to prove Proposition 1.

Proposition 1. If \((\hat{\delta}(X), \text{supp}(\beta))\) satisfies the compatibility condition with constant \( \phi_0 \) and \( \|X_t\|_{\infty} \leq x_{\text{max}} \), the following oracle inequality holds for \( \chi \equiv 4\|\beta\|_0 \lambda / \phi_0^2 \):

\[ \text{Pr}\left[ \|\hat{\beta}_{X,Y} - \beta\|_1 > \chi \right] \leq \exp\left[-C_1 n \chi^2 + \log d \right], \]

where we define \( C_1 \equiv \phi_0^4/(512 \|\beta\|_0^2 \sigma^2 x_{\text{max}}^2) \).

Proof of Proposition 1. On the event \( \mathcal{F} \), we have that

\[ 2\|X(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta} - \beta\|_1 = 2\|X(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1 + \lambda\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1 \]
\[ \leq 4\lambda\|\hat{\beta}_{\text{supp}(\beta)} - \beta_{\text{supp}(\beta)}\|_1 \]
\[ \leq 4\lambda\sqrt{\|\beta\|_0 \|X(\hat{\beta} - \beta)\|_2^2/n \sqrt{n} \phi_0^2} \]
\[ \leq \|X(\hat{\beta} - \beta)\|_2^2/n + 4\lambda^2 \|\beta\|_0 \phi_0^2, \]

where we have used Lemma 5 and the definition of the compatibility condition (Definition 2). Thus, we have shown that for \( \lambda \geq 4\sigma \sqrt{(t^2 + 2 \log d)/n} \), the inequality \( \|\beta - \hat{\beta}\|_1 > 4\lambda\|\beta\|_0 / \phi_0^2 \) holds with probability at most \( \exp[-t^2/2] \) (Lemma 4). The result follows by taking \( t = \sqrt{2n C_1 \chi^2 / x_{\text{max}}^2 - 2 \log d} \).
B. Proof of LASSO Oracle Inequality for Non-i.i.d. Data

Consider the linear model \( W = Z\beta + \varepsilon \) where \( \varepsilon \) is a \( \sigma \)-subgaussian error term. Let the rows \( Z_t \) be non-iid samples from a bounded distribution \( P_Z \) with \( \|Z_t\|_\infty \leq z_{\max} \). Furthermore, let \( \Sigma \equiv E_{Z \sim P(Z)} [ZZ^T] \) and assume that \( (\Sigma, \text{supp}(\beta)) \) satisfies the compatibility condition with constant \( \phi_0 \). Let \( A \subset [n] \) such that \( |A| \geq \frac{4 \max \log d}{pC_2} \).

Furthermore, let \( A' \subset A \) such that \( \frac{|A'|}{|A|} \geq \frac{p}{2} \) for some \( p > 0 \), and \( \{Z_t \mid t \in A'\} \) is a subset of random variables drawn i.i.d. from \( P_Z \). We will prove an oracle inequality holds for the estimator \( \hat{\beta}(A, \lambda) \) with high probability. The proof involves showing that \( \|\hat{\Sigma}(A') - \Sigma\|_\infty \) is small with high probability using matrix perturbation theory. Next, we use the Azuma-Hoeffding inequality to show that \( (\hat{\Sigma}(A), \text{supp}(\beta)) \) satisfies the compatibility condition with constant \( \phi_0 \sqrt{p}/2 \) with high probability. This result provides an oracle inequality for LASSO estimates \( \hat{\beta}(A, \lambda) \) although part of the data is not generated i.i.d. from \( P_Z \).

B.1. Matrix perturbations

**Lemma 6.** Given i.i.d. observations \( Z_1, \ldots, Z_n \in \mathbb{R}^d \) from a distribution \( P_Z \) such that the coordinates \( z_{i,t} \) of each \( Z_t \) are bounded \( |z_{i,t}| \leq z_{\max} \), then we have \( \forall w \)

\[
\Pr \left[ \|\hat{\Sigma} - \Sigma\|_\infty \geq 2L^2 w + 2L\sigma_0 \sqrt{2} + 2L\sigma_0 \left( \sqrt{\frac{2\log(d^2 - d)}{n}} + \frac{L\log(d^2 - d)}{n} \right) \right] \leq e^{-nw},
\]

where \( \Sigma \equiv E_{Z \sim P_Z} [ZZ^T] \), \( \hat{\Sigma} \equiv \sum_{t=1}^n Z_t Z_t^T / n \), \( L = z_{max} \), and \( w = z_{max} \sqrt{e-1} \).

**Proof of Lemma 6** See pg. 535 in Bühlmann and Van De Geer (2011). Note that any bounded random variable \( z \) satisfies their definition of subgaussian (see pg. 483) with parameters \( L = z_{max} \) and \( w = z_{max} \sqrt{e-1} \). \( \Box \)

B.2. Compatibility condition for non-i.i.d. samples

Recall \( A, A', \Sigma, \beta, Z, W \) and assumptions on them from §4.1. We will first show that \( (\hat{\Sigma}(A'), \text{supp}(\beta)) \) satisfies the compatibility condition with high probability.

**Lemma 7.** If the compatibility condition holds for \( (\Sigma_0, \text{supp}(\beta)) \) with constant \( \phi_0 \) and

\[
\|\Sigma_0 - \Sigma_1\|_\infty \leq \frac{\phi_0^2}{32 \|\beta\|_0}
\]

holds, then the compatibility condition holds for \( (\Sigma_1, \text{supp}(\beta)) \) with constant \( \phi_0/\sqrt{2} \).

**Proof of Lemma 7** The proof follows directly from Corollary 6.8 in page 152 of (Bühlmann and Van De Geer 2011) \( \Box \)

**Lemma 8.** The pair \( (\hat{\Sigma}(A'), \text{supp}(\beta)) \) satisfies the compatibility condition with constant \( \phi_0/\sqrt{2} \) when

\[
|A'| \geq \frac{2z_{\max} \log d}{C_2}
\]

with probability \( 1 - e^{-C_2 |A'|} \) where

\[
C_2 \equiv \frac{\phi_0^2}{384 \|\beta\|_0^2 z_{\max}^2}.
\]
Proof of Lemma 8  We can check that if \( w = C_2 \) and \(|\mathcal{A}'|\) is lower-bounded as defined above, then
\[
2L^2w + 2L\sigma_0\sqrt{2w} + 2L\sigma_0\left(\frac{\sqrt{2\log(d^2 - d)}}{|\mathcal{A}'|} + \frac{L\log(d^2 - d)}{|\mathcal{A}'|}\right) \leq \frac{\phi_0^2}{32\|\beta\|_0}
\]
Thus, it follows from Lemma 6 that
\[
\Pr \left[ \|\Sigma - \hat{\Sigma}(\mathcal{A}')\|_\infty \geq \frac{\phi_0^2}{32\|\beta\|_0} \right] \leq e^{-C_2|\mathcal{A}'|}
\]
The result then follows directly from Lemma 7. \(\square\)

Lemma 9. Let \(\mathcal{A}, \mathcal{A}'\) be as in \(\S 4.1\). If the compatibility condition holds with constant \(\phi_0\) for \((\hat{\Sigma}(\mathcal{A}'), \text{supp}(\beta))\), then the compatibility condition holds for \((\hat{\Sigma}(\mathcal{A}), \text{supp}(\beta))\) with constant \(\phi_0\sqrt{|\mathcal{A}'|/|\mathcal{A}|}\).

Proof of Lemma 9 By definition, we can write
\[
\hat{\Sigma}(\mathcal{A}) = \frac{|\mathcal{A}'|}{|\mathcal{A}|} \hat{\Sigma}(\mathcal{A}') + \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}\setminus\mathcal{A}'} Z_tZ_t^T
= \frac{|\mathcal{A}'|}{|\mathcal{A}|} \hat{\Sigma}(\mathcal{A}') + \frac{|\mathcal{A}\setminus\mathcal{A}'|}{|\mathcal{A}|} \hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}').
\]
Then, for all \(v\) satisfying \(\|v_{\text{supp}(\beta)\setminus\mathcal{A}'}\|_1 \leq 3\|v_{\text{supp}(\beta)}\|_1\),
\[
v^T \hat{\Sigma}(\mathcal{A})v = \frac{|\mathcal{A}'|}{|\mathcal{A}|} v^T \hat{\Sigma}(\mathcal{A}')v + \frac{|\mathcal{A}\setminus\mathcal{A}'|}{|\mathcal{A}|} v^T \hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')v
\geq \frac{|\mathcal{A}'|}{|\mathcal{A}|} \frac{\phi_0^2}{16\|\beta\|_0} \|v_{\text{supp}(\beta)}\|_1^2
\]
from the fact that \(\hat{\Sigma}(\mathcal{A}\setminus\mathcal{A}')\) is a covariance matrix, and is therefore positive semi-definite \(\square\)

Now we have everything to finalize proof of Lemma 1.

Lemma 1 Consider all the above assumptions on \(\mathcal{A}, \mathcal{A}', \Sigma, \beta, Z, W\). Then if
\[
|\mathcal{A}| \geq \frac{4z_{\text{max}}\log d}{pC_2} \quad \text{and} \quad \lambda = \frac{\phi_0^2 \beta}{16\|\beta\|_0} \chi,
\]
then the following oracle inequality holds \(\forall \chi > 0\)
\[
\Pr \left[ \|\hat{\beta}(\mathcal{A}, \lambda) - \beta\|_1 > \chi \right] \leq e^{-\frac{|\mathcal{A}|C_1^2\chi^2}{16z_{\text{max}}} + \log d}.
\]
with probability at least \(1 - \exp[-pC_2|\mathcal{A}|/2]\).

Proof of Lemma 1: Applying Lemmas 8 and 9 implies that the compatibility condition holds for \((\hat{\Sigma}(\mathcal{A}), \text{supp}(\beta))\) with constant
\[
\frac{\phi_0}{\sqrt{2}} \sqrt{\frac{|\mathcal{A}'|}{|\mathcal{A}|}}
\]
with probability \(1 - e^{-C_2|\mathcal{A}'|}\) if \(|\mathcal{A}'| \geq 2z_{\text{max}}\log d/C_2\). Therefore, \((\hat{\Sigma}(\mathcal{A}), \text{supp}(\beta))\) satisfies the compatibility condition with constant \(\phi_0\sqrt{p}/2\) when \(|\mathcal{A}| \geq 4z_{\text{max}}\log d/(pC_2)\) with probability at least
\[
1 - e^{-pC_2|\mathcal{A}|/2}.
\]
Applying Proposition 1 in the event that the compatibility condition holds gives us the result. Note that we also used the fact that \(\hat{\Sigma}(\mathcal{A})_{r,r} \leq z_{\text{max}}^2\) \(\square\)
C. Proof of LASSO Oracle Inequality for Forced-Sample Estimator

In this section, we prove an oracle inequality for the forced sample estimator \( \hat{\beta}(T_{i,t}, \lambda_t) \) by applying Lemma 1. Recall that at each \( t \in T_{i,t} \), we draw a context \( X_t \in \mathcal{P}_X \) i.i.d. and play arm \( i \). Moreover, we assumed that the compatibility condition holds with constant \( \phi_0 \) for \( (\Sigma_i, \text{supp}(\beta_i)) \) where \( \Sigma_i = \mathbb{E}_{X \sim \mathcal{P}_X | X \in U_i} [XX^T] \) and also that \( \Pr[X_t \in U_i] \geq p_* \). Also recall that \( q_0 \) is a constant satisfying \( q \geq 4[q_0] \).

**Lemma 10.** If \( t \geq (Kq)^2 \), then \( q_0 \log t \leq |T_{i,t}| \leq 6q \log t \).

**Proof of Lemma 10** Define the \( n^{th} \) round of forced sampling of all the arms

\[
L_n = \{(2^n - 1)Kq + 1, \ldots, (2^n)Kq\}
\]

for \( n \geq 0 \). By construction, arm \( i \) is sampled \( |T_i \cap L_n| = q \) times during \( L_n \), so

\[
|T_i \cap \left( \bigcup_{r=0}^{n-1} L_r \right) | = nq.
\]

Therefore for each \( t \in L_n \), \( nq \leq |T_{i,t}| \leq (n + 1)q \). To show the lower bound, note that for \( t \in L_n \), we have \( t \leq (2^n)Kq \), i.e. \( n \geq \log_2 \left( \frac{t}{Kq} \right) \), so

\[
|T_{i,t}| \geq nq \geq q \log_2 \frac{t}{Kq} \geq q (\log t - \log Kq).
\]

By our assumption that \( q \geq 4[q_0] \), for \( t \geq (Kq)^2 \),

\[
|T_{i,t}| \geq 4[q_0] (\log t - \log Kq)
\]

\[
\geq [q_0] \log t + 2[q_0] (\log t - \log((Kq)^2))
\]

\[
\geq q_0 \log t
\]

where we have used the fact that \( t \geq (Kq)^2 \).

To show the upper bound, note that for \( t \in L_n \), \( t \geq (2^n - 1)Kq \), i.e. \( n \leq \log_2 \left( \frac{t}{Kq} + 1 \right) \), so

\[
|T_{i,t}| \leq (n + 1)q \leq \left( \log_2 \left( \frac{t}{Kq} + 1 \right) + 1 \right) q \leq \frac{3q \log t}{\log 2} \leq 6q \log t \quad \Box
\]

**Lemma 11.** Let \( T_{i,t}' \subseteq T_{i,t} \) be the set of all \( t \in T_{i,t} \) such that \( X_t \in U_i \). Then \( \forall t \in T_{i,t}' \) we have \( X_t \) is i.i.d. from \( \mathcal{P}_{X|X \in U_i} \). In addition for each \( t \in T_{i,t} \), \( t \in T_{i,t}' \) independently with probability at least \( p_* \).

**Proof of Lemma 11** By construction, for each \( t \in T_{i,t} \), \( X_t \) is drawn i.i.d. from \( \mathcal{P}_X \) and therefore with probability at least \( p_* \), \( X_t \in U_i \), i.e. \( t \in T_{i,t}' \). Note that we are doing rejection sampling to construct \( T_{i,t}' \), and so for each \( t \in T_{i,t}' \), \( X_t \) is an i.i.d. sample of \( \mathcal{P}_{X|X \in U_i} \) \( \Box \)

Using Lemma 11 we see that the inclusion of each member of \( T_{i,t} \) in \( T_{i,t}' \) is a Bernoulli i.i.d. random variable with mean at least \( p_* \). Therefore, we get the following result using Chernoff bound.

**Lemma 12.** If \( t \geq (Kq)^2 \), for \( T_{i,t}, T_{i,t}' \) defined as in Lemma 11 the following holds

\[
\Pr \left[ \frac{|T_{i,t}'|}{|T_{i,t}|} \geq \frac{p_*}{2} \right] \geq 1 - \frac{1}{t}.
\]
Therefore, we apply this to indicator random variables \( I \) from random variables with \( \mu \). In this section, we prove the oracle inequality for the all-sample estimator \( \hat{D} \). Proof of LASSO Oracle Inequality for All-Sample Estimator

The forced sample estimator \( \hat{\beta}(T_{t,i}, \lambda) \) satisfies the oracle inequality

\[
\Pr \left[ \| \hat{\beta}(T_{t,i}, \lambda_1) - \beta_i \|_1 > \frac{h}{4\max} \right] \leq e^{-q_0 \log t \cdot \frac{p^2 \log^d}{256 C_{\max}^2} + \log d + \frac{2}{\ell},}
\]

when \( \lambda_1 = \frac{\phi^2 p \cdot h}{64 p \cdot \max} \) and \( t \geq (Kq)^2 \).

**Proof of Proposition 2:** By construction, \( |T_{t,i}| \geq q_0 \log t \). Then, Lemma 11 allows us to apply Lemma 1 (with \( \chi = h/4\max \)) to show that the following inequality holds (assuming \( |T_{t,i}|/|T_{t,i}| \geq (p_. / 2) \) holds)

\[
\Pr \left[ \| \hat{\beta}(T_{t,i}, \lambda_1) - \beta_i \|_1 > \frac{h}{4\max} \right] \leq e^{-q_0 \log t \cdot \frac{p^2 \log^d}{256 C_{\max}^2} + \log d + \frac{2}{\ell},}
\]

with probability at least

\[
1 - \exp \left[ - \frac{p \cdot C_2 |T_{t,i}|}{2} \right] \geq 1 - \exp \left[ - \frac{q_0 p \cdot C_2 \log t}{2} \right] \geq 1 - \frac{1}{\ell}
\]

when \( t \geq (Kq)^2 \) using definition of \( q_0 \) in §3.3. Note that the assumption

\[
|T_{t,i}| \geq q_0 \log t \geq \frac{4\max \log d}{p \cdot C_2}
\]

is also satisfied via \( \log t \geq 1 \) and another use of the definition of \( q_0 \). Now using union bound, and Lemma 12, the result follows \( \square \)

**D. Proof of LASSO Oracle Inequality for All-Sample Estimator**

In this section, we prove the oracle inequality for the all-sample estimator \( \hat{\beta}(S_{t,i}, \lambda_2, i) \) for arms in \( K_{opt} \). The approach mirrors the steps taken in Appendix C. However, there is an additional complication due to the correlation between rows of \( X(S_{t,i}) \) that was discussed in §4.3. Recall the events \( A_t \) defined in Eq. (3).

**Lemma 13.** For each \( i \in [K] \), if \( X_t \in U_i \) and \( A_{t-1} \) holds, LASSO Bandit uses the forced-sample estimator \( \hat{\beta}(T_{t-1}, \lambda_1) \) to select only the optimal arm \( \hat{K} = \{i\} \) at time \( t \).
Proof of Lemma 13  For simplicity throughout this proof we drop the reference to \( \lambda_1 \) in forced-sample estimators. Since \( X_t \in U_i \), we know
\[
X_t^T \beta_i \geq h + \max_{j \neq i} X_t^T \beta_j .
\]
Then, for any \( j \in [K] \setminus \{ i \} \),
\[
X_t^T \left( \hat{\beta}(T_{i,t-1}) - \hat{\beta}(T_{j,t-1}) \right) = X_t^T \left( \hat{\beta}(T_{i,t-1}) - \beta_i \right) - X_t^T \left( \hat{\beta}(T_{j,t-1}) - \beta_j \right) + X_t^T \left( \beta_i - \beta_j \right)
\geq -x_{\max} \frac{h}{4x_{\max}} - x_{\max} \frac{h}{4x_{\max}} + h
\geq h/2
\]
since \( A_{t-1} \) holds. Thus, at time \( t \), LASSO Bandit will use the forced-sample estimator and play arm \( i \). □

Lemma 14. At all time periods \( t \) with \( t \geq (Kq)^2 \) the event \( A_t \) occurs with probability at least
\[
1 - K \left(e^{-\log t \log d + \log d} + \frac{2}{t}\right).
\]

Proof of Lemma 14  For each \( i \in [K] \) and all \( t \geq (Kq)^2 \), we have from Proposition 2,
\[
\Pr \left[ \| \hat{\beta}(T_{i,t}, \lambda_1) - \beta_i \|_1 > \frac{h}{4x_{\max}} \right] \leq \exp \left[ - q_0 h^2 p_0^2 C_1 \log t + \log d \right] + \frac{2}{t}
\leq e^{-\log t \log d + \log d} + \frac{2}{t}
\]
where the second inequality follows from our definition of \( q_0 \) in §3.3. Taking a union bound over all \( K \) arms gives us the result. □

Lemma 15. Let \( i \in [K] \). Recall from §4.3 that \( S_{i,t} \subset [t] \) is the set of all time periods \( r \) such that \( X_r \in U_i \), \( r \notin T_{i,t} \), and the event \( A_{r-1} \) holds at time \( r \). Then the following properties hold.
(1) The set of random variables \( \{X_r \mid r \in S_{i,t}\} \) are i.i.d. from distribution \( P_{X \mid X \in U_i} \).
(2) For each \( r \in [t] / T_{i,t} \), we have \( r \in S_{i,t} \) with probability at least \( p_* / 2 \) when \( t \geq (Kq)^2 \).
(3) \( S_{i,t} \subset S_{i,t} \)

Proof of Lemma 15  For (1), since \( A_{r-1} \) is only a function of samples in \( T_{i,r-1} \), \( A_{r-1} \) is independent of all samples \( X \notin T_{i,r} \) (which are distributed as \( P_X \)). Then, presence of \( X_r \) in \( U_i \) is simply rejection sampling; thus, each \( r \in S_{i,t} \), \( X_r \) is distributed i.i.d. from \( P_{X \mid X \in U_i} \). For (2), we know that \( X \in U_i \) with probability at least \( p_* \) and Lemma 14 implies that \( A_{r-1} \) holds with probability at least \( 1 - K \left( e^{-\log (r - 1) \log d + \log d} + 2/(r - 1) \right) \) when \( (r - 1) \geq (Kq)^2 \). Note that \( (r - 1) \geq (Kq)^2 \geq 16K^2 \) (since \( q \geq 4[q_0 \geq 4] \)), which implies that \( (r - 1) \geq 8K \geq 4K + 4 \), so we can write
\[
\exp \left[ - \log (r - 1) \log d + \log d \right] \leq \exp \left[ - \log (4K + 4) \log d + \log d \right]
\leq \exp \left[ - \log 4K + \log 4 - 1 \log d \right]
\leq \exp \left[ - \log 4K \right] = \frac{1}{4K}
\]
Thus, \( A_{r-1} \) holds with probability at least
\[
1 - K \left( e^{-\log (r - 1) \log d + \log d} + \frac{2}{(r - 1)} \right) \geq 1 - K \left( \frac{1}{4K} + \frac{2}{8K} \right) = \frac{1}{2}.
\]
Then, \( r \in S_{i,t} \) with probability at least \( p_* / 2 \). Finally, for (3), from Lemma 13, we know that for \( X_r \sim P_X \), if \( X_r \in U_i \) and event \( A_{r-1} \) holds, then \( r \in S_{i,t} \), so \( S_{i,t} \subset S_{i,t} \). □
**Lemma 16.** If \( t \geq (Kq)^2 \), for \( S_{t,i}^r \) defined as in §4.3 the following holds

\[
\Pr \left[ |S_{t,i}^r| \geq \frac{tp_*}{4} \right] \geq 1 - e^{tp_*^2/32}.
\]

**Proof of Lemma 16** By definition of \( S_{t,i}^r \) we have for all \( r \in \{t\}/\mathcal{T}_{i,t} \),

\[
\mathbb{I}(r \in S_{t,i}^r) = \mathbb{I}(A_{r-1}) \cdot \mathbb{I}(X_r \in U_i).
\]

Let \( \mathcal{G}_t \) be the sigma algebra generated by the random variables in the first \( t \) rows of the design matrix \( \mathbf{X} \) and the first \( r \) entries of the noise vector \( \varepsilon \), and let \( \mathcal{G}_0 \) be the empty set. Clearly \( \mathbb{I}(A_{r-1}) \) is \( \mathcal{G}_{r-1} \) measurable. Similarly, \( \mathbb{I}(X_r \in U_i) \) is \( \mathcal{G}_t \) measurable while it is independent of \( \mathcal{G}_{r-1} \). Note that \( |S_{t,i}^r| = \sum_{r=1}^{t} \mathbb{I}(r \in S_{t,i}^r) \).

Now define for all \( s \in [t] \cup \{0\} \),

\[
M_s \equiv \mathbb{E} \left[ \sum_{r=1}^{t} \mathbb{I}(r \in S_{t,i}^r) \left| \mathcal{G}_s \right. \right].
\]

It is straightforward to see that \( M_0, M_1, \ldots, M_t \) is a martingale adapted to the filtration \( \mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots \subset \mathcal{G}_t \) (this is the famous Doob’s martingale construction) with \( M_0 = \mathbb{E}(|S_{t,i}^r|) \) and \( M_t = |S_{t,i}^r| \). Now since the martingale differences \( |M_r - M_{r-1}| \) are bounded by 1 we can use Azuma’s inequality, Theorem 7.2.1 from (Alon and Spencer 1992), to get for all \( \eta > 0 \)

\[
\Pr \left[ |S_{t,i}^r| < \mathbb{E}(|S_{t,i}^r|) - \eta \right] \leq e^{\eta^2/(2t)}.
\]

On the other hand by Lemma 15 we have

\[
\mathbb{E}(|S_{t,i}^r|) = \sum_{r=1}^{t} \Pr(r \in S_{t,i}^r) \geq tp_*/2
\]

which means using \( \eta = tp_*/4 \) in Eq. (6) we obtain

\[
\Pr \left[ |S_{t,i}^r| < \frac{tp_*}{4} \right] \leq e^{tp_*^2/32} \square
\]

**Proposition 3** The all-sample estimator \( \hat{\beta}(S_{t,i}, \lambda_{2,i}) \) for \( i \in \mathcal{K}_{\text{opt}} \) satisfies the oracle inequality

\[
\Pr \left[ \left| \|\hat{\beta}(S_{t,i}, \lambda_{2,i}) - \beta_i\|_1 \right| > 16x_{\text{max}} \sqrt{\frac{\log t + \log d}{p^2 C_1 t}} \right] \leq \frac{1}{t},
\]

with probability at least \( 1 - 2 \exp \left[ - \frac{p^2 (C_2 \wedge [1/2])}{16} \cdot t \right] \) when \( \lambda_{2,i} = \frac{\kappa_{2,i} \max_{\lambda} \sqrt{\log t + \log d}}{p C_1 t} \) and \( t \geq (Kq)^2 \).

**Proof of Proposition 3:** Lemma 16 states that at time \( t \geq (Kq)^2 \) we have \( |S_{t,i}^r| \geq p_* t/4 \) with probability \( 1 - \exp[-tp_*^2/32] \). In this event, we can bound \( |S_{t,i}^r| \geq |S_{t,i}^0| \geq p_* t/4 \). Applying Lemma 1 with \( p = p_*/2 \) and \( |A| \geq p_* t/4 \), we get

\[
\Pr \left[ \left| \|\hat{\beta}(S_{t,i}, \lambda) - \beta_i\|_1 \right| > \chi \right] \leq \exp \left[ -t \chi^2 \cdot \frac{p^2 C_1}{256x_{\text{max}}^2} + \log d \right]
\]

when \( \lambda = \frac{\kappa_{2,i} \cdot \chi}{32x_{\text{max}}} \) with probability at least

\[
1 - \exp \left[ - \frac{p^2 C_2}{16} t \right] \cdot \exp[-p^2 t/32] \geq 1 - 2 \exp \left[ - \frac{p^2 (C_2 \wedge [1/2])}{16} \cdot t \right]
\]

where we have used a union bound. There is an assumption from Lemma 1 that

\[
|S_{t,i}^r| \geq \frac{p_* t}{4} \geq \frac{16x_{\text{max}} \log d}{p_* C_2},
\]

but this is satisfied by the assumption that \( t \geq (Kq)^2 \) and the definition of \( q \). Taking

\[
\chi = 16x_{\text{max}} \sqrt{\frac{\log t + \log d}{p^2 C_1 t}}
\]

gives us the desired result \( \square \).
E. 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 \((t \leq (Kq)^2)\) and forced sampling \((t \in \mathcal{T}_{i,T} \text{ for some } i \in [K])\).

(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.

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 from time periods in (a) at time \(T\):

**Lemma 17.** The cumulative expected regret of LASSO Bandit from initialization \((t < (Kq)^2)\) and forced sampling \((t \in \mathcal{T}_{i,T} \text{ for some } i \in [K])\) up to time \(T\) is at most

\[
2qKbx_{\max}(6 \log T + Kq) .
\]

**Proof of Lemma 17:** From Lemma 10, we know we have at most \(6Kq \log T\) forced samples up to time \(T\). We also have \((Kq)^2\) initialization samples. Using Cauchy-Schwarz, we can bound the worst-case regret in each time period by \(\max_{i,j} X^T (\beta_i - \beta_j) \leq 2bx_{\max}\). The result follows directly.

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

**Lemma 18.** If \(f\) is a monotone decreasing function on the range \([r-1, s]\), then

\[
\sum_{t=r}^{s} f(t) \leq \int_{r-1}^{s} f(t) dt .
\]

**Proof of Lemma 18:**

\[
\sum_{t=r}^{s} f(t) \leq \sum_{t=r}^{s} \int_{t-1}^{t} f(t') dt' = \int_{r-1}^{s} f(t) dt .
\]

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

**Lemma 19.** The cumulative expected regret of LASSO Bandit from time periods \((Kq)^2 < t \leq T\) where \(A_{t-1}\) does not hold is at most \(2Kbx_{\max}(2 \log T + 3)\).

**Proof of Lemma 19** From Lemma 14, the probability that \(A_{t-1}\) does not hold is at most

\[
K \exp \left[ -\log t \log d + \log d \right] + 2K/t .
\]

Now we can sum this quantity for \(t \in [(Kq)^2, T - 1]\). Using Lemma 18, the first term yields

\[
\sum_{t=(Kq)^2}^{T-1} K \exp \left[ -\log t \log d + \log d \right] \leq Kd \int_{(Kq)^2-1}^{T} t^{-\log d} dt \\
\leq Kd \int_{e}^{\infty} t^{-\log d} dt = \frac{Ke}{-1 + \log d} < 3K .
\]

Similarly, the second term yields

\[
\sum_{t=(Kq)^2}^{T-1} \frac{2K}{t} \leq \int_{1}^{T} \frac{2K}{t} dt \leq 2K \log T .
\]

Using Cauchy Schwarz, the worst-case regret at any time \(t\) is at most \(2bx_{\max}\), and the result follows.
Before analyzing the regret from group (c), we show that if the event $A_{i-1}$ holds, then the set $\hat{K}$ 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 $\hat{K}$.

**Lemma 20.** If $A_{i-1}$ holds, then the set $\hat{K}$ contains the optimal arm $i^* = \arg\max_{i \in [K]} X_i^T \beta_i$ and no sub-optimal arms from the set $K_{\text{sub}}$.

**Proof of Lemma 20** To simplify notation, we call our forced-sample arm estimators $\hat{\beta}(T_{i,t-1}, \lambda_1)$ at time $t$ as $\hat{\beta}_i$. Since $A_{i-1}$ holds, we have that for any pair of arms $i, j \in [K]$,

$$X_i^T \hat{\beta}_i - X_j^T \hat{\beta}_j = X_i^T (\hat{\beta}_i - \beta_i) + X_j^T (\beta_j - \hat{\beta}_j) + X_j^T (\beta_j - \beta_i).$$

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

On the other hand, consider $i = \arg\max_{r \in [K]} X_r^T \hat{\beta}_r$ and any sub-optimal arm $j \in K_{\text{sub}}$. Then, $X_i^T \hat{\beta}_i - X_j^T \hat{\beta}_j \geq X_i^T \hat{\beta}_{i^*} - X_j^T \hat{\beta}_{i^*}$, and furthermore, since $A_{i-1}$ holds:

$$X_i^T \hat{\beta}_{i^*} - X_j^T \hat{\beta}_{i^*} = X_i^T (\hat{\beta}_{i^*} - \beta_{i^*}) + X_j^T (\beta_j - \hat{\beta}_{i^*}) + X_j^T (\beta_j - \beta_{i^*}) \geq -h/2 + X_j^T (\beta_j - \beta_{i^*}).$$

Recall that for every sub-optimal arm $i \in K_{\text{sub}}$, we have $X_i^T \beta_j < X_i^T \beta_{i^*} - h$. Then, we can write

$$X_i^T (\hat{\beta}_i - \hat{\beta}_j) \geq X_i^T \hat{\beta}_{i^*} - X_i^T \hat{\beta}_j > -h/2 + h = h/2.$$

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

Finally, the next two lemmas bound the regret from time periods in (c) by separately summing over expected regret when the all-sample oracle inequality does and does not hold. We simplify our notation by calling our all-sample estimators $\hat{\beta}(S_{t,t-1}, \lambda_2)$ at time $t$ as $\hat{\beta}_i$, where we recall $\lambda_{2,t} = \frac{2h}{2\log \log d + \log t} \sqrt{\frac{\log \log d + \log t}{K_{\text{opt}}C_3}}$.

**Lemma 21.** If $t > (Kq)^2$, the event $A_t$ holds, and the oracle inequality (Eq. (7)) holds for the all-sample estimators $\hat{\beta}_i$ for $i \in \hat{K}$, then the expected regret at time $t+1$ is bounded by $(2Kbx_{\text{max}})/t + C_3 \cdot (\log t + \log d)/t$, where $C_3 = \frac{1024K C_{\text{opt}}}{2C_1^{\frac{4}{3}}}$.

**Proof of Lemma 21** Without loss of generality, assume that arm 1 is optimal: $1 = \arg\max_{i \in [K]} X_i^T \beta_i$. Then, the expected regret at time $t + 1$ is given by

$$r_{t+1} = \mathbb{E} \left( \sum_{i \in \hat{K}} \mathbb{I} \left[ \text{choose arm } i \right] \cdot [X_{t+1}^T (\hat{\beta}_i - \beta_i)] \right) \leq \sum_{i \in \hat{K}} \mathbb{E} \left( \mathbb{I} \left[ X_{t+1}^T \hat{\beta}_i > X_{t+1}^T \hat{\beta}_1 \right] \cdot [X_{t+1}^T (\beta_1 - \beta_i)] \right)$$

where the last inequality uses the fact that event $\{ i = \arg\max_{j \in [K]} X_j^T \hat{\beta}_j \}$ is a subset of the event $\{ X_{t+1}^T \hat{\beta}_i > X_{t+1}^T \hat{\beta}_1 \}$, and that $X_{t+1}^T (\beta_1 - \beta_i) \geq 0$ (since we have assumed that arm 1 is optimal). Thus, we can bound $r_{t+1}$ through the regret incurred by each arm in $\hat{K}$ with respect to the optimal arm independently of the
other arms. We now define the event $B_i \equiv \{X_{t+1}^T(\beta_1 - \beta_i) > 2\delta x_{\text{max}}\}$, where we take $\delta = 16x_{\text{max}} \sqrt{\log(t + \log d) / p t}$. Then, we can write

$$r_{t+1} \leq \sum_{i \in \mathcal{K}} \mathbb{E} \left[ I \left[ (X_{t+1}^T \hat{\beta}_i > X_{t+1}^T \hat{\beta}_1) \cap B_i \right] \cdot [X_{t+1}^T(\beta_1 - \beta_i)] \right] + \sum_{i \in \mathcal{K}} \mathbb{E} \left[ I \left[ (X_{t+1}^T \hat{\beta}_i > X_{t+1}^T \hat{\beta}_1) \cap B_i \right] \cdot [X_{t+1}^T(\beta_1 - \beta_i)] \right],$$

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

$$r_{t+1} \leq \sum_{i \in \mathcal{K}} 2b x_{\text{max}} \Pr \left[ (X_{t+1}^T \hat{\beta}_i > X_{t+1}^T \hat{\beta}_1) \mid B_i \right] + \sum_{i \in \mathcal{K}} 2\delta x_{\text{max}} \Pr [B_i], \quad (8)$$

Note that choosing arm $i \neq 1$ conditioning on event of $B_i$ implies that

$$0 > X_{t+1}^T \hat{\beta}_1 - X_{t+1}^T \hat{\beta}_i \geq X_{t+1}^T \left( \hat{\beta}_1 - \beta_1 \right) + X_{t+1}^T \left( \beta_1 - \hat{\beta}_i \right) + 2\delta x_{\text{max}},$$

and thus, it must be that either $X_{t+1}^T \left( \hat{\beta}_1 - \beta_1 \right) < -\delta x_{\text{max}}$ or $X_{t+1}^T \left( \beta_1 - \hat{\beta}_i \right) < -\delta x_{\text{max}}$. Therefore,

$$\Pr \left[ (X_{t+1}^T \hat{\beta}_i > X_{t+1}^T \hat{\beta}_1) \mid B_i \right] \leq \Pr \left[ \|\beta_1 - \hat{\beta}_1\|_1 > \delta \right] + \Pr \left[ \|\hat{\beta}_1 - \beta_1\|_1 > \delta \right] \leq \frac{2}{t}, \quad (9)$$

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

We can also bound $\Pr [B_i]$ using Assumption 2 on the margin condition: $\Pr [B_i] = \Pr [X_{t+1}^T (\beta_1 - \beta_i) \leq 2\delta x_{\text{max}}] \leq 2C_0 \delta x_{\text{max}}$. Using this and Eq. (9) in Eq. (8) we obtain

$$r_{t+1} \leq \sum_{i \in \mathcal{K}} \left\{ \frac{4b x_{\text{max}}}{t} + 4C_0 \delta^2 x_{\text{max}}^2 \right\} \leq \frac{4Kbx_{\text{max}}}{t} + C_3 \cdot \frac{\log t + \log d}{t} \quad \square$$

**Lemma 22.** The cumulative expected regret from using the all-sample estimator up to time $T$ is bounded by $(4Kbx_{\text{max}} + C_3 \log d) \log T + C_3 (\log T)^2 + C_4$, where $C_4 = \frac{4b x_{\text{max}}}{1 - \exp \left[ - \frac{\log(2) + 1/2}{\log(2)} \right]}$.

**Proof of Lemma 22** We first sum regret from Lemma 21 (oracle inequality holds): 

$$\sum_{t=(Kq)^2}^{T-1} \frac{4Kbx_{\text{max}}}{t} + C_3 \cdot \frac{\log t + \log d}{t} \leq (4Kbx_{\text{max}} + C_3 \log d) \log T + C_3 (\log T)^2.$$

From Proposition 3, the oracle inequality does not hold with probability $2 \exp \left[ - \frac{\log(2) + 1/2}{16} t \right]$ so we accrue cumulative expected regret $\sum_{t=(Kq)^2}^{T-1} 4b x_{\text{max}} \exp \left[ - \frac{\log(2) + 1/2}{16} t \right] \leq \frac{4b x_{\text{max}}}{1 - \exp \left[ - \frac{\log(2) + 1/2}{16} \right]} \quad \square$

**F. Proof of OLS Tail Inequality for Non-i.i.d. Data**

**F.1. OLS Tail Inequality for Adapted Observations**

**Proposition 4** Given an adapted sequence of observations $\{X_t : t = 1, \ldots, n\}$, if the minimum eigenvalue of $\hat{\Sigma} = X^T X / n$ is bounded below by $\phi_0 > 0$, then $\forall \chi > 0$

$$\Pr \left[ \|\hat{\beta} - \beta\|_1 \leq \chi \right] \geq 1 - \exp \left[ -D_1 n \chi^2 + \log 2d \right],$$

where $D_1 := \phi_0^2 / (2d^2 x_{\text{max}}^2 \sigma^2)$. 
Proof of Proposition 4  For simplicity, we start with the $L_2$ norm. Note that

$$
\|\hat{\beta} - \beta\|_2 = \|(X^T X)^{-1} X^T \varepsilon\|_2 \\
\leq \|(X^T X)^{-1}\|_2 \cdot \|X^T \varepsilon\|_2 \\
= \frac{1}{n\phi_0} \|X^T \varepsilon\|_2.
$$

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

$$
\Pr \left( \|\hat{\beta} - \beta\|_2 \leq \chi \right) \geq \Pr \left( \|X^T \varepsilon\|_2 \leq n\chi \phi_0 \right) \\
\geq 1 - d \Pr \left( |X^{(r)}\varepsilon| > \frac{n\chi \phi_0}{\sqrt{d}} \right),
$$

where we have let $X^{(r)}$ denote the $r$-th column of $X$. We can expand $X^{(r)}\varepsilon = \sum_{t \in [n]} \varepsilon_t X_{t,r}$, where we note that $\varepsilon^T X^{(r)}$ is a $(x_{max} \sigma)$-subgaussian random variable by Definition 1. Then, $D_0, D_1, ... D_n$ is a martingale adapted to the filtration $\mathcal{G}_1 \subset ... \subset \mathcal{G}_n$ since $\mathbb{E}[\varepsilon_t X_{t,r}, X_1, ..., X_{t-1}, y_1, ..., y_{t-1}] = 0$. Using Lemma 3,

$$
\Pr \left( \|\hat{\beta} - \beta\|_2 \leq \chi \right) \geq 1 - 2d \exp \left( -\frac{n\chi^2 \phi_0^2}{2d^2 x_{max}^2 \sigma^2} \right).$$

Now, to bound the $L_1$ norm, we can write

$$
\Pr \left( \|\hat{\beta} - \beta\|_1 \leq \chi \right) \geq \Pr \left( \|\hat{\beta} - \beta\|_2 \leq \frac{\chi}{\sqrt{d}} \right) \geq 1 - 2d \exp \left( -\frac{n\chi^2 \phi_0^2}{2d^2 x_{max}^2 \sigma^2} \right). \Box
$$

F.2. Positive-Definiteness for non-i.i.d. samples

We will first show that $\hat{\Sigma}(A')$ has minimum eigenvalue bounded below with high probability.

**Lemma 23.** The minimum eigenvalue of $\hat{\Sigma}(A')$ is bounded below by $\phi_0/2$ with probability $1 - \exp \left( -D_2 |A'| + \log d \right)$ where $D_2 := \frac{\phi_0}{8x_{max}}$.

**Proof of Lemma 23** First, note that

$$
\lambda_{\max}(\Sigma(A')) = \max_{\|\beta\|_1 = 1} \beta^T \Sigma(A') \beta \\
= \max_{\|\beta\|_1 = 1} (Z^T \beta)^2 \leq z_{\max}^2
$$

Then, it follows from the matrix Chernoff inequality (given by Theorem 5.1.1 in Tropp (2015)) that

$$
\Pr \left[ \lambda_{\min}(\hat{\Sigma}(A')) \leq \frac{\phi_0}{2} \right] \leq d \cdot \exp \left[ -\frac{|A'| \phi_0}{8z_{\max}^2} \right]
$$

if we take $t = 1/2$ and $R = z_{\max}^2$. \Box

**Lemma 24.** If the minimum eigenvalue of $\hat{\Sigma}(A')$ is bounded below by $\phi_0$, then the minimum eigenvalue of $\hat{\Sigma}(A)$ is bounded below by $\phi_0 \cdot |A'|/|A|$.
Proof of Lemma 24  From our definition, we can write
\[ \hat{\Sigma}(A) = \frac{|A|}{|A|} \hat{\Sigma}(A') + 1 \sum_{t \in A/A'} Z_t Z_t^\top, \]
\[ = \frac{|A|}{|A|} \hat{\Sigma}(A') + \frac{|A/A'|}{|A|} \hat{\Sigma}(A/A'). \]
It immediately follows that
\[ \lambda_{\min}(\hat{\Sigma}(A)) \geq \frac{|A'|}{|A|} \lambda_{\min}(\hat{\Sigma}(A')) + \frac{|A/A'|}{|A|} \lambda_{\min}(\hat{\Sigma}(A/A')) \]
from the fact that the minimum eigenvalue is a concave function and \( \hat{\Sigma}(A/A') \) is a covariance matrix (and is therefore positive semi-definite). □

We can combine these results to prove a tail inequality for the OLS estimator with an unknown fraction of non-i.i.d. samples (Lemma 2).

**Lemma 2**  Under the assumptions above, the following tail inequality holds \( \forall \chi > 0: \)
\[ \Pr\left[ \|\hat{\beta}(A) - \beta\|_1 \leq \chi \right] \geq 1 - \exp\left[ -|A| \chi^2 \cdot \frac{p D_1}{4} + \log d \right], \]
with probability at least
\[ 1 - \exp\left[ -|A'\|_2 + \log d \right]. \]

**Proof of Lemma 2**  Applying Lemma 23 to Lemma 24 implies that the minimum eigenvalue of \( \hat{\Sigma}(A) \) is bounded below by
\[ \phi_0 \left( \frac{|A'|}{|A|} \right), \]
with probability \( 1 - \exp\left[ -|A'\|_2 + \log d \right]. \) Applying Proposition 4 gives us the result. □

**G. Proof of Tail Inequalities for OLS Force-Sample and All-Sample Estimators**

**Proposition 5.**  When \( t \geq (Kq)^2 \), the forced sample estimator \( \hat{\beta}(T_{i,t}) \) satisfies the tail inequality
\[ \Pr\left[ \|\hat{\beta}(T_{i,t}) - \beta_i\|_1 > \frac{h}{4x_{\max}} \right] \leq \exp\left[ -q_0 \log t \cdot \frac{p D_1 h^2}{64 x_{\max}^2} + \log d \right] + \frac{2}{t}. \]

**Proof of Proposition 5**  By construction, \( |T_{i,t}| \geq \log t \) and \( \Sigma_i \) has minimum eigenvalue bounded below by \( \phi_0 \). If \( |T_{i,t}'|/|T_{i,t}| \geq p_{s}/2 \), Lemma 11 allows us to apply Lemma 2 (with \( \chi = h/4x_{\max} \)) to show that the following inequality holds
\[ \Pr\left[ \|\hat{\beta}(T_{i,t}) - \beta_i\|_1 > \frac{h}{4x_{\max}} \right] \leq \exp\left[ -q_0 \log t \cdot \frac{p_{s} D_1 h^2}{64 x_{\max}^2} + \log d \right], \]
with probability at least
\[ 1 - \exp\left[ -p_{s} D_2 |T_{i,t}|/2 + \log d \right] \geq 1 - 2 \exp\left[ -q_0 p_{s} D_2 \log t/2 + \log d \right] \geq 1 - \frac{1}{t}, \]
when \( t \geq (Kq)^2 \) by definition of \( q_0 \). Combining this with the probability that \( |T_{i,t}'|/|T_{i,t}| \geq p_{s}/2 \) (Lemma 12) using a union bound gives the result. □

We again define the event \( A_t \) in the same way in order to prove the tail inequality for the all-sample OLS estimator.
proposition 6. For $i \in K_{opt}$, the all-sample estimator $\hat{\beta}(S_{t,t})$ satisfies the tail inequality
\[
\Pr \left[ \| \hat{\beta}(S_{t,t}) - \beta_i \|_1 \leq \xi \right] \geq 1 - \exp \left[ -t \chi^2 \cdot \frac{p_t^2 D_1}{32} + \log d \right],
\]
with probability at least
\[
1 - 3 \exp \left[ - \frac{p_t^2 (D_2 \wedge p_t/2)}{64} t \right] \quad \text{when } t \geq (Kq)^2.
\]

Proof of Proposition 6. First, we note that Lemma 14 holds for the OLS estimator as well since the forced-sample tail inequality for the OLS estimator (Proposition 5) is stronger than the forced-sample oracle inequality for the LASSO estimator (Proposition 2).

From Lemma 16, we have that at time $t \geq (Kq)^2$, each of $\{1, ..., t\} / T_{i,t}$ belongs to $S_{t,t}$ with probability at least $p_t/2$. Applying Lemma 2 with $p = p_t/2$ and $|A| \geq p_t t/4$, we get
\[
\Pr \left[ \| \hat{\beta}(S_{t,t}) - \beta_i \|_1 > \xi \right] \leq \exp \left[ -t \chi^2 \cdot \frac{p_t^2 D_1}{32} + \log d \right]
\]
with probability at least
\[
1 - 2 \exp \left[ - \frac{p_t^2 (D_2 \wedge p_t/2)}{64} t \right] - \exp[\frac{-p_t^2 t/32}{2}] \geq 1 - 3 \exp \left[ - \frac{p_t^2 (D_2 \wedge p_t/2)}{64} t \right]
\]
where we have used a union bound. There is an assumption from Lemma 2 that $|S_{t,t}| \geq \frac{p_t t}{4} \geq \frac{16 \max \log d}{p_t \epsilon_2}$, but this is satisfied by the assumption that $t \geq (Kq)^2$ by our definition of $q$. □

H. Bounding the Regret in the Low-Dimensional Setting

We first note that the regret from 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 oracle 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.

We first bound the expected regret when the tail inequality for the all-sample estimator holds (Proposition 6). In this section, we simplify our notation by letting $\hat{\beta}_i = \hat{\beta}(S_{i,t}) \quad \forall i \in \{1, ..., K\}$ and $D_3 := p_t^2 D_1/32$.

Lemma 25.
\[
1 - \frac{2}{\sqrt{\pi}} \cdot \frac{\exp(\frac{-x^2}{2})}{x + \sqrt{x^2 + 2}} \leq \text{erf}(x) < 1 - \frac{2}{\sqrt{\pi}} \cdot \frac{\exp(\frac{-x^2}{2})}{x + \sqrt{x^2 + 2}}
\]

Proof of Lemma 25. The result follows directly from formula 7.1.13 from Abramowitz and Stegun (Handbook of Mathematical Functions, Dover). □

Lemma 26. If Algorithm 2 uses the all-sample estimator and the tail inequality holds (Proposition 6), then the expected regret at time $t$ is bounded by $212 K D_0 x_{\max}^2 (\log d)^{3/2} / (D_3 t)$.

Proof of Lemma 26. Recall from Lemma 20 that since $A_{t-1}$ holds, the set $\hat{K}$ contains the optimal arm $i^* = \arg \max_{i \in [K]} X_i^T \beta_i$ and no sub-optimal arms from the set $K_{sub}$. Without loss of generality, assume that arm 1 is optimal, i.e., $1 = \arg \max_{i \in \{1, ..., K\}} X_i^T \beta_i$. Then, the expected regret at time $t$ is given by
\[
E[r_t] = \sum_{i \in \hat{K}, i \neq 1} \Pr[\text{choose arm } i] \cdot E[X_i^T \beta_i - X_i^T \beta_1 | \text{choose arm } i]
\]
where we have let $p_{\hat{\beta}_1, \ldots, \hat{\beta}_K}$ denote the joint probability density of the estimators $\hat{\beta}_1, \ldots, \hat{\beta}_K$ at time $t$. The inequality follows from the fact that the event where $i = \arg \max_{j \in \{1, \ldots, K\}} X^T \hat{\beta}_j$ is a subset of the event $X^T \hat{\beta}_i > X^T \hat{\beta}_1$, and that $\mathbb{E}[X^T (\beta_1 - \beta_i)] \geq 0$ (since we have assumed that arm 1 is optimal). Thus, we can bound $r_i$ through the regret incurred by each arm with respect to the optimal arm independently of the other arms. We now define the event

$$B^i_h = \{ X^T (\beta_1 - \beta_i) \in [2x_{max} \cdot h\delta, 2x_{max} \cdot (h + 1)\delta] \}$$

where $\delta$ is a parameter we will choose later to minimize regret. Then, we can write

$$\mathbb{E}[r_i] \leq \sum_{i=2}^K \sum_{h=0}^{\infty} \Pr[B^i_h] \Pr[X^T \hat{\beta}_i > X^T \hat{\beta}_1 \mid B^i_h] : 2x_{max}(h + 1)\delta$$

by the definition of $B^i_h$.

Note that choosing arm $i \neq 1$ in the event of $B^i_h$ implies that

$$0 > X^T \hat{\beta}_1 - X^T \hat{\beta}_i = (X^T \hat{\beta}_1 - X^T \hat{\beta}_i) + (X^T \beta_1 - X^T \hat{\beta}_i) + (X^T \hat{\beta}_i - X^T \beta_i) \geq 2x_{max} h\delta$$

and thus, it must be that either $X^T (\hat{\beta}_1 - \beta_1) > x_{max} h\delta$ or $X^T (\beta_1 - \hat{\beta}_i) > x_{max} h\delta$.

Thus, we can write

$$\Pr[X^T \hat{\beta}_i > X^T \hat{\beta}_1 \mid B^i_h] \leq \Pr[X^T (\hat{\beta}_1 - \beta_1) > x_{max} h\delta] + \Pr[X^T (\beta_1 - \hat{\beta}_i) > x_{max} h\delta] \leq \Pr[\|\hat{\beta}_i - \beta_1\|_1 > h\delta] + \Pr[\|\hat{\beta}_i - \beta_i\|_1 > h\delta]$$

using a union bound. Recall that the tail inequality implies that $\forall h, \delta \geq 0$,

$$\Pr[\|\hat{\beta}_i - \beta_i\|_1 > h\delta] \leq \begin{cases} \exp[-D3h^2\delta^2 + \log d] & \text{if } h > \alpha_i \\ 1 & \text{if } h \leq \alpha_i \end{cases}$$

where

$$\alpha_i := \sqrt{\frac{\log d}{D3h^2\delta^2}} + 1$$

We can also bound $\Pr[B^i_h]$ using our assumption on the margin condition

$$\Pr[X^T \beta - X^T \beta' \leq \kappa] \leq D_0 \kappa \quad \forall \rho \in (0, \kappa_0]$$
for any allowable choices of $X, \beta, \beta'$. Without loss of generality, let $D_0 \geq 1/\kappa_0$ (if not, we can re-define $D_0$ to be $\max(D_0, 1/\kappa_0)$ and obtain a looser bound). Then, the condition holds trivially for all $\kappa > \kappa_0$ as well, and we can write

$$\Pr[B_i] \leq \Pr[X^T(\beta_1 - \beta_i) < 2x_{\max}(h + 1)\delta]$$

$$\leq 2D_0x_{\max}(h + 1)\delta$$

Now we obtain

$$\mathbb{E}[r] \leq \sum_{i=2}^{K} \sum_{h=0}^{\infty} 4D_0x_{\max}^2(h + 1)^2\delta^2 \cdot \left( \Pr[||\beta_1 - \hat{\beta}_1||_1 > h\delta] + \Pr[||\hat{\beta}_i - \beta_i||_1 > h\delta] \right)$$

$$\leq 4D_0x_{\max}^2\delta^2 \left( KQ_{1,t} + \sum_{i=2}^{K} Q_{i,t} \right)$$

where we define

$$Q_{i,t} := \sum_{h=0}^{\alpha_i} (h + 1)^2 + \sum_{h=\alpha_i + 1}^{\infty} (1 + 2h + h^2) \cdot \exp[-t \cdot D_3h^2\delta + \log d]$$

We can balance the terms by taking $\delta = \frac{1}{\sqrt{tD_3}}$, so $\alpha_i = \lfloor \sqrt{\log d} \rfloor + 1$

$$Q_{i,t} = \sum_{h=0}^{\alpha_i} (h + 1)^2 + \sum_{h=\alpha_i + 1}^{\infty} (1 + 2h + h^2) \cdot \exp[-h^2 + \log d]$$

Also note that $\sqrt{\log d} \leq \alpha_i \leq \left(1 + \frac{1}{\sqrt{2}}\right)\sqrt{\log d}$ since we have assumed that $\log d \geq 2$.

**The first term:**

$$\sum_{h=0}^{\alpha_i} (h + 1)^2 \leq \frac{1}{6} \cdot (2\alpha_i^3 + 9\alpha_i^2 + 13\alpha_i + 6) \leq 5\alpha_i^3 = 5 \left(1 + \frac{1}{\sqrt{2}}\right)^3 (\log d)^{3/2} \leq 25(\log d)^{3/2}$$

**The second term:** We first note that the summand is monotonically decreasing in $h$, so we can write

$$d \sum_{h=\alpha_i + 1}^{\infty} \exp[-h^2] \leq d \int_{\alpha_i}^{\infty} \exp[-h^2] dh$$

$$= d\frac{\sqrt{\pi}}{2} \cdot \left[ 1 - \text{erf} \left( \frac{\alpha_i}{2x_{\max}} \right) \right]$$

$$\leq d \cdot \frac{\exp[-\alpha_i^2]}{\alpha_i + \sqrt{\alpha_i^2 + 4/\pi}} \leq \frac{1}{2}$$

using Lemmas 12 and 14.

**The third term:** We note that the summand is monotonically decreasing in $h$ within the range $[\alpha_i, \infty]$. This is because the summand is concave and reaches a global maximum at

$$h^* = \frac{1}{\sqrt{2}} < \alpha_i$$

for $\log d > 1$. Thus, we can upper bound the sum with an integral:

$$d \sum_{h=\alpha_i + 1}^{\infty} h \exp[-h^2] \leq d \int_{\alpha_i}^{\infty} h \exp[-h^2] dh$$

$$= d \frac{\exp[-\alpha_i^2]}{2} \leq \frac{1}{2}$$
The fourth term: Once again, we note that the summand is monotonically decreasing in $h$ within the range $[\alpha_i, \infty]$ because the summand reaches a global maximum at $h^* = 1 < \alpha_i$ for $\log d > 1$. So, we upper bound the sum with an integral and integration by parts gives us:

$$d \sum_{h=\alpha_i+1}^{\infty} k^2 \exp \left[ -h^2 \right] \leq d \int_{\alpha_i}^{\infty} h^2 \exp \left[ -h^2 \right] dh$$

$$= \frac{d}{8} \left[ 2\alpha_i \exp \left[ -\alpha_i^2 \right] + \sqrt{\pi} (1 - \text{erf}(\alpha_i)) \right]$$

$$\leq \frac{d}{4} \left[ \alpha_i \exp \left[ -\alpha_i^2 \right] + \frac{\exp \left[ -\alpha_i^2 \right]}{\alpha_i + \sqrt{\alpha_i^2 + 4/\pi}} \right]$$

$$\leq \frac{\sqrt{\log d} + 2}{4}$$

Now we obtain

$$Q_{i,t} \leq 25 \left( \sqrt{\log d} \right)^{3/2} + \frac{1}{2} + \frac{1}{2} + \frac{\sqrt{\log d} + 2}{4} \leq 27(\log d)^{3/2}$$

so

$$\mathbb{E}[r_t] \leq 4D_0x_{\max}^2 \left( \frac{27K(\log d)^{3/2} + \sum_{i=2}^{K} 27(\log d)^{3/2}}{D_3} \right)$$

$$\leq \frac{212KD_0x_{\max}^2(\log d)^{3/2}}{D_3} \quad \square$$

**Lemma 27.** The cumulative expected regret from the all-sample estimator at time $T$ is bounded by

$$\frac{212KD_0x_{\max}^2(\log d)^{3/2}}{D_3} \log T + D_4$$

$$D_4 := \frac{3}{1 - \exp \left[ -\frac{p_2^2(D_2^{\land p_2}/2)}{64} t \right]}$$

**Proof of Lemma 27** We first sum the regret from when the all-sample estimator tail inequality does hold:

$$\sum_{t=(Kq)^2+1}^{T} \frac{212KD_0x_{\max}^2(\log d)^{3/2}}{D_3} \leq \frac{212KD_0x_{\max}^2(\log d)^{3/2}}{D_3} \log T$$

On the other hand, the all-sample estimator tail inequality does not hold with probability

$$3 \exp \left[ -\frac{p_2^2(D_2^{\land p_2}/2)}{64} t \right]$$

and so we accumulate cumulative regret

$$\sum_{t=(Kq)^2+1}^{T} 3 \exp \left[ -\frac{p_2^2(D_2^{\land p_2}/2)}{64} t \right] \leq \sum_{t=0}^{\infty} 3 \exp \left[ -\frac{p_2^2(D_2^{\land p_2}/4)}{128} t \right]$$

$$\leq 1 - \exp \left[ -\frac{3}{p_2^2(D_2^{\land p_2}/64)} \right] \quad \square$$

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

**Proof of Theorem 2** The total expected cumulative regret of the OLS Bandit up to time $T$ is upper-bounded by summing all the terms from Lemmas 17, 19, and 27):

$$R_T \leq 2qKbx_{\max}(6 \log T + Kq) + 2Kbx_{\max}(2 \log T + 3) + \frac{212KD_0x_{\max}^2(\log d)^{3/2}}{D_3} \log T + D_4$$

$$= O \left( d^2 \log^2 d \cdot \log T \right) \quad \square$$