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 choices at the individual-level. This involves learning a model of decision rewards conditional on individual-specific covariates. In domains such as medical decision-making and personalized advertising, these covariates are often high-dimensional; however, typically only a small subset of these observed features are predictive of each decision’s success. We formulate this problem as a contextual 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: contextual bandits, online learning, high-dimensional statistics, LASSO, treatment allocation, medical decision-making, warfarin dosing

1. Introduction

Consider a decision-maker choosing sequentially among \( K \) decisions (arms) that each have uncertain rewards. For example, a marketing campaign may be considering a set of new promotions, or a clinical trial may be considering several uncertain treatments. As studied in the classic multi-armed bandit (Lai and Robbins 1985), the decision-maker faces a tradeoff between exploration (learning the rewards for each arm through random trials) and exploitation (leveraging current knowledge to play the estimated-best arm).

However, the choice is complicated by the fact that individuals may respond differently to each decision depending on their specific characteristics: for instance, promotions for trendy apparel may be more effective when targeted at younger customers, and some medical treatments have been found to be more effective on patients with certain biomarkers (e.g., among lung cancer patients, Kim et al. (2011) find that choosing chemotherapeutic agents based on patient-specific molecular
biomarker profiles can significantly improve disease control rates). In order to tailor the choice to a specific individual, the decision-maker can leverage user-level data such as purchase history or medical records. This setting is considered in the contextual multi-armed bandit literature, which extends the traditional multi-armed bandit model to include user-specific covariates (Auer 2003, Langford and Zhang 2008, Slivkins 2014, Agarwal et al. 2014, Goldenshluger et al. 2013).

In the contextual bandit framework, the decision-maker typically learns a model predicting user-specific rewards for each choice conditioned on the observed covariates. In particular, under a linear model, each arm $i$ is associated with an unknown parameter $\beta_i \in \mathbb{R}^d$ and the observed reward for playing arm $j$ after observing covariates $X_t \in \mathbb{R}^d$ at time $t$ is

$$X_t^T \beta_j + \epsilon$$

where $\epsilon$ is zero-mean noise. The decision-maker can leverage estimates of the unknown arm parameters from past decisions to optimize future decisions. Again, we face an exploration-exploitation tradeoff: the quality of each estimate $\hat{\beta}_i$ improves as arm $i$ is played for random samples of $X$, but the estimated current-best arm must also be exploited to maximize reward at each time $t$. This problem is tackled by Goldenshluger et al. (2013); however, their algorithm is impractical in settings where the dimension $d$ of the observed covariates $X_t$ is large (in particular, their regret bounds scale as $O(d^3)$).

Yet, in many applications, the observed covariates are high-dimensional. For instance, in web marketing, a user’s covariates may correspond to cookie data, which includes clicks, purchases, and websites visited. In healthcare applications, patient covariates may include genetic or molecular biomarker profiles and electronic health record data containing information on past diagnoses, procedures, and medications. Moreover, oftentimes only a small subset of the observed covariates are relevant to evaluating the benefits of a given arm. In particular, only a small number of a patient’s genetic biomarkers or past diagnoses (corresponding to certain pre-existing conditions) may be relevant to the success of a specific treatment. For example, Kim et al. (2011) found that only one or two of the many available patient biomarkers were predictive of the success of each of four possible chemotherapy choices among lung cancer patients. The challenge is that the identities of the relevant biomarkers or covariates among the high-dimensional vector of observed characteristics are a priori unknown. Thus, we propose using $L_1$-regularized techniques such as the LASSO, which are well-suited for identifying a sparse subset of predictive covariates in the high-dimensional setting (Bühlmann and Van De Geer 2011).

We present a new bandit algorithm (the LASSO Bandit) that efficiently leverages high-dimensional data in the contextual bandit setting by learning LASSO-type estimates. Our main
result is an upper bound on the cumulative regret of this algorithm. The key challenge in our proofs is showing that the LASSO estimates converge despite the non-i.i.d. data induced by the bandit policy; we tackle this by proving a new oracle inequality for the LASSO which holds with high probability when an unknown portion of the data is generated by a non-i.i.d. process. Using this result, we prove that the LASSO Bandit achieves at most $O(s_0^2 \log T)^2 + s_0 \log T \log d)$ regret, where $s_0 \ll d$ is the number of relevant covariates and $T$ is the number of time steps. In particular, we improve over the existing regret bounds of $O(d^3 \log T)$ in the regime where $T = O(d)$. Experiments on synthetic data suggest that our algorithm significantly outperforms the algorithm presented in Goldenshluger et al. (2013) even in the regime where $T \gg d$.

Furthermore, we demonstrate the practical usefulness of our algorithm by evaluating it on the clinically relevant issue of warfarin dosing. Warfarin is a widely used oral anticoagulant agent; in 2004, more than 30 million prescriptions were written for this drug in the United States (Wysowski et al. 2007). The appropriate dose of warfarin is highly variable among individuals (by up to a factor of 10) due to patient-specific factors; yet, an incorrect initial dosage can result in highly adverse consequences such as stroke or internal bleeding. We use a publicly available patient dataset (Consortium 2009) that contains clinical factors, demographic variables, and genetic information that have been found to be predictive of the optimal warfarin dosage. We show that our algorithm outperforms existing bandit methods as well as physicians to correctly dose a majority of patients.

1.1. Related Literature

Several different types of contextual bandit problems have been tackled in the literature (we refer the reader to Chapter 4 of Bubeck (2012) for an informative review). For example, Slivkins (2014), Perchet et al. (2013), Rigollet and Zeevi (2010) prove regret bounds where the arm rewards are smooth, non-parametric functions of the contexts, Langford and Zhang (2008) considers a practical epoch-greedy algorithm whose regret scales with the complexity of the hypothesis class (consisting of mappings from the context space to arm space), Kakade et al. (2008) looks specifically at the setting where the hypothesis class consists of margin-based linear functions for multiclass prediction, and Agarwal et al. (2014) looks at efficient algorithms for choosing the best policy from a finite, known hypothesis class. The specific case where arm rewards are given by a linear function of arm-specific contexts was first introduced by Auer through the LinRel algorithm (Auer 2003) and was subsequently improved through the LinUCB algorithm in Chu et al. (2011). However, all of these settings have regret that is lower-bounded by $O(\sqrt{T})$ for even very simple joint distributions or policy classes.

We focus on an important sub-class of the contextual bandit where much stronger performance guarantees can be achieved. In particular, we consider the case where the contexts are shared
among all arms, and each arm reward is given by a linear function of the shared context and an unknown arm-specific parameter. This setting is introduced in Li et al. (2010) for the application of personalized news article recommendation; however, Li et al. (2010) does not provide a new regret analysis of the setting and simply maps the problem back to the original upper bound of $O(\sqrt{T})$ from Auer (2003). Our work is most related to Goldenshluger et al. (2013), which presents an algorithm and proves a corresponding upper bound of $O(d^3 \log T)$ on its regret. Our algorithm significantly improves upon Goldenshluger et al. (2013) when $d = O(T)$.

We note that high-dimensional arms have been tackled in the linear bandit setting (Carpentier and Munos 2012, Deshpande and Montanari 2012, Abbasi-Yadkori et al. 2012). However, this is a different problem: our algorithm chooses among a finite set of arms in response to exogenously arriving contexts, while linear bandit algorithms choose a context that maximizes the inner product with a fixed unknown parameter. Moreover, to the best of our knowledge, existing algorithms do not use LASSO techniques. In particular, Carpentier and Munos (2012) uses random projections, and Deshpande and Montanari (2012), Abbasi-Yadkori et al. (2012) use $L_2$-regularized regression. In fact, Abbasi-Yadkori et al. (2012) notes that their regret bounds would be significantly worse for LASSO-type procedures under their proof technique.

Finally, our work is also related to the literature on oracle inequalities for LASSO estimators (Bickel et al. 2009, Bühlmann and Van De Geer 2011, Negahban et al. 2009). We use these results to develop new oracle inequalities when an unknown part of the data may be non-i.i.d. There has also been work developing $O(\sqrt{T})$ regret bounds for the online LASSO but this setting does not incorporate arms or bandit feedback (e.g., see Chapter 11 in Cesa-Bianchi and Lugosi (2006)).

Structure of the Paper: We detail the problem formulation in §2 and describe our algorithm in §3. In §4, we state our main result, which upper bounds the cumulative regret of the LASSO Bandit, along with our assumptions and an outline of the key steps in the proof. We present experiments on synthetic data in §5 and evaluate our algorithm on the real clinical problem of warfarin dosing in §6. We conclude and identify opportunities for future work in §7. All complete proofs can be found in the Appendix.

2. Problem Formulation

Let $T$ be the number of time steps and $K$ be the number of arms. Each arm $i$ is associated with an unknown parameter $\beta_i \in \mathbb{R}^d$. For any integer $k$, we will let $[k]$ denote the set $\{1, ..., k\}$. At time $t$, we observe context $X_t \in \mathbb{R}^d$ (where $X_t \sim \mathcal{P}_X$ i.i.d.) and pull an arm $i \in [K]$, which yields reward $X_t^T \beta_i + \epsilon_{i,t}$, where $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$ i.i.d. We then incur regret

$$r_t := \mathbb{E} \left[ \max_j X_t^T \beta_j - X_t^T \beta_i \right],$$
and we seek to minimize the cumulative expected regret $R_T := \sum_{t=1}^{T} r_t$.

We consider the setting where the $\beta_i$ are sparse. The sparsity parameter $s_0 \in [d]$ satisfies $\|\beta_i\|_0 \leq s_0$ for all $i \in [K]$. Our algorithm has strong performance guarantees when $s_0 \ll d$. In other words, the arm rewards are a function of a small set of components of the context $X$.

3. Algorithm

Let $\pi_t \in [K]$ denote the arm chosen by our algorithm at time $t \in [T]$. At a high level, our proposed policy estimates the parameter vector $\hat{\beta}_i$ for each arm $i \in [K]$ based on past samples $X_t$ where arm $i$ was played. The algorithm addresses the exploration-exploitation tradeoff by performing forced sampling from each of the $K$ arms at prescribed time instances, and playing myopically otherwise (i.e., choosing the arm with highest estimated reward based on the current arm parameter estimates).

We use the LASSO estimator (described in §3.1) to take advantage of the sparsity of $\beta_i$. In particular, LASSO estimators require only $O(s_0 \log d)$ samples to estimate the true parameter $\hat{\beta}_i$ while traditional linear regression requires $O(d)$ samples Bühlmann and Van De Geer (2011). The key challenge is ensuring that the estimated parameters $\hat{\beta}_i$ converge to the true parameters $\beta_i$ despite the non-i.i.d. nature of the samples induced by the myopic action of our policy. As described in §3.2, we tackle this by maintaining two sets of estimators: the forced-sampling estimates based only on the forced-samples and the all-sample estimates based on all past samples when arm $i$ was played. The former offers an unbiased estimator while the latter is trained on a larger sample size that grows linearly in $T$ (with high probability). Thus, we use the forced-sampling estimator when we are sufficiently confident of our estimates (i.e. the estimated optimal arm reward is at least some fixed margin higher than the estimated rewards of the other arms), and we take advantage of the more accurate all-sample estimator otherwise.

3.1. LASSO Estimation

First, we describe our strategy for estimating $\beta_i$ for a fixed arm $i \in [K]$. Let the design matrix $X$ be the $T \times d$ matrix whose rows are $X_t$. Similarly, let $Y_i$ be the length $T$ vector of observations $X_t^T \beta_i + \epsilon_{i,t}$. Since we only obtain feedback when arm $i$ is played, entries of $Y_i$ may be missing.

Let $S_i = \{ t \mid \pi_t = i \} \subset [T]$ be the set of times when arm $i$ is played. For any subset $S' \subset S_i$, let $X(S')$ be the $|S'| \times d$ sub-matrix of $X$ whose rows are $X_t$ for each $t \in S'$. Similarly, let $Y_i(S')$ be the length $|S'|$ vector of corresponding observed rewards $(Y_t)_i$ for each $t \in S'$. Since $\pi_t = i$ for each $t \in S'$, the vector $Y_i(S')$ has no missing entries.

We estimate $\beta_i$ using the LASSO estimator

$$\hat{\beta}(S', \lambda) := \arg \min_\beta \left\{ \frac{\|Y_i(S') - X(S')\beta\|_2^2}{|S'|} + \lambda \|\beta\|_1 \right\}.$$
The LASSO estimator satisfies an oracle inequality (see Lemma 1), which guarantees with high probability that the estimator has small $L_1$ error with respect to the true parameter if $X(S')$ satisfies the compatibility condition (Definition 3.1).

For any index set $I \subset [d]$, let $\beta(I)$ be the vector obtained by setting the elements of $\beta$ that are not in $I$ to zero, i.e. $(\beta(I))_j = \beta_j \mathbb{1}[j \in I]$. Let $I_i$ be the index set corresponding to the nonzero elements of $\beta_i$.

**DEFINITION 1.** For each $i \in [K]$ and any subset $S' \subset S_i$, let $\hat{\Sigma}(S') = X(S')^T X(S')/|S'|$. The compatibility condition is satisfied for $\hat{\Sigma}(S')$ with constant $\phi_0 > 0$ if for all $\beta$ satisfying $\|\beta(I_i)\|_1 \leq 3\|\beta(I_i)\|_1$, it holds that $\|\beta(I_i)\|_1 \leq (\beta^T \hat{\Sigma}(S') \beta)^{1/2}/\phi_0$.

Then we have the following oracle inequality from Bühlmann and Van De Geer (2011) for $\hat{\beta}(S', \lambda)$:

**LEMMA 1.** If the compatibility condition holds for $\hat{\Sigma}(S')$ with constant $\phi_0$, then $\forall \chi > 0$,

$$\Pr \left[ \|\hat{\beta}(S', \lambda) - \beta_i\|_1 \leq \chi \right] \geq 1 - \exp \left[ -C_1 n \chi^2 + \log d \right],$$

where $\lambda := \frac{\phi_0^2 \chi}{4 \delta}$ and $C_1 := \frac{\phi_0^2}{128 \delta \sigma^2}$.

### 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 $h > 0$, as well as the regularization parameters $\lambda_1, \{\lambda_{2,t}\}_{t \in [T]}$. These parameters will be specified in Theorem 1.

**Forced-Sample Sets:** We define the set of times when we force-sample arm $i$,

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

Then, the set of forced samples from arm $i$ up to time $t$ is $T_{i,t} := T_i \cap [t]$. By construction, the $T_i$ are disjoint for $i \in [K]$, and we have $|T_{i,t}| \geq q \log \left( \frac{1}{\sqrt{q}} + 1 \right)$. Thus, we have $O(q \log t)$ samples from each arm at time $t$ (see Lemma 8 in appendix for exact values).

**All-Sample Sets:** Let $S_{i,t}$ denote the set of times we play arm $i$

$$S_{i,t} := \{ j \mid \pi_j = i \text{ and } 1 \leq j \leq t \}.$$

**Estimates:** At any time $t$, Algorithm 1 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 $i$, and (2) the all-sample estimate $\hat{\beta}(S_{i,t-1}, \lambda_2,t)$ based on all samples observed from arm $i$.

**Execution:** If the current time $t \in T_i$ for some arm $i$, then arm $i$ is played. Otherwise, two actions are possible. First, we use the forced-sample estimates for each arm to check if there exists some arm $i$ whose estimated reward is higher than the estimated rewards of the other arms by a margin of at least $h/2$. If this is the case, then arm $i$ is played. Otherwise, we use the all-sample estimates to play the arm with highest estimated reward.
Algorithm 1 LASSO Bandit

Input parameters: $q, h, \lambda_1, \{\lambda_2,t\}_{t \in [T]}$

Let $\hat{\beta}(T_i,0,\lambda_1), \hat{\beta}(S_i,0,\lambda_2) \leftarrow 0 \ \forall i \in [K]$

for $t \in [T]$ do

Observe $X_t \in \mathcal{P}_X$

if $t \in T_i$ for any $i$ then

$\pi_t \leftarrow i$

else

if $X_t^T \hat{\beta}(T_i,t-1,\lambda_1) \geq h/2 + \max_{j \neq i} X_t^T \hat{\beta}(T_j,t-1,\lambda_1)$ for any $i$ then

$\pi_t \leftarrow i$

else

$\pi_t \leftarrow \arg \max_i X_t^T \hat{\beta}(S_i,t-1,\lambda_2,t-1)$

end if

end if

$S_{\pi_t,t} \leftarrow S_{\pi_t,t-1} \cup \{t\}$

Play arm $\pi_t$, observe $y_t = X_t^T \hat{\beta}_{\pi_t} + \epsilon_{i,t}$

end for

4. Regret Analysis

We list our assumptions in §4.1, state our main result in §4.2, and describe the key steps for our proof at a high level in §4.3. All complete proofs are relegated to the appendix.

4.1. Assumptions

We first assume that the support of the context distribution $\mathcal{P}_X$ as well as the arm parameters $\beta_i$ are bounded. This ensures that the maximum regret at any time step is bounded.

**Assumption 1.** (parameter set) $\|X_t\|_{\infty} \leq x_{\text{max}}$ for all $t$ and $\|\beta_i\|_1 \leq b$ for all $i \in [K]$.

Second, we assume a margin condition that ensures that $\mathcal{P}_X$ cannot become unbounded “near” a decision boundary $\{X^T \beta_i = X^T \beta_j\}$ for any $i \neq j \in [K]$. This ensures that we do not get too many contexts near the boundary, where our parameter estimates are the most uncertain. This condition is widely used in the classification literature (e.g. Tsybakov (2004)) as well as in Goldenshluger et al. (2013). We also note that any uniformly bounded distribution $\mathcal{P}_X$ satisfies this condition.

**Assumption 2.** (margin condition) There exist $\kappa_0, C_0 \in \mathbb{R}^+$ such that $\forall i \neq j \in [K]$,

$$\Pr \left[ |X^T (\beta_i - \beta_j) | \leq \kappa \right] \leq C_0 \kappa$$

for $\forall \kappa \in (0, \kappa_0]$.

Next, we define sets of contexts $U_i$ such that for any $X \in U_i$, arm $i$ is optimal relative to the other arms by the localization parameter $h > 0$ (input to Algorithm 1). We assume that each $U_i$ has nonzero support in $\mathcal{P}_X$, i.e. each arm will be optimal at least $p_* T$ times in expectation (for some $p_* > 0$). This ensures that we can improve each arm’s parameter estimates over time while still playing the optimal arm with high probability.
Assumption 3. (arm optimality) There exists \( p_* > 0 \) such that if we define the set
\[
U_i := \left\{ X \left| \min_{j \neq i} \left[ X^T (\beta_i - \beta_j) \right] \geq h \right. \right\},
\]
for each \( i \in [K] \), then \( \min_i \Pr[U_i(X) = 1] \geq p^* \).

Our fourth and final assumption will be about the compatibility condition (Definition 3.1), which is required for LASSO estimates to satisfy the oracle inequality (Lemma 1).

Assumption 4. (compatibility condition) There exists \( \phi_0 > 0 \) such that for all \( \beta \) satisfying
\[
\| \beta(I_i) \|_1 \leq 3 \| \beta(I_i) \|_1,
\]
\[
\| \beta(I_i) \|_2^2 \leq (\beta^T \Sigma_i \beta) s_0 / \phi_0^2,
\]
for each \( i \in [K] \), where we define \( \Sigma_i := \mathbb{E}[XX^T | X \in U_i] \).

4.2. Main Result

Our main result establishes an upper bound of \( \mathcal{O}\left(Ks_0^2 \sigma^2 \left[ (\log T)^2 + \log T \log d \right]\right) \) on the cumulative expected regret at time \( T \).

Theorem 1. When \( q \geq 4[q_0] \), \( K \geq 2 \), \( \log d > 1 \), and we take
\[
\lambda_1 = \frac{\phi_0^2 p_* h}{64 s_0 x_{\text{max}}} \quad \text{and} \quad \lambda_{2,t} = \frac{\phi_0^2}{2 s_0} \sqrt{\frac{\log t + \log d}{p_* C_1 t}} \quad \text{and} \quad t \geq (Kq)^2,
\]
we have an upper bound on the expected cumulative regret at time \( T \):
\[
R_T \leq \frac{C_3}{2} (\log T)^2 + 2Kb x_{\text{max}} (6q + 1) + C_3 \log d \log T + (2q^2 K^2 b x_{\text{max}} + 3K + C_4)
\]
\[
= \mathcal{O}\left(Ks_0^2 \sigma^2 \left[ (\log T)^2 + \log T \log d \right]\right),
\]
where we define the constants
\[
C_1 := \frac{\phi_0^2}{128 s_0^2 \sigma^2}, \quad C_2 := \frac{\phi_0^2}{384 s_0 x_{\text{max}}^2}, \quad C_3 := \frac{1024 K C_0 x_{\text{max}}^2}{p_*^4 C_1^2}, \quad \text{and} \quad C_4 := \frac{3}{1 - \exp \left[-p_*^2 (C_2^2 / 128) \right]},
\]
and we take
\[
q_0 = \max \left\{ \frac{8}{p_*(C_2 \wedge p_*)}, \frac{4x_{\text{max}} \log d}{p_* C_2}, \frac{256 x_{\text{max}}^2 \log d}{h^2 p_*^2 C_1} \right\} = \mathcal{O}\left(s_0^2 \sigma^2 \log d\right).
\]

4.3. Key Steps in the Proof of Theorem 1

We first prove a general oracle inequality when an unknown portion of the data is generated by a non-i.i.d. process (§4.3.1). We then use this result to derive oracle inequalities for both the forced- and all-sample estimators (§4.3.2). Finally, in §4.3.3, we state the expected cumulative regret from the various actions in Algorithm 1, which sum to the bound given in Theorem 1.
4.3.1. An Oracle Inequality for non-i.i.d. Data Consider a random variable $Z \in \mathbb{R}^d$ that is bounded in magnitude component-wise by $z_{\text{max}}$. Assume $\Sigma := \mathbb{E}_{Z \sim p_Z}[ZZ^T]$ satisfies the compatibility condition with constant $\phi_0$ with respect to a fixed distribution $\mathcal{P}_Z$.

We define $S$ to be the index set corresponding to samples $Z_1, \ldots, Z_{|S|}$, and $S' \subset S$ such that \{ $Z_i \mid i \in S'$ \} is an i.i.d. subset of random variables with distribution $\mathcal{P}_Z$. Suppose that $S'$ was constructed using a procedure such that if $i \in S$, then $i \in S'$ with at least probability $p$. We will first show that $\hat{\Sigma}(S') = \frac{1}{|S'|} \sum_{i \in S'} Z_i Z_i^T$ satisfies the compatibility condition with constant $\phi_0 / \sqrt{2}$ with high probability. The proof involves showing that (1) $\|\hat{\Sigma}(S') - \Sigma\|_\infty$ is small with high probability using results on matrix perturbations, and (2) invoking oracle inequality results for random design matrices. Next, we use this fact along with a martingale concentration inequality to show that $\hat{\Sigma}(S) = \frac{1}{|S|} \sum_{i \in S} Z_i Z_i^T$ satisfies the compatibility condition with constant $\phi_0 \sqrt{p} / 2$ with high probability. This result implies that the oracle inequality (Lemma 1) holds with high probability for LASSO estimates $\hat{\beta}(S, \lambda)$ although part of the data is not generated i.i.d. from $\mathcal{P}_Z$.

**Theorem 2.** If $\hat{\beta}(S, \lambda)$ is the LASSO solution given observations in $S$ and

$$|S| \geq \frac{4z_{\text{max}} \log d}{pC_2}$$ and $\lambda = \frac{\phi_0^2 p}{16s_0^2} \chi$,

then the following oracle inequality holds $\forall \chi > 0$:

$$\Pr \left[ \|\hat{\beta}(S, \lambda) - \beta\|_1 \leq \chi \right] \geq 1 - \exp \left[ -n \chi^2 \cdot \frac{p^2 C_1}{16} + \log d \right],$$

with probability at least

$$1 - 2 \exp \left[ - \frac{(C_2 \wedge p)p|S|}{8} \right].$$

4.3.2. Oracle Inequalities for the Forced- and All-Sample Estimators In order to bound the regret for the forced- and all-sample estimators, we use oracle inequalities that guarantee good estimates of each $\beta_i$ with high probability. Whenever we use the forced-sample estimator, we play the optimal arm at time $t+1$ as long as the following event holds:

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

The following oracle inequality for the forced sample estimator $\hat{\beta}(T_{t,t}, \lambda_1)$ shows that $A_t$ holds with high probability (which will, in turn, ensure that there are sufficiently many unbiased samples for the all-sample estimator oracle inequality to hold):

**Proposition 1.** The forced sample estimator $\hat{\beta}(T_{t,t}, \lambda)$ satisfies the oracle inequality

$$\Pr \left[ \|\hat{\beta}(T_{t,t}, \lambda) - \beta_i\|_1 \leq \frac{h}{4x_{\text{max}}} \right] \geq 1 - \exp \left[ -q_0 \log t \cdot \frac{p^2 h^2 C_1}{256 x_{\text{max}}^2} + \log d \right] - \frac{2}{t},$$

when $\lambda_1 := \frac{\phi_0^2 p \cdot h}{64s_0 x_{\text{max}}}$ and $t \geq (Kq)^2$. 
Proposition 1 does not follow directly from Lemma 1 because we need to prove that the compatibility condition holds for \( \Sigma(T_{i,t}) \approx \mathbb{E}_{X \sim P_X} [XX^T] \), whereas we have only assumed that it holds for \( \Sigma_i = \mathbb{E}_{X \sim P_X} [XX^T | X \in U_i] \). Instead, we show that \( T_{i,t}' = \{ t' \in T_{i,t} | X_{t'} \in U_i \} \) is a set of i.i.d. samples from \( P_{X;X \in U_i} \), and then apply Theorem 2 with \( S = T_{i,t} \) and \( S' = T_{i,t}' \).

Recall that we use the all-sample estimator instead of the forced-sample estimator when we are not sufficiently confident about the estimated optimal arm. This is because the oracle inequality we establish for the all-sample estimator is stronger due to the increased sample size:

**Proposition 2.** The all-sample estimator \( \hat{\beta}(S_{i,t}, \lambda_2,t) \) satisfies the oracle inequality

\[
\Pr \left[ \| \hat{\beta}(S_{i,t}, \lambda_2,t) - \beta_i \|_1 \leq 16 \left( \frac{\log t + \log d}{p^2 C_1 t} \right) \right] \geq 1 - \frac{1}{t},
\]

with probability at least

\[
1 - 3 \exp \left[ - \frac{p^2 (C_2 \cap p_* / 4)}{128} t \right]
\]

when \( \lambda_2,t := \frac{\phi^2}{2s_0} \sqrt{\log t + \log d} \frac{1}{p_* C_1 t} \) and \( t \geq (Kq)^2 \).

In particular, Proposition 2 guarantees \( \| \hat{\beta}(S_{i,t}, \lambda_2,t) - \beta_i \|_1 = O(\sqrt{\log t / t}) \) with high probability while Proposition 1 only guarantees \( \| \hat{\beta}(T_{i,t}, \lambda_1) - \beta_i \|_1 = O(1) \) with high probability.

To prove Proposition 2, we need to account for the fact 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 \( \{ X_{t'} | t' \in S_{i,t} \text{ and } X_{t'} \in U_i \} \) may be correlated. Suppose we restrict to \( t' \) for which \( A_{t'} \) holds. Since \( A_{t'} \) depends only on past observations, the random variables \( \{ X_{t'} | A_{t'} \text{ holds} \} \) are independent (distributed as \( P_X \)). Furthermore, if we let \( S_{i,t}' = \{ t' \in [t] | A_{t'} \text{ holds and } X_{t'} \in U_i \} \), then the random variables \( \{ X_{t'} | t' \in S_{i,t}' \} \) are independent (distributed as \( P_{X;X \in U_i} \)). Finally, the event \( A_{t'} \) ensures that we play arm \( i \) when \( X_{t'} \in U_i \), so \( S_{i,t}' \subseteq S_{i,t} \). Proposition 1 ensures that \( |S_{i,t}'| \) is sufficiently large, so the result follows from Theorem 2 with \( S = S_{i,t} \text{ and } S' = S_{i,t}' \).

**Table 1 Sources of Regret**

<table>
<thead>
<tr>
<th>Source</th>
<th>Upper Bound on Regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization from ( t \leq (Kq)^2 )</td>
<td>( 2q^2 K^2 \max(bx) )</td>
</tr>
<tr>
<td>Forced-Sampling</td>
<td>( 12qK \max(bx) \log T )</td>
</tr>
<tr>
<td>Use of Forced-Sample Estimator</td>
<td>( 2K \log T + 3K )</td>
</tr>
<tr>
<td>Use of All-Sample Estimator</td>
<td>( (2K \max(bx) + C_3 \log d) \log T + \frac{C_4}{t} (\log T)^2 + C_4 )</td>
</tr>
</tbody>
</table>

**4.3.3. Bounds on the Cumulative Expected Regret** We obtain regret from four sources: (1) “initialization” for \( t \leq (Kq)^2 \), since our oracle inequalities do not hold, (2) forced-sampling,
(3) estimation error when using the forced-sample estimator, and (4) estimation error when using
the all-sample estimator. Table 1 lists regret bounds for each source; their sum gives the result in
Theorem 1.

5. Experiments on Synthetic Data
We simulate our algorithm (which we term LASSO Bandit) on synthetic data. This allows us to
• compare the LASSO Bandit’s performance against the contextual bandit algorithm introduced
  in Goldenshluger et al. (2013) (OLS Bandit), and
• check the algorithm’s robustness to the choices of inputs $q$, $h$, and $\lambda_1, \{\lambda_{2,t}\}_{t \in [T]}$.

**Synthetic Data Generation:** We take $d = 100$, $s_0 = 10$, $\sigma^2 = 0.1$, and $K = 2$. Unless otherwise
specified, $q = 1$, $h = 5$, and $c = 0.05$ (we take $\lambda_1 = c$ and $\lambda_{2,t} = c \sqrt{\log t + \log d} / t$ (as suggested by
Theorem 1). We generate $X_t$ with i.i.d entries from $\mathcal{N}(0, 1)$ and normalize it so that $\|X_t\|_2 = 1$ for
each $t \in [T]$. We also generate $\beta_i$ with $s_0$ randomly selected components set to i.i.d. draws from
$\mathcal{N}(0, 1)$ and the remaining components set to 0 for each $i = 1, 2$.

![Figure 1](image)

**Figure 1**  Cumulative regret for LASSO Bandit vs. OLS Bandit on synthetic data with 2 arms.

Figure 1 compares the cumulative regret (averaged over 50 trials) of the LASSO Bandit against
the OLS Bandit on synthetic data for $T = 10,000$ steps. We see that the LASSO Bandit significantly
outperforms the OLS Bandit in cumulative regret. Furthermore, even after $t \gg d$, the LASSO
Bandit continues to achieve less per-time-step regret than the OLS Bandit (as can be seen by the
slope of the curves at large $t$); this suggests that the OLS Bandit continues to overfit the arm
parameters even after the number of samples greatly exceeds the dimension.
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 regularization parameter $c$. 
Next, we test the robustness of the LASSO Bandit to the choice of input parameters. Figure 2 compares the cumulative regret of the LASSO Bandit while varying any one of

- the forced sampling parameter $q \in \{1, 2, 5\}$
- the localization parameter $h \in \{0.5, 1, 5\}$
- the regularization parameter $c \in \{0.02, 0.05, 0.1, 0.2\}$

The results are computed over $T = 1000$ time steps and averaged over 50 trials. We find that the cumulative regret performance is quite similar despite experimenting with the parameters by up to an order of magnitude, suggesting that the LASSO Bandit is robust to the choice of inputs.

6. Case Study: Warfarin Dosing

We now demonstrate the practical usefulness of the LASSO Bandit by evaluating it on the clinically relevant issue of warfarin dosing. 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-specific factors; moreover, an incorrect initial dosage can result in highly adverse consequences such as stroke or internal bleeding.

6.1. 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. This data contains patient-specific optimal warfarin doses (found through a physician-guided iterative process of trial and error) for 5528 patients, as well as 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:

- gender
- race (white, black, Asian, unknown)
- ethnicity (Hispanic/Latino or not)
- age
- height
- weight
- reason for warfarin treatment (DVT, PE, Afib/flutter, heart valve, cardiomyopathy/LV dilation, stroke, post-orthopedic, other)
- indicator variables for patient conditions (diabetes, congestive heart failure/cardio.myopathy, valve replacement, smoker)
• concomitant medications (aspirin, Tylenol, indication for Tylenol dosage ≤ 1300mg/day, Zocor, Lipitor, Lescol, Mevacor, Pravachol, Crestor, Baycol, Codarone, Tegretol, Dilantin, Rifampin/Rifampicin, Sulfonamide antibiotics, Macrolide antibiotics, Anti-fungal azoles, herbal medications/vitamins/supplements)
• presence of genotype variants of CYP2C9 and VKORC1

We refer the reader to Section 1 of the Supplementary Appendix 1 of Consortium (2009) for further details on the dataset.

6.2. Bandit Formulation

We formulate the problem of assigning initial warfarin doses as a 3-armed contextual bandit.

**Arms:** We bucket the optimal dosages using the “clinically relevant” dosage differences suggested by Consortium (2009):
- Low (under 3mg/day, 33% of our sample)
- Medium (3-7mg/day, 54% of our sample)
- High (over 7mg/day, 13% of our sample)

In particular, patients who require under 3mg/day (low) or over 7mg/day (high) of warfarin would be at risk for excessive or inadequate anti-coagulation under the normal starting dosage of 5mg/day. We also estimate the arm parameters $\beta_i$ using linear regressions on the entire dataset.

**Covariates:** We used the patient data to construct 93 patient-specific covariates, including indicators for NAs when there was missing data. We note that all of these covariates were hand-selected as relevant variables by medical professionals and that there are no extraneously added variables; thus, this truly is a high-dimensional problem.

**Model Choice:** Finally, it is important to note that Consortium (2009) finds that an ordinary least-squares linear model fits the data best (i.e. achieves the best cross-validation accuracy) compared to alternative models (such as support vector regression, regression trees, model trees, multivariate adaptive regression splines, least-angle regression, and LASSO) for the objective of predicting the correct warfarin dosage as a function of the given patient-level variables. Thus, our model assumptions are justified.

6.3. Evaluation & Results

We generate random sequences of patients in each trial and simulate the following policies:
- Oracle, which assigns the optimal estimated dose given the arm parameters $\beta_i$
- LASSO Bandit, described in Algorithm 1
- OLS Bandit, described in Goldenshluger et al. (2013)
- Doctors, who always assign the medium dose of 5mg/day (Consortium 2009)
We average our results over 10 random patient permutations.

Figure 3 compares the fraction of correctly dosed patients under the aforementioned four policies after 100, 500, 1000, and 5000 patients in the warfarin dosing data. We clearly see three data regimes:

- When there is very little information (100 patients), the doctor’s policy of assigning the medium dose (which is optimal for the majority of the patient population) performs the best.
- When there is a moderate amount of information (500 or 1000 patients), the LASSO Bandit effectively learns the arm parameters and outperforms the doctor policy; however, the OLS Bandit still continues to perform poorly compared to both the LASSO Bandit and the doctor policy.
- When there is a lot of information (5000 patients), both bandit policies outperform the doctor policy and begin to look comparable.

Thus, we see that our algorithm can leverage minimal information (∼500 patients) to outperform existing bandit methods as well as physicians to correctly dose a majority of patients. In particular, 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. Our results illustrate the practical relevance of the LASSO Bandit in settings where one must sequentially make decisions under uncertainty.
7. Conclusions and Future Work
We present a contextual bandit algorithm that achieves $O(s_0^2|\log T|^2 + s_0^2|\log T^2\log d)$ regret in the setting where the observed covariates are high-dimensional. We show theoretically and through simulations that our algorithm outperforms existing methods in terms of cumulative regret. Furthermore, we illustrate the practical relevance of our algorithm by evaluate it on the clinically relevant task of warfarin dosing. We find that our algorithm surpasses existing bandit methods as well as physicians to correctly dose a majority of patients.

There are several directions for future work. First, the lower bound in Goldenshluger et al. (2013) can be adapted to our setting to show that no algorithm can achieve regret better than $O(\log T)$ in $T$; however, our analysis provides an upper bound on the regret of $O(\log T^2)$. An immediate question is whether our analysis can be improved to meet the lower bound. Second, our regret bounds scale as $O(s_0^2\log d)$ while the bounds in Goldenshluger et al. (2013) scale as $O(d^3)$ in the dimension. This suggests that our proof technique, if adapted to the low-dimensional setting, may be able to improve the bound in Goldenshluger et al. (2013) to $O(d^2)$. Finally, the results in our work 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).

Appendix. Proof of Main Result
The first part of the proof (§A) will establish oracle inequalities for the forced-sample and all-sample estimators. §B uses these inequalities to bound the expected cumulative regret of our proposed Algorithm 1.

A. Oracle Inequalities
We first prove a general oracle inequality when an unknown fraction of the data is generated by a non-i.i.d. process (Theorem 2 in §1.3). §1.1 and §1.2 collect results on the LASSO oracle inequality and matrix perturbations that are used in the proof.

We then use Theorem 2 to derive oracle inequalities for both the forced-sample estimator (Proposition 1 in §1.4) and the all-sample estimator (Proposition 2 in §1.5).

A.1. LASSO Theory
Assume the underlying true model is linear

$$Y = X\beta + \epsilon$$

where $Y_{n \times 1}$ is the vector of responses, $X_{n \times d}$ is the design matrix, $\epsilon_{n \times 1}$ is the vector of measurement errors with distribution $N(0, \sigma^2 I)$, and $\beta_{d \times 1}$ is the true parameter. We further assume that $X$ has been scaled such that $\text{diag}(X^TX/n) = 1$.

The LASSO chooses

$$\hat{\beta}(\lambda) = \arg\min_\beta \left\{ \frac{\|Y - X\beta\|^2}{n} + \lambda\|\beta\|_1 \right\}$$
Let $I \subset \{1, \ldots, d\}$ denote the subset of indices corresponding to non-zero coefficients of $\beta \in \mathbb{R}^d$. We further let $\beta(I)$ be the vector whose components are $\beta_j I[j \in I]$. The sparsity parameter is the smallest number $s_0 \in [0, d]$ such that $|I| \leq s_0$.

**Definition 2.** The compatibility condition is satisfied on $\hat{\Sigma}$ with constant $\phi_0 > 0$ if for all $\beta$ satisfying $\|\beta(I^c)\|_1 \leq 3\|\beta(I)\|_1$, it holds that

$$
\|\beta(I)\|_2^2 \leq \left(\beta^T \hat{\Sigma} \beta\right) s_0 / \phi_0^2.
$$

**Lemma 2.** If the compatibility condition holds on $\hat{\Sigma}$ with constant $\phi_0$, then $\forall \chi > 0$

$$
\Pr \left[\|\hat{\beta}(\lambda) - \beta\|_1 \leq \chi\right] \geq 1 - \exp \left[-C_1 n \chi^2 + \log d\right]
$$

where

$$
\lambda := \phi_0^2 \sqrt{\frac{4s_0}{\sigma^2}} \quad \text{and} \quad C_1 := \frac{\phi_0^4}{128s_0^2 \sigma^2}.
$$


**A.2. Matrix perturbations**

**Definition 3.** A random variable $Z$ is sub-Gaussian if $\exists L, \sigma_0$ such that

$$
L^2 \left(\mathbb{E} [e^{|Z|^2/L^2}] - 1\right) \leq \sigma_0^2.
$$

**Lemma 3.** Suppose a random variable $Z$ is bounded with $|Z| \leq z_{\max}$. Then $Z$ is sub-Gaussian with $L = z_{\max}$ and $\sigma_0 = z_{\max} \sqrt{e - 1}$ for each $i \in \{1, \ldots, d\}$

*Proof of Lemma 3* The proof follows from the definition of sub-Gaussian.

**Lemma 4.** Given i.i.d. observations $Z_1, \ldots, Z_n \in \mathbb{R}^d$ such that $Z_{i,j}$ are uniformly sub-Gaussian with parameters $L, \sigma_0$, we have $\forall w$

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


**A.3. Compatibility condition for non-i.i.d. samples**

Throughout this section, we take $Z \in \mathbb{R}^d$ to be a random variable that is bounded in magnitude component-wise by $z_{\max}$. Furthermore, we assume

$$
\Sigma := \mathbb{E}_{Z \in \mathcal{P}_Z} [ZZ^T]
$$
satisfies the compatibility condition with constant $\phi_0$ with respect to a fixed distribution $\mathcal{P}_Z$.

We define $S$ to be the index set corresponding to samples $Z_1, \ldots, Z_{|S|}$, and $S' \subset S$ to be a subset of indices such that $\forall i \in S'$, $Z_i \sim \mathcal{P}_Z$ i.i.d. We suppose that $S'$ was constructed using some procedure such that for each $i \in \{1, \ldots, |S|\}$, $Z_i \in S'$ with at least probability $p$. We will first show that

$$
\hat{\Sigma}(S') = \frac{1}{|S'|} \sum_{i \in S'} Z_i Z_i^T
$$
satisfies the compatibility condition with high probability. Next, we will use this fact to show that

$$
\hat{\Sigma}(S) = \frac{1}{|S|} \sum_{i \in S} Z_i Z_i^T
$$
also satisfies the compatibility condition with high probability.

This result implies that the oracle inequality (Lemma 1) holds with high probability for LASSO estimates \( \hat{\beta}(S, \lambda) \) although part of the data is not generated i.i.d. from \( P_Z \).

**Lemma 5.** If the compatibility condition holds for \( \Sigma_0 \) with constant \( \phi_0 \) and

\[
\| \Sigma_0 - \Sigma_1 \|_\infty \leq \frac{\phi_0^2}{32s_0}
\]

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

**Proof of Lemma 5** The proof follows directly from Corollary 6.8 in Buhlmann and Van de Geer (2011).

\( \square \)

**Lemma 6.** \( \hat{\Sigma}(S') \) satisfies the compatibility condition with constant \( \phi_0/\sqrt{2} \) when

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

with probability \( 1 - \exp [-C_2|S'|] \) where

\[
C_2 := \frac{\phi_0^2}{384s_0 z_{\max}^2}
\]

**Proof of Lemma 6** From Lemma 2, \( Z \) is sub-Gaussian with \( L = z_{\max} \) and \( \sigma_0 = z_{\max} \sqrt{e - 1} \). Then, if \( w = C_2 \) and there is a lower bound on \( |S'| \) as defined above, we can check that

\[
2L^2w + 2L\sigma_0 \sqrt{2w} + 2L\sigma_0 \left( \sqrt{\frac{2 \log (d^2 - d)}{|S'|}} + \frac{L \log (d^2 - d)}{|S'|} \right) \leq \frac{\phi_0^2}{32s_0}
\]

Thus, it follows from Lemma 3 that

\[
\Pr \left[ \| \Sigma - \hat{\Sigma}(S') \|_\infty \geq \frac{\phi_0^2}{32s_0} \right] \leq \exp [-C_2|S'|]
\]

The result then follows directly from Lemma 4. \( \square \)

**Lemma 7.** If the compatibility condition holds with constant \( \phi_0 \) for \( \hat{\Sigma}(S') \), then the compatibility condition holds with constant \( \phi_0 \sqrt{\frac{|S'|}{|S|}} \) for \( \hat{\Sigma}(S) \)

**Proof of Lemma 7** From our definition, we can write

\[
\hat{\Sigma}(S) = \frac{|S'|}{|S|} \hat{\Sigma}(S') + \frac{1}{|S|} \sum_{i \in S - S'} Z_i Z_i^T
\]

\[
= \frac{|S'|}{|S|} \hat{\Sigma}(S') + \frac{|S - S'|}{|S|} \hat{\Sigma}(S - S')
\]

Then, for all \( \beta \) satisfying \( \| \beta \|_1 \leq 3 \| \beta_{S_0} \|_1 \),

\[
\beta^T \Sigma_1 \beta = \frac{|S'|}{|S|} \beta^T \hat{\Sigma}(S') \beta + \frac{|S - S'|}{|S|} \beta^T \hat{\Sigma}(S - S') \beta 
\geq \frac{|S'|}{|S|} \left( \| \beta_{S_0} \|_2^2 \phi_0^2 \right) \frac{1}{s_0}
\]

from the fact that \( \hat{\Sigma}(S - S') \) is a covariance matrix, and is therefore positive semi-definite. \( \square \)
Lemma 8. Recall that for each $i \in S$, $i \in S'$ with probability at least $p$. Then, for all $c > 0$,

$$\Pr[|S'| \leq (p - c)|S|] \leq \exp\left(\frac{-c^2|S|}{2}\right)$$

Proof of Lemma 8 We can write $|S'| \geq p|S| + \tilde{N}_{|S|}$ where we define

$$\tilde{N}_i := \tilde{N}_{i-1} + \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

for $i \in \{1, \ldots, |S|\}$ and $\tilde{N}_0 = 0$. Then $\tilde{N}_i$ is a martingale with bounded differences, so we can apply the Azuma-Hoeffding inequality to obtain

$$\Pr\left[|S'| \leq -c|S|\right] \leq \exp\left(-c^2|S|^2\right)$$

The result follows directly. □

A.3.1. Proof of Theorem 2 Recall that Theorem 2 provides a general oracle inequality when an unknown fraction of the data is generated by a non-i.i.d. process.

Proof of Theorem 2 Applying Lemma 6 to Lemma 5 implies that the compatibility condition holds for $\hat{\Sigma}(S)$ with constant

$$\frac{\phi_0}{\sqrt{2}} \sqrt{\frac{|S'|}{|S|}}$$

with probability $1 - \exp\left[-C|S'|ight]$ if $|S'| \geq \frac{2^{|S|}\log d}{C_2}$. Lemma 7 with $c = p/2$ implies that

$$\Pr\left[|S'| \geq \frac{p|S|}{2}\right] \geq 1 - \exp\left[-\frac{p^2|S|}{8}\right]$$

Conditioning on this event and applying a union bound gives us that $\hat{\Sigma}(S)$ satisfies the compatibility condition with constant

$$\frac{\phi_0 \sqrt{p}}{2}$$

when $|S| \geq \frac{4^{\max\log d}}{pC_2}$ with probability

$$1 - \exp\left[-\frac{pC_2|S|}{2}\right] - \exp\left[-\frac{p^2|S|}{8}\right] \geq 1 - 2\exp\left[-\frac{p(C_2 \wedge p)|S|}{8}\right]$$

Applying Lemma 1 in the event that the compatibility condition holds gives us the result. □

A.4. Oracle inequality for the forced sample estimator

In this section, we prove an oracle inequality for the forced sample estimator $\hat{\beta}(T_{i,t}, \lambda_1)$ by applying Theorem 2. Recall that at each $t \in T_{i,t}$, we draw a context $X_t \in P_X$ i.i.d. and play arm $i$. Moreover, we assumed that the compatibility condition holds with constant $\phi_0$ for $\Sigma_i = \mathbb{E}_{X \sim P_{X_i}}[XX^T]$ (where $P_X = P_{X,X \in U_i}$) and also that $\Pr[X_t \in U_i] \geq p_*$. 

Lemma 9. If $t \geq (Kq)^2$, then $q_0 \log t \leq |T_{i,t}| \leq 6q_0 \log t$.

Proof of Lemma 9 Define the $n^{th}$ round of forced sampling of all the arms

$$L_n := \{(2^n - 1)Kq + 1, \ldots, (2^n + 1)Kq\}$$
for \(n \geq 0\). By construction, arm \(i\) is sampled \(|T_{i,t} \cap L_i| = q\) times, so

\[
|T_{i,t} \cap \left( \bigcup_{j=0}^{n-1} L_j \right) | = nq
\]

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 \(t \leq (2^n+1)Kq\), i.e. \(n \leq \log_2 \left( \frac{t}{Kq} + 1 \right) - 1\), so

\[
|T_{i,t}| \geq nq \geq q \left( \log_2 \left( \frac{t}{Kq} + 1 \right) - 1 \right) \geq q \left( \log t - \log Kq - 1 \right)
\]

since \(\log 2 < 1\). By our assumption that \(q \geq 4[q_0]\), for \(t \geq (Kq)^2\),

\[
|T_{i,t}| \geq 4[q_0] \left( \log t - \log Kq - 1 \right)
\]

\[
\geq [q_0] \log t + 2[q_0] \left( \log t - \log((Kq)^2) \right) + [q_0] \left( \log t - 4 \right)
\]

\[
\geq q_0 \log t
\]

where we have used the fact that \(t \geq (Kq)^2 \geq 4\) (since \(K \geq 2\) and \(q \geq 1\)).

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

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

\(\square\)

**Lemma 10.** Let \(T'_{i,t} \subset T_{i,t}\) be the set of all \(t \in T_{i,t}\) such that \(X_i \in U_t\). Then \(\forall t \in T'_{i,t}, X_i \sim P_X\) i.i.d., and for each \(t \in T'_{i,t}, t \in T'_{i,t}\) with probability at least \(p_\ast\).

**Proof of Lemma 10** By construction, for each \(t \in T_{i,t}\), \(X_i\) is drawn i.i.d. from \(P_X\) and therefore with probability at least \(p_\ast\), \(X_i \in U_t\), i.e. \(t \in T'_{i,t}\). Note that we are doing rejection sampling to construct \(T'_{i,t}\), and so for each \(t \in T'_{i,t}, X_i\) is an i.i.d. sample of \(P_{X|X \in U_t} = P_X\). \(\square\)

**Note:** for the rest of the appendix, we replace \(z_{\max}\) with \(x_{\max}\) in the definition of

\[
C_2 := \frac{\phi_0^2}{384s_0x_{\max}^2}
\]

**A.4.1. Proof of Proposition 1** Recall that Proposition 1 provides an oracle inequality for the forced sample estimator \(\hat{\beta}(T_{i,t}, \lambda_1)\).

**Proof of Proposition 1** By construction, \(|T_{i,t}| \geq q_0 \log t\). Lemma 9 allows us to apply Theorem 2 that the oracle inequality (with \(\chi = h/4x_{\max}\)) holds with probability at least

\[
1 - 2\exp \left[ - \frac{p_\ast (C_2 \wedge p_\ast) |T_{i,t}|}{8} \right] \geq 1 - 2\exp \left[ - \frac{q_0 p_\ast (C_2 \wedge p_\ast) \log t}{8} \right] \geq 1 - \frac{2}{t}
\]

(by definition of \(q_0\)) when \(t \geq (Kq)^2\). Note that the assumption

\[
|T_{i,t}| \geq q_0 \log t \geq \frac{4x_{\max} \log d}{p_\ast C_2}
\]

implies \(\log t \geq 1\) (by definition of \(q_0\)), which is already satisfied by our assumption that \(t \geq (Kq)^2 \geq 4\). A union bound then gives the result. \(\square\)
A.5. Oracle inequality for the all-sample estimator

In this section, we prove an oracle inequality for the all-sample estimator $\hat{\beta}(S_{i,t}, \lambda_t)$. The approach mirrors the approach taken in Section A.4. However, there is additional bias in the sets $S_{i,t}$ introduced by the policies $\pi_t$. In other words, at every step we observe $X \sim \mathcal{P}_X$, but we selectively sample $Y = X^T \beta_i + \epsilon$ where $i_t = \pi_t(X)$ depends on our policy.

In order to handle this, we show that an expected constant fraction of our samples are still i.i.d. from $\mathcal{P}_i$ from the use of the forced-sample estimator outside the margin when the following event holds:

**Definition 4.**
$$A_t = \left\{ \|\tilde{\beta}(T_{i,t}, \lambda_1) - \beta_i\|_1 \leq \frac{h}{4x_{\text{max}}} \quad \forall i \in \{1, ..., K\} \right\}$$

Using this, we can again prove that the oracle inequality holds with high probability for the all-sample estimator.

**Lemma 11.** For each $i \in \{1, ..., K\}$, if $X \in U_i$ and $A_t$ holds, Algorithm 1 uses the forced-sample estimator $\tilde{\beta}(T_{i,t}, \lambda_1)$ and pulls the optimal arm $i$.

**Proof of Lemma 11** Since $X \in U_i$, we know
$$X^T \beta_i \geq h + \max_{j \neq i} X^T \beta_j$$

Then, for any $j \neq i$,
$$X^T \left( \tilde{\beta}(T_{i,t}) - \beta_i \right) = X^T \left( \tilde{\beta}(T_{i,t}) - \beta_i \right) + X^T \left( \beta_j - \tilde{\beta}(T_{j,t}) \right) + X^T \left( \beta_i - \beta_j \right)$$
$$\geq -x_{\text{max}} \frac{h}{4x_{\text{max}}} - x_{\text{max}} \frac{h}{4x_{\text{max}}} + h$$
$$\geq h/2$$

since $A_t$ holds. Thus, Algorithm 1 will use the forced-sample estimator and play arm $i$. □

**Lemma 12.** At time $t$, the event $A_t$ occurs with probability
$$1 - K \left( \exp \left[ -\log t \log d + \log d \right] + \frac{2}{t} \right)$$
when $t \geq (Kq)^2$.

**Proof of Lemma 12** For each $i \in \{1, ..., K\}$, we have from Proposition 1 that
$$\Pr \left[ \|\tilde{\beta}(T_{i,t}, \lambda_1) - \beta_i\|_1 \leq \frac{h}{4x_{\text{max}}} \right] \geq 1 - \exp \left[ -\frac{qh^2 \log t}{16x_{\text{max}}^2} \cdot \frac{p^2 C_1}{16} + \log d \right] - \frac{2}{t}$$
$$\geq 1 - \exp \left[ -\log t \log d + \log d \right] - \frac{2}{t}$$

where the second inequality follows from our definition of $q_0$. Taking a union bound over all $K$ arms gives us the result. By Proposition 2, this happens as long as $t \geq (Kq)^2$. □

**Lemma 13.** Let $S'_{i,t} \subset \{1, ..., t\}$ be the set of all $j$ such that $X_j \in U_i$ and the event $A_t$ holds at time $t$. Then, (1) $X_j \sim \mathcal{P}_X$ i.i.d. for all $j \in S'_{i,t}$, (2) for each $j \in \{1, ..., t\}$, $j \in S'_{i,t}$ with probability at least $p_* / 2$ when $t \geq (Kq)^2$, and (3) $S'_{i,t} \subset S_{i,t}$.
Proof of Lemma 13  For (1), since \( A_i \) is only a function of past samples, we are again doing rejection sampling, so for each \( j \in S'_{i,t} \), \( X_j \) is distributed i.i.d. in \( \mathcal{P}_X \). For (2), we know that \( X \in U_i \) with probability at least \( p_* \), and Lemma 11 implies that \( A_i \) holds with probability at least \( 1 - K \left( \exp \left[ -\log t \log d + \log d \right] + 2/t \right) \) when \( t \geq (Kq)^2 \). Note that \( t \geq (Kq)^2 \geq 16K^2 \) (since \( q \geq [q_0] \geq 4 \)), which implies that \( t \geq 8K \geq 4K + 4 \), so we can bound

\[
\exp \left[ -\log t \log d + \log d \right] \leq \exp \left[ -\log (4K + 4) \log d + \log d \right] \\
\leq \exp \left[ -(\log 4K + \log 4 - 1) \log d \right] \\
\leq \exp \left[ -\log 4K \right] = \frac{1}{4K}
\]

Then, \( A_i \) holds with probability at least

\[
1 - K \left( \exp \left[ -\log t \log d + \log d \right] + \frac{2}{t} \right) \geq 1 - K \left( \frac{1}{4K} + \frac{2}{8K} \right) = \frac{1}{2}
\]

Then, \( j \in S'_{i,t} \) with probability at least \( p_*/2 \). Finally, for (3), from Lemma 10, we know that for \( X_j \sim \mathcal{P}_X \), if \( X \in U_i \) and event \( A_i \) holds, then \( j \in S_{i,t} \), so \( S'_{i,t} \subseteq S_{i,t} \). □

A.5.1. Proof of Proposition 2  Recall that Proposition 2 provides an oracle inequality for the all-sample estimator \( \hat{\beta}(S_{i,t}, \lambda_2,t) \).

Proof of Proposition 2  Lemma 12 states that at time \( t \geq (Kq)^2 \), each of \( \{1,...,t\} \) belongs to \( S'_{i,t} \) with probability at least \( p_*/2 \). Applying Lemma 7 with \( c = p_*/4 \) implies that \( |S'_{i,t}| \geq p_* t/4 \) with probability \( 1 - \exp[-p_*^2 t/32] \). In this event, we can bound \( |S_{i,t}| \geq |S'_{i,t}| \geq t/4 \). Applying Theorem 2 with \( p = p_*/2 \) and \( n = p_* t/4 \), we get

\[
\Pr \left[ \| \hat{\beta}(S_{i,t},\lambda) - \beta_i \|_1 \leq \chi \right] \geq 1 - \exp \left[ -t \chi^2 \frac{p_*^2 C_1}{256} + \log d \right]
\]

when \( \lambda = \frac{\chi p_* t}{42}\) with probability at least

\[
1 - 2 \exp \left[ -\frac{p_*^2 (C_2 \wedge p_*^2/2)}{64} t \right] \leq 1 - 3 \exp \left[ -\frac{p_*^2 (C_2 \wedge p_*^2/2)}{64} t \right]
\]

where we have used a union bound. There is an assumption from Theorem 2 that \( |S_{i,t}| \geq \frac{p_* t}{4} \geq \frac{16 e_{max} \log d}{p_*^2 C_2 t} \), but this is satisfied by the assumption that \( t \geq (Kq)^2 \) by our definition of \( q \). Taking

\[
\chi = 16 \sqrt{\frac{\log t + \log d}{p_*^2 C_1 t}}
\]

gives us the desired result. □

B. Bounding the Regret

At each time \( t \geq (Kq)^2 \), Algorithm 1 either (1) performs a forced sample \( t \in T_i \) for some \( i \), or (2) uses the forced-sample estimator, or (3) uses the all-sample estimator. Moreover, our oracle inequalities require \( t \geq (Kq)^2 \) so we will add a separate term for the regret accrued from this “initialization”. We will separately bound the expected cumulative regret in each of these cases, and sum them to obtain a bound on the total expected cumulative regret \( R_T \).
B.1. Regret from Forced Sampling & Initialization

**Lemma 14.** The cumulative expected regret from forced sampling and initialization at time $T$ is bounded by

$$2qK \alpha \max(6 \log T + Kq)$$

**Proof of Lemma 14** From Lemma 8, we know we have at most $6Kq \log T$ forced samples at time $T$. We also need $(Kq)^2$ initialization samples. Each time, the regret is bounded by $\max_{i,j} X^T (\beta_i - \beta_j) \leq 2\alpha \max$. The result follows directly. □

For the remainder of this section, we consider the case $t > (Kq)^2$.

B.2. Regret from Using Forced-Sample Estimator

**Lemma 15.** If Algorithm 1 uses the forced-sample estimator at time $t$ and $A_t$ holds, then Algorithm 1 plays the optimal arm $i$ and no regret is accrued.

**Proof of Lemma 15** Since arm $i$ was played, we know that $\forall j \neq i$,

$$h/2 \leq X^T_i \left( \hat{\beta}(T_{i,t}) - \hat{\beta}(T_{j,t}) \right)$$

$$= X^T_i \left( \hat{\beta}(T_{i,t}) - \beta_i \right) + X^T_i \left( \beta_j - \hat{\beta}(T_{j,t}) \right) + X^T_i (\beta_i - \beta_j)$$

$$\leq \alpha \max \left( \frac{h}{4\alpha \max} + \alpha \max \frac{h}{4\alpha \max} + X^T_i (\beta_i - \beta_j) \right)$$

if $A_t$ holds. In particular, we get that

$$X^T_i (\beta_i - \beta_j) \geq 0$$

Thus, arm $i$ is optimal and no regret is accrued. □

**Lemma 16.** If $f$ is a monotone decreasing function on the range $[a - 1, b]$, then

$$\sum_{y=a}^{b} f(y) \leq \int_{a-1}^{b} f(y)dy$$

**Proof of Lemma 16**

$$\sum_{y=a}^{b} f(y) \leq \sum_{y=a}^{b} \int_{y-1}^{y} f(y')dy' = \int_{a-1}^{b} f(y)dy$$

□

**Lemma 17.** The cumulative expected regret from the forced sampling estimator at time $T$ is bounded by

$$2K \log T + 3K$$

**Proof of Lemma 17** From Lemma 14, we only accrue regret at time $t+1$ if $A_t$ does not hold, which occurs with probability at most

$$K \left( \exp \left[ -t \log d + \log d \right] + \frac{2}{t} \right)$$
We first use Lemma 15 to sum
\[
\sum_{t=(Kq)^2}^{T-1} (\exp [-t \log d + \log d]) \leq d \int_{(Kq)^2-1}^{T} t^{-\log d} dt \\
\leq d \int_{c}^{\infty} t^{-\log d} dt \\
\leq \frac{1}{\log d - 1} \\
\leq 3
\]

Next, we sum
\[
\sum_{t=(Kq)^2}^{T-1} \frac{2}{t} \leq \int_{1}^{T} \frac{2}{t} dt \leq 2 \log T
\]
The result follows directly. □

B.3. Regret from Using All-Sample Estimator

We first bound the expected regret when the oracle inequality does hold (Proposition 2). In this section, we simplify our notation by letting
\[
\hat{\beta}_i = \hat{\beta}(S_{i,t}, \lambda_{2,t}) \quad \forall i \in \{1, \ldots, K\}
\]
where we recall that
\[
\lambda_{2,t} := \frac{s_0^2}{2s_0} \sqrt{\log t + \log d} p \bar{C}_1 t
\]

**Lemma 18.** If Algorithm 1 uses the all-sample estimator and the oracle inequality holds, then the expected regret at time \( t + 1 \) is bounded by
\[
\frac{2Kb_{x_{\max}}}{t} + C_3 \cdot \frac{\log t + \log d}{t}
\]
where we define
\[
C_3 := \frac{1024KC_0b_{x_{\max}}^2}{p^2\bar{C}_1}
\]

**Proof of Lemma 18** Without loss of generality, assume that arm 1 is optimal, i.e.
\[
1 = \arg \max_{i \in \{1, \ldots, K\}} X^T \beta_i
\]
Then, the expected regret at time \( t + 1 \) is given by
\[
r_{t+1} = \sum_{i=2}^{K} \Pr[\text{choose arm } i] \cdot \mathbb{E}[X^T \beta_1 - X^T \beta_i | \text{choose arm } i]
\]
\[
= \sum_{i=2}^{K} \int \mathbb{E} [X^T (\beta_1 - \beta_i)] \cdot \mathbb{E} \left[ \text{arg max}_{j \in \{1, \ldots, K\}} X^T \hat{\beta}_j = \hat{\beta}_i \right] dp_{\beta_1, \ldots, \beta_K}
\]
\[
\leq \sum_{i=2}^{K} \int \mathbb{E} [X^T (\beta_1 - \beta_i)] \cdot \mathbb{E} \left[ X^T \hat{\beta}_1 > X^T \hat{\beta}_i \right] dp_{\beta_1, \ldots, \beta_K}
\]
\[
= \sum_{i=2}^{K} \Pr \left[ X^T \hat{\beta}_1 > X^T \hat{\beta}_i \right] \cdot \mathbb{E} \left[ X^T (\beta_1 - \beta_i) \right] X^T \hat{\beta}_1 > X^T \hat{\beta}_1
\]
where we have let \( p_{\beta_1, \ldots, \beta_K} \) denote the joint probability density of the estimators \( \hat{\beta}_1, \ldots, \hat{\beta}_K \) at time \( t + 1 \).
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
\( \mathbf{X}^T \hat{\beta}_i > \mathbf{X}^T \hat{\beta}_1 \), and that \( \mathbb{E} [\mathbf{X}^T (\beta_1 - \beta_i)] \geq 0 \) (since we have assumed that arm 1 is optimal). Thus, we can bound \( r_{t+1} \) through the regret incurred by each arm with respect to the optimal arm independently of the other arms.

We now define the event

\[
B_i = \{ \mathbf{X}^T (\beta_1 - \beta_i) > 2\delta x_{\max} \}
\]

where we take

\[
\delta := 16 \sqrt{\frac{\log t + \log d}{p^2 C_1 t}}
\]

Then, we can write

\[
r_{t+1} \leq \sum_{i=2}^{K} \Pr [B_i] \Pr [X^T \hat{\beta}_i > X^T \hat{\beta}_1 | B_i] \cdot \max_i X^T (\beta_1 - \beta_i)
\]

\[
+ (1 - \Pr [B_i]) \Pr [X^T \hat{\beta}_i > X^T \hat{\beta}_1 | \neg B_i] \cdot \max_i X^T (\beta_1 - \beta_i) | \neg B_i]
\]

\[
\leq \sum_{i=2}^{K} \Pr [X^T \hat{\beta}_i > X^T \hat{\beta}_1 | B_i] \cdot b x_{\max} + (1 - \Pr [B_i]) \cdot 2\delta x_{\max}
\]

by the definition of \( B_i \).

Note that choosing arm \( i \neq 1 \) in the event of \( B_i \) implies that

\[
0 > X^T \hat{\beta}_1 - X^T \hat{\beta}_i = (X^T \hat{\beta}_1 - X^T \beta_1) + (X^T \hat{\beta}_i - X^T \hat{\beta}_1) + (X^T \hat{\beta}_i - X^T \beta_i)
\]

\[
\geq (X^T (\hat{\beta}_1 - \beta_1)) + (X^T (\hat{\beta}_i - \beta_1)) + 2x_{\max} \delta
\]

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

Thus, we can write

\[
\Pr [X^T \hat{\beta}_i > X^T \hat{\beta}_1 | B_i] \leq \Pr [X^T (\hat{\beta}_1 - \beta_1) > x_{\max} \delta] + \Pr [X^T (\hat{\beta}_i - \beta_1) > x_{\max} \delta]
\]

\[
\leq \Pr [||\beta_1 - \hat{\beta}_1||_1 > \delta] + \Pr [||\beta_i - \hat{\beta}_i|| > \delta]
\]

\[
\leq \frac{2}{t}
\]

using a union bound and the oracle inequality.

We can also bound \( \Pr [B_i] \) using our assumption on the margin condition

\[
\Pr [X^T \beta - X^T \beta' \leq \kappa] \leq C_0 \kappa \quad \forall \rho \in (0, \kappa_0]
\]

for any allowable choices of \( X, \beta, \beta' \). Without loss of generality, let \( C_0 \geq 1/\kappa_0 \) (if not, we can re-define \( C_0 \) to be \( \max (C_0, 1/\kappa_0) \) and obtain a looser bound). Then, the condition holds trivially for all \( \kappa > \kappa_0 \) as well, and we can write

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

\[
\leq 1 - 2C_0 x_{\max} \delta
\]
Now we obtain
\[ r_{t+1} \leq \sum_{i=2}^{K} 2b x_{\text{max}} t + 4C_3 \delta^2 x_{\text{max}}^2 \]
\[ \leq 2K b x_{\text{max}} + C_3 \cdot \frac{\log t + \log d}{t} \]
where \( C_3 \) is as defined above. \( \square \)

**Lemma 19.** The cumulative expected regret from the all-sample estimator at time \( T \) is bounded by
\[ (2K b x_{\text{max}} + C_3 \log d) \log T + \frac{C_3}{2} (\log T)^2 + C_4 \]
\[ C_4 := \frac{3}{1 - \exp\left[-\frac{p^2 (C_2 \wedge p_*)}{128}\right]} \]

**Proof of Lemma 19** We first sum the regret from when the oracle inequality does not hold. First, note that
\[ \sum_{t=(Kq)^2}^{T-1} \frac{1}{t} \leq \int_{(Kq)^2-1}^{T-1} \frac{dt}{t} \leq \log T \]
from Lemma 15. Similarly,
\[ \sum_{t=(Kq)^2}^{T-1} \frac{\log t}{t} \leq \int_{(Kq)^2-1}^{T-1} \frac{\log t \cdot dt}{t} \leq \frac{1}{2} (\log T)^2 \]
Using these inequalities, we can sum the regret from Lemma 17
\[ \sum_{t=(Kq)^2}^{T-1} \frac{2K b x_{\text{max}}}{t} + C_3 \cdot \frac{\log t + \log d}{t} \leq (2K b x_{\text{max}} + C_3 \log d) \log T + \frac{C_3}{2} (\log T)^2 \]
On the other hand, from Proposition 2, the oracle inequality does not hold with probability
\[ 3 \exp\left[-\frac{p^2 (C_2 \wedge p_*)}{128}\right] \leq \sum_{t=0}^{\infty} \frac{3}{1 - \exp\left[-\frac{p^2 (C_2 \wedge p_*)}{128}\right]} \]
and so we accumulate cumulative regret
\[ \sum_{t=(Kq)^2}^{T-1} 3 \exp\left[-\frac{p^2 (C_2 \wedge p_*)}{128}\right] \leq 3 \exp\left[-\frac{p^2 (C_2 \wedge p_*)}{128}\right] \]
\[ \leq \frac{3}{1 - \exp\left[-\frac{p^2 (C_2 \wedge p_*)}{128}\right]} \]
\[ \square \]

**B.4. Proof of Theorem 1**
We sum all the terms from Lemmas 13, 16, and 18 to get
\[ 2q K b x_{\text{max}} (6 \log T + K q) + 2K \log T + 3K + (2K b x_{\text{max}} + C_3 \log d) \log T + \frac{C_3}{2} (\log T)^2 + C_4 \]
\[ = \frac{C_3}{2} (\log T)^2 + (2K b x_{\text{max}} (6q + 1) + 2K + C_3 \log d) \log T + (2q^2 K^2 b x_{\text{max}} + 3K + C_4) \]
References


