Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Online Decision Making with High-Dimensional Covariates

Hamsa Bastani, Mohsen Bayati

To cite this article:
Hamsa Bastani, Mohsen Bayati (2019) Online Decision Making with High-Dimensional Covariates. Operations Research

Published online in Articles in Advance 07 Nov 2019

. https://doi.org/10.1287/opre.2019.1902

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–19
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Online Decision Making with High-Dimensional Covariates

Hamsa Bastani,[a] **Mohsen Bayati**[b]

[a] Wharton School, Operations Information and Decisions, University of Pennsylvania, Philadelphia, Pennsylvania 19104;
[b] Stanford Graduate School of Business, Stanford University, Stanford, California 94305

Contact: hamsab@wharton.upenn.edu, https://orcid.org/0000-0002-8793-4732 (HB); bayati@stanford.edu, https://orcid.org/0000-0002-7280-912X (MB)

Received: March 31, 2017
Revised: December 9, 2017; March 20, 2019
Accepted: June 6, 2019
Published Online in Articles in Advance: November 7, 2019

Subject Classifications: statistics; simulation
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2019.1902

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

History: First Place Winner (Hamsa Bastani), George Nicholson Student Paper Competition Award, 2016.
Funding: The authors gratefully acknowledge the National Science Foundation [Graduate Research Fellowship Grant DGE-114747, NSF EAGER award CMMI:1451037, NSF CAREER award CMMI: 1554140, and NSF Grant CCF:1216011] and the Stanford Cyber Security Initiative for financial support.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2019.1902.

Keywords: contextual bandits • adaptive treatment allocation • online learning • high-dimensional statistics • LASSO • personalized decision making

1. Introduction

The growing availability of user-specific data provides a unique opportunity for decision makers to *personalize* service decisions for individuals. In healthcare, doctors can personalize treatment choices based on patient biomarkers and clinical history. For example, the BATTLE trial demonstrated that the effectiveness of different chemotherapeutic agents on a cancer patient depends on the molecular biomarkers found in the patient's tumor biopsy; thus, personalizing the chemotherapy regimen led to increased treatment success rates (Kim et al. 2011). Similarly, in marketing, companies may achieve greater conversion rates by targeting advertisements or promotions based on user demographics and search key words. Personalization is typically achieved by (a) learning a model that predicts a user's outcome for each available decision as a function of the user's observed covariates and (b) using this model to inform the chosen decision for subsequent new users (see, e.g., He et al. 2012, Bertsimas and Kallus 2014, Chen et al. 2015, Ban and Rudin 2019).

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 laboratory tests, diagnoses, procedures, and medications) or genetic or molecular biomarker profiles. The resulting number of covariates in medical decision-making problems can be as many as a few thousand (in Bayati et al. 2014) or tens of thousands (in Razavian et al. 2015). Similarly, user covariates in web marketing are often high dimensional, because 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 (Tibshirani 1996, Chen et al. 1998), help alleviate this issue by producing good estimates using far fewer user samples than traditional statistical models (Candes and Tao 2007, Bickel et al. 2009, Bühlmann and Van De Geer 2011). In particular, the LASSO identifies a *sparse* subset of predictive covariates, which is an effective approach for treatment effect estimation in practice (Belloni et al. 2014, Athey et al. 2016). For example, the BATTLE cancer trial found that only a few of many available patient biomarkers were predictive of the success of any given treatment (Kim et al. 2011). Similarly, variable selection is often used to predict Internet users' click-through rates in online advertising (see, e.g., Yan et al. 2014).

However, we must be careful not to sacrifice asymptotic performance when using such techniques. They create substantial bias in our estimates to increase predictive accuracy for small sample sizes. Thus, it is valuable to incorporate new observations and carefully tune the bias-variance trade-off over time to ensure good performance for both initial users (data-poor regime) and later users (data-rich regime). This can be performed *online*: after making a decision, we learn from the resulting reward, for example, how well a treatment worked for a patient or the profit from an advertisement. This process suffers from *bandit feedback*; that is, 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, because 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 trade-off has been studied in the framework of contextual bandits (Auer 2003, Langford and Zhang 2008). Although many algorithms have been proposed and analyzed in the literature, they typically optimize asymptotic performance (when the number of users T grows large) and may not perform well in the data-poor regime. In particular, the performance of all existing algorithms scales polynomially in the number of covariates d, and provide no theoretical guarantees when the number of users T is of order d (see, e.g., Goldenshluger and Zeevi 2013), even when the underlying model is known to be sparse (Abbasi-Yadkori et al. 2012). Thus, such algorithms may essentially randomize on the initial $\mathcal{O}(d)$ individuals, which as discussed earlier, may be prohibitively costly in high-dimensional settings.

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

1.1. Main Contributions

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

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

1.1.2. Theory. We measure performance using the standard notion of *expected cumulative regret*, which is the total expected deficit in reward achieved by our algorithm compared with an oracle that knows all the problem parameters. Our main result establishes that the LASSO Bandit asymptotically achieves expected cumulative regret that scales polylogarithmically with the dimension of the covariates. The technical

challenge is that the bandit policy induces non-i.i.d. samples from each arm during the exploitation phase. In particular, even though the sequence of all covariates consists of i.i.d. samples from a fixed distribution, the subset of covariates for which the outcome of a fixed arm is observed may not be i.i.d. In low-dimensional settings, this is typically addressed using martingale matrix Chernoff inequalities (Tropp 2015). We prove analogous results in the high-dimensional setting for the convergence of the LASSO estimator using matrix perturbation theory and martingale concentration results. In particular, we prove a new tail inequality for the LASSO (that may be of independent interest) that holds with high probability even when an unknown portion of the samples are generated by a non-i.i.d. process.

We further derive an optimal specification for the LASSO regularization parameters, and prove that the resulting cumulative regret of the LASSO Bandit over T users is at most $\mathbb{O}(s_0^2[\log T + \log d]^2)$, 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 polylogarithmically 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 also can 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.

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

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

1.2. Related Literature

As discussed earlier, there is a significant OR/MS literature on learning predictive models from historical data and using such models to inform context-specific decision making (e.g., Bertsimas and Kallus 2014, Ban and Rudin 2019). 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 trade-off in the contextual bandit framework (also known as contextual bandits or linear bandits with changing action space) from OR/MS, computer science, and statistics. One approach is to make no parametric assumptions about arm rewards. For example, Rigollet and Zeevi (2010), Perchet and Rigollet (2013), and Slivkins (2014) analyze settings in which the arm rewards are given by any smooth, nonparametric function of the observed covariates. However, these algorithms perform very poorly in high dimension as the cumulative regret depends exponentially on the covariate dimension d.

Thus, much of the bandit literature (including the present paper) has focused on the case in which 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), Abbasi-Yadkori et al. (2011), Chu et al. (2011), and Deshpande and Montanari (2012). (Note that some of these papers study the linear bandit, which is different from a contextual bandit; however, the theoretical guarantees of a linear bandit can be mapped to theoretical guarantees for a contextual bandit *if* the feasible action set for the linear bandit is allowed to change exogenously over time (Abbasi-Yadkori 2012).) These algorithms use the idea of optimism-in-the-face-of-uncertainty (OFU), which elegantly solves the exploration-exploitation trade-off by maintaining confidence sets for arm parameter estimates and choosing arms optimistically from within these confidence sets. Follow-up work demonstrated that similar guarantees can be achieved using a posterior sampling algorithm (Agrawal and Goyal 2013, Russo and Van Roy 2014b). We also note that Carpentier and Munos (2012) tackle a linear bandit in the high-dimensional sparse setting, but

they use a nonstandard definition of regret and do not consider the relevant case in which the action set changes over time.

However, this literature typically does not make any assumptions on how the user covariates X_t are generated. In particular, they allow for arbitrarily constructed covariate sequences that may be generated by an adversary to make learning difficult; chapter 3 of Bubeck and Cesa-Bianchi (2012) provides a detailed survey of "adversarial bandits." For example, if X_t is equal to a fixed vector X that does not change over time, it is impossible to learn more than one parameter per arm. This may explain why the current-best cumulative regret bounds are given by: $\mathbb{O}(d\sqrt{T})$ in the low-dimensional setting (Dani et al. 2008, Abbasi-Yadkori et al. 2011) and $\mathbb{O}(\sqrt{ds_0T})$ in the high-dimensional sparse setting (Abbasi-Yadkori et al. 2012). Note that such algorithms still achieve regret that is polynomial in d and T, implying slow rates of convergence. In particular, when $T = \mathbb{O}(d)$ (the regime of interest here), these regret bounds are no longer sublinear in T.

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

Yet assuming covariate sequences can be selected completely arbitrarily constitutes a pessimistic environment that is unlikely to occur in practical settings. For example, in healthcare, the treatment choices made for one patient do not directly affect the health status of the next patient, suggesting that covariates are roughly i.i.d. Thus, we focus on the case in which 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 $\mathbb{O}(d^3 \log T)$ on cumulative regret, which significantly improves the $\mathbb{O}(d\sqrt{T})$ bound for arbitrary covariate sequences as T grows large. We adapt this idea to the high-dimensional setting using LASSO estimators. However, we require a much tighter regret analysis as well as new convergence results on LASSO estimators, which we use to prove a regret bound of $\mathbb{O}(s_0^2[\log T + \log d]^2)$. Note that we relax the polynomial dependence on d to a polylogarithmic factor by leveraging sparsity. As a consequence of our new proof technique, we also improve the regret bound in the low-dimensional setting from $\mathbb{O}(d^3 \log T)$ (Goldenshluger and Zeevi 2013) to $\mathbb{O}(d^2 \log^{\frac{3}{2}} d \cdot \log T)$. These results hold while allowing for some arms to be uniformly suboptimal; in contrast, the formulation in Goldenshluger and Zeevi (2013) requires the assumption that every arm is optimal for some subset of users.

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

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

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

Finally, our paper is also related to recent papers in the operations management literature at the intersection of machine learning and multiarmed bandits. Kallus and Udell (2016) use low-rank matrix completion for dynamic assortment optimization with a large number of customers, and Elmachtoub et al. (2017) introduce a novel bootstrap-inspired method for performing Thompson sampling using decision

trees. In contrast, our work focuses on developing provable guarantees for bandits with covariates under the LASSO estimator; to that end, we introduce new theoretical results for the LASSO with adapted sequences of (possibly non-i.i.d.) observations.

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

2. Problem Formulation

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

2.1. Notation

For any integer n, we will let $[n]$ denote the set $\{1,\dots,n\}$. For any index set $I \subset [d]$, and a vector $\beta \in \mathbb{R}^d$, let $\beta_I \in \mathbb{R}^d$ be the vector obtained by setting the elements of β that are not in I to zero. For a vector $v \in \mathbb{R}^m$, let the support of v (denoted $\text{supp}(v)$) to be the set of indices corresponding to nonzero entries of v. For any vector X or matrix $\mathbf{X}$, the infinity norm (i.e., $\|\cdot\|_\infty$) is the maximum absolute value of its entries. We also use $\mathbb{R}^+$ and $\mathbb{Z}^+$ to refer to positive reals and integers respectively, and use $\mathbb{R}^{d\times d}_{\geq 0}$ for the set of d by d positive semidefinite matrices.

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

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

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

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

This definition implies $\mathbb{E}[z] = 0$ 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.

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

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

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

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

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

2.2. Assumptions

We now describe the assumptions we require on the problem parameters for our regret analysis. These assumptions are adapted from the bandit literature and will be attributed in the text below. For simplicity,

we introduce a specific example and show how each assumption translates to the example. Later, we describe more generic examples that are encompassed by our formulation.

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

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

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

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

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

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

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

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

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

a. Suboptimal arms $\mathcal{K}_{sub}$ that are *strictly* suboptimal for all covariate vectors in $\mathcal{X}$, that is, there exists a constant $h_{sub} > 0$ such that for each $i \in \mathcal{K}_{sub}$, $x^\top\beta_i < \max_{j\ne i} x^\top\beta_j - h_{sub}$ for every $x \in \mathcal{X}$.

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

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

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

Remark 4. We emphasize that Assumption 3 differs from the "gap" assumption made in problem-dependent bounds in the bandit literature (see, e.g., Abbasi-Yadkori et al. 2011), which assumes that there exists some gap $\Delta > 0$ between the rewards of the optimal arm i_* and the next best arm, that is, $\Delta \le \min_{j, x\in\mathcal{X}} x^\top(\beta_{i^*} - \beta_j)$. In a general contextual bandit, no $\Delta > 0$ satisfies the gap assumption, because the user covariate vector X can be drawn arbitrarily close to the decision boundary for some β_k (i.e., arbitrarily close to the set $\{x \in \mathcal{X} \mid x^\top\beta_{i^*} = x^\top\beta_k\}$). Rather, Assumption 3 posits that such a gap exists ($\Delta = h$) only with some probability $p_* > 0$. While the "gap" assumption

does not hold for most covariate distributions (e.g., uniform), our assumption holds for a very wide class of continuous and discrete covariate distributions (as we will discuss below).

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

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

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

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

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

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

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

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

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

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

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

3. LASSO Bandit Algorithm

We begin by providing some brief intuition about the LASSO Bandit algorithm. Our policy produces LASSO estimates $\hat{\beta}_i$ for the parameter of each arm $i \in [K]$ based on past samples X_t where arm i was played. A typical approach for addressing the exploration-exploitation trade-off 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 probably incur at least $\Omega(\sqrt{T})$ regret, because we will require many forced samples for the estimates to converge fast enough.

Instead, our estimates may converge faster if we use *all* past samples (including non-i.i.d. samples from myopic play) from arm i to estimate β_i. However, because these samples are not i.i.d., standard convergence guarantees for LASSO estimators do not apply, and we cannot ensure that the estimated parameters $\hat{\beta}_i$ converge to the true parameters β_i. We tackle this by adapting an idea from the low-dimensional bandit algorithm by Goldenshluger and Zeevi (2013), that is, maintaining two sets of estimators for each arm: (1) *forced sampling estimates* trained on only forced samples and (2) *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 preprocessing 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 preprocessing step guarantees convergence of the all-sample estimator. A key novel ingredient of our algorithm is specifying the regularization paths to control the convergence of our LASSO estimators by carefully trading off bias and variance over time. Intuitively, we build low-dimensional linear models in the data-poor regime by limiting the number of allowed covariates; this allows us to make reasonably good decisions even with limited data. As we collect more data, we allow for increasingly complex models (consisting of more covariates), eventually recovering the standard OLS model.

3.1. Additional Notation

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

3.2. LASSO Estimation

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

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

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

The LASSO estimator satisfies the following *tail inequality*.

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

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

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

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

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

proof of the standard LASSO oracle inequality[2] using martingale theory and is provided in Section EC.1 in the online appendix.

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

3.3. Description of Algorithm

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

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

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

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

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

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

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

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

Algorithm (LASSO Bandit)

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

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

3.4. Main Result: Regret Analysis of LASSO Bandit

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

Theorem 1. *When* $q \ge 4\lceil q_0 \rceil$, $K \ge 2$, $d > 2, t \ge C_5$, *and we take* $\lambda_1 = (\phi_0^2 p_* h)/(64 s_0 x_{\max})$ *and* $\lambda_{2,0} = [\phi_0^2/(2s_0)]\sqrt{1/(p_* C_1)}$, *we have the following (nonasymptotic)*

upper bound on the expected cumulative regret of the LASSO Bandit at time T by

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

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

$$\begin{aligned} C_1(\phi_0) &\equiv \frac{\phi_0^4}{512 s_0^2 \sigma^2 x_{\max}^2}, \\ C_2(\phi_0) &\equiv \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 x_{\max}^2}\right), \\ C_3(\phi_0, p_*) &\equiv \frac{1024 K C_0 x_{\max}^2}{p_*^3 C_1}, \\ C_4(\phi_0, p_*) &\equiv \frac{8Kbx_{\max}}{1 - \exp\left[-\frac{p_*^2 C_2^2}{32}\right]}, \\ C_5 &\equiv \min\left\{t \in \mathbb{Z}^+ | t \ge 24Kq\log t + 4(Kq)^2\right\}, \end{aligned}$$

and we take

$$\begin{aligned} q_0 &\equiv \max\left\{\frac{20}{p_*}, \frac{4}{p_* C_2^2}, \frac{12\log d}{p_* C_2^2}, \frac{1024 x_{\max}^2 \log d}{h^2 p_*^2 C_1}\right\} \\ &= \mathbb{O}\left(s_0^2 \log d\right). \end{aligned}$$

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

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

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

4. Key Steps of the Analysis of LASSO Bandit

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

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

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

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

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

Then, for any $\lambda \geq 0$ we have a LASSO estimator trained on samples in $\mathcal{A}$:

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

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

Lemma 1. *For any $\chi > 0$, if $d > 1$, $|\mathcal{A}'|/|\mathcal{A}| \geq p/2$, $|\mathcal{A}| \geq 6 \log d/(p\, C_2(\phi_1)^2)$, and $\lambda = \lambda(\chi, \phi_1\sqrt{p}/2) = \chi\, \phi_1{}^2\, p/(16 s_0)$, then the following tail inequality holds:*

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

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

4.2. LASSO Tail Inequality for the Forced Sample Estimator

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

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

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

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

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

4.3. LASSO Tail Inequality for the All-Sample Estimator

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.5. Proof of the Main Result

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

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

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

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

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

5. Empirical Results

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

5.1. Synthetic Data

We evaluate the LASSO Bandit on a synthetically generated data set to address two questions: (1) how does the LASSO Bandit's performance compare against existing algorithms empirically, and (2) is the LASSO Bandit robust to the choice of input parameters?

We compare the LASSO Bandit against (a) 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), (b) a sparse variant OFUL-EG for high-dimensional settings (Abbasi-Yadkori 2012, Abbasi-Yadkori et al. 2012), and (c) the OLS Bandit by Goldenshluger and Zeevi (2013). Our results demonstrate that the LASSO Bandit significantly outperforms these benchmarks. Separately, we find that the LASSO Bandit is robust to changes in input parameters by even an order of magnitude.

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

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

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

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

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

5.1.3. 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.

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

Figure 1(b) considers a larger number of covariates d. As expected, we see that the performance gap between the LASSO Bandit and the other algorithms

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

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

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

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

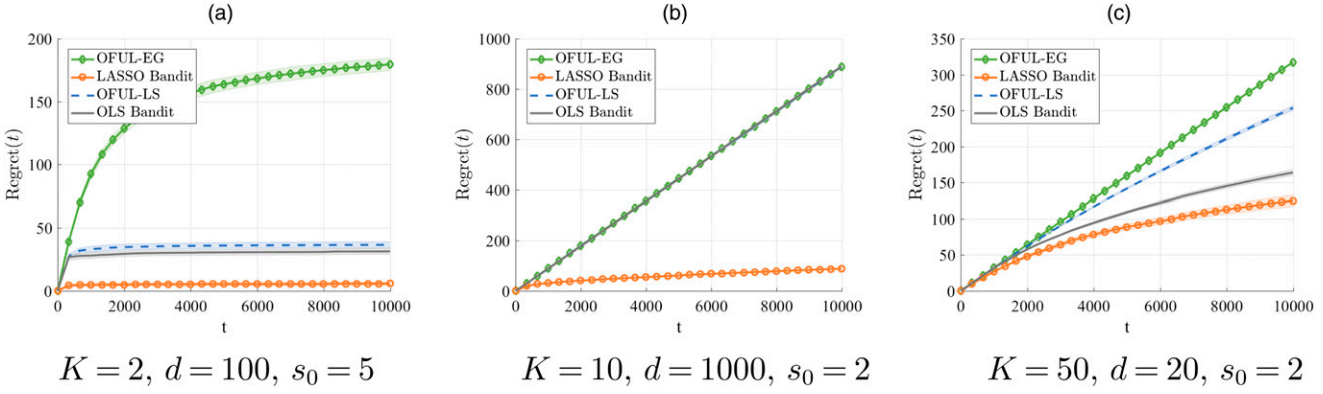
5.2. Case Study: Warfarin Dosing

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

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

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

Figure 1. Comparison of the Cumulative Regret of the LASSO Bandit Against Other Bandit Algorithms on Synthetic Data for Different Values of K, d, and s_0



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

Physicians currently follow a fixed-dose strategy: they start patients on 5 mg/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.

5.2.2. Data Set. We use a publicly available patient data set that was collected by staff at the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) for 5,700 patients who were treated with warfarin from 21 research groups spanning 9 countries and 4 continents. Importantly, these data contain 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 5,528 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 (International Warfarin Pharmacogenetics Consortium et al. 2009). These covariates include the following:

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

Details on the data set can be found in supplementary appendix 1 of International Warfarin Pharmacogenetics Consortium et al. (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 International Warfarin Pharmacogenetics Consortium et al. (2009) report that an ordinary least squares linear model fits the data best (i.e., achieves the best cross-validation accuracy) compared with alternative models (such as support vector regression, regression trees, model trees, multivariate adaptive regression splines, least-angle regression, LASSO) for the objective of predicting the correct warfarin dosage as a function of the given patient-level variables. The results of International Warfarin Pharmacogenetics Consortium et al. (2009) suggest no underlying sparsity in these data. Thus, one might expect low-dimensional algorithms like the OLS Bandit or OFUL-LS to perform no worse than the LASSO Bandit; surprisingly, we find that this is not the case in the online setting.

5.2.3. 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 International Warfarin Pharmacogenetics Consortium et al. (2009): (1) low: under 3 mg/day (33% of cases), (2) medium: 3–7 mg/day (54% of cases), and (3) high: over 7 mg/day (13% of cases). In particular, patients who require a low (high) dose would be at risk for excessive (inadequate) anticoagulation under the physician's medium starting dose. We estimate the true arm parameters β_i using linear regressions on the entire data set.

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

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

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

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

1. **LASSO Bandit**, described in Section 3 of this paper,
2. **OLS Bandit**, described in Goldenshluger and Zeevi (2013),

3. **OFUL-LS**, described in Abbasi-Yadkori et al. (2011),
4. **OFUL-EG**, described in Abbasi-Yadkori et al. (2012),[5]
5. **Doctors**, who currently always assign an initial medium dose (International Warfarin Pharmacogenetics Consortium et al. 2009), and
6. **Oracle**, which assigns the optimal estimated dose given the true arm parameters β_i.

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

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

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

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

***Moderate Data*:** When there are a moderate number of samples (200–1,500 patients), the LASSO Bandit effectively learns the arm parameters and outperforms the doctor's policy; however, the remaining bandit algorithms still perform worse than physicians.

***Big Data*:** For a large number of samples (1,500–5,000 patients), both the LASSO and OLS bandit policies outperform the physician's policy and begin to look comparable. However, the OFUL-LS and OFUL-EG algorithms still perform worse than do doctors.

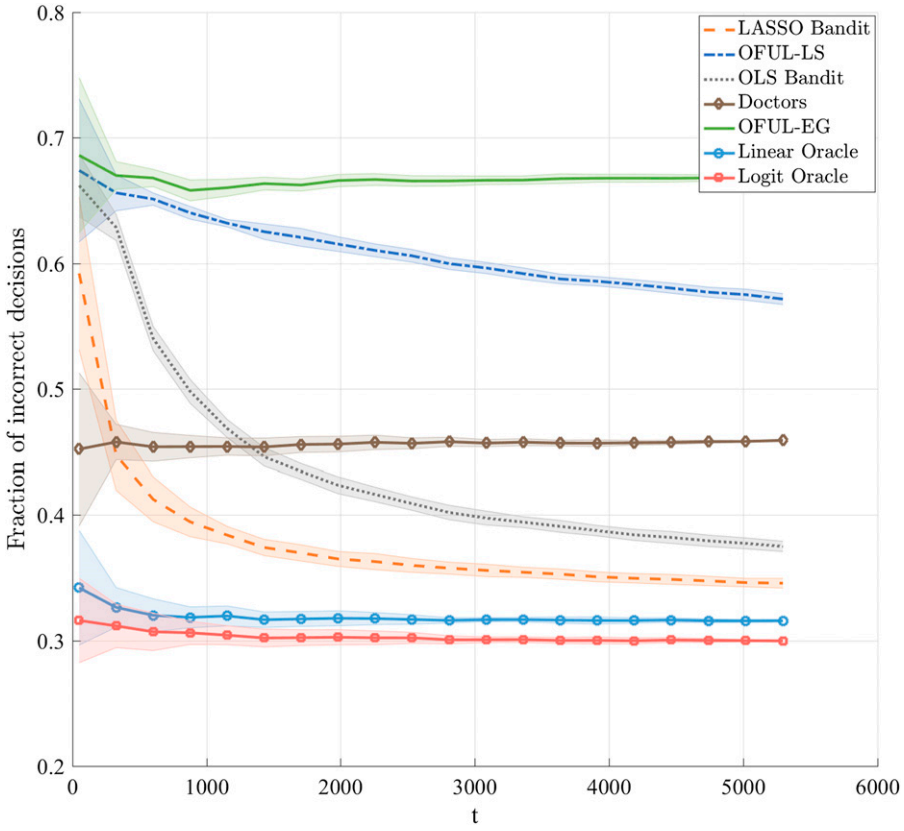
Note that all three existing bandit algorithms required more than 1,500 patient samples before outperforming the doctor's static policy; this may be prohibitively costly in a healthcare setting and may hinder adoption of learning strategies in practice. In contrast, we see that the LASSO Bandit starts 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 data set better than a LASSO model, it may be more effective to use the LASSO Bandit in an online setting in order to more efficiently use information as it is collected. In particular, the LASSO Bandit uses regularization to first build simple predictive models (with few covariates), and gradually builds more complex predictive models (by including more covariates over time); this helps us make reasonable decisions in the small-data regime without sacrificing performance in the big-data regime.

5.2.6. Risk Implications. One concern that arose in conversations with clinicians is that although the LASSO Bandit policy achieves a higher dosing accuracy overall (compared with doctors), it may assign a "significantly worse" dose to some patients. In particular, the bandit algorithm may assign a low dose to a patient whose true dose is high (or vice versa); on the other hand, the doctor always hedges her bet by assigning the medium dose.

To better illustrate the risk consequences, we tabulate the assigned versus 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 anticoagulation) although they account for a third of the patient population. This trade-off can be explored further by adjusting the reward function; in particular, we have used a binary loss for misdosing, but the loss can be a function of the magnitude of misdosing.

Remark 10. Several simplifying assumptions were made in this case study. For example, warfarin dosing is not a truly bandit problem, because we always observe the optimal arm (patient's true dose) even if we play the wrong arm (assign the wrong dose initially) as the doctor tunes the dosage over time. Yet we use this setting as a case study to evaluate bandit policies because the data contains the true counterfactual outcomes without performing an experiment. For problems with true

Figure 2. Comparison of the Fraction of Incorrectly Dosed Patients Under the Oracle, LASSO Bandit, OLS Bandit, OFUL-LS, OFUL-EG, and Doctor Policies as a Function of the Number of Patients in the Warfarin Data



bandit feedback, we do not observe counterfactual rewards for actions that were not chosen in the data, so we cannot evaluate the counterfactual performance of the LASSO Bandit. In practice, the LASSO Bandit would be most useful for bandit settings where the patient can only receive one treatment and the counterfactual outcomes under other treatments cannot be observed, for example, the problem of choosing chemotherapy agents as described in the introduction (Kim et al. 2011).

6. Conclusions

We present the LASSO Bandit algorithm for contextual bandit problems with high-dimensional covariates, and we prove the first regret bound that grows only polylogarithmically 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 *nonsparse* problems. We illustrate the LASSO Bandit's practical relevance by evaluating it on a medical decision-making problem of warfarin dosing; we find that it surpasses existing bandit methods as well as physicians to correctly dose a majority of patients and thereby improve overall patient outcomes. We note that several simplifying assumptions were made in

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

	LASSO bandit policy assigned dosage (%)			Physician policy assigned dosage (%)			
True dosage	Low	Medium	High	Low	Medium	High	% of patients
Low	57	42	1	0	100	0	33
Medium	14	83	3	0	100	0	54
High	3	90	7	0	100	0	13

Note. Blue numbers represent the fraction of patients who were dosed correctly; red numbers represent the fraction of patients who were dosed incorrectly by two buckets.

this evaluation, and thus, modeling the full complexity of the problem would be a valuable direction to pursue in future work.

6.1. Limitations and Future Directions

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

Acknowledgments

This paper benefitted from valuable feedback from Stephen Chick, Nima Hamidi, Hamid Nazerzadeh, and Stefanos Zenios; anonymous referees; and various seminar participants, all of whom were instrumental in guiding the authors in improving the paper.

Endnotes

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

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

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

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

[5] The original OFUL-EG requires the assumption that $\sum_{i=1}^{k} \|\beta_i\|_1 = 1$ (Abbasi-Yadkori 2012); however, there is no way to guarantee that this holds on a real data set, 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.

References

Abbasi-Yadkori Y (2012) Online learning for linearly parametrized control problems. PhD thesis, University of Alberta, Edmonton, AB, Canada.

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 24 (Curran Associates, Red Hook, NY), 2312–2320.

Abbasi-Yadkori Y, Pal D, Szepesvari C (2012) Online-to-confidence-set conversions and application to sparse stochastic bandits. *Proc. Machine Learn. Res.* 22:1–9.

Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. *Proc. Machine Learn. Res.* 28:127–135.

Athey S, Imbens GW, Wager S (2016) Approximate residual balancing: de-biased inference of average treatment effects in high dimensions. Working paper, Stanford University, Stanford, CA.

Auer P (2003) Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learn. Res.* 3:397–422.

Ban G-Y, Rudin C (2019) The big data newsvendor: practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Bayati M, Braverman M, Gillam M, Mack K, Ruiz G, Smith M, Horvitz E (2014) Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One* 9(10):e109264.

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econom. Stud.* 81(2):608–650.

Bertsimas D, Kallus N (2014) From predictive to prescriptive analytics. Working paper, Massachusetts Institute of Technology, Cambridge.

Bickel P, Ya'acov R, Tsybakov A (2009) Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* 37(4):1705–1732.

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.

Budnitz DS, Pollock DA, Weidenbach KN, Mendelson AB, Schroeder TJ, Annest JL (2006) National surveillance of emergency department visits for outpatient adverse drug events. *JAMA* 296: 1858–1866.

Bühlmann P, Van De Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Springer Science & Business Media, New York).

Candes E, Tao T (2007) The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* 35(6):2313–2351.

Carpentier A, Munos R (2012) Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. *15th International Conference on Artificial Intelligence and Statistics (AISTATS), La Palma, Canary Islands*, 190–198.

Chen, SS, Donoho DL, Saunders MA (1998) Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20(1):33–61.

Chen X, Owen Z, Pixton C, Simchi-Levi D (2015) A statistical learning approach to personalization in revenue management. Working paper, New York University, New York.

Chu W, Li L, Reyzin L, Schapire R (2011) Contextual bandits with linear payoff functions. *Proc. Machine Learn. Res.* 15:208–214.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under Bandit feedback. Servedio RA, Zhang T, eds. *Proc. Conf. Learn. Theory* (Omnipress, Madison, WI), 355–366.

Deshpande Y, Montanari A (2012) Linear bandits in high dimension and recommendation systems. Preprint, submitted January 8, https://arxiv.org/abs/1301.1722.

Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap* (Chapman Hall, New York).

Elmachtoub AN, McNellis R, Oh S, Petrik M (2017) A practical method for solving contextual bandit problems using decision trees. *Proc. 33rd Conf. Uncertainty Artificial Intelligence (UAI), Sydney, Australia*, 11–15.

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

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning* (Springer, New York).

He B, Dexter F, Macario A, Zenios S (2012) The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manufacturing Service Oper. Management* 14(1):99–114.

International Warfarin Pharmacogenetics Consortium, Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT, et al. (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England J. Medicine* 360(8):753–764.

Kallus N, Udell M (2016) Dynamic assortment personalization in high dimensions. Preprint, arXiv:1610.05604.

Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, Stewart DJ, et al. (2011) The battle trial: personalizing therapy for lung cancer. *Cancer Discovery* 1(1):44–53.

Kivinen J, Warmuth MK (1997) Exponentiated gradient vs. gradient descent for linear predictors. *Inform. Comput.* 132(1):1–63.

Langford J, Zhang T (2008) The epoch-greedy algorithm for multi-armed bandits with side information. Platt JC, Koller D, Singer Y, Roweis ST, eds. *Advances in Neural Information Processing Systems*, vol. 20 (Curran Associates, Red Hook, NY), 817–824.

Naik P, Wedel M, Bacon L, Bodapati A, Bradlow E, Kamakura W, Kreulen J, Lenk P, Madigan DM, Montgomery A (2008) Challenges and opportunities in high-dimensional choice data analyses. *Marketing Lett.* 19(3–4):201–213.

Negahban SN, Ravikumar P, Wainwright MJ, Yu B (2012) A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers. *Statist. Sci.* 27(4):538–557.

Perchet V, Rigollet P (2013) The multi-armed bandit problem with covariates. *Ann. Statist.* 41(2):693–721.

Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D (2015) Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* 3(4): 277–287.

Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. Kalai AT, Mohri M, eds. *Proc. Conf. Learn. Theory* (Omnipress, Madison, WI), 54–66.

Rusmevichientong, P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Math. Oper. Res.* 35(2):395–411.

Russo D, Van Roy B (2014a) *Learning to optimize via information-directed sampling*, Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Red Hook, NY), 1583–1591.

Russo D, Van Roy B (2014b) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.

Slivkins A (2014) Contextual bandits with similarity information. *J. Mach. Learn. Res.* 15(1):2533–2568.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. Series B (Methodological)* 58(1): 267–288.

Tropp J (2015) An introduction to matrix concentration inequalities. *Foundations Trends Machine Learn.* 8(1–2):1– 230.

Tsybakov AB (2004) Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* 32(1):135–166.

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

Yan L, Li W-J, Xue G-R, Han D (2014) Coupled group lasso for web-scale CTR prediction in display advertising. *Proc. Machine Learn. Res.* 32(2):802–810.

Hamsa Bastani is an assistant professor of operations, information, and decisions at the Wharton School, University of Pennsylvania. Her research focuses on data-driven decision making and its applications to healthcare, revenue management, and social good.

Mohsen Bayati is an associate professor of operations, information, and technology at the Graduate School of Business, Stanford University. His research interests are message-passing algorithms, network models, and personalized decision making with healthcare applications.