# Adaptive Algorithm for Multi-armed Bandit Problem with High-dimensional Covariates

Wei Qian[a], Ching-Kang Ing[b], and Ji Liu[c]

[a]Department of Applied Economics and Statistics, University of Delaware, Newark, DE
[b]Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan
[c]Meta Platforms, Inc., Seattle, WA

**Abstract**

This paper studies an important sequential decision making problem known as the multi-armed stochastic bandit problem with covariates. Under a linear bandit framework with high-dimensional covariates, we propose a general multi-stage arm allocation algorithm that integrates both arm elimination and randomized assignment strategies. By employing a class of high-dimensional regression methods for coefficient estimation, the proposed algorithm is shown to have near optimal finite-time regret performance under a new study scope that requires neither a margin condition nor a reward gap condition for competitive arms. Based on the synergistically verified benefit of the margin, our algorithm exhibits adaptive performance that automatically adapts to the margin and gap conditions, and attains optimal regret rates simultaneously for both study scopes, without or with the margin, up to a logarithmic factor. Besides the desirable regret performance, the proposed algorithm simultaneously generates useful coefficient estimation output for competitive arms and is shown to achieve both estimation consistency and variable selection consistency. Promising empirical performance is demonstrated through extensive simulation and two real data evaluation examples.

**Key Words:** contextual bandits, exploration-exploitation tradeoff, high-dimensional regression model, sequential decision making, stepwise regression procedure

## 1. Introduction

Sequential decision making problems are commonly encountered optimization tasks with important modern applications. For example, in medical service, a physician must decide the appropriate dose level for prescriptions, with the hope of maximizing patients' well-being and preventing adverse effects; in online service, a news website must recommend "top" news articles from multiple candidate news articles to upcoming website visitors to attract more readings; in financial service, a lending firm seeks to decide whether and under what terms they should approve upcoming applicants' loan requests and to reduce overall default rates. These decision

making problems can be formulated as the multi-armed stochastic bandit problem: at each user visit, an agent must choose one of the candidate decision arms (e.g., news articles) and then observe a reward (e.g., 1 for reading and 0 for non-reading) from the chosen arm, where the reward follows some unknown distribution; the primary target is to maximize the overall reward over a certain number of visits.

The classic settings (Robbins, 1954; Lai and Robbins, 1985; Berry and Fristedt, 1985; Lai, 1987; Gittins, 1989; Auer et al., 2002) typically assume that the reward distribution of each arm is homogeneous. See, e.g., Bubeck and Cesa-Bianchi (2012), Lattimore and Szepesvári (2020), Chan (2020), and references therein for a recent overview on algorithm efficiencies under related settings. In many real applications, we have access to extra covariate information from users of the service, which holds promise for personalized service. In personalized medical service, for example, the treatment effect can be dependent on a patient's medical profiles such as age, medical history, and genetic information; in personalized online service, a reader's interest in news article contents may also be associated with information such as location and browsing history. This promising variation of sequential decision making problems that incorporate user-space covariates is known as the multi-armed bandit problem with covariates.

Initialized by Woodroofe (1979), bandit problems with covariates tend to be classified into two categories according to assumptions on the mean reward functions. The first category is referred to as the nonparametric bandit problem with covariates, in which the mean reward functions are assumed to satisfy mild smoothness conditions. Notably, Yang and Zhu (2002) studied strong consistency properties of a class of randomized allocation algorithms. Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) proposed arm-elimination type algorithms and established their near minimax rates for cumulative regrets. Some related recent work in this category can also be found in Qian and Yang (2016*a*,*b*), Guan and Jiang (2018), and Reeve et al. (2018).

The second category is called the parametric linear bandit problem with covariates, where the mean reward functions take a linear form with unknown *arm-specific* parameters. In this category, Goldenshluger and Zeevi (2009, 2013) and Bastani and Bayati (2020) considered fixed dimensions and high-dimensional covariates, respectively, and showed that their forced sampling algorithms with exploitation achieve (near) minimax rates when a margin condition (Tsybakov, 2004) and a constant gap condition are imposed. However, the performance of their algorithms remains unknown in more general scenarios where these two conditions are possibly violated. A detailed discussion involving these conditions is given in Section 6 to exhibit the valuable

connection and critical difference between our work and the literature.

In this paper, we propose a multi-stage arm allocation algorithm with arm elimination and randomized allocation to solve the linear bandit problem with high-dimensional covariates. We particularly study the integration of a class of stepwise-type high-dimensional regression methods into the proposed approach and develop new technical tools to analyze non-i.i.d. samples inherited from arm allocation of the bandit algorithm. Our work significantly extends the theoretical understanding under the parametric framework; *the main contribution is outlined as follows.*

First, this paper investigates a new study scope that does not necessarily require the margin condition or the constant gap condition of competitive arms (the arms with positive probabilities of being optimal), and demonstrates a finite-time regret analysis that shows near minimax optimal performance of the proposed algorithm (Section 5.2). To our knowledge, no other existing algorithm is known to work under this new study scope (see also the discussion in Section 6.1). By the discovery of an intriguing connection between the margin and the gap conditions, our new results on regret analysis also synergistically complement the existing literature and together verify the "benefit" of margin conditions in a minimax sense that, if satisfied, can lead to significantly improved regret rates. Second, our algorithm enjoys adaptive performance, in that it automatically captures the regret benefit under the margin and the constant gap conditions and always maintains near-optimal performance regardless of whether these conditions are satisfied (Section 6). This seems to be the first study to exhibit such an adaptive phenomenon for linear bandits with high-dimensional covariates. Third, we show that the outputs of our bandit algorithm possess desired statistical properties, including parameter estimation consistency and variable selection consistency for competitive arms (Section 5.3). Note that variable selection consistency with simultaneous optimal regret guarantees (without or with the margin and constant gap conditions) has not been reported elsewhere in the literature. Also, promising applications of our proposal are demonstrated through two real data examples on drug dose assignment and news article recommendation.

It is worth noting that bandit problems have been studied under other related settings. The examples include best policy matching (e.g., Langford and Zhang, 2008; Agarwal et al., 2014), arm-space (with or without user-space) contextual bandits (e.g., Auer et al., 2007; Abbasi-Yadkori et al., 2011), difficulty links on simple and cumulative regret minimization (Bubeck et al., 2011), the multi-class banditron (e.g., Kakade et al., 2008; Beygelzimer et al., 2017), Bayesian-type approaches (e.g., May et al., 2012; Laber et al., 2018), and bandits with delayed feedback

(e.g., Bistritz et al., 2019; Arya and Yang, 2020), among many others (see, e.g., Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012; Zhou, 2015; Lattimore and Szepesvári, 2020 for bibliographic remarks, surveys and references therein). However, these alternative settings and the corresponding algorithms do not address the main issue of this study. For example, Lattimore and Szepesvári (2020, Ch.23) studied a general arm-space setting for sparse contextual linear bandits, where the (possibly infinitely many) arms share the same unknown sparse coefficient vector. The cumulative regret of the algorithm designed for this setting increases at a polynomial rate with respect to the arm feature dimension. In constrast, our study framework focuses on a user-space setting with a finite and relatively small number of arms, which have their own individual sparse coefficients. As will be seen, the optimal arm depends on the user covariates, and the corresponding cumulative regret has the desirable logarithmic rate in terms of the user covariate dimension.

In fact, our study is in line with the very fruitful research topic known as dynamic treatment regimes (DTR; e.g., Murphy, 2003; Qian and Murphy, 2011; Goldberg and Kosorok, 2012; McKeague and Qian, 2014; Laber et al., 2014; Shi et al., 2018, and many important others). Rather than considering an i.i.d. sample with multi-time point decision rules, this paper focuses on the single-time point decision for sequentially coming users and intends to achieve guaranteed near optimal cumulative rewards for all these users as a whole.

In the remainder of the paper, we provide the basic settings of the bandit problem with high-dimensional covariates in Section 2. The main algorithm and the integrated stepwise-type coefficient estimation are described in Sections 3 and 4, followed by a theoretical investigation in Section 5. The benefit of the margin condition and the algorithm's adaptive performance are studied in Section 6. Simulation and real data evaluation are given in Sections 7 and 8, respectively.

We close this section by briefly summarizing the notation consistently used in this article: $n$ for the user visit index and $N$ for the total number of visits; $k$ for the stage index and $K$ for the total number of stages; $i$ for the arm index, $I$ for a chosen arm, and $l$ for the total number of arms.

## 2. Setting for linear bandits with high-dimensional covariates

In many applications, as opposed to the classical setting with homogeneous distributions, the reward from a decision arm often depends on many user covariates. In the following, we propose

developing a new algorithm to solve the sequential decision making problem with linear mean reward structures in high-dimensional settings. Suppose there are $l$ candidate decision arms ($l \geq 2$) and let $N$ be the total number of user visits. Given user covariate vector $\mathbf{X} \in \mathbb{R}^p$ and arm $i$ ($1 \leq i \leq l$), we consider linear model structures in which the observed reward $Y_i$ has the conditional mean $f_i(\mathbf{X}) := \mathrm{E}(Y_i \,|\, \mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}_i$, where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \cdots, \beta_{ip})^T \in \mathbb{R}^p$ is the true coefficient vector for arm $i$. We assume the sparsity condition in which only a subset of elements in $\mathbf{X}$ is associated with $Y_i$. Define the set of relevant variables for arm $i$ to be $\mathcal{V}_i = \{1 \leq j \leq p : \, |\beta_{ij}| > 0\}$ and its size $q_i := |\mathcal{V}_i| < p$.

Our problem of interest works like the classical setting but with the necessary incorporation of the covariates. At each user visit $n$ ($1 \leq n \leq N$), a user covariate vector $\mathbf{X}_n \in \mathbb{R}^p$ is first revealed, where the $\mathbf{X}_n$'s are i.i.d. from some unknown distribution (same as $\mathbf{X}$) with domain $\mathcal{X} \subset \mathbb{R}^p$. Let $I_j$ be the chosen arm at each visit point $j$ ($1 \leq j < N$), and let $Y_{i,j}$ be the reward if arm $i$ is chosen. Then given the observable information $\{(\mathbf{X}_j, I_j, Y_{I_j,j}),\, 1 \leq j \leq n-1\}$ and current covariate vector $\mathbf{X}_n$, a bandit algorithm is applied to choose an arm $I_n$ and receive the corresponding reward $Y_{I_n,n} = \mathbf{X}_n^T\boldsymbol{\beta}_{I_n} + \varepsilon_{I_n,n}$, where $\varepsilon_{i,n}$ is the random error of arm $i$ and is not necessarily independent of $\mathbf{X}_n$.

### 2.1. Definitions and assumptions

Before introducing the algorithm evaluation, we first give key assumptions. For $\mathbf{x} \in \mathcal{X}$, define the optimal mean reward $f^*(\mathbf{x}) = \max_{1\leq i\leq l} \mathbf{x}^T\boldsymbol{\beta}_i$. Assume that the set $\mathcal{I} = \{1, \cdots, l\}$ of all candidate arms can be partitioned into a set of competitive arms $\mathcal{I}_o$ and a set of non-competitive arms $\mathcal{I}_u$. Let $\mathcal{T}_i$ be the competitive region where arm $i \in \mathcal{I}$ is optimal:

$$\mathcal{T}_i = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}^T\boldsymbol{\beta}_i - \max_{j\neq i} \mathbf{x}^T\boldsymbol{\beta}_j > 0\}. \tag{1}$$

As given in Assumption 1, we define that arm $i$ is a competitive arm in $\mathcal{I}_o$ if it is an optimal arm with a positive probability bounded away from zero.

**Assumption 1.** (Competitive arms) There is a positive constant $c_1$ such that for each arm $i \in \mathcal{I}_o$, $P(\mathbf{X} \in \mathcal{T}_i) > c_1$.

As given in Assumption 2, we define that arm $i$ is a non-competitive arm in $\mathcal{I}_u$ if it is always a sub-optimal arm with a gap of $\tilde{\zeta}_N$ from the optimal reward. Here we allow $\mathcal{I}_u$ to be an empty set. If $\mathcal{I}_u = \varnothing$, then Assumption 2 simply reduces to a null assumption, which is also the case in the settings of Goldenshluger and Zeevi (2013). If $\mathcal{I}_u \neq \varnothing$, $\tilde{\zeta}_N$ is allowed to approach zero as

$N \to \infty$.

**Assumption 2.** (Non-competitive arms) Each arm $i \in \mathcal{I}_u$ satisfies that with probability 1, $\max_{1\le j\le l} \mathbf{X}^T\boldsymbol{\beta}_j - \mathbf{X}^T\boldsymbol{\beta}_i > \tilde{\zeta}_N$, where $\tilde{\zeta}_N \ge \frac{c_2}{N^{\psi}\vee(\log N)^{1/2}}$ for some constants $c_2 > 0$ and $0 \le \psi \le 1/4$.

We also assume in Assumption 3 that the covariates satisfy a version of the restricted isometry property (RIP; Candes and Tao, 2005). The RIP condition and its related variants have often been used in the analysis of high-dimensional linear regression methods (e.g., Meinshausen and Yu, 2009; Zhang, 2010, 2011*b*). By the nature of our targeted bandit problem with covariates, an "oracle" allocation strategy (the benchmark in regret definition that knows the competitive regions for all the competitive arms) is to always deliver a competitive arm at this arm's own competitive region; it is then natural to have conditions that use the arms' own competitive regions, since under the "oracle" benchmark, each competitive arm's data points must all fall within its own competitive region. Specifically, for each arm $i \in \mathcal{I}_o$, define the conditional second moment on the competitive region in which $\Sigma_i = \mathrm{E}(\mathbf{X}\mathbf{X}^T \,|\, \mathbf{X} \in \mathcal{T}_i)$; for each arm $i \in \mathcal{I}_u$, define $\Sigma_i = \Sigma = \mathrm{E}(\mathbf{X}\mathbf{X}^T)$. Given any arm $i \in \mathcal{I}$ and positive integer $s$, define $\lambda_i(s) = \min\{\mathbf{v}^T\Sigma_i\mathbf{v} : \|\mathbf{v}\|_2 = 1, \|\mathbf{v}\|_0 \le s\}$.

**Assumption 3.** There exists a constant $c_* > 0$ such that for each arm $i \in \mathcal{I}$, $\lambda_i(q_*) > c_*$, where $q_* := C_1 \max_{1\le i\le l} q_i$ for some constant $C_1 > 1$.

In Assumption 3, $q_*$ serves as an upper bound of all $q_i$'s at the same order of $\max_{i\in\mathcal{I}} q_i$; a sufficient condition of Assumption 3 is that the minimum eigenvalues of the $\Sigma_i$'s, denoted by $\lambda_{\min}(\Sigma_i)$, are bounded away from zero.

In addition, we assume bounded reward coefficients such that $\|\boldsymbol{\beta}_i\|_1 \le b$ for some constant $b > 0$, and the sub-Gaussian condition for random errors such that $\mathrm{E}(e^{v\varepsilon_{i,n}} \,|\, \mathbf{X}_n) \le \exp(v^2\sigma^2/2)$ for all $v \in \mathbb{R}$. For simplicity, we consider bounded domain $\mathcal{X}$ with $\|\mathbf{X}_n\|_\infty \le \theta$ for some constant $\theta > 0$, but it may be extended to covariates with a sub-Gaussian distribution.

### 2.2. Algorithm evaluation

Let $i^*(\mathbf{x}) = \mathrm{argmax}_{i\in\mathcal{I}} f_i(\mathbf{x})$ be the arm that has the maximum mean reward given $\mathbf{x}$, and define $f^*(\mathbf{x}) = f_{i^*}(\mathbf{x})$. Without knowledge of random error, the "oracle" (but clearly not applicable) benchmark is to choose the optimal arm $I_n^* := i^*(\mathbf{X}_n)$ at each visit point $n$. To evaluate the algorithm performance, define the cumulative regret $R_N$ that measures the shortfall of the algorithm

in cumulative mean reward compared to the "oracle" benchmark:

$$R_N = \sum_{n=1}^{N} \big( f^*(\mathbf{X}_n) - f_{I_n}(\mathbf{X}_n) \big). \tag{2}$$

It is desirable for an allocation strategy to have a guaranteed finite-time upper bound on cumulative regret. Note that for each visit point $n$, only the reward of the chosen arm can be observed while the rewards of all the other arms are not observable: we inevitably encounter incomplete information under the bandit settings.

In addition, a useful but less discussed question of interest in the linear bandit problem is whether the devised algorithm outputs meaningful variable selection results for the competitive arms. Suppose at the end of running an allocation strategy, the algorithm output gives a set of estimated competitive arms $\hat{\mathcal{I}}_o$, and for each arm $i \in \hat{\mathcal{I}}_o$, there is an associated estimate $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i1}, \hat{\beta}_{i2}, \cdots, \hat{\beta}_{ip})$ for $\boldsymbol{\beta}_i$; the estimated set of important variables is defined as $\hat{\mathcal{V}}_i = \{1 \leq j \leq p : |\hat{\beta}_{ij}| > 0\}$. Then we say an algorithm is variable selection consistent if

$$P(\hat{\mathcal{I}}_o = \mathcal{I}_o) \to 1 \text{ and } P(\hat{\mathcal{V}}_i = \mathcal{V}_i \text{ for all } i \in \mathcal{I}_o) \to 1 \text{ as } N \to \infty. \tag{3}$$

It is also desirable to establish that the algorithm is coefficient estimation consistent. That is, for each competitive arm $i \in \mathcal{I}_o$, $\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2 = O_p(\vartheta_N)$, where $\vartheta_N$ is the (preferably fast) convergence rate with $\vartheta_N \to 0$ as $N \to \infty$. Both variable selection consistency and coefficient estimation consistency (e.g., Zou, 2006; Meinshausen and Yu, 2009; Fan and Lv, 2010; Qian et al., 2019a and references therein) are widely studied in the statistics literature for high-dimensional regression problems. In our bandit problem setting, these results provide some asymptotic theoretical guarantees on the algorithm output for an analyst who may want to subsequently use the output for understanding relevant variables and designing new offline policies.

### 2.3. A useful example

In our following study, we will first focus on the study scope from Section 2.1, that is, the class of $l$-armed bandit reward function (or coefficient) sets with joint distributions $P_{\mathbf{X},\boldsymbol{\varepsilon}}$ of $(\mathbf{X}_n, \varepsilon_{1,n}, \cdots, \varepsilon_{l,n})$ that satisfy all the conditions in Section 2.1. Each member in the class is characterized by a set of coefficients $\{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_l\}$ with a distribution $P_{\mathbf{X},\boldsymbol{\varepsilon}}$. Later on in Section 6, we will present another study scope that imposes two additional assumptions including a margin condition and a constant gap condition of competitive arms. In general, more assumptions lead to smaller class size and a potentially lower (minimax) optimal regret rate; as will be seen, the different study scopes lead to different optimality results (and different algorithmic design).

To facilitate an appreciation of the generality and challenges of the study scope in Section 2.1, we next present a useful example. Given $l = 2$ and $q$, define a subclass consisting of all the two-armed bandit pairs of coefficients $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ with $P_{\mathbf{X},\varepsilon}$ that satisfy the following scenarios. Treating the first elements in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as intercept terms, we define $\boldsymbol{\beta}_1 = (0, \frac{\kappa}{\sqrt{q}}, \cdots, \frac{\kappa}{\sqrt{q}}, \cdots, 0)^T \in \mathbb{R}^p$, $\boldsymbol{\beta}_2 = (\omega, -\frac{\kappa}{\sqrt{q}}, \cdots, -\frac{\kappa}{\sqrt{q}}, \cdots, 0)^T \in \mathbb{R}^p$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have $q$ nonzero elements besides the intercept, $\kappa > 0$, $\omega \in (-\kappa, \kappa)$, and $\kappa\sqrt{q}$ is upper bounded by a positive constant. Also denote the covariates by $\mathbf{X} = (1, X_1, \cdots, X_{p-1})$, where $X_1, \cdots, X_{p-1}$ are i.i.d. with Uniform$[-1, 1]$; conditioning on $\mathbf{X}_n$, the random errors $\varepsilon_{1,n}$ and $\varepsilon_{2,n}$ satisfy the sub-Gaussian condition. This gives the simple scenarios in which $f_1(\mathbf{X}) = \frac{\kappa}{\sqrt{q}} \sum_{j=1}^{q} X_j$ and $f_2(\mathbf{X}) = \omega - \frac{\kappa}{\sqrt{q}} \sum_{j=1}^{q} X_j$; the competitive region for arm $i$ $(i = 1, 2)$ is $\mathcal{T}_i = \{\mathbf{x} \in \mathcal{X} : f_i(\mathbf{x}) - f_j(\mathbf{x}) > 0,\ j \neq i\}$. For convenience, we denote this bandit subclass as $\mathcal{P}$. Then all the members in $\mathcal{P}$ satisfy the assumptions in Section 2.1 and indeed fall within the intended study scope (as shown by Propositions 7 and 8 in Supplement A.1). We can then construct a sequence of its members with both coefficient parameters $\kappa$ and $\omega$ indexed by $N$: let $\kappa = \kappa_N = N^{-\alpha}$ for some constant $\alpha > 0$ and $\omega = \omega_N \in (-\kappa_N, \kappa_N)$; we denote the corresponding mean reward function pairs as $\{f_{1,N}(\cdot), f_{2,N}(\cdot)\}$. This example gives the properties in Proposition 1.

**Proposition 1.** *Consider the sequence of the class members constructed above from $\mathcal{P}$. Then given any constants $\alpha > \alpha' > 0$ with $\tilde{\delta}_N = N^{-\alpha'}$, we have*

$$P(0 < f_N^*(\mathbf{X}) - f_N^\sharp(\mathbf{X}) < \tilde{\delta}_N) \to 1 \quad \text{as } N \to \infty, \tag{4}$$

*where $f_N^*(\mathbf{X}) = \max(f_{1,N}(\mathbf{X}), f_{1,N}(\mathbf{X}))$ $f_N^\sharp(\mathbf{X}) = \min(f_{1,N}(\mathbf{X}), f_{2,N}(\mathbf{X}))$; equivalently,*

$$P(f_{2,N}(\mathbf{X}) - f_{1,N}(\mathbf{X}) > \tilde{\delta}_N) + P(f_{1,N}(\mathbf{X}) - f_{2,N}(\mathbf{X}) > \tilde{\delta}_N) \to 0 \quad \text{as } N \to \infty. \tag{5}$$

Proposition 1 reflects a philosophy for our proposed study in which a newly designed algorithm may ideally be able to handle increasingly closer competitive arms as $N$ gets larger, so that to some extent, it parallels the statistical thinking that larger sample size allows for the finding of increasingly smaller treatment effects. The class $\mathcal{P}$ will also be helpful to establish a regret lower bound (to be shown in Section 5.2).

Noting the polynomially decreasing $\tilde{\delta}_N$ in (4) and (5), it will be seen in Section 6.1 that the study scope of Section 2.1 and the associated algorithm design are deemed different from the existing literature. On one hand, Bastani and Bayati (2020) novelly designed algorithms that are well-suited with provable optimality under the additional margin condition and constant gap condition for competitive arms. On the other hand, neither of these two additional conditions

are necessarily satisfied for Section 2.1, and the literature has not yet shown how to design a generally near optimal algorithm. We will defer the detailed discussion to Section 6.1 on the connection between the different study scopes, without or with the two conditions.

Furthermore, it would be interesting for a newly designed algorithm to simultaneously perform optimally when these additional conditions are imposed: that is, can an algorithm adaptively achieve near optimality in both worlds of the different study scopes, and attain potential regret "benefit" if the additional conditions are satisfied? The efforts to address this issue will be presented in Section 6.2.

## 3. A multi-stage algorithm in high dimensions

Our proposed algorithm divides the total visit points into $K+1$ stages, with stage 0 being the initial forced sampling stage. Here $\tilde{N}_k$ $(1 \leq k \leq K)$ is the end visit point of stage $k$, and $N_k = \tilde{N}_k - \tilde{N}_{k-1}$ is the sample size of stage $k$. Set $N_0 = l\tau_0$, $\tau_0 = c_0 q_*^2 \log p_N (N^{2\psi} \vee \log N)$, $N_k = 2N_{k-1}$, and $K = \lceil \log_2(1+N/N_0) - 1 \rceil$, where $p_N = p \vee N$, $c_0$ is some positive constant, $\lceil \cdot \rceil$ is the ceiling function, and stage $K$ may have a sample size less than $2N_{K-1}$. We set $c_0 = 32\theta^2 c_\rho c_2^{-2}$ (or its upper bound) for Section 5, where $c_\rho > 0$ is a constant (to be given in Theorem 1). Given stage $k$, define $\mathcal{A}_{k,i} = \{n : \tilde{N}_{k-1} + 1 \leq n \leq \tilde{N}_k,\, I_n = i\}$ to be the set of visit points where arm $i$ is chosen; similarly, define $\mathcal{B}_{k,i} = \{n : 1 \leq n \leq \tilde{N}_k,\, I_n = i\}$.

Let $\mathbb{X}_N = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N)^T$ be the $N \times p$ matrix containing all the user covariates, and let $\mathbf{y}_N = (y_1, y_2, \cdots, y_N)^T$ be the vector containing the reward responses from the chosen arms with $y_n = Y_{I_n,n}$ $(1 \leq n \leq N)$. Then given any visit index set $\mathcal{A} = \{j_1, j_2, \cdots, j_{|\mathcal{A}|}\}$ with $1 \leq j_1 < \cdots < j_{|\mathcal{A}|} \leq N$, define $\mathbb{X}_\mathcal{A} \in \mathbb{R}^{|\mathcal{A}| \times p}$ and $\mathbf{y}_\mathcal{A} \in \mathbb{R}^{|\mathcal{A}|}$ to be the corresponding covariate design sub-matrix from $\mathbb{X}_N$ and the reward response sub-vector from $\mathbf{y}_N$, respectively; that is, $\text{row}_n(\mathbb{X}_\mathcal{A}) = \text{row}_{j_n}(\mathbb{X}_N)$ and $\text{row}_n(\mathbf{y}_\mathcal{A}) = \text{row}_{j_n}(\mathbf{y}_N)$ for $1 \leq n \leq |\mathcal{A}|$. We can apply a specified high-dimensional linear regression method with tuning parameter $\xi$ to obtain the coefficient estimator $\hat{\boldsymbol{\beta}}(\mathbb{X}_\mathcal{A}, \mathbf{y}_\mathcal{A}, \xi)$. In our following discussion, unless stated otherwise we will use the high-dimensional Interactive Greedy Algorithm (IGA, Qian et al., 2019b), which is a generalized method from stepwise-type regression (e.g., Zhang, 2011*a*,*b*; Ing and Lai, 2011). Here, $\xi$ represents the tuning parameter for IGA and regulates the estimator sparsity from the solution path. It is closely related to the penalty term of the high-dimensional information criterion (Ing and Lai, 2011), which is used to overcome potential overfitting problems associated with the orthogonal greedy algorithm. We offer a brief description of the coefficient estimation by IGA in Section 4.

**Algorithm 1** Stage-wise arm elimination with randomized allocation.

1. Set initial sampling stage with sample size $N_0$. Choose each arm an equal number of times $\tau_0$. For each arm $i \in \mathcal{I}$, compute the initial estimated coefficient $\tilde{\boldsymbol{\beta}}_i$. Set $k = 1$.

2. At stage $k$, perform the following substeps at $n = \tilde{N}_{k-1} + 1, \cdots, \tilde{N}_k$.

   - Reveal covariate $\mathbf{X}_n \in \mathbb{R}^p$.
   - Pre-screen arms using the initial sampling data to generate the arm set
   $$\tilde{\mathcal{S}}_n := \{i \in \mathcal{I} : \max_{j \in \mathcal{I}} \mathbf{X}_n^T \tilde{\boldsymbol{\beta}}_j - \mathbf{X}_n^T \tilde{\boldsymbol{\beta}}_i \leq \delta_N\}. \tag{6}$$
   - If $k > 1$, eliminate arms on $\tilde{\mathcal{S}}_n$ to generate the set of "promising" arms
   $$\hat{\mathcal{S}}_n := \{i \in \tilde{\mathcal{S}}_n : \max_{j \in \tilde{\mathcal{S}}_n} \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{j,k} - \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{i,k} \leq \Delta_k\}; \tag{7}$$
   otherwise, set $\hat{\mathcal{S}}_n = \tilde{\mathcal{S}}_n$.
   - Define $\hat{I}_n = \text{argmax}_{i \in \hat{\mathcal{S}}_n} \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{i,k}$. Perform randomized allocation to choose an arm $I_n$ from $\hat{\mathcal{S}}_n$ with $h \geq 1$ and receive reward $Y_{I_n,n}$:
   $$I_n = \begin{cases} \hat{I}_n, & \text{with probability } \frac{h}{h+|\hat{\mathcal{S}}_n|-1}, \\ i, & \text{with probability } \frac{1}{h+|\hat{\mathcal{S}}_n|-1},\ i \neq \hat{I}_n,\ i \in \hat{\mathcal{S}}_n. \end{cases}$$

3. Find the estimated coefficient for next stage by computing $\hat{\boldsymbol{\beta}}_{i,k+1}$ for each $i \in \mathcal{I}$.

4. Set $k = k + 1$. Repeat steps 2–4 until the end of $N$ user visits.

5. Obtain an estimated set of competitive arms $\hat{\mathcal{I}}_N = \bigcup_{n=\tilde{N}_{K-2}+1}^{N} \hat{\mathcal{S}}_n$ and output the estimated coefficient $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{i,K}$ for all $i \in \hat{\mathcal{I}}_N$.

Then, given arm $i$, $\tilde{\boldsymbol{\beta}}_i := \hat{\boldsymbol{\beta}}(\mathbb{X}_{\mathcal{A}_{0,i}}, \mathbf{y}_{\mathcal{A}_{0,i}}, \xi_0)$ are the estimated coefficients from stage 0; we set $\hat{\boldsymbol{\beta}}_{i,k} := \hat{\boldsymbol{\beta}}(\mathbb{X}_{\mathcal{A}_{k-1,i}}, \mathbf{y}_{\mathcal{A}_{k-1,i}}, \xi_k)$ to be the coefficients used by stage $k$ and estimated from the data of its previous stage, where the $\xi_k$'s are their respective tuning parameters. If $\mathcal{A}_{k-1,i} = \varnothing$, we set $\hat{\boldsymbol{\beta}}_{i,k} := \hat{\boldsymbol{\beta}}(\mathbb{X}_{\mathcal{B}_{k-1,i}}, \mathbf{y}_{\mathcal{B}_{k-1,i}}, \xi_k)$, where the alternative choice of estimated coefficients with the larger sample $\mathcal{B}_{k-1,i}$ (that includes all historical data of arm $i$) is given in Remark 2 of Section 4.

We are now ready to describe the details of the proposed multi-stage algorithm as shown in Algorithm 1. Specifically, **Step 1** is the initial sampling of stage 0 that allocates each arm an equal number of times. **Step 2** shows that for each visit point $n$ of a given stage $k$, after the observation of covariate $\mathbf{X}_n \in \mathbb{R}^p$, there are two substeps of arm screening procedures: (6) pre-screens out uncompetitive arms, and (7) performs an extra elimination step to generate "promising" arms for use in the subsequent randomized allocation substep. We set the parameters $\delta_N = 2\theta b_0$ and $\Delta_k = 2\theta b_k$ with $b_0 = q_* \sqrt{2c_\rho \log p_N / \tau_0}$ and $b_k = q_* \sqrt{2\tilde{c}_\rho \log p_N / N_k}$, $k \geq 2$, for Section 5, where

$c_\rho$ and $\tilde{c}_\rho$ are positive constants (to be given in Theorems 1 and 2). Here $q_*$ can also be replaced by a general upper bound $s_*$ ($s_* \geq q_*$); its implication w.r.t. the analysis is given in Remark 6 of Section 6.2.

In the last substep of Step 2, define $\hat{I}_n = \text{argmax}_{i \in \hat{\mathcal{S}}_n} \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{i,k}$ where any tie-breaking rule may apply. Let $h \geq 1$ be a randomization parameter. Then, under the randomized allocation scheme, we choose an arm $i$ from $\hat{\mathcal{S}}_n$ with probability $0 < p_{n,i} \leq 1$, where $\sum_{i \in \hat{\mathcal{S}}_n} p_{n,i} = 1$ and $\frac{p_{n,\hat{I}_n}}{p_{n,i}} = h$ for all $i \neq \hat{I}_n$; that is, $p_{n,\hat{I}_n} = \frac{h}{h+|\hat{\mathcal{S}}_n|-1}$ and $p_{n,i} = \frac{1}{h+|\hat{\mathcal{S}}_n|-1}$ for $i \neq \hat{I}_n$ in $\hat{\mathcal{S}}_n$. In particular, $h = 1$ corresponds to simple randomization among arms in $\hat{\mathcal{S}}_n$. We use $h = 1$ in theoretical development for simplicity.

**Step 3** updates the coefficient estimation after the current stage. In **Step 4**, the algorithm moves to the next stage, and continues in a stage-wise fashion until the end of $N$ user visits. Then **Step 5** outputs the estimated set of competitive arms and their associated coefficient estimates. Considering the scenario in which the last stage $K$ has a small sample size, we use the last two stages to estimate $\hat{\mathcal{I}}_N$.

*Remark* 1. Algorithm 1 includes the arm pre-screening substep (6) for all stages. If $\mathcal{I}_u = \varnothing$, the algorithm can be further simplified by removing this substep. However, if $\mathcal{I}_u \neq \varnothing$, the optimal arm may be eliminated by a non-competitive arm, and the analysis argument (to be outlined in Section 5.1 and Proposition 3 for having "good" events) may not hold without this substep. The use of randomized allocation with $h > 1$ (as opposed to $h = 1$) is mainly motivated by the potentially more efficient exploitation of the estimated promising arms in practice. A similar empirical idea for randomization has also been used for the nonparametric bandit problem with covariates (e.g., Qian and Yang, 2016*b*); the feature of (non-uniform) randomized allocation, together with the embedded key arm-elimination technique (Perchet and Rigollet, 2013), can be practically useful to provide additional flexibility for an algorithm to further utilize the reward function estimation; all theoretical results of our proposed algorithm remain the same for upper bounded $h$; we will demonstrate its empirical performance with $h > 1$ in the numerical studies.

## 4. Coefficient estimation

As IGA is embedded into Algorithm 1 and plays an important role in coefficient estimation, we next briefly describe main steps of IGA summarized in Algorithm 2 to keep the paper self-contained.

Given the input design matrix $\mathbb{X} \in \mathbb{R}^{m \times p}$ and response vector $\mathbf{y} \in \mathbb{R}^m$, define the objective

**Algorithm 2** Stepwise coefficient estimation.

1. Initialize $r = 0$, $\boldsymbol{\beta}^{(r)} = 0$, $G^{(0)} = \varnothing$, $0 < \rho \leq 1$ and $\xi > 0$. Set $\phi^{(0)} = Q(\boldsymbol{\beta}^{(r)}) - \min_{1 \leq j \leq p,\, \alpha \in \mathbb{R}} Q(\boldsymbol{\beta}^{(r)} + \alpha \mathbf{e}_j)$.

2. Perform forward selection with the following substeps.

   (a) Find candidate variable set
   $$G_\rho = \{g \notin G^{(r)} : Q(\boldsymbol{\beta}^{(r)}) - \min_{\alpha \in \mathbb{R}} Q(\boldsymbol{\beta}^{(r)} + \alpha \mathbf{e}_g) \geq \rho \phi^{(r)}\}. \tag{8}$$

   (b) Select element $g^{(r)} \in G_\rho$ and set $G^{(r+1)} = G^{(r)} \cup \{g^{(r)}\}$.

   (c) Compute $\boldsymbol{\beta}^{(r+1)} = \arg\min_{\text{supp}(\boldsymbol{\beta}) \in G^{(r+1)}} Q(\boldsymbol{\beta})$, and find $\xi^{(r+1)} = Q(\boldsymbol{\beta}^{(r)}) - Q(\boldsymbol{\beta}^{(r+1)})$.

   (d) Set $r = r + 1$.

3. Set $\tilde{\phi}^{(r)} = \min_{j \in G^{(r)}} Q(\boldsymbol{\beta}^{(r)} - \mathbf{e}_j^T \boldsymbol{\beta}^{(r)} \mathbf{e}_j) - Q(\boldsymbol{\beta}^{(r)})$. If $\tilde{\phi}^{(r)} < \xi^{(r)}/2$, perform backward selection with following substeps.

   (a) Find $g^{(r)} = \arg\min_{j \in G^{(r)}} Q(\boldsymbol{\beta}^{(r)} - \mathbf{e}_j^T \boldsymbol{\beta}^{(r)} \mathbf{e}_j)$.

   (b) Set $r = r - 1$ and $G^{(r)} = G^{(r+1)} \backslash \{g^{(r+1)}\}$.

   (c) Compute $\boldsymbol{\beta}^{(r)} = \arg\min_{\text{supp}(\boldsymbol{\beta}) \in G^{(r)}} Q(\boldsymbol{\beta})$.

   (d) Update $\tilde{\phi}^{(r)} = \min_{j \in G^{(r)}} Q(\boldsymbol{\beta}^{(r)} - \mathbf{e}_j^T \boldsymbol{\beta}^{(r)} \mathbf{e}_j) - Q(\boldsymbol{\beta}^{(r)})$.

   (e) If $\tilde{\phi}^{(r)} < \xi^{(r)}/2$, repeat backward selection substeps above.

4. Find $\phi^{(r)} = Q(\boldsymbol{\beta}^{(r)}) - \min_{1 \leq j \leq p,\, \alpha \in \mathbb{R}} Q(\boldsymbol{\beta}^{(r)} + \alpha \mathbf{e}_j)$. If $\phi^{(r)} \geq \xi$, repeat Steps 2–4; otherwise, output $\boldsymbol{\beta}^{(r)}$.

function $Q(\boldsymbol{\beta}) = \frac{1}{m}\|\mathbf{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$. Let $\mathbf{e}_j \in \mathbb{R}^p$ be the unit vector with the $j$-th element being zero. Then from Algorithm 2, following initialization (Step 1), the forward selection in Step 2 selects one variable into the active set $G^{(r)}$ and drives down the objective function $Q(\boldsymbol{\beta})$ in a stepwise fashion, that is, (8) essentially considers all the candidate variables one by one and finds those that rank high in reduction of $Q(\boldsymbol{\beta})$. Alternatively, to avoid repeated optimization tasks on the objective function and to significantly reduce computation time, we can also replace (8) and $\phi^{(r)}$ by gradient-based criterion:

$$\phi^{(r)} = \|\nabla Q(\boldsymbol{\beta}^{(r)})\|_\infty \text{ and } G_\rho = \{g \notin G^{(r)} : |\nabla_g Q(\boldsymbol{\beta}^{(r)})| \geq \rho\phi^{(r)}\}, \tag{9}$$

where $\nabla Q(\boldsymbol{\beta})$ is the gradient vector and $\nabla_g Q(\boldsymbol{\beta})$ is its $g$-th element. Without additional information on true variables, it suffices that we set $\rho = 1$. Step 3 is the backward elimination step that checks if some variables may become redundant after the new variable is included from forward selection. This forward-backward iteration scheme continues until the addition of any new variables does not significantly reduce the objective function as shown in Step 4.

*Remark* 2. Given $\mathbb{X}$, $\mathbf{y}$, and $\xi$, the output of Algorithm 2 gives the coefficient estimator $\hat{\boldsymbol{\beta}}(\mathbb{X}, \mathbf{y}, \xi)$. The parameter $\xi$ regulates the solution sparsity: a larger $\xi$ tends to provide a sparser solution. In empirical studies, instead of giving explicit values for $\xi$, we use the number of steps to determine solution sparsity, which is automatically selected by ten-fold cross validation (CV) on $(\mathbb{X}, \mathbf{y})$ under the mean square error criterion. The package that implements the IGA method with CV is publicly available on GitHub. Also, in the description of Algorithm 1, we use the stage-specific sample $\mathcal{A}_{k,i}$ for coefficient estimation to make the proofs more concise. In practice, we recommend using the sample choice of including all historical data from previous stages so that $\hat{\boldsymbol{\beta}}_{i,k+1} := \hat{\boldsymbol{\beta}}(\mathbb{X}_{\mathcal{B}_{k,i}}, \mathbf{y}_{\mathcal{B}_{k,i}}, \xi_{k+1})$.

## 5. Understanding algorithm performance

To understand the performance of the proposed algorithm, it is helpful to study how the algorithm estimates the conditional mean rewards and the coefficients and how these estimates are associated with "good" events on arm selection. In Section 5.1, we outline the analysis strategy for the cumulative regret upper bounds, which consist of four main steps. We provide the upper and lower bounds on the cumulative regret in Section 5.2, and establish the variable selection and coefficient estimation consistency properties in Section 5.3.

### 5.1. Outline of main analysis steps

**The first main step** is regret decomposition via the partitioning of the sample space into properly defined events. Specifically, let $R_{N0}$ and $R_{N1}$ be the regrets accumulated in Stage 0 and the following stages, respectively. Then we see that $R_N = R_{N0} + R_{N1}$. Also define the following events on coefficient estimation errors. For $2 \leq k \leq K$, define

$$F_0 = U_1 = \{\forall i \in \mathcal{I},\ \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i\|_1 \leq b_0\},\ F_k = \{\forall i \in \mathcal{I}_o,\ \|\hat{\boldsymbol{\beta}}_{i,k} - \boldsymbol{\beta}_i\|_1 \leq b_k\}, \tag{10}$$

and $U_k = F_0 \cap \left(\bigcap_{j=2}^{k} F_j\right)$. The whole sample space can be partitioned into the events

$$U_1^c,\ U_k \cap F_{k+1}^c,\ U_K \quad \text{ for } 1 \leq k \leq K-1 \tag{11}$$

to further decompose the cumulative regret, so that

$$R_{N1} = R_{N1} I(U_1^c) + \sum_{k=1}^{K-1} R_{N1} I(U_k \cap F_{k+1}^c) + R_{N1} I(U_K) =: R_0 + \sum_{k=1}^{K-1} R_k + R_K. \tag{12}$$

To provide upper bounds for the decomposed regrets, we need to understand the properties and implications of these associated events to be shown in the next two main steps.

**In the second main step**, we intend to achieve the following specific objective (1): under "good" events, via connection with coefficient/reward estimation errors, the regret can be upper-bounded. We further divide the analysis effort of this step into two substeps, which include studying (1a) *arm pre-screening behavior* and (1b) *arm elimination behavior*. Steps (1a) and (1b) are summarized in Propositions 2 and 3, respectively, whose proofs are relegated to Supplement A.2.

**Proposition 2.** *Given stage $k$ ($k \geq 1$), if the event $U_k$ holds, then at any visit point $n$ ($\tilde{N}_{k-1}+1 \leq n \leq \tilde{N}_k$), the optimal arm $I_n^*$ remains in $\tilde{\mathcal{S}}_n$, and any non-competitive arm $i \in \mathcal{I}_u$ is excluded from $\tilde{\mathcal{S}}_n$.*

**Proposition 3.** *Given stage $k$ ($k \geq 2$), if the event $U_k$ holds, then at any visit point $n$ ($\tilde{N}_{k-1}+1 \leq n \leq \tilde{N}_k$), the optimal arm $I_n^*$ remains in $\hat{\mathcal{S}}_n$; in addition, any "promising" arm $i \in \hat{\mathcal{S}}_n$ belongs to the arm set $\mathcal{U}_{n,k} = \{j \in \mathcal{I}_o : \mathbf{X}_n^T \boldsymbol{\beta}_{I_n^*} - \mathbf{X}_n^T \boldsymbol{\beta}_j \leq 2\Delta_k\}$.*

The two propositions above suggest that with the arm pre-screening and elimination procedures, the event $U_k$ regarding the coefficient estimation errors leads to the "good" event that the algorithm always keeps the optimal arm while all the other remaining arms must be in the arm set $\mathcal{U}_{n,k}$, thereby restricting the regret of each step within $2\Delta_k$ to achieve objective (1). Therefore, to study the maintenance of "good" events for arm selection, it is important to understand

the coefficient estimation errors.

Due to the nature of necessarily evolving arm allocation in sequential decision making, only one response from the selected arm is revealed while responses from all the other arms are not available; the accumulated data for each arm are *not* i.i.d. random samples anymore (as opposed to regular settings in high-dimensional regression problems), which poses unique challenges in studying the statistical properties of the estimated coefficients. With the multi-stage approach and stage-wise arm elimination, we also employ randomized arm allocation to help partly overcome the technical issues (besides empirical performance considerations, to achieve a balance between exploration and exploitation).

**In the third main step**, we intend to achieve the specific objective (2): the (conditional) probabilities of violating the "good" events are relatively small. For this purpose, we establish Theorems 1 and 2 (see below). These theorems are proved through four substeps (2a) randomized allocation with "random" samples, (2b) sample size determination, (2c) covariate "design matrix" properties, and (2d) coefficient estimation upper bounds, details of which are also relegated to Supplement A.2. Note that $\xi_0$ and the $\xi_k$'s correspond to the tuning parameter $\xi$ in Algorithm 2, which computes $\tilde{\boldsymbol{\beta}}_i$ and the $\hat{\boldsymbol{\beta}}_{i,k}$'s, respectively; recall that $p_N = p \vee N$.

**Theorem 1.** *Suppose Assumptions 1–3 hold. Then there exists a positive constant $c_r$ such that given $\xi_0 = \frac{c_r \log p_N}{\tau_0}$, it holds with probability less than $l/N^3$ that*

$$\|\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_1 > \sqrt{\frac{c_\rho q_* (q_i + \log N + q_{i,0} \log p_N)}{\tau_0}}$$

*for some $i \in \mathcal{I}$, where $q_{i,0} = |J_{i,0}|$, $J_{i,0} = \{j \in \mathcal{V}_i : |\beta_{i,j}| < \sqrt{c_\beta \log p_N / \tau_0}\}$, and $c_\rho, c_\beta > 0$ are some constants.*

**Theorem 2.** *Suppose Assumptions 1–3 hold. Then there exists a positive constant $c_r'$ such that given $\xi_{k+1} = \frac{c_r' \log p_N}{N_k}$ and $U_k$ $(1 \le k \le K-1)$, it holds with probability less than $3l/N^3$ that*

$$\|\hat{\boldsymbol{\beta}}_{i,k+1} - \boldsymbol{\beta}_i\|_1 > \sqrt{\frac{\tilde{c}_\rho q_* (q_i + \log N + q_{i,k} \log p_N)}{N_k}}$$

*for some $i \in \mathcal{I}_o$, where $q_{i,k} = |J_{i,k}|$, $J_{i,k} = \{j \in \mathcal{V}_i : |\beta_{i,j}| < \sqrt{\tilde{c}_\beta \log p_N / N_k}\}$, and $\tilde{c}_\rho, \tilde{c}_\beta > 0$ are some constants.*

These two theorems suggest that with the proposed algorithm, given $U_k$, the probability of violating $F_{k+1}$ (or $U_{k+1}$) on the coefficient estimation errors is small; consequently, since $U_{k+1}$ always implies the "good" arm selection events on the next stage as shown in the propositions

for objective (1), the same probability bound applies to violating these "good" events, thereby achieving objective (2).

**As the last main step**, we obtain the decomposed regrets by Propositions 2 and 3 from objective (1) and Theorems 1 and 2 from objective (2), and subsequently assemble the cumulative regret upper bounds to be shown next in Section 5.2.1.

## 5.2. Upper and lower bounds on cumulative regret

We demonstrate here the near minimax optimal regret performance of the proposed algorithm, where the upper bound and the lower bound are given in Sections 5.2.1 and 5.2.2, respectively.

### 5.2.1 Upper bound

The analysis efforts briefly summarized in Section 5.1 enable us to provide the following finite-time regret analysis for (2).

**Theorem 3.** *Suppose Assumptions 1–3 hold. Then there exist positive constants $C_{21}$ and $C_{22}$ such that the cumulative regret of Algorithm 1 satisfies*

$$\mathrm{E}(R_N) \leq C_{21} l q_*^2 \log p_N (N^{2\psi} \vee \log N) + C_{22} q_* \sqrt{N \log p_N} \tag{13}$$

*with $C_{21} = 4\theta b c_0 + 6\theta b$ and $C_{22} = 8\theta \tilde{c}_\rho^{1/2}$; in particular, if $\psi = 0$ and $p = o(N^\zeta)$ for some constant $\zeta > 0$ with fixed $l$ and $q_*$, then for any large enough $N$,*

$$\mathrm{E}(R_N) \leq 2 C_{22} q_* \sqrt{N \log p_N}. \tag{14}$$

In Theorem 3, the upper bound of (13) consists of two components. Roughly speaking, the first component is mainly attributed to the initial forced sampling, which generates initial crude estimates for the coefficients and ensures good performance for the pre-screening of the uncompetitive arms; mainly from the much more refined arm elimination stages for the competitive arms, the second component is usually a dominating term as shown by (14).

Note that under additional conditions (to be introduced in Section 6.1), existing algorithms (Goldenshluger and Zeevi, 2013; Bastani and Bayati, 2020) indicate that by an exploitation-based strategy, it is ensured for regret analysis that the optimal arm in its competitive region with a certain constant reward gap can be exclusively selected with high probability. However, such analysis argument is not technically feasible here. To overcome this difficulty, we employ arm elimination and randomized allocation to carefully control regret accumulation in a stagewise fashion, thereby circumventing the need for these additional conditions. The inherited new

technical challenges in regret analysis are naturally shared with the simultaneous establishment of variable selection consistency to be shown in Section 5.3.

### 5.2.2 Lower bound

We then seek to address whether it is possible for any alternative algorithm to achieve a regret rate much slower than that of (14). For this purpose, recall the bandit subclass $\mathcal{P}$ defined from the example of Section 2.3, which has been verified to satisfy all the conditions of Section 2.1.

**Theorem 4.** *For any admissible bandit strategy, there is a positive constant $C_3$ such that with any large enough $N$, we can always find some class member in $\mathcal{P}$ under which its cumulative regret satisfies*

$$\mathrm{E}(R_N) > C_3\sqrt{N}.$$

The regret lower bound in Theorem 4 implies that the upper bound in Theorem 3 is almost not improvable for $N$ (up to a logarithmic factor), and that our proposed algorithm has near minimax optimal performance under the study scope of Section 2.1.

*Remark* 3. In the upper-bound regret analysis, it is assumed that $\|\mathbf{X}_n\|_\infty$ is bounded above by a constant $\theta > 0$, which is involved in setting the coefficients of algorithm parameters. This condition can be relaxed to allow element-wise sub-Gaussian conditions on the covariates. Specifically, assume that for all covariates $\mathbf{X}_n = (X_{n,1}, X_{n,2}, \cdots, X_{n,p})^T$, there exists some constant $\sigma_X > 0$ such that $\mathrm{E}(e^{vX_{n,j}}) \leq \exp(v^2\sigma_X^2/2)$ for $v \in \mathbb{R}$ and $1 \leq j \leq p$. Define the event $A = \{\|\mathbf{X}_n\|_\infty \leq c_x\sigma_X\sqrt{\log p_N} \text{ for all } 1 \leq n \leq N\}$ with some constant $c_x \geq 2\sqrt{2}$. Then the following Proposition 4 shows that the regret contributed by $A^c$ is relatively negligible.

**Proposition 4.** *Given the sub-Gaussian conditions on covariates, it is satisfied that*

$$\mathrm{E}\big(R_N I(A^c)\big) \leq 4bc_x\sigma_X p_N^{-1}\sqrt{\log p_N}.$$

By treating $A^c$ as a "bad" event in our regret decomposition, Proposition 4 suggests that we can just focus on the "good" event in which all covariates are bounded by $\tilde{\theta}_N = c_x\sigma_X\sqrt{\log p_N}$ and replace the constant $\theta$ by $\tilde{\theta}_N$ instead; as a result, the algorithm analysis under event $A$ can be performed similarly, with the mild price on regret rate by extra multiplicative factors of $\log p_N$.

## 5.3. Variable selection and coefficient estimation consistency

The proposed algorithm also generates consistently estimated competitive arms $\hat{\mathcal{I}}_N$ and their consistently estimated coefficients as shown in Theorem 5. Here, $\bar{q}_i$ is the size of variables with

relatively weak signals. Note that the coefficient estimation error bound of $\hat{\boldsymbol{\beta}}_i$ in Theorem 2 includes the slight price of an extra additive $\log N$ term; this reflects the subtle need for the bandit algorithm to simultaneously achieve the desired finite-time regret guarantees. However, this extra $\log N$ term can be removed for the coefficient estimation consistency in Theorem 5, which matches a known result of a regular sparse high-dimensional regression setting (that is, $O_p(\sqrt{(q_i + \bar{q}_i \log p_N)/N})$).

**Theorem 5.** *Under the same conditions of Theorem 3, the algorithm output of the estimated competitive arms satisfies* $P(\hat{\mathcal{I}}_N = \mathcal{I}_o) \to 1$ *as* $N \to \infty$. *In addition, the output of coefficient estimation for each arm* $i \in \mathcal{I}_o$ *satisfies* $\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2 = O_p(\sqrt{\frac{q_i + \bar{q}_i \log p_N}{N}})$, *where* $\bar{q}_i = |\bar{J}_i|$, *and* $\bar{J}_i = \{j \in \mathcal{V}_i : |\beta_{i,j}| < \sqrt{\frac{4\tilde{c}_\beta \log p_N}{N}}\}$.

Combined with a beta-min condition, we further establish coefficient estimation and variable selection consistency simultaneously for the competitive arms in Theorem 6. Therefore, the proposed bandit algorithm also achieves the desired property (3).

**Theorem 6.** *Suppose an arm* $i \in \mathcal{I}_o$ *satisfies* $\min_{j \in \mathcal{V}_i} |\beta_{i,j}| \geq \sqrt{\frac{4\tilde{c}_\beta \log p_N}{N}}$. *Then under the same conditions of Theorem 3, the output of coefficient estimation for arm* $i \in \mathcal{I}_o$ *satisfies*
*1) coefficient estimation consistency:* $\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2 = O_p(\sqrt{\frac{q_i}{N}})$;
*2) variable selection consistency:* $P(\hat{\mathcal{V}}_i = \mathcal{V}_i) \to 1$ *as* $N \to \infty$.
*In particular, if* $\min_{i \in \mathcal{I}_o, j \in \mathcal{V}_i} |\beta_{i,j}| \geq \sqrt{\frac{4\tilde{c}_\beta \log p_N}{N}}$, *Algorithm 1 is variable selection consistent.*

The variable selection consistency of Theorems 5 and 6 also uses results from finite-time analysis, which shows the desired sparsity recovery with high probability. Indeed, it is shown in Supplement A.4 that for any large enough $N$, $P(\mathcal{I}_N \neq \mathcal{I}_o) \leq 3K/N$ and for every $i \in \mathcal{I}_o$, $P(\hat{\mathcal{V}}_i \neq \mathcal{V}_i) \leq 4K/N$.

*Remark* 4. From the proofs of Theorems 1 and 2, we can see that the positive constants $c_r', \tilde{c}_\rho, \tilde{c}_\beta, c_r, c_\rho, c_\beta$ exist. Given that there are constants $c_d, c_f > 0$ associated with the IGA method as shown in Lemma 1 of Supplement B, we can set

$$c_r' = \frac{128\theta^2\sigma^2 c_f}{c_1^4 c_*}(2 + \frac{1}{8\theta^2}),\ \tilde{c}_\rho = \frac{64\sigma^2}{c_1^4 c_*}(c_d + 4c_f) + \frac{32\theta^2 c_r'}{c_1^2 c_*},\ \tilde{c}_\beta = \frac{512\theta^2 c_r'}{c_1^2 c_*},$$
$$c_r = 16\theta^2\sigma^2 c_f c_*^{-1}(2 + \frac{1}{8\theta^2}),\ c_\rho = 8\sigma^2 c_*^{-1}(c_d + 4c_f) + 8\theta^2 c_r c_*^{-1},\ c_\beta = 128\theta^2 c_r c_*^{-1}.$$

## 6. Adaptive performance

### 6.1. Benefit of margin condition

A margin condition is known as an assumption that regulates the complexity and rates of convergence for classification and estimation problems (Mammen and Tsybakov, 1999; Tsybakov, 2004; Audibert and Tsybakov, 2007). To fully appreciate the contribution of our new algorithm design in this work and discern its distinction from the existing literature, it is helpful to consider and discuss a margin condition under linear bandits with covariates. In particular, a margin condition has been assumed and carefully studied in earlier work under both the fixed-dimension setting (Goldenshluger and Zeevi, 2013) and the targeted high-dimensional setting (Bastani and Bayati, 2020); their corresponding bandit algorithms are well-designed to optimally solve the problem under both a margin condition and a constant gap condition.

We next define these conditions. For $\mathbf{x} \in \mathcal{X}$, let $\mathcal{I}^\sharp(\mathbf{x}) = \{i \in \mathcal{I}_o : f_i(\mathbf{x}) < f^*(\mathbf{x})\}$ be the set of sub-optimal arms. Then define $f^\sharp(\mathbf{x}) = \max_{i \in \mathcal{I}^\sharp(\mathbf{x})} \mathbf{x}^T \boldsymbol{\beta}_i$ if $I^\sharp(\mathbf{x}) \neq \varnothing$, and $f^\sharp(\mathbf{x}) = f^*(\mathbf{x})$ otherwise.

**Assumption 4.** There exists a positive constant $L$ such that given any $\delta > 0$,

$$P\big(0 < f^*(\mathbf{X}) - f^\sharp(\mathbf{X}) < \delta\big) \leq L\delta.$$

Assumption 4 requires that except for a subset of the domain with small probability close to the decision boundary, the optimal mean reward can be separated from sub-optimal rewards by arbitrarily small $\delta$. Alongside the margin condition, earlier work also assumes the following constant gap condition.

**Assumption 5.** There are positive constants $\varpi, \tilde{c}_1 > 0$ such that for each arm $i \in \mathcal{I}_o$, $P(\mathbf{X} \in \tilde{\mathcal{T}}_i) > \tilde{c}_1$, where

$$\tilde{\mathcal{T}}_i = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}^T \boldsymbol{\beta}_i - \max_{j \neq i} \mathbf{x}^T \boldsymbol{\beta}_j > \varpi\}.$$

First, we discover that the margin condition of Assumption 4 and the gap condition of Assumption 5 are closely related. Indeed, as shown in the following first statement of Proposition 5, if we impose the margin condition in addition to those of Section 2.1, then the resulting study scope becomes largely equivalent to that of Bastani and Bayati (2020) since it is guaranteed that Assumption 5 is also satisfied.

**Proposition 5.** *If Assumption 1 holds, then Assumption 4 implies Assumption 5. On the other hand, Assumption 5 implies Assumption 1.*

The second statement of Proposition 5 implies that the study scope of Bastani and Bayati (2020) is subsumed in (and is smaller than) that of Section 2.1. In particular, neither Assumption 4 nor Assumption 5 are necessarily satisfied under the study scope of Section 2.1 with Assumption 1: indeed, as an example, the bandit class $\mathcal{P}$ of the example in Section 2.3 together with Proposition 1 implies the following results.

**Proposition 6.** *Assumptions 1–3 are satisfied for all the class members in $\mathcal{P}$, but neither Assumption 4 nor Assumption 5 holds for all the members in $\mathcal{P}$.*

Consequently, in light of the connection illustrated by Proposition 5, the key difference in the study scopes and the regret bounds for Section 2.1 from the existing literature lies in the margin condition. In a synergistic manner, our regret bounds in Section 5.2 complement earlier results with the margin condition (Goldenshluger and Zeevi, 2013; Bastani and Bayati, 2020), and together verify the benefit of a margin condition to achieve a significantly improved regret rate (from polynomial to logarithmic).

*Remark* 5. The discussion above resolves the seemingly contradictory optimal regret rates for the bandit problem with high-dimensional covariates: In Section 5.2, we show that the near $N^{1/2}$ rate is optimal and is achievable by Algorithm 1, but the existing literature (Bastani and Bayati, 2020) shows that the near $\log N$ rate is optimal and is achievable by an exploitation-based algorithm. There is no conflict here since the study scope of Section 2.1 imposes no assumption on the margin (or the related constant gap condition); hence under this more "difficult" situation without assuming the margin, it is natural that the optimal regret rate is higher than the logarithmic rate; Theorem 4 has shown that no algorithm is able to give a regret rate lower than $N^{1/2}$. To some extent, this observation of different optimal regret rates is reminiscent of the intriguing debates on the optimal convergence rates (and their associated classifier rules) for nonparametric classification in the statistics literature as discussed by Tsybakov (2004, p.146):

> *How fast can the convergence of classifiers be and how does one construct the classifiers that have optimal convergence rates? ... Yang (1999) claims that the optimal rates are quite slow (substantially slower than $n^{-1/2}$) and they are attained with plug-in rules; Mammen and Tsybakov (1999) claim that the rates are fast (between $n^{-1/2}$ and $n^{-1}$) and they are attained by ERM (empirical risk minimization rules) and related classifiers. ...* **In fact, there is no contradiction since different classes of joint distributions of $(X, Y)$ are considered.** *Yang (1999) ... do not impose*

*assumption on the margin. Therefore, it is not surprising that they get rates slower than $n^{-1/2}$: one cannot obtain a rate faster than $n^{-1/2}$ with no assumptions on the margin. ... On the contrary, Mammen and Tsybakov (1999) ... show what can be achieved when ... assumption on the margin holds. In this case the fast rates (up to $n^{-1}$) are realizable.*

Therefore, the results presented in this subsection for the targeted bandit problem with covariates pleasantly join the celebrated group of known benefits by margin conditions (if satisfied) as exhibited in nonparametric estimation and nonparametric bandit problems (Tsybakov, 2004; Audibert and Tsybakov, 2007; Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013).

### 6.2. Achieving regret benefit adaptively

An important question naturally arises from our discussion in Section 6.1: since it is usually unknown whether the margin condition (or the closely related constant gap condition) holds, is it possible to design a bandit algorithm to adaptively achieve the regret benefit from the margin condition? That is, does there exist an algorithm that can simultaneously perform optimally under both of the study scopes, without or with assuming the margin, and automatically take advantage of the desirable regret benefit if the margin condition is satisfied? To a large extent, this question also resembles the spirit of adaptive performance to the margin proposed for classical classification and estimation problems (Tsybakov, 2004). In the following, we provide an affirmative answer and show that our proposed algorithm indeed adapts to the two different study scopes, and always attains near optimal regret rates (up to a logarithmic factor) regardless of whether the margin condition holds.

**Assumption 6.** If $\mathcal{I}_u \neq \varnothing$, Assumption 2 holds with $\psi = 0$.

Like Assumptions 4 and 5, Assumption 6 above for non-competitive arms was also used in Bastani and Bayati (2020), which considers a special case of Assumption 2. Now our study scope in this subsection, similar to that of Bastani and Bayati (2020), is devised to be the bandit class that imposes Assumptions 4 and 6 in addition to those of Section 2.

**Theorem 7.** *Suppose Assumptions 4 and 6 and the conditions of Theorem 3 hold. Then there exists a positive constant $\tilde{C}_2$ such that the cumulative regret of Algorithm 1 satisfies*

$$\mathrm{E}(R_N) \leq \tilde{C}_2 l q_*^2 \log p_N \log N, \tag{15}$$

*with $\tilde{C}_2 = 4\theta b c_0 + 6\theta b + 32\theta^2 \tilde{c}_\rho$.*

Using the same algorithm designed in Section 3, Theorem 7 shows that under the margin condition, our algorithm also enjoys a nearly optimal regret rate up to a logarithmic factor (the lower bound is given by Goldenshluger and Zeevi, 2013); for example, if $l$ and $q_*$ are upper bounded and $p = o(N^\zeta)$ with some constant $\zeta > 0$, then the regret upper bound in Theorem 7 is simplified to $O((\log N)^2)$. The upper bound here slightly improves on the result in Bastani and Bayati (2020) by removing an additive term of $O((\log p)^2)$. This result together with Theorem 3 and Theorem 4 confirms that our proposed algorithm simultaneously enjoys near optimal performance under both study scopes given in Section 2.1 and Section 6.

In addition, as the conditions of Theorem 3 are still satisfied here, the variable selection consistency results of Theorem 6 for the proposed algorithm continue to hold under the margin.

*Remark* 6. For studying Algorithm 1 in the previous two sections, to help maintain the "good" events of arm elimination and selection required by Propositions 2 and 3 with high probabilities, the coefficients used in parameters $\tau_0$, $\delta_N$, and $\Delta_k$ involve $q_*$, an upper bound of $\max_{i\in\mathcal{I}} q_i$ at the same order. We can also replace $q_*$ with a general upper bound $s_*$ ($s_* \geq q_*$) in setting these coefficients; then the proofs remain largely the same, although as a mild compromise, in the regret upper bounds of Theorems 3 and 7, $q_*$ should be replaced by $s_*$ as well. We note that the use of a general upper bound $s_*$ in setting algorithm parameter coefficients for theoretical development was also required in the related literature; for example, the regret bound in Theorem 7 becomes $O(ls_*^2 \log p_N \log N)$, and the quadratic rate of $s_*$ matches the result of Bastani and Bayati (2020), which required both Assumption 4 and Assumption 5. In addition, the regret lower bounds with the margin (Goldenshluger and Zeevi, 2013) and without the margin (Theorem 4) are both in respect of $N$ only. It remains unclear whether $s_*$ can be unknown to an algorithm and whether a matching bound for $s_*$ can be obtained. We leave these as open challenging questions for future investigation.

## 7. Simulation

We next evaluate the performance of the proposed bandit algorithms on simulated data. For brevity, the **m**ulti-**s**tage type algorithms described in Section 3 are abbreviated as "MS". We considered IGA and lasso as the methods for coefficient estimation and denote the corresponding bandit algorithms by MS-IGA and MS-lasso. For comparison, we used the MS algorithm without any covariates (denoted by MS-simple), that is, the mean reward estimates in Algorithm 1 were replaced by the simple average of the accumulated response values of each arm. We also

considered the bandit algorithm in Bastani and Bayati (2020) as a useful benchmark (denoted by B-lasso). Due to the page limit, all simulation settings and results are relegated to Supplement C, where we evaluate the performance of the proposed algorithms in Supplement C.1 and perform a sensitivity analysis on parameter choice in Supplement C.2.

## 8. Real data evaluation

We next use two real data sets to evaluate the performance of the proposed algorithm. One challenge naturally arises due to the incomplete nature of the data sets for the bandit setting: unlike simulation, for each user visit, we only observe the user response to one selected arm. To account for such limited feedback, the following two data sets require different evaluation strategies, which will be described in their respective subsections. In addition, to achieve faster computation for MS-IGA, we used the gradient-version of Algorithm 2 that replaces criterion (8) with (9). The parameters were chosen the same way as discussed in Supplement C.2.

### 8.1. Warfarin dose assignment

Warfarin is a widely used anticoagulant, and its appropriate dosing is important for the prevention of adverse events (International Warfarin Pharmacogenetics Consortium, 2009). The warfarin data set (available from `https://www.pharmgkb.org`) contains 6922 patient records, each of which has covariate information including demographic variables (e.g., gender, ethnicity, age), clinical background variables (e.g., height, weight, comorbidities, medication, smoking), and genotypic variables (CYP2C9 and VKORC1 genetic variants). We converted categorical variables to corresponding binary indicators and replaced missing values by the respective sample means, which resulted in 127 covariates for each patient. In addition, the continuous outcome variable was the stable therapeutic dose of warfarin, and we included 6037 patients for bandit algorithm evaluation after removing records with missing dose values.

To generate bandit arms, we categorized the outcome variable by grouping it to $l$ ($l = 2, 3, 4$) categories, using the $l$-quantiles as breaking points (that is, we used median for $l = 2$, tertiles for $l = 3$, and quartiles for $l = 4$) so that each arm (or category) in the data set corresponds to approximately the same number of patients. Since the outcome variable is the doctor-prescribed steady-state dose values that gave stable anticoagulation levels, if the therapeutic dose value fell in the category of an arm $i^*$, we set this arm $i^*$ to be the patient's optimal arm with reward 1, while all the other arms $j$ ($j \neq i^*$) were considered sub-optimal with reward 0. This setting

allowed us to evaluate any bandit algorithm: an algorithm incurs no regret if it chooses $i^*$ for the patient, and incurs unit regret otherwise. We randomized the order of patient visits and ran the bandit algorithms sequentially over the whole data set to record the final per-round regret $r_N$, the sample size of each chosen arm $n_i$, and the number of selected variables $\text{nVar}_i$ $(i = 1, \cdots, l)$. The experiment was repeated 100 times with permuted visit orders; the averaged results are summarized in Figure 1 and Table 1.

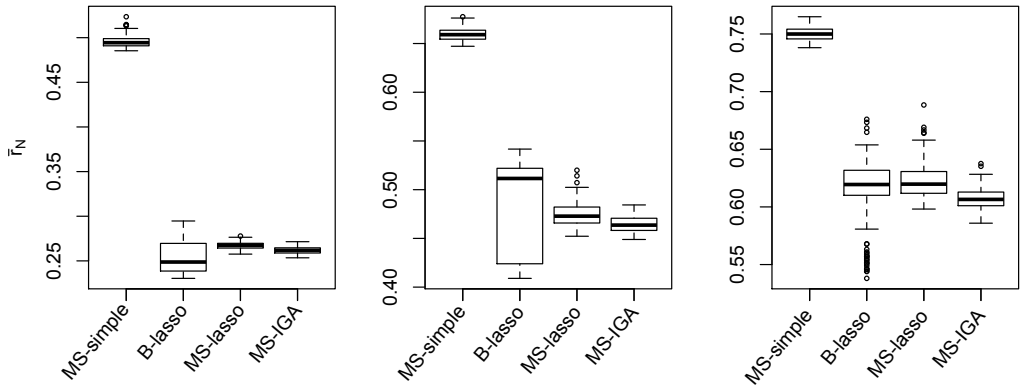


Figure 1: Boxplots of per-round regret from different bandit algorithms using warfarin dose data with 100 random permutations. Left panel: 2 arms; middle panel: 3 arms; right panel: 4 arms.

Table 1: Averaged algorithm performance using warfarin dose data with 100 random permutations.

| | 2 arms | | 3 arms | | | 4 arms | | | |
|---|---|---|---|---|---|---|---|---|---|
| Arm $i$ | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 3 | 4 |
| $\bar{n}_i$ | | | | | | | | | |
| MS-simple | 4493 | 1544 | 2004 | 3225 | 808 | 1799 | 1825 | 1299 | 1114 |
| B-lasso | 3124 | 2913 | 2361 | 2621 | 1055 | 2334 | 1120 | 2073 | 510 |
| MS-lasso | 3025 | 3012 | 2242 | 1685 | 2110 | 2001 | 1047 | 1079 | 1910 |
| MS-IGA | 3041 | 2996 | 2194 | 1744 | 2099 | 1905 | 1123 | 1148 | 1861 |
| $\overline{\text{nVar}_i}$ | | | | | | | | | |
| B-lasso | 28.45 | 27.66 | 29.29 | 23.47 | 15.92 | 29.40 | 25.37 | 23.12 | 7.41 |
| MS-lasso | 27.57 | 28.70 | 25.31 | 6.07 | 24.94 | 24.41 | 1.06 | 1.52 | 24.72 |
| MS-IGA | 16.17 | 20.81 | 15.77 | 4.73 | 19.55 | 16.59 | 5.60 | 4.51 | 19.58 |
| $\bar{r}_N$ | | | | | | | | | |
| MS-simple | 0.495 | (0.001) | | 0.659 | (0.001) | | 0.750 | (0.001) | |
| B-lasso | 0.254 | (0.003) | | 0.476 | (0.005) | | 0.611 | (0.004) | |
| MS-lasso | 0.267 | (0.001) | | 0.474 | (0.001) | | 0.623 | (0.002) | |
| MS-IGA | 0.261 | (0.001) | | 0.464 | (0.001) | | 0.607 | (0.001) | |

The boxplots from Figure 1 show that MS-simple without considering covariates yielded the least favorable performance in all three scenarios, indicating the effectiveness of using covariate information in choosing warfarin dose. Together with Table 1, we observe that MS-IGA

performed better than MS-lasso in these scenarios; MS-IGA also performed very competitively compared to the benchmark and had reduced variability in per-round regret. In addition, the averaged sample sizes of different arms appear more balanced for MS-IGA than for the benchmark, particularly under the 3-arm and 4-arm scenarios; to some extent, this may reflect the less greedy nature of the proposed algorithm. MS-IGA often selected fewer variables than the benchmark; the exceptions come from arm 3 of the 3-arm scenario and arm 4 of the 4-arm scenario as these arms were chosen less often than the other candidate arms by the benchmark.

### 8.2. News article recommendation

In the following, we use the Yahoo! front page user click log data set (version 2.0; Yahoo! Academic Relations, 2011; available from `http://webscope.sandbox.yahoo.com`). The complete set includes about 28 million user visits to the news front page from October 2 to 16, 2011, and each user visit record has 135 binary user covariates and a pool of candidate news articles. One article is chosen uniformly at random from the pool and is displayed to the user; the binary user response to the selected article is also recorded, with 1 for click and 0 for non-click. As the candidate pools of news articles are dynamic and the popularity of a news article can change in the long run, to account for these complications in algorithm evaluation, we adopted a screening strategy similar to May et al. (2012) and only considered short-term performance using data collected on the first day (October 2, 2011) with a three-article (id 563115, 563846, 565822) set as the stationary candidate arms. Accordingly, we retained the user visit records where the candidate pool contained all three articles and the displayed article was one of them. The resulting reduced data set contained 148,341 user visits for subsequent bandit algorithm evaluation.

Unlike the warfarin dose data, since a randomly selected news article is displayed at each visit, we should not assume the optimal arm is known. Instead, we applied the unbiased offline evaluation strategy developed in Li et al. (2010) to evaluate a bandit algorithm. That is, for each user visit, if the arm chosen by the algorithm matched the displayed arm, we kept this visit as a "valid" data point for algorithm use; otherwise, this visit record was ignored and not accessible by the algorithm. Accordingly, each algorithm ran through the data set sequentially until $N$ "valid" data points were obtained with $N = 30{,}000$; the resulting "valid" data was used to calculate the click through rate (CTR) as an unbiased evaluation of the bandit algorithm performance. We ran the MS-simple, B-lasso, and MS-IGA algorithms over a random permutation of the reduced data set and repeated the experiment 100 times. We used the averaged CTR from a complete

random strategy (that chose arms uniformly at random) to generate each algorithm's relative CRT by computing the ratio between the algorithm's CRT and that of the complete random strategy. We then summarized the numerical results in Figure 2 and Table 2.

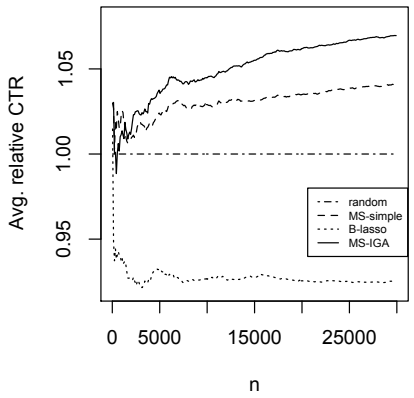


Figure 2: Averaged relative CRT with news article recommendation data.

Table 2: Averaged algorithm performance with news article recommendation data.

| | | MS-simple | B-lasso | MS-IGA | MS-B-lasso |
|---|---|---|---|---|---|
| Avg. relative $\text{CTR}_N$ | | 1.040 (0.003) | 0.924 (0.003) | 1.070 (0.003) | 1.070 (0.003) |
| $\bar{n}_i$ | | | | | |
| | arm 1 | 4358 | 29235 | 7373 | 6760 |
| | arm 2 | 7092 | 526 | 8960 | 8869 |
| | arm 3 | 18550 | 239 | 13667 | 14371 |
| $\overline{\text{nVar}_i}$ | | | | | |
| | arm 1 | - | 8.34 | 4.78 | 9.27 |
| | arm 2 | - | 0.26 | 3.99 | 7.89 |
| | arm 3 | - | 0.04 | 7.38 | 8.76 |

Compared to the complete random strategy, we observe from the plots in Figure 2 that MS-simple (without considering covariates) significantly improves the averaged CTR by about 4%. MS-IGA further improves the averaged CTR, which can be attributed to the user covariates in the reward modeling, while the benchmark surprisingly underperforms. The very unbalanced arm sample sizes from the benchmark suggest that its observed result could be again due to the more greedy nature of the benchmark designed to emphasize arm exploitation more than the MS-type algorithms; as a numerical check, we then revised the benchmark by keeping the lasso as the coefficient estimation method (with the same tuning parameter setting as B-lasso) but adopting our MS-type algorithm instead (thus we denote it by MS-B-lasso). Interestingly, as shown in Table 2, MS-B-lasso performs competitively in this case compared to MS-IGA, with less sparse variable selection outcomes and reasonably balanced sample sizes.

## 9. Discussion

We study the bandit problem with high-dimensional covariates by designing an adaptive algorithm with arm elimination and randomized allocation. The algorithm enjoys near minimax optimal regret performance under both study scopes (without or with the margin), and demonstrates adaptive performance by one unified algorithm. We also establish simultaneous coefficient estimation and variable selection consistencies for the output of the proposed algorithm. The

extensive numerical studies indicate that our proposal holds promise in real applications on personalized medical and online services. The previous discussion implicitly assumes that the total number of visits $N$ is known *a priori*; if $N$ is unknown, the proposed approach can be extended by employing the "doubling argument" (e.g., Cesa-Bianchi and Lugosi, 2006; Perchet and Rigollet, 2013). Although we only used IGA (as opposed to lasso) for Algorithm 1 to help achieve variable selection consistency with improved coefficient estimation consistency, we expect that popular shrinkage-type regression methods such as the adaptive lasso, SCAD, and MCP (Zou, 2006; Fan and Li, 2001; Zhang, 2010) could be other promising coefficient estimation candidates to be integrated for the bandit problem algorithms; a comprehensive and rigorous investigation on their theoretical and numerical properties could be of independent interest and is left for future studies.

## Supplementary Materials

Supplement to "Adaptive Algorithm for Multi-armed Bandit Problem with High-dimensional Covariates" (supplement.pdf): Supplement A provides the proofs of the propositions and the main theorems. The technical ancillary lemmas for the theorems are relegated to Supplement B. Our simulation studies are given in Supplement C.
MATLAB package for the IGA method is available at `https://github.com/weiqian1/IGA`.


## Acknowledgments

The authors sincerely thank the Editor, the Associate Editor, and three anonymous referees for their valuable comments that helped improve this manuscript significantly. Ching-Kang Ing is partially supported by the Science Vanguard Research Program of the Ministry of Science and Technology, Taiwan. Wei Qian is partially supported by NSF DMS-1916376, NIH R21NS122033A, and JPMC Faculty Fellowship.


## Disclosure Statement

The authors report there are no competing interests to declare.

## References


Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011), Improved algorithms for linear stochastic bandits, *in* 'Advances in Neural Information Processing Systems', pp. 2312–2320.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L. and Schapire, R. (2014), Taming the monster: A fast and simple algorithm for contextual bandits, *in* 'International Conference on Machine Learning', pp. 1638–1646.

Arya, S. and Yang, Y. (2020), 'Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards', *Statistics & Probability Letters* **164**, 1–9.

Audibert, J.-Y. and Tsybakov, A. B. (2007), 'Fast learning rates for plug-in classifiers', *The Annals of Statistics* **35**(2), 608–633.

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002), 'Finite-time analysis of the multiarmed bandit problem', *Machine Learning* **47**, 235–256.

Auer, P., Ortner, R. and Szepesvári, C. (2007), Improved rates for the stochastic continuum-armed bandit problem, *in* 'Proceedings of 20th Annual Conference on Learning Theory'.

Bastani, H. and Bayati, M. (2020), 'Online decision making with high-dimensional covariates', *Operations Research* **68**(1), 276–294.

Berry, D. A. and Fristedt, B. (1985), *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, New York.

Beygelzimer, A., Orabona, F. and Zhang, C. (2017), Efficient online bandit multiclass learning with order root T regret, *in* 'Proceedings of the 34th International Conference on Machine Learning', pp. 488–497.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N. and Blanchet, J. (2019), Online EXP3 learning in adversarial bandits with delayed feedback, *in* 'Advances in Neural Information Processing Systems', pp. 11349–11358.

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

Bubeck, S., Munos, R. and Stoltz, G. (2011), 'Pure exploration in finitely-armed and continuous-armed bandits', *Theoretical Computer Science* **412**(19), 1832–1852.

Candes, E. J. and Tao, T. (2005), 'Decoding by linear programming', *IEEE Transactions on Information Theory* **51**(12), 4203–4215.

Cesa-Bianchi, N. and Lugosi, G. (2006), *Prediction, Learning and Games*, Cambridge University Press, Cambridge, UK.

Chan, H. P. (2020), 'The multi-armed bandit problem: An efficient nonparametric solution', *The Annals of Statistics* **48**(1), 346–373.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. and Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101.

Gittins, J. C. (1989), *Multi-Armed Bandit Allocation Indices*, Wiley, New York.

Goldberg, Y. and Kosorok, M. R. (2012), 'Q-learning with censored data', *The Annals of Statistics* **40**(1), 529–560.

Goldenshluger, A. and Zeevi, A. (2009), 'Woodroofe's one-armed bandit problem revisited', *The Annals of Applied*

*Probability* **19**, 1603–1633.

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

Guan, M. Y. and Jiang, H. (2018), Nonparametric stochastic contextual bandits, *in* 'Proceedings of Association for the Advancement of Artificial Intelligence'.

Ing, C.-K. and Lai, T. L. (2011), 'A stepwise regression method and consistent model selection for high-dimensional sparse linear models', *Statistica Sinica* **21**(4), 1473–1513.

International Warfarin Pharmacogenetics Consortium (2009), 'Estimation of the warfarin dose with clinical and pharmacogenetic data', *New England Journal of Medicine* **360**(8), 753–764.

Kakade, S. M., Shalev-Shwartz, S. and Tewari, A. (2008), Efficient bandit algorithms for online multiclass prediction, *in* 'Proceedings of the 25th International Conference on Machine Learning', ACM, pp. 440–447.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E. and Murphy, S. A. (2014), 'Dynamic treatment regimes: Technical challenges and applications', *Electronic Journal of Statistics* **8**(1), 1225.

Laber, E. B., Meyer, N. J., Reich, B. J., Pacifici, K., Collazo, J. A. and Drake, J. M. (2018), 'Optimal treatment allocations in space and time for on-line control of an emerging infectious disease', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(4), 743–789.

Lai, T. L. (1987), 'Adaptive treatment allocation and the multi-armed bandit problem', *The Annals of Statistics* **15**, 1091–1114.

Lai, T. L. and Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**, 4–22.

Langford, J. and Zhang, T. (2008), The Epoch-Greedy algorithm for contextual multi-armed bandits, *in* 'Advances in Neural Information Processing Systems'.

Lattimore, T. and Szepesvári, C. (2020), *Bandit Algorithms*, Cambridge University Press.

Li, L., Chu, W., Langford, J. and Schapire, R. E. (2010), A contextual-bandit approach to personalized news article recommendation, *in* 'Proceedings of the 19th International World Wide Web Conference'.

Mammen, E. and Tsybakov, A. B. (1999), 'Smooth discrimination analysis', *The Annals of Statistics* **27**(6), 1808–1829.

May, B. C., Korda, N., Lee, A. and Leslie, D. S. (2012), 'Optimistic Bayesian sampling in contextual-bandit problems', *Journal of Machine Learning Research* **13**, 2069–2106.

McKeague, I. W. and Qian, M. (2014), 'Estimation of treatment policies based on functional predictors', *Statistica Sinica* **24**(3), 1461–1485.

Meinshausen, N. and Yu, B. (2009), 'Lasso-type recovery of sparse representations for high-dimensional data', *The Annals of Statistics* **37**(1), 246–270.

Murphy, S. A. (2003), 'Optimal dynamic treatment regimes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2), 331–355.

Perchet, V. and Rigollet, P. (2013), 'The multi-armed bandit problem with covariates', *The Annals of Statistics* **41**, 693–721.

Qian, M. and Murphy, S. A. (2011), 'Performance guarantees for individualized treatment rules', *The Annals of Statistics* **39**(2), 1180.

Qian, W., Ding, S. and Cook, R. D. (2019a), 'Sparse minimum discrepancy approach to sufficient dimension

reduction with simultaneous variable selection in ultrahigh dimension', *Journal of the American Statistical Association* **114**(527), 1277–1290.

Qian, W., Li, W., Sogawa, Y., Fujimaki, R., Yang, X. and Liu, J. (2019b), 'An interactive greedy approach to group sparsity in high dimensions', *Technometrics* **61**(3), 409–421.

Qian, W. and Yang, Y. (2016*a*), 'Kernel estimation and model combination in a bandit problem with covariates', *Journal of Machine Learning Research* **17**(149), 1–37.

Qian, W. and Yang, Y. (2016*b*), 'Randomized allocation with arm elimination in a bandit problem with covariates', *Electronic Journal of Statistics* **10**(1), 242–270.

Reeve, H., Mellor, J. and Brown, G. (2018), The K-nearest neighbour UCB algorithm for multi-armed bandits with covariates, *in* F. Janoos, M. Mohri and K. Sridharan, eds, 'Proceedings of Machine Learning Research', Vol. 83, pp. 725–752.

Rigollet, P. and Zeevi, A. (2010), Nonparametric bandits with covariates, *in* 'Proceedings of the 23rd International Conference on Learning Theory', Omnipress, pp. 54–66.

Robbins, H. (1954), 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society* **58**, 527–535.

Shi, C., Fan, A., Song, R. and Lu, W. (2018), 'High-dimensional A-learning for optimal dynamic treatment regimes', *The Annals of Statistics* **46**(3), 925–957.

Tsybakov, A. B. (2004), 'Optimal aggregation of classifiers in statistical learning', *The Annals of Statistics* **32**, 135–166.

Woodroofe, M. (1979), 'A one-armed bandit problem with a concomitant variable', *Journal of the American Statistical Association* **74**, 799–806.

Yahoo! Academic Relations (2011), 'Yahoo! front page today module user click log dataset, version 2.0'. Available from http://webscope.sandbox.yahoo.com.

Yang, Y. (1999), 'Minimax nonparametric classification—Part I. Rates of convergence', *IEEE Transactions on Information Theory* **45**(7), 2271–2284.

Yang, Y. and Zhu, D. (2002), 'Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates', *The Annals of Statistics* **30**, 100–121.

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**(2), 894–942.

Zhang, T. (2011*a*), 'Adaptive forward-backward greedy algorithm for learning sparse representations', *IEEE Transactions on Information Theory* **57**(7), 4689–4708.

Zhang, T. (2011*b*), 'Sparse recovery with orthogonal matching pursuit under RIP', *IEEE Transactions on Information Theory* **57**(9), 6215–6221.

Zhou, L. (2015), 'A survey on contextual multi-armed bandits', *arXiv preprint arXiv:1508.03326* .

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.