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

Wei Qian, Ching-Kang Ing and Ji Liu *

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

*Wei Qian is Associate Professor, Department of Applied Economics and Statistics, University of Delaware. Ching-Kang Ing is Professor, Institute of Statistics, National Tsing Hua University, Taiwan. Ji Liu is Assistant Professor, Department of Computer Science, University of Rochester. C.-K. Ing is partially supported by Vanguard Research Program of the Ministry of Science and Technology, Taiwan. W. Qian is partially supported by NSF DMS-1916376, NIH R21NS122033A, and JPMC Faculty Fellowship.
Correspondence: cking@stat.nthu.edu.tw.edu

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. The proofs of propositions and main theorems as well as technical lemmas, ancillary expositions, and additional numerical results are all relegated to the Supplement. In particular, we provide the general reader with an illustrative exposition in Supplement A for a multi-stage algorithm and its analysis under the classical stochastic bandits, which may yield a better intuition of the main contents of our high-dimensional counterparts (see Remark 7 in Supplement A).

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 $f_i(\mathbf{X}) := \mathrm{E}(Y_i \mid \mathbf{X}) = \mathbf{X}^T \boldsymbol{\beta}_i$, 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 $\tilde{\zeta}_N \geq \frac{c_2}{N^{\psi} \vee (\log N)^{1/2}}$ 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 $\Sigma_i = \mathrm{E}(\mathbf{X}\mathbf{X}^T \mid \mathbf{X} \in \mathcal{T}_i)$; 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 Σ_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 $\mathrm{E}(e^{v\varepsilon_{i,n}} \mid \mathbf{X}_n) \le \exp(v^2\sigma^2/2)$ 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 $i^*(\mathbf{x}) = \operatorname{argmax}_{i\in\mathcal{I}} f_i(\mathbf{x})$ 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 $\hat{\mathcal{V}}_i = \{1 \le j \le p : |\hat{\beta}_{ij}| > 0\}$. 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 ϑ_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 10 and 11 in Supplement B.2). We can then construct a sequence of its members with both coefficient parameters κ and ω 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*

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

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,

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

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, $\mathrm{row}_n(\mathbb{X}_\mathcal{A}) = \mathrm{row}_{j_n}(\mathbb{X}_N)$ and $\mathrm{row}_n(\mathbf{y}_\mathcal{A}) = \mathrm{row}_{j_n}(\mathbf{y}_N)$ for $1 \leq n \leq |\mathcal{A}|$. We can apply a specified high-dimensional linear regression method with tuning parameter ξ 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

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
         $\tilde{\mathcal{S}}_n := \{i \in \mathcal{I} : \max_{j \in \mathcal{I}} \mathbf{X}_n^T \tilde{\boldsymbol{\beta}}_j - \mathbf{X}_n^T \tilde{\boldsymbol{\beta}}_i \leq \delta_N\}.$   (6)
   • If $k > 1$, eliminate arms on $\tilde{\mathcal{S}}_n$ to generate the set of "promising" arms
         $\hat{\mathcal{S}}_n := \{i \in \tilde{\mathcal{S}}_n : \max_{j \in \tilde{\mathcal{S}}_n} \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{j,k} - \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{i,k} \leq \Delta_k\};$   (7)
     otherwise, set $\hat{\mathcal{S}}_n = \tilde{\mathcal{S}}_n$.
   • Define $\hat{I}_n = \mathrm{argmax}_{i \in \hat{\mathcal{S}}_n} \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_{i,k}$. 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 $i \in \hat{\mathcal{I}}_N$.
```

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, ξ 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. 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 ξ_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_ρ 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.

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
        $G_\rho = \{g \notin G^{(r)} : Q(\boldsymbol{\beta}^{(r)}) - \min_{\alpha \in \mathbb{R}} Q(\boldsymbol{\beta}^{(r)} + \alpha \mathbf{e}_g) \geq \rho \phi^{(r)}\}.$    (8)
    (b) Select element $g^{(r)} \in G_\rho$ and set $G^{(r+1)} = G^{(r)} \cup \{g^{(r)}\}$.
    (c) Compute $\boldsymbol{\beta}^{(r+1)} = \arg\min_{\mathrm{supp}(\boldsymbol{\beta}) \in G^{(r+1)}} Q(\boldsymbol{\beta})$, 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 $\boldsymbol{\beta}^{(r)} = \arg\min_{\mathrm{supp}(\boldsymbol{\beta}) \in G^{(r)}} Q(\boldsymbol{\beta})$.
    (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 $\boldsymbol{\beta}^{(r)}$.
```

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 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 ξ, the output of Algorithm 2 gives the coefficient estimator $\hat{\boldsymbol{\beta}}(\mathbb{X}, \mathbf{y}, \xi)$. The parameter ξ regulates the solution sparsity: a larger ξ tends to provide a sparser solution. In empirical studies, instead of giving explicit values for ξ, 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 and express the algorithm description as parallel as possible with Algorithm 3 of Supplement A. 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 $2 \le k \le K$, define

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

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

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

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 B.3.

Proposition 2. *Given stage k ($k \ge 1$), if the event U_k holds, then at any visit point n ($\tilde{N}_{k-1}+1 \le n \le \tilde{N}_k$), 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 ($k \ge 2$), if the event U_k holds, then at any visit point n ($\tilde{N}_{k-1}+1 \le n \le \tilde{N}_k$), 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 $\mathcal{U}_{n,k} = \{j \in \mathcal{I}_o : \mathbf{X}_n^T \boldsymbol{\beta}_{I_n^*} - \mathbf{X}_n^T \boldsymbol{\beta}_j \le 2\Delta_k\}$.*

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 B.3. Note that ξ_0 and the ξ_k's correspond to the tuning parameter ξ 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*

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

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,

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

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 θ 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 B.5 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 C, 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 δ. 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 τ_0, δ_N, and Δ_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 D, where we evaluate the performance of the proposed algorithms in Supplement D.1 and perform a sensitivity analysis on parameter choice in Supplement D.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 D.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 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.

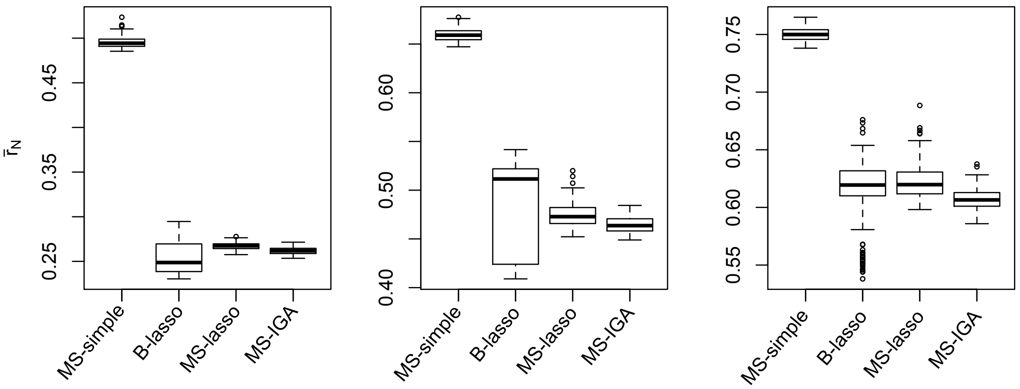

Figure 1: Boxplots of per-round regret from different bandit algorithms using warfarin dose data with 100 random permutations. Left panel: 2 arms; middle panel: 3 arms; right panel: 4 arms.

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.

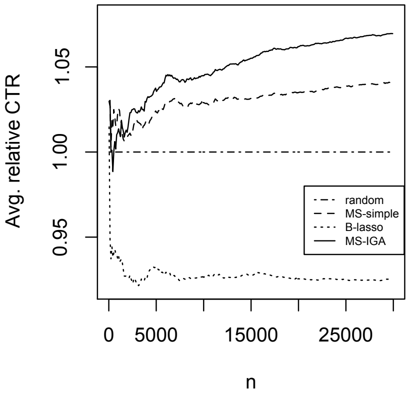

Figure 2: Averaged relative CRT with news article recommendation data.

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

	MS-simple	B-lasso	MS-IGA	MS-B-lasso
Avg. relative 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.

ACKNOWLEDGMENTS

The authors sincerely thank the Editor, the Associate Editor, and three anonymous referees for their valuable comments that helped improve this manuscript significantly.

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.

Allasia, G. (1981), 'Approximation of the normal distribution function by means of a spline function', *Statistica* **41**(2), 325–332.

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. and Ortner, R. (2010), 'UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem', *Periodica Mathematica Hungarica* **61**(1-2), 55–65.

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.

Beck, A. and Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

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.

Bickel, P. J. and Levina, E. (2008), 'Covariance regularization by thresholding', *The Annals of Statistics* **36**(6), 2577–2604.

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.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

Gill, R. D. and Levit, B. Y. (1995), 'Applications of the van Trees inequality: a Bayesian Cramér-Rao bound', *Bernoulli* **1**(1-2), 59–79.

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.

Marengo, J. E., Farnsworth, D. L. and Stefanic, L. (2017), 'A geometric derivation of the Irwin-Hall distribution', *International Journal of Mathematics and Mathematical Sciences* pp. 1–6.

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.

Supplement to “Adaptive Algorithm for Multi-armed Bandit Problem with High-dimensional Covariates”

A. Illustrative exposition with classical stochastic bandits

As an ancillary exposition, we use the classical setting of stochastic bandits (without considering covariates) to help gain intuition for a general reader on useful elements of our proposal for the bandits with high-dimensional covariates. Suppose an agent is faced with l ($l \geq 2$) candidate arms from a set $\mathcal{I} = \{1, \cdots, l\}$, where each arm represents a candidate decision. At each visit point n, if arm i is chosen, then the reward $Y_{i,n}$ is revealed from some unknown distribution F_i. The mean of $Y_{i,n}$ is denoted by μ_i. We allow the distribution to be sub-Gaussian, that is, there is a constant σ (with known upper bound) such that $\mathrm{E}(e^{v(Y_{i,n}-\mu_i)}) \leq \exp(v^2\sigma^2/2)$ for all $i \in \mathcal{I}$ and $v \in \mathbb{R}$; this subsumes binary outcomes as a special case.

With the aim of achieving the maximum (mean) reward, it would be ideal to always choose the optimal arm $i^* = \arg\max_{i\in\mathcal{I}} \mu_i$ with the optimal mean reward $\mu^* := \mu_{i^*}$; however, this “oracle” strategy is impractical due to the lack of knowledge in μ_i. Define $\Lambda_i = \mu^* - \mu_i$ to be the (*unknown*) mean reward difference between the optimal arm and arm i. Then if the agent chooses an arm i other than the optimal arm, we say that a positive *regret* of Λ_i is incurred; otherwise, the regret is 0. For technical convenience of the exposition, assume i^* is unique and Λ_i’s are upper bounded by a positive constant $c_{\bar{\Lambda}}$.

Given a finite number N of user visits, the agent must make sequential decisions: at each visit point n ($1 \leq n \leq N$), the agent chooses an arm I_n and observes the reward $Y_{I_n,n}$ realized from the unknown distribution F_{I_n}. For any sequential arm allocation rule, the arms $I_2, I_3, \cdots, I_n, \cdots$ chosen can only depend on I_1 and $Y_{I_1,1}$, on (I_1, I_2) and $(Y_{I_1,1}, Y_{I_2,2}), \cdots$, on $(I_1, \cdots, I_{n-1})$ and $(Y_{I_1,1}, \cdots, Y_{I_{n-1},n-1}), \cdots$ respectively. We define the *cumulative regret* to be the sum of all the regrets incurred within the N visits:

$$R_N = \sum_{n=1}^{N} (\mu^* - \mu_{I_n}). \tag{A.1}$$

The main goal of a stochastic bandit problem is to devise a sequential decision making algorithm to achieve low cumulative regret.

A.1. Multi-stage algorithm with arm elimination

Next, we use the classical stochastic bandit setting to design a multi-stage algorithm with arm elimination and analyze how it performs in a stage-wise fashion. The multi-stage algorithm is summarized in Algorithm 3. Specifically, we divide the N visits into multiple stages and define $\tilde{N}_k$ to be the end visit point of stage k with $k = 0, 1, \cdots, K$, where K denotes the last stage. **Step 1** is the algorithm initialization by stage $k = 0$ whose sample size is $N_0 := \tilde{N}_0 = l\tau_0$, where $\tau_0 \asymp \log N$. The set of candidate arms is $\hat{\mathcal{S}}_0 = \mathcal{I}$.

Step 2 performs arm elimination for the subsequent stage k ($1 \leq k \leq K$), and the stage sample size is denoted by $N_k := \tilde{N}_k - \tilde{N}_{k-1}$; at stage k, given l_k "promising" arms (to be defined in (A.2)), we set $N_k = l_k\tau_k$ and $\tau_k = c\tau_{k-1}$ with multiplicative factor $c = 2$ to simply double the allocated sample size for each "promising" arm. The key parameters ζ_k's ($1 \leq k \leq K$) are used in the arm elimination step of each stage k. At the beginning of each stage $k \geq 1$, the algorithm generates the set $\hat{\mathcal{S}}_k$ of "promising" arms so that $\hat{\mathcal{S}}_k$ serves as the new set of candidate arms for the current stage k: based on a data sample generated from the previous stage $k-1$, define $\hat{\mu}_{i,k-1}$ to be the sample mean of arm $i \in \hat{\mathcal{S}}_{k-1}$; then the set of "promising" arms is defined as

$$\hat{\mathcal{S}}_k := \Big\{ i \in \hat{\mathcal{S}}_{k-1} : \max_{j \in \hat{\mathcal{S}}_{k-1}} \hat{\mu}_{j,k-1} - \hat{\mu}_{i,k-1} \leq \zeta_k/2 \Big\}, \tag{A.2}$$

where we set $\zeta_{k+1} = \lambda\sqrt{\frac{\log(N/\tau_{k,N})}{\tau_k}}$ and $\tau_{k,N} = \min(\tau_k, N/2)$, and λ is some specified constant with $\lambda \geq \max(8\sigma, \sqrt{2/\log 2})$; here, $\tau_{k,N}$ and τ_k are empirically equivalent, but we use $\tau_{k,N}$ in the definition of ζ_{k+1} due to technical convenience for the analysis in Section A.2 to ensure that ζ_{k+1} remains well-defined for an arbitrarily large integer $k \geq 1$. All the arms in $\hat{\mathcal{S}}_{k-1} \backslash \hat{\mathcal{S}}_k$ are literally eliminated from stage k, and there remain $l_k = |\hat{\mathcal{S}}_k|$ arms.

Then in **Step 3**, the algorithm repeatedly cycles over the l_k "promising" arms; the sample size of each arm is τ_k. In practice, it is often preferable to choose arms using randomization, as shown in Section 3.

A.2. Understanding algorithm performance

We next provide some analysis for Algorithm 3 in terms of the cumulative regret defined in (A.1). Roughly speaking, the regret analysis lies in understanding the maintenance of "good" events as follows.

Define arm sets $\mathcal{M}_0 = \{i \in \mathcal{I} : \Lambda_i \geq \zeta_1\}$ and $\mathcal{M}_k = \{i \in \mathcal{I} : \zeta_{k+1} < \Lambda_i \leq \zeta_k\}$ with $k \geq 1$. For each arm i ($i \neq i^*$), define k_i to be the unique stage number associated with arm

Algorithm 3 A multi-stage approach to classical stochastic bandits.

1. Set the initial stage $k = 0$ with sample size $N_0 = l\tau_0$ and arm set $\hat{\mathcal{S}}_0 = \mathcal{I}$. Choose each arm τ_0 times. Then set the next stage $k = 1$.
2. At stage k, find the set $\hat{\mathcal{S}}_k$ of "promising" arms by (A.2), that is,
$$\hat{\mathcal{S}}_k := \Big\{ i \in \hat{\mathcal{S}}_{k-1} : \max_{j \in \hat{\mathcal{S}}_{k-1}} \hat{\mu}_{j,k-1} - \hat{\mu}_{i,k-1} \leq \zeta_k/2 \Big\},$$
where $\hat{\mu}_{i,k-1}$ is the sample mean of arm $i \in \hat{\mathcal{S}}_{k-1}$ and ζ_k is a user-specified arm elimination parameter.
3. For $n = N_{k-1}+1, N_{k-1}+2, \cdots, N_k$, choose arms by repeatedly cycling over the "promising" arms so that each of them is sampled $\tau_k = 2\tau_{k-1}$ times during stage k.
4. Set the next stage $k = k + 1$. Repeat steps 2–3 until the end of N visits.

i such that $i \in \mathcal{M}_{k_i}$. Without loss of generality, we assume ordered arm indices such that $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_l = \mu^*$, which implies that $k_1 \leq k_2 \leq \cdots \leq k_{l-1}$. Then for each (non-optimal) arm i, define the "good" events

$$G_{i,1} := \{\text{the optimal arm } i^* \text{ remains a "promising" arm in } \hat{\mathcal{S}}_{k_i+1} \text{ at stage } k_i + 1\},$$

$$G_{i,2} := \{\text{each arm } j \text{ (for } j \leq i) \text{ is } \textit{not} \text{ a "promising" arm in } \hat{\mathcal{S}}_{k_j+1} \text{ at stage } k_j + 1\},$$

and $G_i := G_{i,1} \cap G_{i,2}$. Here, G_i is considered to be "good" events since any arm j ($j \leq i$) in $\cup_{k=0}^{k_i} \mathcal{M}_k$ is eliminated after $\tilde{N}_{k_j}$ while the optimal arm i^* remains "promising". Then the following two propositions provide insight into conditions for the maintenance of $G_{i,1}$ and $G_{i,2}$, respectively. Define G_0 to be the sample space and set $k_0 = 0$; also define events

$$\tilde{F}_{i,k} := \big\{ |\hat{\mu}_{i,k} - \mu_i| < \zeta_k/4 \text{ and } |\hat{\mu}_{i^*,k} - \mu^*| < \zeta_k/4 \big\} \text{ for } i^*, i \in \hat{\mathcal{S}}_k; \tag{A.3}$$

$$\tilde{F}_i := \big\{ |\hat{\mu}_{i,k_i} - \mu_i| < \zeta_k/4 \text{ and } |\hat{\mu}_{i^*,k_i} - \mu^*| < \zeta_k/4 \big\} \text{ for } i^*, i \in \hat{\mathcal{S}}_{k_i}. \tag{A.4}$$

Proposition 7. *Suppose G_{i-1} holds. Given any arm j with $i \leq j < l$, if at some stage k ($k_{i-1} + 1 \leq k \leq k_i$), both arm i^* and arm j remain "promising" in $\hat{\mathcal{S}}_k$ and the event $\tilde{F}_{j,k}$ holds, then at stage $k + 1$, the optimal arm i^* cannot be eliminated by arm j, that is, we have $\hat{\mu}_{j,k} - \hat{\mu}_{i^*,k} \leq \zeta_k/2$.*

Proposition 8. *Suppose G_{i-1} and $G_{i,1}$ hold. Then if arm i remains "promising" in $\hat{\mathcal{S}}_{k_i}$ and the event $\tilde{F}_i$ holds, then the event $G_{i,2}$ also holds.*

The two propositions above demonstrate that understanding the estimation errors for the mean rewards μ_i shown in (A.3) and (A.4) is important for analysis of "good" events. Using

Bernstein-type inequalities for the estimation errors, the probability upper bounds for "bad" events can be found as follows.

Proposition 9. *Given arm i ($i \neq i^*$), we have $P(G_{i-1} \cap G_{i,1}^c) \leq 8(l-i)(\tau_{k_i} - \tau_{k_{i-1}})/N$ and $P(G_{i-1} \cap G_{i,1} \cap G_{i,2}^c) \leq 2\tau_{k_i}/N$.*

Then note that given any arm l_0 with $l_0 < l$, the whole sample space can be partitioned into the events

$$G_{i-1} \cap G_i^c = (G_{i-1} \cap G_{i,1}^c) \cup (G_{i-1} \cap G_{i,1} \cap G_{i,2}^c) \text{ and } G_{l_0} \quad \text{for } 1 \leq i \leq l_0, \tag{A.5}$$

which allows us to decompose the expected cumulative regret accordingly. Together with Proposition 9 and the definition of k_i in which $\zeta_{k_i+1} \leq \Lambda_i < \zeta_{k_i}$, additional algebra provides the following cumulative regret under the classical setting. Recall that we assume Λ_i's are upper bounded by a constant $c_{\bar{\Lambda}}$.

Theorem 8. *There exist positive constants $\bar{c}_1, \bar{c}_2$ and $C > 4$ such that for any arm l_0 with $l_0 < l$, the cumulative regret satisfies*

$$\mathrm{E}(R_N) \leq \bar{c}_1 \sum_{i=1}^{l_0} \frac{\log(N\Lambda_i^2 + C)}{\Lambda_i} + \frac{\bar{c}_2(l-l_0)\log(N\Lambda_{l_0}^2 + C)}{\Lambda_{l_0}} + N\Lambda_{l_0+1}, \tag{A.6}$$

where $\bar{c}_1 = 66\tilde{C}$, $\bar{c}_2 = 64\tilde{C}$ with some positive constant $\tilde{C} \geq 4\lambda^2 + 2c_{\bar{\Lambda}}^2$.

The obtained regret bound of (A.6) matches that of the Successive Elimination algorithm seminally proposed and rigorously analyzed in Perchet and Rigollet (2013). In particular, with $l_0 = l-1$ and denoting Λ to be a lower bound of all Λ_i's ($i \neq i^*$), we can see that $\mathrm{E}(R_N) \preceq \frac{\tilde{C} l \log(N\Lambda^2 + C)}{\Lambda}$, which shows the well-known logarithmic rate expected for classical stochastic bandits (Lai and Robbins, 1985). Also, by choosing a proper l_0 (e.g., $\Lambda_{l_0} \asymp \sqrt{\tilde{C} l \log l/N}$ when such an arm exists), we get the upper bound that $\mathrm{E}(R_N) \preceq \sqrt{\tilde{C}(l \log l)N}$. Algorithm 3 is similar to the Improved UCB algorithm in Auer and Ortner (2010) but unbounded rewards along with somewhat different stage-specific sampling sizes and elimination criterion are considered; a different analysis strategy is given here in order to illustrate the relevance and potential connection with our analysis efforts for the challenging scenarios of high-dimensional covariates in the main sections.

Remark 7. We have managed to organize and deliver the relevant concepts for analysis in Section A by a mostly parallel fashion to their counterparts in Section 3 and Section 5. These relevant concepts include the multi-stage algorithm structure with embedded arm elimination;

the definition of "good" and "bad" events and the associated regret decomposition by partitioning the whole sample space with events (A.5) vs. (11); the connection between "good" events and reward (function) estimation by Propositions 7 and 8 vs. Propositions 2 and 3; probability upper bounds for "bad" events by Proposition 9 vs. Theorems 1 and 2; and assembly of cumulative regret upper bounds in Theorem 8 vs. Theorems 3 and 7.

B. Proofs of main propositions and theorems

B.1. Proofs for Supplement A

Proof of Proposition 7. Since arms i^* and j are both "promising" at stage k, it follows that

$$\hat{\mu}_{j,k} - \hat{\mu}_{i^*,k} = \hat{\mu}_{j,k} - \mu_j + \mu_j - \mu^* + \mu^* - \hat{\mu}_{i^*,k} \leq \zeta_k/4 + \zeta_k/4 = \zeta_k/2,$$

where the inequality follows by (A.3) and $\mu^* \geq \mu_j$. This completes the proof of Proposition 7. □

Proof of Proposition 8. Under $G_{i,1}$, arm i^* is retained in $\hat{\mathcal{S}}_{k_i+1}$. This implies that

$$\begin{aligned}\hat{\mu}_{i^*,k_i} - \hat{\mu}_{i,k_i} &= \hat{\mu}_{i^*,k_i} - \mu^* + \mu^* - \mu_i + \mu_i - \hat{\mu}_{i,k_i} \\ &> \Lambda_i - \zeta_{k_i+1}/2 \geq \zeta_{k_i+1} - \zeta_{k_i+1}/2 = \zeta_{k_i+1}/2,\end{aligned}$$

where the first inequality follows by (A.4) and the second inequality follows by the definition of M_{k_i}. Therefore, arm i is not a "promising" arm in $\hat{\mathcal{S}}_{k_j+1}$ and $G_{i,2}$ holds. □

Proof of Proposition 9. By Proposition 7 and the definition of G_{i-1} and $G_{i,1}^c$, we have

$$\begin{aligned}P(G_{i-1} \cap G_{i,1}^c) \leq P\big(&\text{there is a stage } k+1 \text{ with } k_{i-1}+1 \leq k \leq k_i \text{ such that the optimal arm } i^* \\ &\text{is eliminated by some arm } j \text{ with } i \leq j \leq l-1\big) \\ &\leq \sum_{k=k_{i-1}+1}^{k_i} \sum_{j=i}^{l-1} P(\tilde{F}_{j,k}^c) \leq \sum_{k=k_{i-1}+1}^{k_i} \sum_{j=i}^{l-1} \frac{4\tau_k}{N} \\ &= \sum_{j=i}^{l-1} \frac{8\tau_{k_i}}{N}\big(1 - 2^{-(k_i - k_{i-1})}\big) \leq \frac{8(l-i)(\tau_{k_i} - \tau_{k_{i-1}})}{N},\end{aligned}$$

where the second and third inequalities follow by our choice of $\lambda \geq 8\sigma$, the union bound, and Hoeffding's inequality. Also, by Proposition 8, we have

$$P(G_{i-1} \cap G_{i,1} \cap G_{i,2}^c) \leq P(\tilde{F}_i^c) \leq \frac{4\tau_{k_i}}{N}.$$

This completes the proof of Proposition 9. □

Proof of Theorem 8. By the partition of the sample space, we note that

$$
\begin{aligned}
\mathrm{E}(R_N) &= \mathrm{E}\Big(\sum_{i=1}^{l_0} R_N I(G_{i-1}\cap G_i^c)\Big) + \mathrm{E}\big(R_N I(G_{l_0})\big)\\
&\le \mathrm{E}\Big(\sum_{i=1}^{l_0}\big(N\Lambda_i + 2\sum_{j=1}^{i-1}\tau_{k_j}\Lambda_j\big)I(G_{i-1}\cap G_i^c)\Big) + \mathrm{E}\Big(\big(N\Lambda_{l_0+1} + 2\sum_{j=1}^{l_0}\tau_{k_j}\Lambda_j\big)I(G_{l_0})\Big)\\
&\le \sum_{i=1}^{l_0} N\Lambda_i P(G_{i-1}\cap G_i^c) + N\Lambda_{l_0+1} + 2\sum_{j=1}^{l_0}\tau_{k_j}\Lambda_j\\
&\le \sum_{i=1}^{l_0} N\Lambda_i\big(P(G_{i-1}\cap G_{i,1}^c) + P(G_{i-1}\cap G_{i,1}\cap G_{i,2}^c)\big) + N\Lambda_{l_0+1} + 2\sum_{j=1}^{l_0}\Lambda_j\tau_{k_j}. \qquad (A.7)
\end{aligned}
$$

Then by Proposition 9,

$$
\begin{aligned}
&\sum_{i=1}^{l_0} N\Lambda_i P(G_{i-1}\cap G_{i,1}^c) \le \sum_{i=1}^{l_0} 8\Lambda_i\sum_{j=i}^{l-1}(\tau_{k_i}-\tau_{k_{i-1}}) = 8\sum_{j=1}^{l-1}\sum_{i=1}^{l_0\wedge j}\Lambda_i(\tau_{k_i}-\tau_{k_{i-1}})\\
=&8\sum_{j=1}^{l-1}\sum_{i=1}^{l_0\wedge j}(\Lambda_i\tau_{k_i} - \Lambda_i\tau_{k_{i-1}} + \Lambda_{i+1}\tau_{k_i} - \Lambda_{i+1}\tau_{k_i}) \le 8\sum_{j=1}^{l-1}\sum_{i=1}^{l_0\wedge j}(\Lambda_i - \Lambda_{i+1})\tau_{k_i} + 8\sum_{j=1}^{l-1}\Lambda_{l_0\wedge j+1}\tau_{k_{l_0\wedge j}}.
\end{aligned}
\qquad (A.8)
$$

Also, by definition of the $\mathcal{M}_k$'s and ζ_k's, if $k_i > 0$, it is not hard to see that there is a constant $C > 4$ such that $\tau_{k_i} \le \frac{4\lambda^2}{\Lambda_i^2}\log(N\Lambda_i^2 + C)$ for $\lambda > \sqrt{2/\log 2}$. If $k_i = 0$, we note that $\tau_0 \asymp \log N$, $\Lambda_i \ge \lambda\sqrt{\frac{\log(N/\tau_0)}{\tau_0}}$, and Λ_i is upper bounded. Consequently, there exists a positive constant $\tilde{C} \ge 4\lambda^2 + 2c_{\bar{\Lambda}}^2$ such that for all k_i's,

$$
\tau_{k_i} \le \frac{\tilde{C}}{\Lambda_i^2}\log(N\Lambda_i^2 + C). \qquad (A.9)
$$

By (A.9) and integration by parts, we obtain

$$
\sum_{i=1}^{l_0\wedge j}(\Lambda_i - \Lambda_{i+1})\tau_{k_i} \le \tilde{C}\sum_{i=1}^{l_0\wedge j}(\Lambda_i - \Lambda_{i+1})\frac{\log(N\Lambda_i^2 + C)}{\Lambda_i^2} \le \frac{4\tilde{C}\log(N\Lambda_{l_0\wedge j}^2 + C)}{\Lambda_{l_0\wedge j}}.
$$

Together with (A.8), the display above implies that

$$
\sum_{i=1}^{l_0} N\Lambda_i P(G_{i-1}\cap G_{i,1}^c) \le 64\tilde{C}\sum_{j=1}^{l-1}\frac{\log(N\Lambda_{l_0\wedge j}^2 + C)}{\Lambda_{l_0\wedge j}}. \qquad (A.10)
$$

In addition, by Proposition 9, (A.9) and similar arguments from above, we have

$$
\sum_{i=1}^{l_0} N\Lambda_i P(G_{i-1}\cap G_{i,1}\cap G_{i,2}^c) + 2\sum_{j=1}^{l_0}\Lambda_j\tau_{k_j} \le 2\tilde{C}\sum_{i=1}^{l_0}\frac{\log(N\Lambda_i^2 + C)}{\Lambda_i}. \qquad (A.11)
$$

Then (A.7), (A.10), and (A.11) together imply that

$$
\mathrm{E}(R_N) \le 66\tilde{C}\sum_{i=1}^{l_0}\frac{\log(N\Lambda_i^2 + C)}{\Lambda_i} + \frac{64\tilde{C}(l - l_0)\log(N\Lambda_{l_0}^2 + C)}{\Lambda_{l_0}} + N\Lambda_{l_0+1}.
$$

Setting $\bar{c}_1 = 66\tilde{C}$ and $\bar{c}_2 = 64\tilde{C}$ from above, we complete the proof of Theorem 8. □

B.2. Proofs and propositions for Section 2.3

Proof of Proposition 1. Note that

$$P(|f_{1,N}(\mathbf{X}) - f_{2,N}(\mathbf{X})| > \tilde{\delta}_N) \le P(\frac{1}{\sqrt{q}}\sum_{j=1}^{q} X_j > \frac{N^{\alpha-\alpha'}-1}{2}) + P(\frac{1}{\sqrt{q}}\sum_{j=1}^{q} X_j < \frac{-N^{\alpha-\alpha'}+1}{2}).$$

Then (4) holds immediately by Chebyshev's inequality and $\mathrm{Var}(\frac{1}{\sqrt{q}}\sum_{j=1}^{q} X_j) = \frac{1}{3}$. □

Proposition 10. *For all the class members in $\mathcal{P}$, Assumptions 1 and 2 are both satisfied such that $\mathcal{I}_u = \varnothing$, and $P(\mathbf{X} \in \mathcal{T}_i)$ are bounded away from zero by a positive constant ($i = 1, 2$).*

Proof of Proposition 10. For any member in $\mathcal{P}$, it is clear that both arms are competitive arms; thus Assumption 2 becomes void and trivially holds. It remains to verify Assumption 1: that $P(\mathbf{X} \in \mathcal{T}_i)$ is bounded away from zero ($i = 1, 2$). Define $Z = \sqrt{\frac{3}{q}}\sum_{j=1}^{q} X_j$ and $z = \frac{\sqrt{3}}{2}$. Let $F_Z(\cdot)$ be the cumulative distribution function (CDF) of Z. Then

$$P(\mathbf{X} \in \mathcal{T}_1) = P\Big(\frac{2\kappa}{\sqrt{q}}\sum_{j=1}^{q} X_j > \omega\Big) \ge P(Z > z) \ge \Phi(-z) - |F_Z(-z) - \Phi(-z)| > 1/10,$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, and the last inequality holds by a uniform error bound for the Irwin–Hall distribution (Allasia, 1981; Marengo et al., 2017) with an approximating normal distribution such that $|F_Z(-z) - \Phi(-z)| \le \frac{\sqrt{3/q}}{20}$. The lower bound for $P(\mathbf{X} \in \mathcal{T}_2)$ can be derived similarly as above. Therefore, $P(\mathbf{X} \in \mathcal{T}_i)$ are bounded away from zero, which completes the proof of Proposition 10. □

Proposition 11. *For all the class members in $\mathcal{P}$, Assumption 3 is satisfied such that the minimum eigenvalues $\lambda_{\min}(\Sigma_i)$ are bounded away from zero by a positive constant ($i = 1, 2$).*

Proof of Proposition 11. To verify Assumption 3 for the class $\mathcal{P}$, it is sufficient to examine $\Sigma_i = (\mathbf{X}\mathbf{X}^T \,|\, \mathbf{X} \in \mathcal{T}_i)$ and verify that $\lambda_{\min}(\Sigma_i)$ is bounded away from zero. For this purpose, we can see that Σ_i has the block diagonal structure: the three diagonal block components of Σ_i are 1, Σ_{qi}, and $\frac{1}{3}I_{p-q-1}$, where Σ_{qi} is a $q \times q$ matrix with a compound symmetry structure. Specifically for Σ_{qi}, the diagonal elements are $a_i := \mathrm{E}(X_1^2 \,|\, \mathbf{X} \in \mathcal{T}_i)$ and the off-diagonal elements are $d_i := \mathrm{E}(X_1 X_2 \,|\, \mathbf{X} \in \mathcal{T}_i)$. Without loss of generality, assume that $\omega \ge 0$ and $q > 2$ (the proof can be similarly done for $\omega < 0$). Then note that

$$\Sigma_{qi} = (a_i + (q-1)d_i)\mathrm{P}_{\mathbf{1}_q} + (a_i - d_i)(I_q - \mathrm{P}_{\mathbf{1}_q}),$$

where $\mathbf{1}_q \in \mathbb{R}^q$ is the one vector, and $\mathrm{P}_{\mathbf{1}_q} = \frac{1}{q}\mathbf{1}_q\mathbf{1}_q^T$ is the projection matrix onto the subspace spanned by $\mathbf{1}_q$. Therefore, eigenvalues of Σ_{qi} are $a_i - d_i$ and $a_i + (q-1)d_i$. Then it suffices to find some positive constant lower bounds for $a_1 - d_1$ and $a_2 + (q-1)d_2$. For the former, we have

$$\begin{aligned} a_1 - d_1 &= \frac{1}{2}\mathrm{E}\big((X_1 - X_2)^2 \,|\, \sum_{j=1}^{q} X_j > \frac{\sqrt{q}\omega}{2\kappa}\big) \\ \geq & \frac{\mathrm{E}\big((X_1 - X_2)^2 I(X_1 + X_2 > \frac{\omega}{\sqrt{q}\kappa})\big)P\big(\sum_{j=3}^{q} X_j > \frac{(q-2)\omega}{2\sqrt{q}\kappa}\big)}{2P\big(\sum_{j=1}^{q} X_j > \frac{\sqrt{q}\omega}{2\kappa}\big)} \\ \geq & \frac{P\big(\frac{1}{\sqrt{q-2}}\sum_{j=3}^{q} X_j > \frac{1}{2}\big)}{2P\big(\frac{1}{\sqrt{q}}\sum_{j=1}^{q} X_j > \frac{1}{2}\big)}\mathrm{E}\big((X_1 - X_2)^2 I(X_1 + X_2 > \frac{1}{\sqrt{3}})\big) \\ \geq & \frac{\Phi(-\frac{\sqrt{3}}{2}) - \frac{1}{20}}{2\Phi(-\frac{\sqrt{3}}{2}) + \frac{1}{10}}\mathrm{E}\big((X_1 - X_2)^2 I(X_1 + X_2 > \frac{1}{\sqrt{3}})\big) > 0, \end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and the inequality of the last line holds by the uniform error bound for the Irwin–Hall distribution with an approximating normal distribution. For the latter, define $Z = \sqrt{\frac{3}{q}}\sum_{j=1}^{q} X_j$. Then we note that

$$a_2 + (q-1)d_2 = \frac{1}{q}(qa_2 + q(q-1)d_2) = \mathrm{E}\big(\frac{1}{q}(\sum_{j=1}^{q} X_j)^2 \,|\, \mathbf{X} \in \mathcal{T}_2\big)$$

$$= \frac{\mathrm{E}\big(Z^2 I(Z < \frac{\sqrt{3}\omega}{2\kappa})\big)}{3P\big(Z < \frac{\sqrt{3}\omega}{2\kappa}\big)} \geq \frac{\mathrm{E}(Z^2)}{6\Phi(\frac{\sqrt{3}}{2}) + \frac{3}{10}} = \frac{1}{6\Phi(\frac{\sqrt{3}}{2}) + \frac{3}{10}} > 0.$$

The two displays above imply that $\lambda_{\min}(\Sigma_{qi})$ are bounded away from zero by a positive constant for all members in the class $\mathcal{P}$, and the proof of Proposition 11 is complete. □

B.3. Proofs for Section 5.1

Proof of Proposition 2 (Arm pre-screening behavior). Given $\mathbf{x} \in \mathcal{X}$, define $\tilde{\boldsymbol{\beta}}_* = \tilde{\boldsymbol{\beta}}_{i^*(\mathbf{x})}$ and $\boldsymbol{\beta}_* = \boldsymbol{\beta}_{i^*(\mathbf{x})}$. Given any arm $i \in \mathcal{I}$,

$$\mathbf{x}^T\tilde{\boldsymbol{\beta}}_i - \mathbf{x}^T\tilde{\boldsymbol{\beta}}_* = \mathbf{x}^T(\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) + \mathbf{x}^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_*) + \mathbf{x}^T(\boldsymbol{\beta}_* - \tilde{\boldsymbol{\beta}}_*) \leq 2\theta b_0 = \delta_N.$$

In addition, given any arm $i \in \mathcal{I}_u$, given $\tau_0 = c_0 q_*^2 \log p_N (N^{2\psi} \vee \log N)$ where $c_0 \geq 32\theta^2 c_\rho c_2^{-2}$, we have

$$\begin{aligned} \mathbf{x}^T\tilde{\boldsymbol{\beta}}_* - \mathbf{x}^T\tilde{\boldsymbol{\beta}}_i &= \mathbf{x}^T(\tilde{\boldsymbol{\beta}}_* - \boldsymbol{\beta}_*) + \mathbf{x}^T(\boldsymbol{\beta}_* - \boldsymbol{\beta}_i) + \mathbf{x}^T(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i) \\ > \zeta_N - 2\theta b_0 &\geq \frac{c_2}{\sqrt{N^{2\psi} \vee \log N}} - 2\theta q_*\sqrt{2c_\rho \log p_N/\tau_0} \geq \delta_N. \end{aligned}$$

Therefore, $I_n^* \in \tilde{\mathcal{S}}_n$ and $i \notin \tilde{\mathcal{S}}_n$ for any $i \in \mathcal{I}_u$. This completes the proof of Proposition 2. □

Proof of Proposition 3 (Arm elimination behavior). Given stage k and $\mathbf{X}_n$, define $\hat{\boldsymbol{\beta}}_* = \hat{\boldsymbol{\beta}}_{i^*(\mathbf{X}_n),k}$

and $\boldsymbol{\beta}_* = \boldsymbol{\beta}_{i^*(\mathbf{X}_n)}$. Then, under U_k, by Proposition 2, $I_n^* \in \tilde{\mathcal{S}}_n$ and for any $i \in \tilde{\mathcal{S}}_n$,

$$\mathbf{X}_n^T \hat{\boldsymbol{\beta}}_i - \mathbf{X}_n^T \hat{\boldsymbol{\beta}}_* = \mathbf{X}_n^T(\hat{\boldsymbol{\beta}}_{i,k} - \boldsymbol{\beta}_i) + \mathbf{X}_n^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_*) + \mathbf{X}_n^T(\boldsymbol{\beta}_* - \hat{\boldsymbol{\beta}}_*) \leq 2\theta b_k = \Delta_k.$$

Therefore, $I_n^* \in \hat{\mathcal{S}}_n$. In addition, for every $i \in \hat{\mathcal{S}}_n$,

$$\mathbf{X}_n^T \boldsymbol{\beta}_* - \mathbf{X}_n^T \boldsymbol{\beta}_i = \mathbf{X}_n^T(\boldsymbol{\beta}_* - \hat{\boldsymbol{\beta}}_*) + \mathbf{X}_n^T(\hat{\boldsymbol{\beta}}_* - \hat{\boldsymbol{\beta}}_i) + \mathbf{X}_n^T(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \leq 2\theta b_k + \Delta_k = 2\Delta_k.$$

This completes the proof of Proposition 3. □

Proof of Theorem 2. We intend to perform an analysis on the sample collected at stage k that corresponds to each arm $i \in \mathcal{I}_o$. Here we follow substeps (2a)–(2d) for objective (2) described in Section 5.1.

(2a) Randomized allocation with "random" samples. By using the randomized allocation scheme, we first account for the difficulty of analyzing the non-i.i.d. data by finding a random sample with "known" covariate properties. Define $l_o = |\mathcal{I}_o|$. Under U_k, by Proposition 3, for any $\mathbf{x} \in \mathcal{T}_i$, we have $i \in \hat{\mathcal{S}}_k(\mathbf{x})$. As a result, if $\mathbf{x} \in \mathcal{T}_i$, $P(I_n = i \,|\, \mathbf{X}_n = \mathbf{x},\, U_k) \geq l_o^{-1}$. Then we artificially divide arm i into two sub-arms $\bar{i}$ and $\tilde{i}$ so that

$$P(Z_n = 1 \,|\, \mathbf{X}_n = \mathbf{x},\, U_k) = l_o^{-1} \text{ and } P(I_n = \tilde{i} \,|\, \mathbf{X}_n = \mathbf{x},\, U_k) = P(I_n = i \,|\, \mathbf{X}_n = \mathbf{x},\, U_k) - l_o^{-1},$$

where $Z_n = I(I_n = \bar{i})$, that is, the probability of selecting arm $\bar{i}$ is a constant given any $\mathbf{x} \in \mathcal{T}_i$. Consequently, $X_n \,|\, X_n \in \mathcal{T}_i, Z_n = 1$ with $\tilde{N}_{k-1} + 1 \leq n \leq N_k$ has the same distribution as $X_n \,|\, X_n \in \mathcal{T}_i$, and a random sample for arm $\bar{i}$ with an identical covariate distribution (following $X_n \,|\, X_n \in \mathcal{T}_i$) can be obtained.

(2b) Sample size determination. To find the corresponding sample size, define $\bar{J}_{i,k} = \{\tilde{N}_{k-1} + 1 \leq n \leq \tilde{N}_k : \mathbf{X}_n \in \mathcal{T}_i,\, Z_n = 1\}$. Note that by Assumption 1, we have $P(\mathbf{X}_n \in \mathcal{T}_i,\, Z_n = 1 | U_k) \geq p_i l_o^{-1}$, where $p_i = P(\mathbf{X} \in \mathcal{T}_i)$. Therefore, by an extended Bernstein inequality (e.g., Lemma 2 in Qian and Yang, 2016*a*) and $c_0 \geq 28 c_1^{-2}$, we have

$$P\Big(|\bar{J}_{i,k}| \leq \frac{N_k p_i l_o^{-1}}{2} \,\Big|\, U_k\Big) \leq \exp\Big(-\frac{3 N_k p_i l_o^{-1}}{28}\Big) \leq \frac{1}{N^3}. \tag{A.12}$$

Denote H_k to be the event that for all $i \in \mathcal{I}_o$, $|\bar{J}_{i,k}| > \frac{N_k p_i l_o^{-1}}{2}$. Then by (A.12), under U_k, we have H_k with probability greater than $1 - l_o/N^3$.

(2c) Covariate "Design matrix" properties. Defining $J_{i,k} = \{\tilde{N}_{k-1} + 1 \leq n \leq \tilde{N}_k : I_n = i\}$ and $\tilde{J}_{i,k} = J_{i,k} \backslash \bar{J}_{i,k}$, we next look at the design matrix properties. Define $\mathbb{X}_{i,k}$, $\bar{\mathbb{X}}_{i,k}$, and $\tilde{\mathbb{X}}_{i,k}$ to be the covariate design matrix associated with $J_{i,k}$, $\bar{J}_{i,k}$, and $\tilde{J}_{i,k}$, respectively. Define $\hat{\Sigma}_{i,k} = \mathbb{X}_{i,k}^T \mathbb{X}_{i,k} / |J_{i,k}|$, $\bar{\Sigma}_{i,k} = \bar{\mathbb{X}}_{i,k}^T \bar{\mathbb{X}}_{i,k} / |\bar{J}_{i,k}|$, and $\tilde{\Sigma}_{i,k} = \tilde{\mathbb{X}}_{i,k} \tilde{\mathbb{X}}_{i,k}^T / |\tilde{J}_{i,k}|$. Define $\bar{\sigma}_{i,k} = |\bar{\Sigma}_{i,k} - \Sigma_i|_\infty$ and $\bar{\sigma}_k = \max_{i \in \mathcal{I}_o} \bar{\sigma}_{i,k}$, where $|A|_\infty$ denotes the maximum element in A. Then by (10) in Bickel and

Levina (2008) and a union bound, given $\epsilon > 0$, there exist constants $c_a, c_b > 0$ such that

$$P(\bar{\sigma}_k > \epsilon,\, H_k \,|\, U_k) \leq l_o \max_{i \in \mathcal{I}_o,\, j \geq \frac{N_k p_i l_o^{-1}}{2l}} P(\bar{\sigma}_{i,k} > \epsilon \,|\, |\bar{J}_{i,k}| = j, U_k) \leq c_a l_o p^2 \exp\Big(-\frac{N_k c_1 \epsilon^2}{2 c_b l_o}\Big).$$

Taking $\epsilon = \sqrt{\frac{C_b \log p_N}{N_k}}$ with $C_b \geq 12 c_b c_1^{-2}$, we have

$$P(D_k^c \cap H_k \,|\, U_k) \leq p_N^{-3}, \tag{A.13}$$

where $D_k = \{\bar{\sigma}_k \leq \sqrt{\frac{C_b \log p_N}{N_k}}\}$. Also, note that since

$$\hat{\Sigma}_{i,k} = \frac{|\bar{J}_{i,k}|}{|J_{i,k}|}\bar{\Sigma}_{i,k} + \frac{|\tilde{J}_{i,k}|}{|J_{i,k}|}\tilde{\Sigma}_{i,k} = \frac{|\bar{J}_{i,k}|}{|J_{i,k}|}\Sigma_i + \frac{|\bar{J}_{i,k}|}{|J_{i,k}|}(\bar{\Sigma}_{i,k} - \Sigma_i) + \frac{|\tilde{J}_{i,k}|}{|J_k|}\tilde{\Sigma}_{i,k},$$

we have under U_k, H_k, and D_k, for $q \leq q_*$, with $\mathbf{v} \in \mathbb{S}^{p-1}$,

$$\tilde{\lambda}_{i,k}(q) := \min_{\|\mathbf{v}\|_0 \leq q} \mathbf{v}^T \hat{\Sigma}_{i,k} \mathbf{v} \geq \frac{c_1^2 \lambda_i(q)}{2} - \bar{\sigma}_k \max_{\|\mathbf{v}\|_0 \leq q} \|\mathbf{v}\|_1^2 \geq \frac{c_1^2 c_*}{2} - q_* \sqrt{\frac{C_b \log p_N}{N_k}} \geq \frac{c_1^2 c_*}{4}. \tag{A.14}$$

(2d) Coefficient estimation upper bounds. To evaluate the coefficient estimation of $\hat{\boldsymbol{\beta}}_{i,k+1}$, note that given $\mathbb{X} = \mathbb{X}_{\mathcal{A}_{k,i}}$, the elements in $\mathbf{y}_{\mathcal{A}_{k,i}}$ are conditionally independent. Suppose U_k, H_k, and D_k hold and $|\mathcal{A}_{k,i}| = m$. Also assume that $|\hat{\mathcal{V}}_i| \leq q_* - q_i$ and

$$\begin{aligned} \|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2^2 \leq{}& m\big(Q(\hat{\boldsymbol{\beta}}_{i,k+1}) - Q(\boldsymbol{\beta}_i)\big) \\ &+ 2\|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2 \sigma \sqrt{c_d q_i + 2 c_f |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \log p + c_f \log(\frac{2e}{\eta})}, \end{aligned} \tag{A.15}$$

where $\eta = 2e/N^4$. Then by Lemmas 9 and 10,

$$Q(\hat{\boldsymbol{\beta}}_{i,k+1}) - Q(\boldsymbol{\beta}_i) \leq \frac{8\theta^2 \xi_{k+1}}{\tilde{\lambda}_{i,k}(q_*)} |J_{i,\tau}|, \tag{A.16}$$

where $J_{i,\tau} = \{j \in \mathcal{V}_i \backslash \hat{\mathcal{V}}_i : \beta_{i,j}^2 < \tau\}$ with $\tau = 128\theta^2 \xi_{k+1} / \tilde{\lambda}_{i,k}(q_*)$. Take $c_r' = \frac{128\theta^2\sigma^2 c_f}{c_1^4 c_* \rho}(2 + \frac{1}{8\theta^2})$. Then by our choice of ξ_{k+1} and Lemma 8,

$$2 c_f \sigma^2 \log p |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \leq \frac{m \tilde{\lambda}_{i,k}(q_*) \rho \xi_{k+1}}{16\theta^2} |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \leq \frac{1}{8} \|\mathbb{X}(\hat{\boldsymbol{\beta}}_{i,k+1} - \boldsymbol{\beta}_i)\|_2^2. \tag{A.17}$$

Then, (A.15) and (A.16) give

$$\begin{aligned} &\frac{1}{2}\|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2^2 \\ \leq{}& -\frac{1}{2}\|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2^2 \\ &\quad + 2\|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2 \sigma \sqrt{c_d q_i + 2 c_f |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \log p + c_f \log(\frac{2e}{\eta})} + \frac{8\theta^2 \xi_{k+1} m}{\tilde{\lambda}_{i,k}(q_*)} |J_{i,\tau}| \\ \leq{}& 2\sigma^2\big(c_d q_i + c_f \log(\frac{2e}{\eta})\big) + 4\sigma^2 c_f |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \log p + \frac{8\theta^2 \xi_{k+1} m}{\tilde{\lambda}_{i,k}(q_*)} |J_{i,\tau}| \\ \leq{}& 2\sigma^2\big(c_d q_i + c_f \log(\frac{2e}{\eta})\big) + \frac{1}{4}\|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2^2 + \frac{8\theta^2 \xi_{k+1} m}{\tilde{\lambda}_{i,k}(q_*)} |J_{i,\tau}|, \end{aligned}$$

where the last inequality follows by (A.17). The display above implies that

$$\tilde{\lambda}_{i,k}(q_*)\|\hat{\boldsymbol{\beta}}_{i,k+1} - \boldsymbol{\beta}_i\|_2^2 \leq \frac{1}{m}\|\mathbb{X}\hat{\boldsymbol{\beta}}_{i,k+1} - \mathbb{X}\boldsymbol{\beta}_i\|_2^2 \leq \frac{8\sigma^2}{m}\big(c_d q_i + c_f \log(\frac{2e}{\eta})\big) + \frac{8\theta^2\xi_{k+1}}{\tilde{\lambda}_{i,k}(q_*)}|J_{i,\tau}|. \quad \text{(A.18)}$$

Consequently,

$$\|\hat{\boldsymbol{\beta}}_{i,k+1} - \boldsymbol{\beta}_i\|_2^2 \leq \frac{64\sigma^2}{c_1^4 c_* N_k}(c_d q_i + 4c_f \log N) + \frac{32\theta^2 c_r' \log p_N}{c_1^2 c_* N_k}|J_{i,\tau}| \leq \frac{\tilde{c}_\rho(q_i + \log N + q_{i,k}\log p_N)}{N_k}$$

with $\tilde{c}_\rho = \frac{64\sigma^2}{c_1^4 c_*}(c_d + 4c_f) + \frac{32\theta^2 c_r'}{c_1^2 c_*}$ and $\tilde{c}_\beta = \frac{512\theta^2 c_r'}{c_1^2 c_*}$. Lastly, by Lemmas 1 and 11, (A.12) and (A.13), together with the Cauchy–Schwarz inequality, we complete the proof of Theorem 2. □

Proof of Theorem 1. The proof is similar to (and simpler due to the forced sampling with random sample) that of Theorem 2, and we can replace c_r' with $c_r = 16\theta^2\sigma^2 c_f c_*^{-1}\rho^{-1}(2+1/(8\theta^2))$, $\tilde{c}_\rho$ with $c_\rho = 8\sigma^2 c_*^{-1}(c_d + 4c_f) + 8\theta^2 c_r c_*^{-1}$, and $\tilde{c}_\beta$ with $c_\beta = 128\theta^2 c_r c_*^{-1}$ to obtain Theorem 1. Thus we omit the proof details. □

B.4. Proofs for Section 5.2

Proof of Theorem 3. First, we describe *regret decomposition.* Recall that R_{N0} and R_{N1} are the regrets accumulated in Stage 0 and the following stages, respectively. Then $R_N = R_{N0} + R_{N1}$. In addition, we partition the sample space into the events as shown in (11):

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

As a result, (12) follows that

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

Then we have $R_{N0} \leq 2\theta b l \tau_0 = 2\theta b c_0 l q_*^2 \log p_N (N^{2\psi} \vee \log N)$.

Next, to provide bounds for R_k $(1 \leq k \leq K)$, it is important to understand the properties and implications regarding these associated events. As summarized in Section 5.1, we intend to achieve two objectives: (1) Under "good" events, the regret can be properly upper bounded via a connection with coefficient/reward estimation errors; (2) The (conditional) probabilities of violating the "good" events are relatively small. To accomplish these two objectives, we further divide the proof into multiple substeps. Specifically, for objective (1), the substeps include studying (1a) *arm pre-screening behavior*; and (1b) *arm elimination behavior.* Steps (1a) and (1b) are summarized in Propositions 2 and 3, respectively. For objective (2), the overall task is summarized in Theorems 1 and 2.

After accomplishing these steps for the two objectives, by the results of Theorems 1 and 2 from objective (2), we obtain

$$\mathrm{E}(R_0) \leq 2\theta b N P(U_1^c) \leq 2\theta bl/N^2 \text{ and } \mathrm{E}(R_k) \leq 2\theta b N P(U_k \cap F_{k+1}^c) \leq 6\theta bl/N^2 \tag{A.19}$$

for $1 \leq k \leq K-1$. Also, by the results of Proposition 3 from objective (1), we have

$$R_K \leq \sum_{k=1}^{K} \sum_{n=\tilde{N}_{k-1}+1}^{\tilde{N}_k} \big(f^*(\mathbf{X}_n) - f_{I_n}(\mathbf{X}_n)\big) I(U_K) \leq 2\theta bl\tau_0 + \sum_{k=2}^{K} \sum_{n=\tilde{N}_{k-1}+1}^{\tilde{N}_k} 4\theta b_k I(U_k),$$

which implies that

$$\begin{aligned} \mathrm{E}(R_K) &\leq 2\theta b c_0 l q_*^2 \log p_N (N^{2\psi} \vee \log N) + \sum_{k=2}^{K} 4\theta b_k N_k \\ &\leq 2\theta b c_0 l q_*^2 \log p_N (N^{2\psi} \vee \log N) + 8\theta \tilde{c}_\rho^{1/2} q_* \sqrt{N \log p_N}. \end{aligned} \tag{A.20}$$

By (12), (A.19), (A.20), and setting $C_{21} = 4\theta b c_0 + 6\theta b$ and $C_{22} = 8\theta \tilde{c}_\rho^{1/2}$, we obtain (13). Lastly, noting that $lq_* \log N = o(\sqrt{\frac{N}{\log p_N}})$ obviously holds with the additional conditions for (14), we complete the proof of Theorem 3. □

Proof of Theorem 4. We prove the lower bound through the two-armed bandit class $\mathcal{P}$ defined in Section 2.3. For simplicity, we only consider $q = 1$ in the following proof since for $q > 1$, our proof leads to basically the same lower bound through normal approximation for the Irwin–Hall distribution. Also assume the random error $\varepsilon_{i,n}$ of each arm i follows $N(0, \sigma^2)$. Given N, let κ be some fixed value whose choice will be given later. We can then apply a Bayes average technique adapted from Goldenshluger and Zeevi (2013). Specifically, assume that ω is randomly drawn according to a continuous prior probability w, and let $p_w(\cdot)$ denote the density function of w and $\mathrm{E}_w(\cdot)$ denote the expectation with respect to the prior w. Then we choose

$$p_w(\omega) = \frac{1}{\kappa} \cos^2\big(\frac{\pi\omega}{2\kappa}\big) I(|\omega| \leq \kappa).$$

Given any admissible bandit strategy, let $R_N(F_\omega)$ be its cumulative regret for a class member F_ω in $\mathcal{P}$. Let Z^n be the set of past observations $\{\mathbf{X}_j, I_j, Y_{I_j,j}\}_{j=1}^{n-1}$ prior to the n-th visit, where $\mathbf{X}_n = (1, X_{n,1}, \cdots, X_{n,p-1})$. Let $\mathcal{F}_n$ and $\tilde{\mathcal{F}}_n$ be the σ-fields generated by Z^n and $(Z^n, \mathbf{X}_n)$, respectively. Then we have

$$
\begin{aligned}
&\sup_{F_\omega \in \mathcal{P}} \mathrm{E}\big(R(F_\omega)\big) \geq \mathrm{E}_w \mathrm{E}\big(R(F_\omega)\big) \\
\geq &\, \mathrm{E}_w \mathrm{E}\Big(\sum_{n=1}^{N} |2\kappa X_{n,1} - \omega| \big(I(2\kappa X_{n,1} > \omega,\, I_n = 2) + I(2\kappa X_{n,1} < \omega,\, I_n = 1)\big)\Big) \\
\geq &\, \mathrm{E}\Big(\sum_{n=1}^{N} \Big(\mathrm{E}_w\big((2\kappa X_{n,1} - \omega) I(2\kappa X_{n,1} > \omega) \,|\, \tilde{\mathcal{F}}_n\big) I(I_n = 2) \\
&\quad - \mathrm{E}_w\big((2\kappa X_{n,1} - \omega) I(2\kappa X_{n,1} < \omega) \,|\, \tilde{\mathcal{F}}_n\big) I(I_n = 1)\Big)\Big).
\end{aligned}
$$

By plugging in the decision rule that chooses arm $I_n = 2$ if $\mathrm{E}_w(2\kappa X_{n,1} - \omega \,|\, \tilde{\mathcal{F}}_n) \leq 0$ (that is, $\hat{\omega}_n \geq 2\kappa X_{n,1}$ with $\hat{\omega}_n = \mathrm{E}_w(\omega \,|\, \mathcal{F}_n)$) and arm $I_n = 1$ otherwise to minimize the display above, we have

$$
\begin{aligned}
&\sup_{F_\omega \in \mathcal{P}} \mathrm{E}\big(R(F_\omega)\big) \geq \mathrm{E}\Big(\sum_{n=1}^{N} \Big(\mathrm{E}_w\big((2\kappa X_{n,1} - \omega) I(\omega < 2\kappa X_{n,1} \leq \hat{\omega}_n) \,|\, \tilde{\mathcal{F}}_n\big) \\
&\qquad - \mathrm{E}_w\big((2\kappa X_{n,1} - \omega) I(\hat{\omega}_n < 2\kappa X_{n,1} \leq \omega) \,|\, \tilde{\mathcal{F}}_n\big)\Big)\Big) \\
\geq &\, \frac{1}{8\kappa} \mathrm{E}\Big(\sum_{n=1}^{N} \mathrm{E}_w\big((\hat{\omega}_n - \omega)^2 \,|\, \mathcal{F}_n\big)\Big) = \frac{1}{8\kappa} \sum_{n=1}^{N} \mathrm{E}_w \mathrm{E}(\hat{\omega}_n - \omega)^2 \\
\geq &\, \frac{1}{8\kappa} \sum_{n=1}^{N} \frac{1}{(n-1)\sigma^{-2} + I(w)},
\end{aligned}
$$

where the last inequality follows by the van Trees inequality (Gill and Levit, 1995; Goldenshluger and Zeevi, 2013), and $I(w) = \mathrm{E}_w(\frac{\partial \log p_w(\omega)}{\partial \omega})^2 = \pi^2/\kappa^2$. Therefore, taking $\kappa = \sigma N^{-1/2}$, we obtain

$$
\sup_{F_\omega \in \mathcal{P}} \mathrm{E}\big(R(F_\omega)\big) \geq \frac{1}{8} \sum_{n=1}^{N} \frac{\sigma^2}{n\kappa + \pi^2\sigma^2/\kappa} \geq C_3\sqrt{N},
$$

where $C_3 = \frac{\sigma}{8(1+\pi^2)}$. This completes the proof of Theorem 4.

□

Proof of Proposition 4. First, note by the union bounds and sub-Gaussian conditions that for any $1 \leq n \leq N$,

$$
\begin{aligned}
&\mathrm{E}\big(\|\mathbf{X}_n\|_\infty I(\|\mathbf{X}_n\|_\infty \geq c_x \sigma_X \sqrt{\log p_N})\big) \leq \int_0^\infty P\big(\|\mathbf{X}_n\|_\infty I(\|\mathbf{X}_n\|_\infty \geq c_x \sigma_X \sqrt{\log p_N}) > \epsilon\big) d\epsilon \\
\leq &\, c_x \sigma_X \sqrt{\log p_N} P(\|\mathbf{X}_n\|_\infty > c_x \sigma_X \sqrt{\log p_N}) + \int_{c_x \sigma_X \sqrt{\log p_N}}^\infty P(\|\mathbf{X}_n\|_\infty > \epsilon) d\epsilon \\
\leq &\, 2 c_x \sigma_X p_N^{-3} \sqrt{\log p_N} + 2p \int_{c_x \sigma_X \sqrt{\log p_N}}^\infty \exp\big(-\frac{\epsilon^2}{2\sigma_X^2}\big) d\epsilon \leq 4 c_x \sigma_X p_N^{-3} \sqrt{\log p_N}.
\end{aligned}
$$

Also note that

$$P(A^c) \leq \sum_{n=1}^{N} P(\|\mathbf{X}_n\|_\infty \geq c_x \sigma_X \sqrt{\log p_N}) \leq p_N^{-2}.$$

The two displays above imply that

$$\begin{aligned}
&\mathrm{E}\big(\|\mathbf{X}_n\|_\infty I(A^c)\big) \\
\leq &\mathrm{E}\big(\|\mathbf{X}_n\|_\infty I(A^c,\, \|\mathbf{X}_n\|_\infty < c_x \sigma_X \sqrt{\log p_N})\big) + \mathrm{E}\big(\|\mathbf{X}_n\|_\infty I(A^c,\, \|\mathbf{X}_n\|_\infty \geq c_x \sigma_X \sqrt{\log p_N})\big) \\
\leq &c_x \sigma_X \sqrt{\log p_N} P(A^c) + \mathrm{E}\big(\|\mathbf{X}_n\|_\infty I(\|\mathbf{X}_n\|_\infty \geq c_x \sigma_X \sqrt{\log p_N})\big) \leq 2 c_x \sigma_X p_N^{-2} \sqrt{\log p_N}.
\end{aligned}$$

Consequently, we obtain

$$\mathrm{E}\big(R_N I(A^c)\big) \leq 2b \sum_{n=1}^{N} \mathrm{E}\big(\|\mathbf{X}_n\|_\infty I(A^c)\big) \leq 4 b c_x \sigma_X p_N^{-1} \sqrt{\log p_N},$$

which completes the proof of Proposition 4.

□

B.5. Proofs for Section 5.3

Proof of Theorem 5. Suppose that $\mathcal{I}_N \neq \mathcal{I}_o$. This implies that either event A or B occurs, where

$$\begin{aligned}
A &= \{\exists\, n \geq \tilde{N}_{k-2} + 1 \text{ such that } i \in \tilde{\mathcal{S}}_n \text{ for some } i \in \mathcal{I}_u\}, \\
B &= \{\exists\, i \in \mathcal{I}_o \text{ such that } \forall \tilde{N}_{k-2} + 1 \leq n \leq \tilde{N}_{k-1},\, i \notin \hat{\mathcal{S}}_n\}.
\end{aligned}$$

By Proposition 2,

$$P(A) \leq \sum_{n=\tilde{N}_{k-2}+1}^{N} P(i \in \tilde{\mathcal{S}}_n \text{ for some } i \in \mathcal{I}_u) \leq N_{K-1} P(U_{K-1}^c) + N_K P(U_K^c). \tag{A.21}$$

In addition, by induction, we have $P(U_k^c) \leq k/N^2$ for all k $(1 \leq k \leq K)$. Indeed, it is known by Theorem 1 that $P(U_1^c) \leq 1/N^2$. If we suppose $P(U_k^c) \leq k/N^2$ holds, then by the arguments in the proof of Theorem 2, we have

$$\begin{aligned}
P(U_{k+1}^c) &\leq P(U_k^c) + P(U_k \cap H_k^c) + P(U_k \cap H_k \cap D_k^c) + P(U_k \cap H_k \cap D_k \cap F_{k+1}^c) \\
&\leq k/N^2 + l/N^3 + 1/p_N^3 + 4el/N^4 \leq \frac{k+1}{N^2}.
\end{aligned} \tag{A.22}$$

Therefore, (A.21) and (A.22) show that $P(A) \leq 2K/N$. Also note that

$$P(B) \leq P(U_{K-1}^c) + P(U_{K-1} \cap H_{K-1}^c) \leq (K-1)/N^2 + l/N^3 \leq K/N.$$

Consequently, $P(\mathcal{I}_N \neq \mathcal{I}_o) \leq P(A) + P(B) \leq \frac{3K}{N} \to 0$ as $N \to \infty$. For coefficient estimation, the consistency is the immediate result of (A.18) for Theorem 2 and (A.22), and we complete the proof of Theorem 5. □

Proof of Theorem 6. We only show variable selection consistency, as coefficient consistency is an immediate result of Theorem 5. Following the proof of Theorem 2, assume that U_K, D_K, and H_K hold, and suppose $|\mathcal{A}_{K,i}| = m$. Also for an arm $i \in \mathcal{I}_o$, assume event W_i holds in which $|\hat{\mathcal{V}}_i| < q_* - q_i$. Let $\mathbb{X} = \mathbb{X}_{\mathcal{A}_{k,i}}$ and $\mathbf{y} = \mathbf{y}_{\mathcal{A}_{k,i}}$. Define $G_i' = \hat{\mathcal{V}}_i \cup \mathcal{V}_i$ and $\hat{\boldsymbol{\beta}}_{G_i'} = \mathrm{argmin}_{\mathrm{supp}(\boldsymbol{\beta}) = G_i'} Q(\boldsymbol{\beta})$, and $\hat{\boldsymbol{\beta}}_{i,0} = \mathrm{argmin}_{\mathrm{supp}(\boldsymbol{\beta}) = \mathcal{V}_i} Q(\boldsymbol{\beta})$. If $\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G_i'} - \hat{\boldsymbol{\beta}}_{i,0})\|_2 \le 3\|\mathbb{X}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{G_i'})\|_2$, then by Lemmas 8 and 10,

$$\frac{\rho \xi_K \tilde{\lambda}_{i,K}(q_*)}{2\theta^2} |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \le \frac{1}{m} \|\mathbb{X}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i,0})\|_2^2 \le \frac{64\theta^2 \xi_K}{\tilde{\lambda}_{i,K}(q_*)} |\mathcal{V}_i \backslash \hat{\mathcal{V}}_i| \le \frac{128\theta^2 \xi_K}{\tilde{\lambda}_{i,K}(q_*)} |J_{i,\tau}|,$$

where $J_{i,\tau} = \{j \in \mathcal{V}_i \backslash \hat{\mathcal{V}}_i : \beta_{i,j}^2 < \tau\}$ with $\tau = 4\tilde{c}_\beta \log p_N / N$. The display above implies that $|\mathcal{V}_i \backslash \hat{\mathcal{V}}_i| \le 2|J_{i,\tau}|$ and

$$|\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \le \frac{256\theta^4}{\rho \tilde{\lambda}_{i,K}(q_*)^2} |J_{i,\tau}| \le \frac{256\theta^4}{\rho c_1^4 c_*^2/16} |J_{i,\tau}| =: c_\theta |J_{i,\tau}|.$$

On the other hand, if $\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G_i'} - \hat{\boldsymbol{\beta}}_{i,0})\|_2 > 3\|\mathbb{X}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{G_i'})\|_2$, then by Lemmas 8 and 9,

$$\frac{\rho \xi_K \tilde{\lambda}_{i,K}(q_*)}{4\theta^2} |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \le \frac{1}{2m} \|\mathbb{X}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i,0})\|_2^2 \le Q(\hat{\boldsymbol{\beta}}_{i,0}) - Q(\hat{\boldsymbol{\beta}}_i). \tag{A.23}$$

Also, suppose it holds that given $\eta = 2e/N^4$,

$$Q(\hat{\boldsymbol{\beta}}_{i,0}) - Q(\hat{\boldsymbol{\beta}}_i) \le \sigma^2 m^{-1} \big(2c_f |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \log p + c_f \log(\frac{2e}{\eta})\big). \tag{A.24}$$

The two displays above imply that

$$|\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| \le \frac{c_f \log(\frac{2e}{\eta})}{\frac{\rho m \xi_K \tilde{\lambda}_{i,K}(q_*)}{4\theta^2 \sigma^2} - 2c_f \log p} < \frac{\log N}{\log p_N} \le 1, \tag{A.25}$$

that is, $|\hat{\mathcal{V}}_i \backslash \mathcal{V}_i| = 0$. Then by (A.23) and (A.24), we have

$$\frac{1}{2m} \|\mathbb{X}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i,0})\|_2^2 \le \frac{c_f \sigma^2}{m} \log(\frac{2e}{\eta}). \tag{A.26}$$

Also suppose that

$$\|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|_\infty \le \sigma \sqrt{\frac{2}{m \tilde{\lambda}_{i,K}(q_*)} \log(\frac{2q_i}{\eta})}. \tag{A.27}$$

Note that

$$\begin{aligned} \frac{1}{m} \|\mathbb{X}(\hat{\boldsymbol{\beta}}_{i,0} - \hat{\boldsymbol{\beta}}_i)\|_2^2 &\ge \tilde{\lambda}_{i,K}(q_*) \|\hat{\boldsymbol{\beta}}_{i,0,\mathcal{V}_i \backslash \hat{\mathcal{V}}_i}\|_2^2 \ge \tilde{\lambda}_{i,K}(q_*) \big(\frac{1}{2} \|\boldsymbol{\beta}_{i,\mathcal{V}_i \backslash \hat{\mathcal{V}}_i}\|_2^2 - \|\hat{\boldsymbol{\beta}}_{i,0,\mathcal{V}_i \backslash \hat{\mathcal{V}}_i} - \boldsymbol{\beta}_{i,\mathcal{V}_i \backslash \hat{\mathcal{V}}_i}\|_2^2\big) \\ &\ge \tilde{\lambda}_{i,K}(q_*) \big(\frac{1}{2} \|\boldsymbol{\beta}_{i,\mathcal{V}_i \backslash \hat{\mathcal{V}}_i}\|_2^2 - |\mathcal{V}_i \backslash \hat{\mathcal{V}}_i| \|\hat{\boldsymbol{\beta}}_{i,0} - \boldsymbol{\beta}_i\|_\infty^2\big). \end{aligned}$$

If $|G_0 \backslash G^{(r^*)}| \ne 0$, the display above together with (A.26) and (A.27) implies that

$$\frac{1}{2} \tilde{\lambda}_{i,K}(q_*) \|\boldsymbol{\beta}_{i,\mathcal{V}_i \backslash \hat{\mathcal{V}}_i}\|_2^2 \le \frac{\sigma^2}{m} \big(2c_f \log(\frac{2e}{\eta}) + 2\log(\frac{2q_i}{\eta})\big) |\mathcal{V}_i \backslash \hat{\mathcal{V}}_i| \le \frac{\tilde{\lambda}_{i,K}(q_*) \rho \xi_K}{\theta^2} |\mathcal{V}_i \backslash \hat{\mathcal{V}}_i|,$$

where the last inequality holds as was derived for (A.25). Also note that

$$\frac{1}{2}\tilde{\lambda}_{i,K}(q_*)\|\boldsymbol{\beta}_{i,\mathcal{V}_i\backslash\hat{\mathcal{V}}_i}\|_2^2 \geq \frac{1}{2}\tilde{\lambda}_{i,K}(q_*)\tau|J'_{i,\tau}|,$$

where $J'_{i,\tau} = \{j \in \mathcal{V}_i\backslash\hat{\mathcal{V}}_i : \beta_{i,j}^2 \geq \tau\}$. The two displays above together with $\log p = o(N)$ show that $|J'_{i,\tau}| \leq \frac{2\rho\xi_K}{\tau\theta^2}|\mathcal{V}_i\backslash\hat{\mathcal{V}}_i| \leq \frac{1}{2}|\mathcal{V}_i\backslash\hat{\mathcal{V}}_i|$, which implies that $|\mathcal{V}_i\backslash\hat{\mathcal{V}}_i| \leq 2|J_{i,\tau}|$. Therefore, by Lemmas 2, 11 and the known probability bound for (A.27), from the beta-min condition, we have

$$\begin{aligned}
&P(\hat{\mathcal{V}}_i \neq \mathcal{V}_i) = P(\mathcal{I}_N \neq \mathcal{I}_o) + P(\mathcal{I}_N = \mathcal{I}_o,\ |\hat{\mathcal{V}}_i\backslash\mathcal{V}_i| > c_\theta|J_{i,\tau}| \text{ or } |\mathcal{V}_i\backslash\hat{\mathcal{V}}_i| \leq 2|J_{i,\tau}|)\\
\leq& 3K/N + P(U_K^c) + P(U_K \cap H_K^c) + P(U_K \cap H_K \cap D_K^c) + P(U_K \cap H_K \cap D_K \cap W_i^c) + 4e/N^4\\
\leq& 3K/N + (K+1)/N^2 \leq 4K/N,
\end{aligned}$$

which approaches 0 as $N \to \infty$. We complete the proof of Theorem 6. □

B.6. Proofs for Section 6

Proof of Proposition 5. We prove the first statement by contradiction. Suppose Assumption 5 does not hold. Then for every $\epsilon, c > 0$, there are some members in the considered bandit class and some $i \in \mathcal{I}_o$ such that

$$P\big(f_i(\mathbf{X}) - \max_{j\neq i} f_j(\mathbf{X}) > \epsilon\big) < c.$$

Together with Assumption 4, this implies that

$$\sum_{\tilde{i}\neq i} P\big(f_{\tilde{i}}(\mathbf{X}) - \max_{j\neq\tilde{i}} f_j(\mathbf{X}) > 0\big) \geq \sum_{\tilde{i}\neq i} P\big(f_{\tilde{i}}(\mathbf{X}) - \max_{j\neq\tilde{i}} f_j(\mathbf{X}) > \epsilon\big) > 1 - L\epsilon - c.$$

Consequently, $P\big(f_i(\mathbf{X}) - \max_{j\neq i} f_j(\mathbf{X}) > 0\big) < L\epsilon + c$. Then with $\epsilon = c/L$, this implies that for every $c > 0$, there are some members and some $i \in \mathcal{I}_o$ such that $P\big(f_i(\mathbf{X}) - \max_{j\neq i} f_j(\mathbf{X}) > 0\big) < 2c$, which is in contradiction with Assumption 1. The second statement holds trivially by noting that under Assumption 5, for any $i \in \mathcal{I}_o$, $P(\mathbf{X} \in \mathcal{T}_i) \geq P(\mathbf{X} \in \tilde{\mathcal{T}}_i) > \tilde{c}_1$. The proof is complete. □

Proof of Proposition 6. The first statement is simply the reiteration of Propositions 10 and 11. For the second statement, it is not hard to see that Assumption 4 and Assumption 5 are in direct contradiction with the statements (4) and (5) of Proposition 1, respectively. This completes the proof of Proposition 6. □

Proof of Theorem 7. The proof of Theorem 7 follows the same proof structure of Theorem 3; we need only modify the proof for the regret upper bound of R_K. Specifically, by Proposition 3, we have

$$\begin{aligned} R_K &\le \sum_{k=1}^{K}\sum_{n=\tilde{N}_{k-1}+1}^{\tilde{N}_k}\big(f^*(\mathbf{X}_n)-f_{I_n}(\mathbf{X}_n)\big)I(U_K)\\ &\le 2\theta bl\tau_0+\sum_{k=2}^{K}\sum_{n=\tilde{N}_{k-1}+1}^{\tilde{N}_k}4\theta b_k I(U_k,\ f^*(\mathbf{X}_n)-f^\sharp(\mathbf{X}_n)\le 4\theta b_k). \end{aligned}$$

Then, by Assumption 4 and Assumption 6,

$$\mathrm{E}(R_K)\le 2\theta bc_0lq_*^2\log p_N\log N+\sum_{k=2}^{K}16\theta^2b_k^2N_k\le 2\theta bc_0lq_*^2\log p_N\log N+32\theta^2\tilde{c}_\rho Kq_*^2\log p_N. \tag{A.28}$$

By (12), (A.19), and (A.28) and setting $\tilde{C}_2=4\theta bc_0+6\theta b+32\theta^2\tilde{c}_\rho$, we obtain the conclusion of Theorem 7. □

C. Ancillary lemmas and proofs

To perform an analysis for $\hat{\boldsymbol{\beta}}_{i,k+1}$, we require some ancillary lemmas. For notational brevity, we omit the subscripts for arm i and stage $k+1$. Using the definitions in the proof of Theorem 2, we assume throughout this section that U_k, D_k, and H_k hold, with sample size $m:=|J_{i,k}|$ ($m>N_kp_il_o^{-1}/2$). Given variable set $G\subset\{1,\cdots,p\}$, without confusion, the true coefficient is $\boldsymbol{\beta}_0:=\boldsymbol{\beta}_i$, the true set of relevant variables is $G_0:=\mathcal{V}_i=\mathrm{supp}(\boldsymbol{\beta}_0)$, least square estimation on set G is $\hat{\boldsymbol{\beta}}_G:=\mathrm{argmin}_{\mathrm{supp}(\boldsymbol{\beta})\in G}\,Q(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_0=\hat{\boldsymbol{\beta}}_{G_0}$, $\tilde{\lambda}(s):=\tilde{\lambda}_{i,k}(s)$, and $\xi:=\xi_{k+1}$, where $Q(\boldsymbol{\beta})$ is defined in Section 4 with response $\mathbf{y}:=\mathbf{y}_{\mathcal{A}_{k,i}}$ and covariate matrix $\mathbb{X}:=\mathbb{X}_{\mathcal{A}_{k,i}}$.

Due to key arguments for the design matrix under the randomized allocation scheme, we are able to prove Lemmas 1–4 in a similar way to that of Lemmas C3, C4, B1, and B2 in Zhang (2011). Their proofs are thus omitted. Let $\boldsymbol{\varepsilon}=(\tilde{\varepsilon}_1,\cdots,\tilde{\varepsilon}_m)^T$ be the random error vector, which has conditionally independent elements given $\mathbb{X}$.

Lemma 1. *With probability greater than* $1-\eta$*, there exist constants* $c_d,c_f>0$ *(e.g.,* $c_d=7.4,c_f=2.7$*) such that given* $\mathbb{X}$*, for all* $G\subset\{1,\cdots,p\}$*,*

$$\|\mathbb{X}\hat{\boldsymbol{\beta}}_G-\mathbb{X}\boldsymbol{\beta}_0\|_2^2\le m[Q(\hat{\boldsymbol{\beta}}_G)-Q(\boldsymbol{\beta}_0)]+2\|\mathbb{X}\hat{\boldsymbol{\beta}}_G-\mathbb{X}\boldsymbol{\beta}_0\|_2\sigma\sqrt{c_d|G_0|+2c_f|G\backslash G_0|\log p+c_f\log(\frac{2e}{\eta})}.$$

Lemma 2. *With probability greater than* $1-\eta$*, given* $\mathbb{X}$*, for all* $G\subset\{1,\cdots,p\}$*,*

$$Q(\hat{\boldsymbol{\beta}}_0)-Q(\hat{\boldsymbol{\beta}}_{G\cup G_0})\le\sigma^2m^{-1}\big(2c_f|G\backslash G_0|\log p+c_f\log(\frac{2e}{\eta})\big).$$

Lemma 3. *(Forward Step) Given $G \subset \{1, \cdots, p\}$, define $G' = G \cup G_0$ and $s = |G'|$. Then*

$$Q(\hat{\boldsymbol{\beta}}_G) - \min_{\alpha\in\mathbb{R},\, j\in G_0\backslash G} Q(\hat{\boldsymbol{\beta}}_G + \alpha \mathbf{e}_j) \geq \frac{\tilde{\lambda}(s)}{4\theta^2|G'\backslash G|}\Big(\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}_{G'})\|_2^2 + Q(\hat{\boldsymbol{\beta}}_G) - Q(\hat{\boldsymbol{\beta}}_{G'})\Big).$$

Lemma 4. *(Backward Step) Given $G \subset \{1, \cdots, p\}$ and $\hat{\boldsymbol{\beta}}_G = (\hat{\beta}_{G,1}, \cdots, \hat{\beta}_{G,p})^T$, we have*

$$\min_{j\in G} Q(\hat{\boldsymbol{\beta}}_G - \hat{\beta}_{G,j}\mathbf{e}_j) - Q(\hat{\boldsymbol{\beta}}_G) \leq \frac{\theta^2}{|G\backslash G_0|}\sum_{j\in G\backslash G_0}\hat{\beta}_{G,j}^2.$$

The following lemmas are related to IGA estimator properties at a certain iteration r with selected variable index set $G^{(r)}$, immediately after the completion of Step 2 in Algorithm 2, and are derived for Lemma 11. Let $G^{(r-1)}$ be the obtained index set at the end of the previous iteration. Define $s = |G^{(r)} \cup G_0|$.

Lemma 5. *Suppose the current iteration has no backward elimination. Then*

$$\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2 \geq \frac{\rho\xi\tilde{\lambda}(s)}{2\theta^2}|G^{(r)}\backslash G_0|.$$

Proof of Lemma 5. Let $G' = G^{(r)} \cup G_0$. Then it is not hard to see that since $G^{(r-1)} \subset G^{(r)} \subset G'$,

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}_{G^{(r-1)}}) - Q(\hat{\boldsymbol{\beta}}_{G'}) &= \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \hat{\boldsymbol{\beta}}_{G^{(r-1)}})\|_2^2, \\ Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) - Q(\hat{\boldsymbol{\beta}}_{G'}) &= \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \hat{\boldsymbol{\beta}}_{G^{(r)}})\|_2^2. \end{aligned} \tag{A.29}$$

Also, note from Step 2 that

$$\begin{aligned} \xi^{(r)} &= Q(\hat{\boldsymbol{\beta}}_{G^{(r-1)}}) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) \\ &\geq Q(\hat{\boldsymbol{\beta}}_{G^{(r-1)}}) - \min_{\alpha\in\mathbb{R}} Q(\hat{\boldsymbol{\beta}}_{G^{(r-1)}} + \alpha\mathbf{e}_j) \quad \text{for some } j \in G_\rho \\ &\geq \rho\Big(Q(\hat{\boldsymbol{\beta}}_{G^{(r-1)}}) - \min_{j\notin G^{(r-1)}, \alpha\in\mathbb{R}} Q(\hat{\boldsymbol{\beta}}_{G^{(r-1)}} + \alpha\mathbf{e}_j)\Big) = \rho\phi^{(r-1)} \qquad \text{(A.30)} \\ &\geq \frac{\rho\tilde{\lambda}(s)}{4\theta^2|G'\backslash G^{(r-1)}|}\big(Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) - Q(\hat{\boldsymbol{\beta}}_{G'})\big) \\ &= \frac{\rho\tilde{\lambda}(s)}{4\theta^2|G'\backslash G^{(r-1)}|m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r-1)}} - \hat{\boldsymbol{\beta}}_{G'})\|_2^2, \qquad \text{(A.31)} \end{aligned}$$

where the last inequality holds by Lemma 3 and (A.29), and is used later by Lemma 6.

By assumption that there is no backward elimination at the current iteration, by Lemma 4,

$$\frac{\xi^{(r)}}{2} \leq \min_{j\in G^{(r)}} Q(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\beta}_{G^{(r)},j}\mathbf{e}_j) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) \leq \frac{\theta^2}{|G^{(r)}\backslash G_0|}\sum_{j\in G^{(r)}\backslash G_0}\hat{\beta}_{G^{(r)},j}^2 =: \frac{\theta^2}{|G^{(r)}\backslash G_0|}\|\hat{\boldsymbol{\beta}}_{G^{(r)}\backslash G_0}^{(r)}\|_2^2. \tag{A.32}$$

Then, by (A.32) and (A.30), we have

$$\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}_{G^{(r)}\backslash G_0}^{(r)}\|_2^2 \geq \frac{\tilde{\lambda}(s)\xi^{(r)}}{2\theta^2}|G^{(r)}\backslash G_0| \geq \frac{\rho\xi\tilde{\lambda}(s)}{2\theta^2}|G^{(r)}\backslash G_0|.$$

This completes the proof of Lemma 5. □

Lemma 6. *Under the same conditions as Lemma 5, if* $\tilde{\lambda}(s)^2 \geq \frac{8\theta^4\gamma^2|G'\backslash G^{(r-1)}|}{|G^{(r)}\backslash G_0|}$ *with some* $\gamma \geq 2$, *then*

$$(\frac{1+2\gamma^{-1}}{m})\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2 \leq Q(\boldsymbol{\beta}_0) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}).$$

Proof of Lemma 6. By (A.31) and (A.32),

$$\frac{2\theta^2}{|G^{(r)}\backslash G_0|}\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2^2 \geq \frac{\rho\tilde{\lambda}(s)}{4\theta^2 m|G'\backslash G^{(r-1)}|}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'})\|_2^2 \geq \frac{\rho\tilde{\lambda}(s)^2}{4\theta^2 m|G'\backslash G^{(r-1)}|}\|\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'}\|_2^2,$$

which implies by the value of γ that

$$\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2 \geq \gamma\|\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'}\|_2 \geq \gamma(\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2 - \|\hat{\boldsymbol{\beta}}'_{G^{(r)}\backslash G_0}\|_2),$$

where $\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}_{G^{(r)}}$ and $\hat{\boldsymbol{\beta}}' = \hat{\boldsymbol{\beta}}_{G'}$. The two displays above imply that

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}'_{G^{(r)}\backslash G_0}\|_2 &\geq (1-\gamma^{-1})\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2 \\ &\geq (1-\gamma^{-1})\sqrt{\frac{\rho\tilde{\lambda}(s)}{4\theta^2 m|G'\backslash G^{(r-1)}|}\frac{|G^{(k)}\backslash G_0|}{2\theta^2}}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'})\|_2 \\ &\geq (\gamma-1)\sqrt{\frac{1}{\tilde{\lambda}(s)m}}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'})\|_2.\end{aligned} \tag{A.33}$$

Therefore,

$$\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}'_{G'\backslash G_0}\|_2^2 \geq \frac{(\gamma-1)^2}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'})\|_2^2, \tag{A.34}$$

which implies that

$$\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2 \geq (\gamma-1)(\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2 - \|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2)$$

and

$$\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2 \leq \frac{\gamma}{\gamma-1}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2. \tag{A.35}$$

Therefore,

$$\begin{aligned}Q(\boldsymbol{\beta}_0) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) =& Q(\boldsymbol{\beta}_0) - Q(\hat{\boldsymbol{\beta}}_{G'}) + Q(\hat{\boldsymbol{\beta}}_{G'}) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) \\ =& \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 - \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \hat{\boldsymbol{\beta}}_{G^{(r)}})\|_2^2 \\ \geq& \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 - \frac{1}{(\gamma-1)^2 m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 \\ \geq& (1+\frac{2}{\gamma})\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2,\end{aligned}$$

where the first inequality follows by (A.34) and the last inequality follows by (A.35). This completes the proof of Lemma 6.

□

Lemma 7. *Under the same conditions as Lemma 6,*

$$Q(\hat{\boldsymbol{\beta}}_0) - Q(\hat{\boldsymbol{\beta}}_{G'}) \geq \frac{\tilde{\lambda}(s)(1-\gamma^{-1})^2\rho\xi|G^{(r)}\backslash G_0|}{2\theta^2}.$$

Proof of Lemma 7. We can see that

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}_0) - Q(\hat{\boldsymbol{\beta}}_{G'}) &= \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \hat{\boldsymbol{\beta}}_0)\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}'_{G^{(r)}\backslash G_0}\|_2^2 \\ &\geq (1-\gamma^{-1})^2\tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2^2 \geq \frac{\tilde{\lambda}(s)(1-\gamma^{-1})^2\rho\xi|G^{(r)}\backslash G_0|}{2\theta^2}, \end{aligned}$$

where the second to last inequality follows by (A.33) and the last inequality follows by (A.30) and (A.32). This completes the proof of Lemma 7. □

In Lemmas 8–10, we assume that IGA obtains the variable index set $G^{(r)}$ when it terminates. Let $s = |G^{(r)} \cup G_0|$. Note that these lemmas still hold if we replace $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_0$.

Lemma 8. *When IGA terminates, we have*

$$\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2 \geq \frac{\rho\xi\tilde{\lambda}(s)}{2\theta^2}|G^{(r)}\backslash G_0|. \tag{A.36}$$

Proof of Lemma 8. By the backward termination condition and (A.30), we know that

$$\min_{j\in G^{(r)}} Q(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\beta}_j^{(r)}\mathbf{e}_j) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) > 0.5\xi^{(r)} \geq 0.5\rho\xi,$$

where $\hat{\beta}_j^{(r)}$ is the j-th element of $\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}_{G^{(r)}}$. Together with Lemma 4, we have

$$\theta^2\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2^2 \geq 0.5\rho\xi|G^{(r)}\backslash G_0|.$$

Then, we obtain (A.36) by noting that

$$\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0\|_2^2 \geq \tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}^{(r)}_{G^{(r)}\backslash G_0}\|_2^2.$$

□

Lemma 9. *When IGA terminates, if* $\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2 > 3\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \hat{\boldsymbol{\beta}}_{G'})\|_2$, *then*

$$Q(\boldsymbol{\beta}_0) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) \geq \frac{1}{2m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2.$$

Proof of Lemma 9. This lemma is proved by noting that

$$\begin{aligned} &Q(\boldsymbol{\beta}_0) - Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) = Q(\boldsymbol{\beta}_0) - Q(\hat{\boldsymbol{\beta}}_{G'}) - \big(Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}) - Q(\hat{\boldsymbol{\beta}}_{G'})\big) \\ =&\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 - \frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - Q(\hat{\boldsymbol{\beta}}_{G'}))\|_2^2 \\ \geq&\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 - \frac{1}{9m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'} - \boldsymbol{\beta}_0)\|_2^2 \geq \frac{1}{2m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}} - \boldsymbol{\beta}_0)\|_2^2, \end{aligned}$$

where the last inequality follows by

$$\begin{aligned}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\boldsymbol{\beta}_0)\|_2 &\le \|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\hat{\boldsymbol{\beta}}_{G'})\|_2+\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'}-\boldsymbol{\beta}_0)\|_2\\ &\le\frac{1}{3}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'}-\boldsymbol{\beta}_0)\|_2+\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'}-\boldsymbol{\beta}_0)\|_2\le\frac{4}{3}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'}-\boldsymbol{\beta}_0)\|_2.\end{aligned}$$

□

Lemma 10. *When IGA terminates, if*

$$\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'}-\boldsymbol{\beta}_0)\|_2<3\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\hat{\boldsymbol{\beta}}_{G'})\|_2,$$

then

$$\begin{aligned}\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\boldsymbol{\beta}_0)\|_2^2&\le\frac{64\theta^2\xi}{\tilde{\lambda}(s)}|G_0\backslash G^{(r)}|\le\frac{128\theta^2\xi}{\tilde{\lambda}(s)}|J_{0,\tau}|,\\ Q(\hat{\boldsymbol{\beta}}_{G^{(r)}})-Q(\boldsymbol{\beta}_0)&\le\frac{4\theta^2\xi}{\tilde{\lambda}(s)}|G_0\backslash G^{(r)}|\le\frac{8\theta^2\xi}{\tilde{\lambda}(s)}|J_{0,\tau}|,\end{aligned}$$

where $J_{0,\tau}=\{j\in G_0\backslash G^{(r)}:\ \beta_{0,j}^2<\tau\}$ *with* $\tau=128\theta^2\xi/\tilde{\lambda}(s)^2$.

Proof of Lemma 10. Note that by the termination condition and Lemma 3,

$$\xi\ge Q(\hat{\boldsymbol{\beta}}_{G^{(r)}})-\min_{\alpha\in\mathbb{R},\,j\in G_0\backslash G^{(r)}}Q(\hat{\boldsymbol{\beta}}_{G^{(r)}}+\alpha\mathbf{e}_j)\ge\frac{\tilde{\lambda}(s)}{4\theta^2|G_0\backslash G^{(r)}|m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\hat{\boldsymbol{\beta}}_{G'})\|_2^2,\qquad\text{(A.37)}$$

which implies that

$$\begin{aligned}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\boldsymbol{\beta}_0)\|_2^2&\le\left(\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\hat{\boldsymbol{\beta}}_{G'})\|_2+\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G'}-\boldsymbol{\beta}_0)\|_2\right)^2\\ &\le16\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\hat{\boldsymbol{\beta}}_{G'})\|_2^2\le\frac{64\theta^2m\xi}{\tilde{\lambda}(s)}|G_0\backslash G^{(r)}|.\qquad\text{(A.38)}\end{aligned}$$

Also, note that

$$Q(\hat{\boldsymbol{\beta}}_{G^{(r)}})-Q(\boldsymbol{\beta}_0)\le Q(\hat{\boldsymbol{\beta}}_{G^{(r)}})-Q(\hat{\boldsymbol{\beta}}_{G'})=\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\hat{\boldsymbol{\beta}}_{G'})\|_2^2\le\frac{4\theta^2\xi}{\tilde{\lambda}(s)}|G_0\backslash G^{(r)}|,\qquad\text{(A.39)}$$

where the last inequality follows by (A.37). In addition,

$$\frac{1}{m}\|\mathbb{X}(\hat{\boldsymbol{\beta}}_{G^{(r)}}-\boldsymbol{\beta}_0)\|_2^2\ge\tilde{\lambda}(s)\|\hat{\boldsymbol{\beta}}_{G^{(r)}}-\boldsymbol{\beta}_0\|_2^2\ge\tilde{\lambda}(s)\|\boldsymbol{\beta}_{0,G_0\backslash G^{(r)}}\|_2^2\ge\tilde{\lambda}(s)\tau|J_{0,\tau}'|,$$

where $J_{0,\tau}'=\{j\in G_0\backslash G^{(r)}:\ \beta_{0,j}^2\ge\tau\}$ with $\tau=128\theta^2\xi/\tilde{\lambda}(s)^2$. The displays above together with (A.38) imply that $\frac{64\theta^2m\xi}{\tilde{\lambda}(s)}|G_0\backslash G^{(r)}|\ge\tilde{\lambda}(s)\tau|J_{0,\tau}'|$. Then, by our choice of τ, we have

$$|G_0\backslash G^{(r)}|\ge2|J_{0,\tau}'|=2(|G_0\backslash G^{(r)}|-|J_{0,\tau}|).$$

Consequently, $|G_0\backslash G^{(r)}|\le2J_{0,\tau}$. Together with (A.38) and (A.39), we obtain the conclusions of this lemma. □

The following lemma provides an upper bound on the size of the selected variables. Let $q=q_i$.

Lemma 11. *Assume that IGA terminates with the variable index set $G^{(r^*)}$ and we set $\xi \geq \frac{16\theta^2\sigma^2 c_f}{m\tilde{\lambda}(q_*)\rho}(2\log p + \frac{\log(2e/\eta)}{32\theta^2 q})$. Given $\mathbb{X}$, with probability greater than $1-\eta$, we have $r^* < q_* - q$.*

Proof of Lemma 11. Suppose that we have $r^* \geq q_* - q$. Then assume that $r = q_* - q$ is first recorded. Then note that there is no backward step in the previous step. Let $G' = G^{(r)} \cap G_0$. We seek to verify that the conditions of Lemma 7 hold. Indeed, we can see that

$$32\theta^2|G'\backslash G^{(r-1)}| \leq 32\theta^2(q+1) \leq \tilde{\lambda}(q_*)^2(q_* - 2q) \leq \tilde{\lambda}(q_*)^2|G^{(r)}\backslash G_0|, \tag{A.40}$$

where the second inequality holds because under (A.14),

$$\tilde{\lambda}(q_*)^2 \geq \frac{c_1^4 c_*^2}{16} \geq \frac{64\theta^2}{(C_1-2)} \geq \frac{32\theta^2(q+1)}{(C_1-2)q}, \tag{A.41}$$

with some large enough constant $C_1 > 2$. Then we can apply Lemma 6 to obtain

$$Q(\hat{\boldsymbol{\beta}}_{G'}) \leq Q(\hat{\boldsymbol{\beta}}_0) - \frac{\tilde{\lambda}(q_*)\rho\xi|G^{(r)}\backslash G_0|}{8\theta^2}. \tag{A.42}$$

Also, suppose that

$$Q(\hat{\boldsymbol{\beta}}_0) - Q(\hat{\boldsymbol{\beta}}_{G'}) \leq \frac{\sigma^2}{m}\big(2c_f|G^{(r)}\backslash G_0|\log p + c_f\log(\frac{2e}{\eta})\big) \tag{A.43}$$

holds. Then (A.42) and (A.43) imply that

$$\xi \leq \frac{8\theta^2\sigma^2}{m\tilde{\lambda}(q_*)\rho}\big(2c_f\log p + \frac{c_f}{|G^{(r)}\backslash G_0|}\log(\frac{2e}{\eta})\big) \leq \frac{8\theta^2\sigma^2}{m\tilde{\lambda}(q_*)\rho}\big(2c_f\log p + \frac{c_f}{32\theta^2(q+1)}\log(\frac{2e}{\eta})\big),$$

where the last inequality follows by (A.40). However, this contradicts our choice of ξ, and thus (A.43) does not hold. Together with Lemma 2, we complete the proof of Lemma 11. □

D. Simulation

In this section, we evaluate the performance of the proposed bandit algorithms on simulated data. We compare the performance of different bandit algorithms in Supplement D.1 and perform a sensitivity analysis on parameter choice in Supplement D.2.

D.1. Performance with different algorithms

For brevity, the **m**ulti-**s**tage type algorithms described in Section 3 are abbreviated as "MS". Throughout the following numerical evaluation, we set the initial sampling size $\tau_0 = 20$ and set the arm screening and elimination parameters to be $\delta_N = c\sqrt{\log p_N/\tau_0}$ and $\Delta_k = c\sqrt{l\log p_N/N_k}$. Unless stated otherwise, we simply set $c = 1$ and $h = 4$. We considered IGA and lasso as the methods for coefficient estimation and denote the corresponding bandit algorithms by MS-IGA and MS-lasso. For MS-IGA, rather than directly setting the IGA parameter ξ, we generated the solution path through the forward-backward selection steps, and applied ten-fold cross validation (CV) to determine the best number of selection steps and find the IGA estimates for each stage. For MS-lasso, we found the solution path with a decreasing sequence of the tuning parameter values (Friedman et al., 2010) via an accelerated proximal gradient descent (Beck and Teboulle, 2009), and then applied ten-fold CV to generate lasso estimates for each stage. 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 LASSO bandit algorithm in Bastani and Bayati (2020) as a useful benchmark (denoted by B-lasso); to avoid having to perform computationally more expensive CV at each user visit point, we adopted the parameter values recommended for B-lasso algorithm in numerical evaluation.

In the simulation, we set the number of arms $l = 3$, the number of covariates $p = 500$ and the total number of visits $N = 10000$. We generated covariate vectors $\mathbf{X}_n$ from a multivariate normal distribution with mean $\mathbf{0}$ and covariance Σ with exponential decay $(\Sigma)_{i,j} = \rho^{|i-j|}$ and $\rho = 0.5$. For each arm i $(i = 1, 2, 3)$, we set the number of relevant variables to be the same for all arms at $q_0 = q_i = 5, 10$ or 15. The index set for nonzero coefficients was $F_{q_0} = \{j : j = 10(k-1)+1,\, k = 1, \cdots, q_0\}$, and the j-th element of $\boldsymbol{\beta}_i$ for $j \in F_{q_0}$ was generated randomly by $P(\beta_{ij} = \vartheta) = P(\beta_{ij} = -\vartheta) = 0.5$, where we set $\vartheta = 0.2$ or 0.4. Then the response followed the linear model $Y_{i,n} = \mathbf{X}_n^T \boldsymbol{\beta}_i + \varepsilon_{i,n}$, where the $\varepsilon_{i,n}$'s are independent $N(0, \sigma^2)$ with $\sigma = 2$. We then ran the aforementioned bandit algorithms over the entire simulated data set in a sequential manner.

For the algorithm performance, we recorded the per-round regret trajectory, that is, $r_n = R_n/n$ $(n = 1, \cdots, N)$. We also evaluated the coefficient estimation and variable selection performance of each arm from the final algorithm output: given arm i and algorithm output $\hat{\boldsymbol{\beta}}_i$, let $\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2$ be the estimation error in the l_2 norm, $C_i = |\mathcal{V}_i \cap \hat{\mathcal{V}}_i|$ be the number of correctly identified variables in $\hat{\boldsymbol{\beta}}_i$, and $IC_i = |\hat{\mathcal{V}}_i \backslash \mathcal{V}_i|$ be the number of incorrectly identified variables. The experiment was repeated 100 times to obtain the averaged results of these measures.

From the averaged per-round regret $\bar{r}_N$ summarized in Table 3 (numbers in parentheses are standard errors), it is not surprising that MS-simple did not perform well since it ignores the covariate information; satisfactorily, MS-IGA performed better or competitively compared to MS-lasso and the benchmark. We also plotted the averaged per-round regrets against the user visit points n in Figure 3 and Figure 4. Except for MS-simple, all three algorithms considering covariates exhibit a decreasing trend in these plots.

Table 3: Averaged per-round regret for different bandit algorithms on simulated data from 100 runs.

	$\vartheta = 0.2$			$\vartheta = 0.4$		
q_0	5	10	15	5	10	15
MS-simple	0.359 (0.006)	0.528 (0.005)	0.645 (0.005)	0.732 (0.012)	1.068 (0.010)	1.298 (0.011)
B-lasso	0.245 (0.003)	0.308 (0.002)	0.335 (0.002)	0.322 (0.004)	0.359 (0.002)	0.364 (0.002)
MS-lasso	0.303 (0.005)	0.358 (0.004)	0.359 (0.004)	0.224 (0.004)	0.249 (0.003)	0.263 (0.005)
MS-IGA	0.202 (0.004)	0.274 (0.004)	0.317 (0.005)	0.177 (0.006)	0.221 (0.008)	0.273 (0.010)

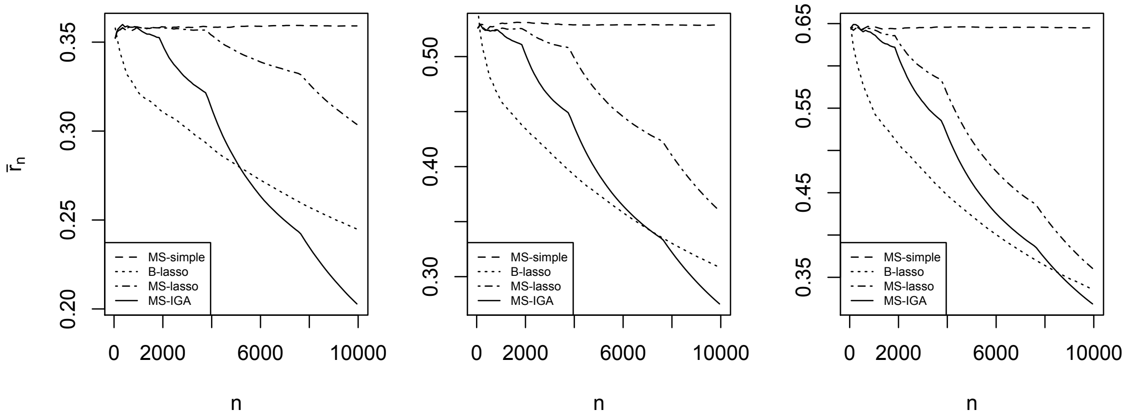

Figure 3: Averaged per-round regret curves of different bandit algorithms ($\vartheta = 0.2$). Left panel: $q_0 = 5$; middle panel: $q_0 = 10$; right panel: $q_0 = 15$.

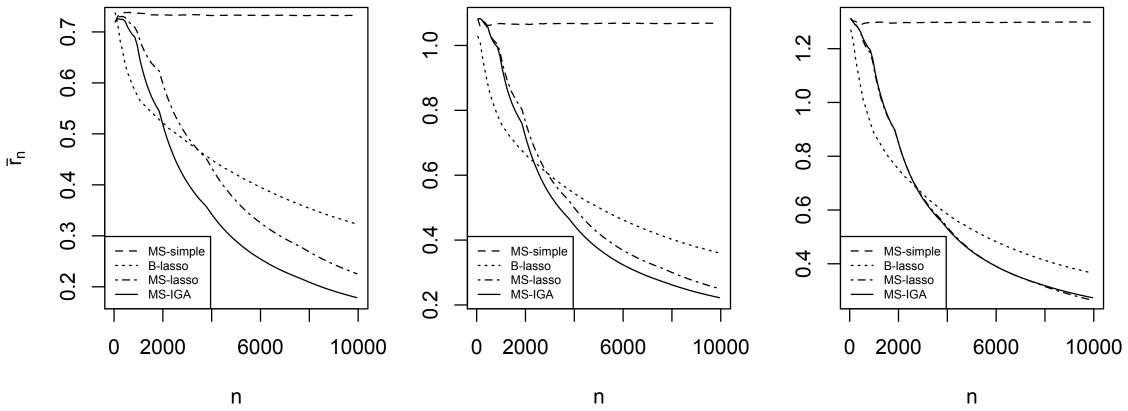

Figure 4: Averaged per-round regret curves of different bandit algorithms ($\vartheta = 0.4$). Left panel: $q_0 = 5$; middle panel: $q_0 = 10$; right panel: $q_0 = 15$.

Besides the regret performance, we summarized the coefficient estimation and variable selection results of the algorithms' final output for $\vartheta = 0.2$ in Table 4. The averaged sample sizes

Table 4: Averaged simulation results of different bandit algorithms based on 100 runs ($\vartheta = 0.2$).

	$q_0 = 5$			$q_0 = 10$			$q_0 = 15$		
Arm i	1	2	3	1	2	3	1	2	3
$\bar{n}_i$									
MS-simple	3697	3234	3069	3055	3635	3310	3444	3232	3324
B-lasso	3394	3337	3269	3253	3376	3371	3371	3317	3312
MS-lasso	3279	3439	3282	3276	3434	3290	3488	3236	3276
MS-IGA	3413	3253	3334	3202	3397	3401	3497	3212	3290
Avg. $\lVert\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\rVert_2$									
B-lasso	1.22	1.246	1.254	1.237	1.205	1.192	1.183	1.2	1.205
	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.011)	(0.010)	(0.010)	(0.010)
MS-lasso	0.364	0.362	0.368	0.394	0.413	0.422	0.433	0.448	0.449
	(0.009)	(0.008)	(0.008)	(0.011)	(0.011)	(0.012)	(0.007)	(0.008)	(0.008)
MS-IGA	0.175	0.206	0.205	0.250	0.256	0.238	0.254	0.311	0.306
	(0.011)	(0.012)	(0.012)	(0.014)	(0.015)	(0.012)	(0.012)	(0.014)	(0.014)
$\bar{C}_i$									
B-lasso	5.00	5.00	5.00	10.00	10.00	10.00	14.99	15.00	15.00
MS-lasso	2.62	2.92	2.63	8.39	8.30	7.88	14.36	14.37	14.34
MS-IGA	4.46	4.26	4.26	8.77	8.60	9.02	14.14	13.56	13.5
$\overline{IC}_i$									
B-lasso	469.66	468.29	470.19	463.67	463.43	463.09	457.64	458.3	459.32
MS-lasso	0.74	1.68	0.60	8.82	6.03	5.61	11.98	13.98	13
MS-IGA	0.75	0.96	1.08	0.92	0.90	0.85	1.09	1.64	1.37

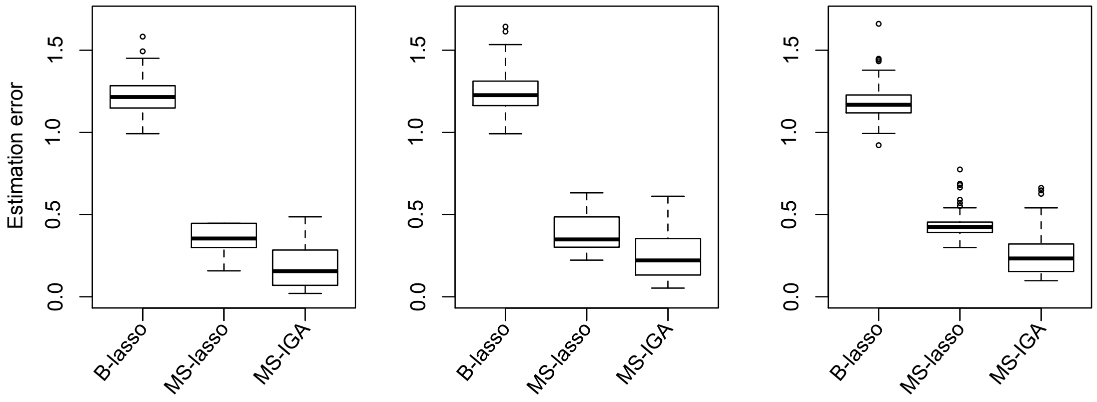

Figure 5: Boxplots for coefficient estimation error of arm 1 from output of different bandit algorithms ($\vartheta = 0.2$). Left panel: $q_0 = 5$; middle panel: $q_0 = 10$; right panel: $q_0 = 15$.

$\bar{n}_i$ appear well-balanced among different arms, which is expected from the randomized generation of the true coefficients. Both MS-lasso and MS-IGA generated much sparser coefficient estimates than that of the benchmark, as we empirically employed the data-driven approach for parameter tuning in the MS algorithms. In particular, for most cases here, MS-IGA resulted in fewer incorrectly identified variables than MS-lasso. Similar patterns on variable selection with

Table 5: Averaged simulation results of different bandit algorithms based on 100 runs ($\vartheta = 0.4$).

	$q_0 = 5$			$q_0 = 10$			$q_0 = 15$		
Arm i	1	2	3	1	2	3	1	2	3
$\bar{n}_i$									
MS-simple	3382	3382	3236	3043	3711	3246	3377	3243	3380
B-lasso	3357	3343	3300	3382	3332	3286	3372	3309	3319
MS-lasso	3271	3362	3367	3319	3427	3253	3287	3369	3344
MS-IGA	3251	3421	3328	3359	3339	3301	3345	3312	3343
Avg. $\Vert\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\Vert_2$									
B-lasso	1.177	1.180	1.186	1.131	1.133	1.154	1.098	1.126	1.123
	(0.012)	(0.013)	(0.011)	(0.011)	(0.010)	(0.011)	(0.010)	(0.011)	(0.010)
MS-lasso	0.388	0.399	0.389	0.443	0.431	0.441	0.523	0.514	0.506
	(0.007)	(0.009)	(0.007)	(0.008)	(0.007)	(0.009)	(0.008)	(0.011)	(0.008)
MS-IGA	0.113	0.112	0.118	0.149	0.159	0.150	0.174	0.177	0.174
	(0.004)	(0.006)	(0.009)	(0.008)	(0.011)	(0.008)	(0.005)	(0.004)	(0.004)
$\bar{C}_i$									
B-lasso	5.00	5.00	5.00	10.00	10.00	10.00	15.00	15.00	15.00
MS-lasso	4.99	4.96	4.99	10.00	10.00	9.99	15.00	14.92	15.00
MS-IGA	5.00	5.00	4.95	10.00	9.92	10.00	15.00	15.00	15.00
$\overline{IC}_i$									
B-lasso	467.37	466.90	468.20	461.21	462.13	462.72	455.16	456.6	457.47
MS-lasso	2.45	2.23	2.22	10.50	11.86	13.13	20.15	19.48	20.42
MS-IGA	0.22	0.33	0.25	0.32	0.40	0.35	0.32	0.20	0.31

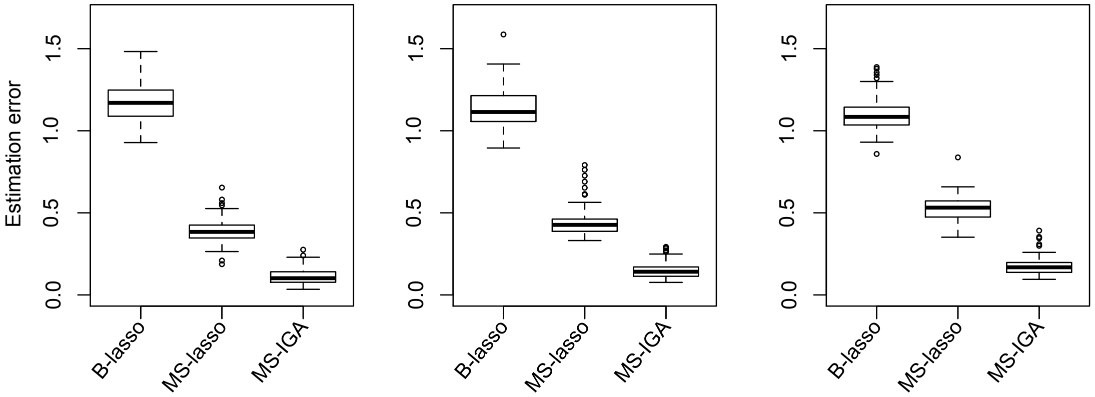

Figure 6: Boxplots for coefficient estimation error of arm 1 from output of different bandit algorithms ($\vartheta = 0.4$). Left panel: $q_0 = 5$; middle panel: $q_0 = 10$; right panel: $q_0 = 15$.

satisfactory performance by MS-IGA were observed under an increased coefficient signal with $\vartheta = 0.4$ (Table 5). In addition, boxplots for arm 1's coefficient estimation errors are given in Figure 5 with $\vartheta = 0.2$ and Figure 6 with $\vartheta = 0.4$. The boxplots for arm 2 and arm 3 are similar to arm 1 (Figures 7–10). The averaged coefficient estimation from MS-IGA outperformed that of the other two alternatives in all cases, and the advantage of MS-IGA appears to be widened with $\vartheta = 0.4$ compared to that of $\vartheta = 0.2$; this observation coincides with Theorem 5, which suggests that MS-IGA may become more favorable with strong signals and small $\bar{q}_i$.

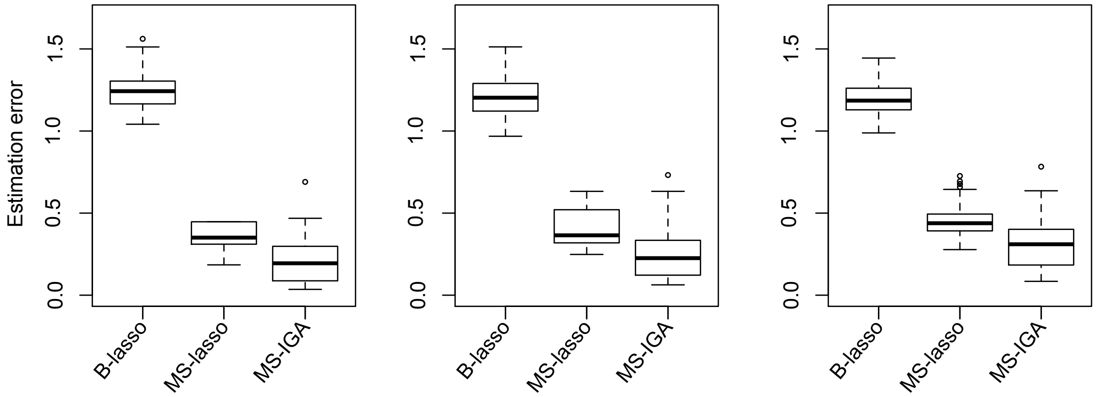

Figure 7: Boxplots for coefficient estimation errors of arm 2 from output of different bandit algorithms ($\vartheta = 0.2$). Left panel: $q = 5$; middle panel: $q = 10$; right panel: $q = 15$.

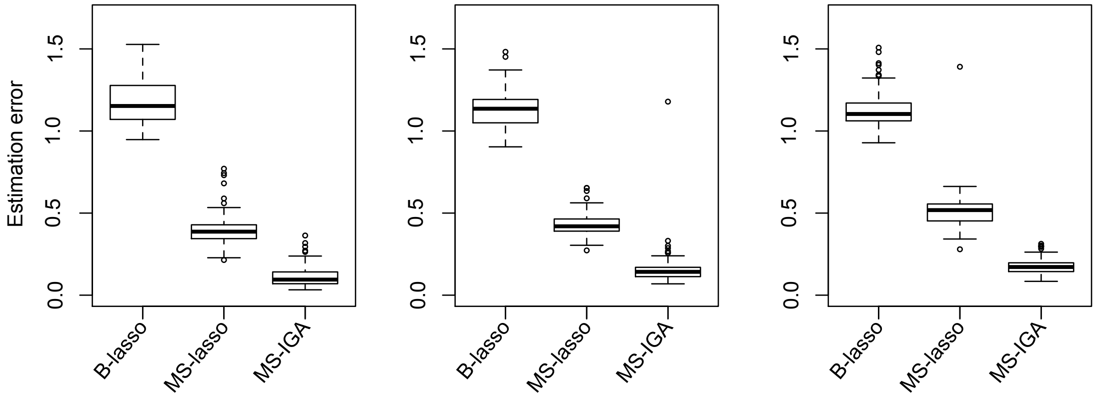

Figure 8: Boxplots for coefficient estimation errors of arm 2 from output of different bandit algorithms ($\vartheta = 0.4$). Left panel: $q_0 = 5$; middle panel: $q_0 = 10$; right panel: $q_0 = 15$.

D.2. Performance with different parameter values

To provide more guidance on our proposal's empirical applications, we performed further evaluation on the sensitivity of the MS-IGA algorithm with different parameter value choices. In particular, note that c is a parameter for arm screening/elimination and h is for randomized allocation. We first considered $c = 0.5$, 0.75, or 1 while keeping all other experimental settings exactly the same as that in Section D.1. The averaged per-round regret of MS-IGA with different values of c is summarized in Table 6, and the averaged results on coefficient estimation are given

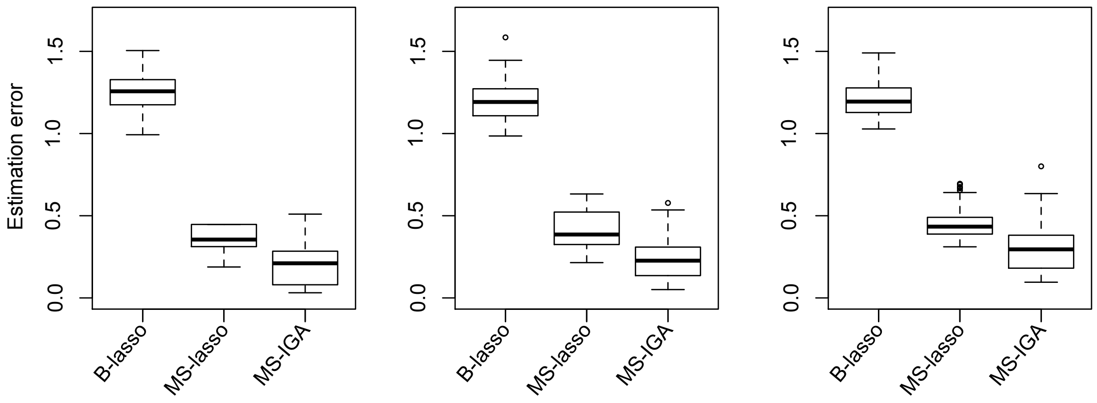

Figure 9: Boxplots for coefficient estimation errors of arm 3 from output of different bandit algorithms ($\vartheta = 0.2$). Left panel: $q = 5$; middle panel: $q = 10$; right panel: $q = 15$.

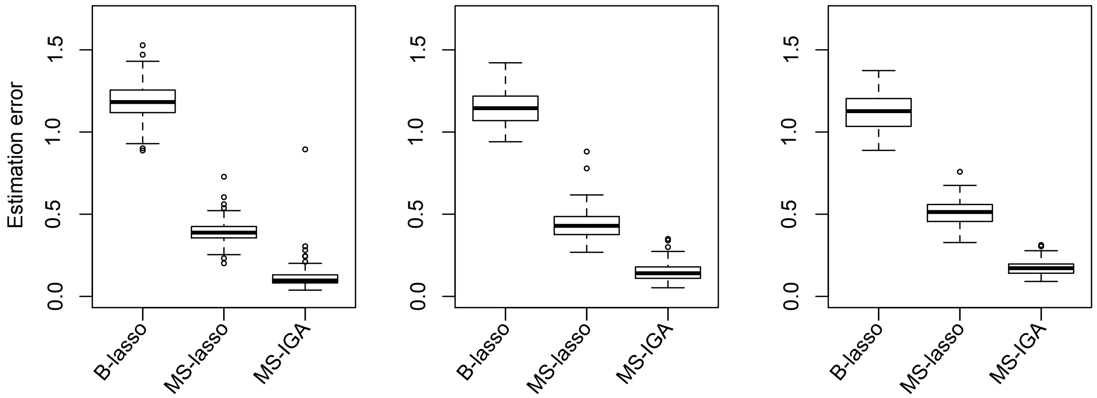

Figure 10: Boxplots for coefficient estimation errors of arm 3 from output of different bandit algorithms ($\vartheta = 0.4$). Left panel: $q_0 = 5$; middle panel: $q_0 = 10$; right panel: $q_0 = 15$.

in Table 7. In addition, we considered different randomization parameters $h = 7$ and $h = 10$ while keeping $c = 1$; the results on regret and coefficient estimation are summarized in Table 8 and Table 9.

Table 6: Averaged per-round regret of MS-IGA with different c values on simulated data.

	$\vartheta = 0.2$			$\vartheta = 0.4$		
q_0	5	10	15	5	10	15
$c = 0.5$	0.203 (0.005)	0.289 (0.007)	0.368 (0.010)	0.230 (0.013)	0.316 (0.018)	0.448 (0.030)
$c = 0.75$	0.199 (0.004)	0.273 (0.004)	0.329 (0.006)	0.193 (0.011)	0.236 (0.011)	0.307 (0.016)
$c = 1$	0.202 (0.004)	0.274 (0.004)	0.317 (0.005)	0.177 (0.006)	0.221 (0.008)	0.273 (0.012)

It can be seen from Table 6 that the regret of MS-IGA often increased if we set the c value too small; this can be explained by the over-elimination of competitive arms, which inaccurately

Table 7: Averaged simulation results of MS-IGA with different c values on simulated data ($\vartheta = 0.2$).

	$q_0 = 5$			$q_0 = 10$			$q_0 = 15$		
Arm i	1	2	3	1	2	3	1	2	3
$\bar{n}_i$									
$c = 0.5$	3403	3371	3226	3059	3572	3369	3487	3333	3180
$c = 0.75$	3450	3293	3257	3163	3435	3402	3471	3206	3323
$c = 1$	3413	3253	3334	3202	3397	3401	3497	3212	3290
Avg. $\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2$									
$c = 0.5$	0.19	0.205	0.213	0.293	0.256	0.280	0.322	0.343	0.340
	(0.013)	(0.014)	(0.014)	(0.018)	(0.015)	(0.023)	(0.021)	(0.019)	(0.024)
$c = 0.75$	0.170	0.200	0.208	0.264	0.252	0.250	0.277	0.331	0.297
	(0.010)	(0.013)	(0.013)	(0.014)	(0.013)	(0.013)	(0.015)	(0.016)	(0.016)
$c = 1$	0.175	0.206	0.205	0.250	0.256	0.238	0.254	0.311	0.306
	(0.011)	(0.012)	(0.012)	(0.014)	(0.015)	(0.012)	(0.012)	(0.014)	(0.014)
$\bar{C}_i$									
$c = 0.5$	4.26	4.02	3.86	7.95	8.56	8.43	12.64	12.64	12.48
$c = 0.75$	4.43	4.33	4.02	8.77	8.84	9.00	13.66	13.04	13.46
$c = 1$	4.46	4.26	4.26	8.77	8.60	9.02	14.14	13.56	13.5
$\overline{IC}_i$									
$c = 0.5$	0.74	0.71	0.64	1.06	0.83	1.03	1.37	1.81	1.46
$c = 0.75$	0.61	1.11	0.66	1.15	0.95	1.19	1.08	1.59	1.29
$c = 1$	0.75	0.96	1.08	0.92	0.90	0.85	1.09	1.64	1.37

Table 8: Averaged per-round regret of MS-IGA with different h values on simulated data.

	$\vartheta = 0.2$			$\vartheta = 0.4$		
q_0	5	10	15	5	10	15
$h = 4$	0.202 (0.004)	0.274 (0.004)	0.317 (0.005)	0.177 (0.006)	0.221 (0.008)	0.273 (0.012)
$h = 7$	0.195 (0.004)	0.265 (0.003)	0.312 (0.005)	0.175 (0.006)	0.218 (0.008)	0.275 (0.012)
$h = 10$	0.190 (0.004)	0.263 (0.004)	0.315 (0.005)	0.177 (0.006)	0.227 (0.008)	0.286 (0.011)

removes “promising” arms from candidate sets for randomized allocation. Table 7 for $\vartheta = 0.2$ (along with Table 10 for $\vartheta = 0.4$) also shows that an overly small c (such as $c = 0.5$) sometimes leads to increased coefficient estimation errors and less ideal variable selection results, and the performance of MS-IGA can be sensitive to small c values. A simple choice of $c = 1$ often gave reasonable performance compared to those smaller alternatives. On the other hand, as shown in Table 8, the use of a larger randomization parameter h with $h = 7$ or 10 did not significantly change the averaged per-round regret in most cases. In addition, we observed no clear patterns in the change of the coefficient estimation and variable selection performance from Table 9 with $\vartheta = 0.2$ (and Table 11 with $\vartheta = 0.4$); differences in averaged estimation errors were mostly not significant in these cases. These seem to suggest that MS-IGA is relatively robust to these

Table 9: Averaged simulation results of MS-IGA with different h values on simulated data ($\vartheta = 0.2$).

	$q_0 = 5$			$q_0 = 10$			$q_0 = 15$		
Arm i	1	2	3	1	2	3	1	2	3
$\bar{n}_i$									
$h = 4$	3413	3253	3334	3202	3397	3401	3497	3212	3290
$h = 7$	3352	3215	3433	3204	3369	3427	3484	3235	3281
$h = 10$	3269	3304	3426	3138	3492	3370	3516	3106	3378
Avg. $\lVert\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\rVert_2$									
$h = 4$	0.175	0.206	0.205	0.250	0.256	0.238	0.254	0.311	0.306
	(0.011)	(0.012)	(0.012)	(0.014)	(0.015)	(0.012)	(0.012)	(0.014)	(0.014)
$h = 7$	0.192	0.203	0.197	0.270	0.261	0.251	0.275	0.322	0.311
	(0.012)	(0.013)	(0.013)	(0.014)	(0.017)	(0.015)	(0.014)	(0.017)	(0.018)
$h = 10$	0.219	0.201	0.190	0.294	0.256	0.265	0.293	0.360	0.299
	(0.014)	(0.013)	(0.014)	(0.019)	(0.016)	(0.015)	(0.017)	(0.019)	(0.015)
$\bar{C}_i$									
$h = 4$	4.46	4.26	4.26	8.77	8.60	9.02	14.14	13.56	13.50
$h = 7$	4.19	4.04	4.15	8.79	8.68	8.81	13.79	13.12	12.97
$h = 10$	4.02	4.00	4.02	8.03	8.53	8.56	13.42	12.34	13.58
$\overline{IC}_i$									
$h = 4$	0.75	0.96	1.08	0.92	0.90	0.85	1.09	1.64	1.37
$h = 7$	0.67	0.58	0.74	1.49	1.30	1.10	1.24	1.67	1.00
$h = 10$	0.79	0.58	0.46	1.10	0.91	0.93	1.38	1.66	1.44

different choices of h. Accordingly, in the real data evaluation studies of Section 8, for any MS algorithm, we simply used the last parameters in Tables 6 and 8 (that is, $c = 1$ and $h = 10$) across all experiments.

Table 10: Averaged simulation results of MS-IGA with different c values on simulated data ($\vartheta = 0.4$).

	$q_0 = 5$			$q_0 = 10$			$q_0 = 15$		
Arm i	1	2	3	1	2	3	1	2	3
$\bar{n}_i$									
$c = 0.5$	3132	3351	3517	3388	3492	3119	3395	3491	3113
$c = 0.75$	3143	3379	3478	3312	3363	3324	3319	3298	3383
$c = 1$	3251	3421	3328	3359	3339	3301	3345	3312	3343
Avg. $\\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\\|_2$									
$c = 0.5$	0.196	0.182	0.157	0.222	0.204	0.293	0.282	0.307	0.347
	(0.030)	(0.023)	(0.018)	(0.029)	(0.022)	(0.042)	(0.038)	(0.042)	(0.050)
$c = 0.75$	0.143	0.125	0.125	0.153	0.155	0.148	0.201	0.222	0.175
	(0.014)	(0.010)	(0.010)	(0.012)	(0.008)	(0.005)	(0.020)	(0.023)	(0.005)
$c = 1$	0.113	0.112	0.118	0.149	0.159	0.150	0.174	0.177	0.174
	(0.004)	(0.006)	(0.009)	(0.008)	(0.011)	(0.008)	(0.005)	(0.004)	(0.004)
$\bar{C}_i$									
$c = 0.5$	4.64	4.71	4.75	9.30	9.62	8.80	13.80	13.49	13.25
$c = 0.75$	4.85	4.95	4.94	9.90	9.96	10.00	14.70	14.59	15.00
$c = 1$	5.00	5.00	4.95	10.00	9.92	10.00	15.00	15.00	15.00
$\overline{IC}_i$									
$c = 0.5$	0.28	0.56	0.35	0.35	0.49	0.28	0.46	0.20	0.21
$c = 0.75$	0.29	0.43	0.29	0.21	0.37	0.22	0.32	0.36	0.36
$c = 1$	0.22	0.33	0.25	0.32	0.40	0.35	0.32	0.20	0.31

Table 11: Averaged simulation results of MS-IGA with different h values on simulated data ($\vartheta = 0.4$).

	$q_0 = 5$			$q_0 = 10$			$q_0 = 15$		
Arm i	1	2	3	1	2	3	1	2	3
$\bar{n}_i$									
$h = 4$	3251	3421	3328	3359	3339	3301	3345	3312	3343
$h = 7$	3267	3413	3320	3339	3354	3306	3314	3329	3357
$h = 10$	3246	3420	3334	3364	3316	3320	3335	3278	3387
Avg. $\\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\\|_2$									
$h = 4$	0.113	0.112	0.118	0.149	0.159	0.150	0.174	0.177	0.174
	(0.004)	(0.006)	(0.009)	(0.008)	(0.011)	(0.008)	(0.005)	(0.004)	(0.004)
$h = 7$	0.135	0.132	0.131	0.148	0.155	0.158	0.185	0.184	0.179
	(0.009)	(0.009)	(0.009)	(0.005)	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)
$h = 10$	0.151	0.141	0.128	0.147	0.170	0.162	0.180	0.190	0.165
	(0.013)	(0.014)	(0.011)	(0.005)	(0.012)	(0.006)	(0.005)	(0.008)	(0.003)
$\bar{C}_i$									
$h = 4$	5.00	5.00	4.95	10.00	9.92	10.00	15.00	15.00	15.00
$h = 7$	4.97	4.96	4.95	10.00	9.97	10.00	15.00	15.00	15.00
$h = 10$	4.90	4.85	4.93	10.00	9.90	10.00	14.99	14.98	15.00
$\overline{IC}_i$									
$h = 4$	0.22	0.33	0.25	0.32	0.40	0.35	0.32	0.20	0.31
$h = 7$	0.48	0.50	0.33	0.32	0.40	0.49	0.49	0.53	0.43
$h = 10$	0.42	0.40	0.22	0.23	0.37	0.41	0.32	0.47	0.15