Self-fulfilling Bandits: Dynamic Selection in Algorithmic Decision-making

Jin Li, Ye Luo, Xiaowei Zhang*
Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, jli1@hku.hk, kurtluo@hku.hk, xiaoweiz@hku.hk

This paper identifies and addresses dynamic selection problems that arise in online learning algorithms with endogenous data. In a contextual multi-armed bandit model, we show that a novel bias (*self-fulfilling bias*) arises because the endogeneity of the data influences the choices of decisions, affecting the distribution of future data to be collected and analyzed. We propose a class of algorithms to correct for the bias by incorporating instrumental variables into leading online learning algorithms. These algorithms lead to the true parameter values and meanwhile attain low (logarithmic-like) regret levels. We further prove a central limit theorem for statistical inference of the parameters of interest. To establish the theoretical properties, we develop a general technique that untangles the interdependence between data and actions.

Key words: Self-fulfilling bias, dynamic selection, endogeneity spillover, contextual multi-armed bandit model

---

* We are grateful for helpful conversations with Chunrong Ai, Xi Chen, Pingyang Gao, Bob Gibbons, John Klopfer, Danielle Li, Wing Suen, Chang Sun, Brian Viard, and seminar participants at University of Hong Kong and Chinese University of Hong Kong (Shenzhen). All remaining errors are ours.

1. Introduction

Due to better data availability and improved computing power in recent years, organizations are increasingly using algorithms to make decisions that depend on the characteristics of individual customers. For example, Netflix recommends movies based on the past browsing history of its users, Amazon suggests products based on individual shopping history, and Facebook displays advertisements that are customer-dependent. Unlike the widely used A/B testing[1], these algorithms do not wait for the final outcome of the experiments before choosing their actions. Instead, they engage in *online* learning so that their choices of actions are constantly adjusted in response to the new information and feedback from the behaviors of the targeted users.

In addition to e-commerce recommendations, there are many business applications for these online learning algorithms (Blattner, Nelson, and Spiess 2021). Banks can use past data to decide whether to approve a loan for applicants of particular characteristics, and if approved, the information on how the applicant repays the loan can be used to update their decision rules in the future. Schools and companies can use existing data to decide whether to interview a candidate of a certain background, and the outcome of the interview can be used to update their future interview criterion. As more organizations are going through digital transformation, these online learning algorithms will become more important. The competitiveness of the organizations will also increasingly depend on the effectiveness of these algorithms because they influence key business policies such as hiring and customer recommendation.

The existing studies on online learning algorithms, carried out mostly by computer scientists (Chu et al. 2011, Agrawal and Goyal 2013, Slivkins 2019), have typically assumed that the model for the data generating process accurately reflects the problem to be studied. In particular, they have not paid attention to the endogeneity problem in data, such as the omitted variables, sample selection, and model mis-specification and etc. The endogeneity problems are, of course, essential for analyzing human behaviors and are prevalent in the applications of these online learning algorithms. The endogeneity issue also introduces a new type of *dynamic selection* problem in online learning algorithms because the data analysis and decision-making interact. Essentially, the endogeneity in the data will influence the decisions—whether to approve a loan application or to interview a candidate—and since the counterfactual outcomes of those who are turned away are never observed, the decisions lead to a selected sample to be analyzed, causing further bias in the data analysis.

[1] The use of A/B testing or randomized controlled experiments in social sciences has received substantial interests in recent years (Blake, Nosko, and Tadelis 2015, Athey and Imbens 2017, Koning, Hasan, and Chatterji 2019, Kohavi, Tang, and Xu 2020, Azevedo et al. 2020.)

The purpose of this paper is to identify and address the dynamic selection problems that arise in online learning algorithms with data endogeneity. Our setup is a contextual multi-armed bandit problem. An agent (e.g., a website) in each period first observes a vector of covariates (e.g., visitor characteristics such as age, gender, browsing history, etc.), and then decides on which arm to select (e.g., banners to display). The decision gives the agent a random reward in each period. The agent does not know the expected rewards of pulling the arms. In deciding which arm to pull, he balances exploitation (pulling the arm with a higher expected reward) with exploration (pulling the arm that generates information that allows him to make better decisions in the future). This bandit problem is contextual in nature because the agent's expected reward from pulling an arm depends on the vector of covariates: the website's choice of which banner to display depends on the characteristics of the visitor.

Unlike typical contextual bandit problems, we assume that there exists endogeneity in the data, which may arise because of a number of reasons, such as measurement error, model mis-specification, sample selection, omitted variables, and so on. For example, suppose that when the consumer sentiment is high, the visitors to a website are more likely to click on the banners. Suppose also that the website is more likely to attract older customers when the consumer sentiment is high. In this case, the age of the customers (covariate) becomes correlated with the clicks (reward) through an omitted variable (consumer sentiment), thereby leading to an endogeneity problem.

Our first contribution is to identify a novel type of bias that arises in this situation. We refer to this bias as *self-fulfilling bias* because actions (choices of arms) and beliefs regarding the rewards of the arms (that arise from the data analysis) are entangled. Wrong beliefs regarding the rewards can lead to wrong actions, and these actions can then generate data that actually vindicate the beliefs. In online learning environments, the actions are not purely random as the decision policy adapts to past data. When the noise is correlated with the covariates, the noise also becomes correlated with the actions, causing them to become endogeneous. We refer to this as the *endogeneity spillover*, which is the key reason underlying the self-fulfilling bias.

Our second contribution is to propose algorithms that not only correct for the self-fulfilling bias but also generate actions that obtain low levels of regret (loss in payoff compared to optimal actions). Our method incorporates the instrumental variable (IV) approach to leading online learning algorithms for contextual bandit problems—that is, algorithms of Greedy or upper-confidence-bound (UCB) types. Our proposed algorithms consist of three phases. In Phase 1, arms are randomly selected. At the end of Phase 1, the two-stage least squares (2SLS) method is implemented on the data associated with each arm to compute the arm-specific coefficient estimates. In Phase 2, the

arm with the highest expected reward (using the coefficient estimates in Phase 1) is selected in each period. The covariance matrix of the estimates are calculated at the end of Phase 2. In Phase 3, arms are selected according to the particular algorithm used—Greedy or UCB—and the coefficients and covariance matrices are updated using "joint-2SLS" in each period. Under these algorithms, the coefficients converge to the true value in the long run. Depending on whether the Greedy or UCB algorithm is used, the regret of our algorithm is of order $\mathcal{O}(\log(T))$ and $\mathcal{O}(\log^2(T))$, respectively.

Our third contribution is to develop a technique that facilitates the theoretical analysis of online learning algorithms with dynamic selection problems. As mentioned above, a key feature of online learning algorithms is that data and actions are intertwined. This complicates the analysis because the effects of past actions can be long-lasting. In particular, when a poor estimate of the coefficients of the covariates causes the algorithm to select a wrong arm, it contaminates the generated data by changing the distribution of actions that are endogenous in the future (because the data associated with the optimal arm is no longer observed). The contaminated data could then worsen the future coefficient estimates that are calculated based on it. The algorithm is potentially trapped in a vicious cycle—the coefficient estimates are inconsistent and the policy is sub-optimal even in the long run.

Mathematically, in our algorithms and in online learning algorithms in general, this complication manifests in a two-way dependence between (i) the estimates of the coefficients and (ii) the estimates of the covariance matrix between the IVs and an augmented vector—that is induced by the actions—of the covariates. On the one hand, in 2SLS, the coefficient estimates depend on the covariance estimates. On the other hand, each covariance estimate depends on *all* the past coefficient estimates because they affect the past actions and the past realization of data. This latter dependence is new and central to the analysis of online learning algorithms with selection problems. Moreover, the dependence is complicated both because the coefficient estimates are serially correlated and because they affect the estimates of the covariance matrix in a nonlinear manner.

Our technique deals with this two-way dependence by performing an induction in a zig-zag matter. In each induction step, we first derive the properties of the covariance estimates (zig), and then use this property to establish the induction hypothesis on coefficient estimates (zag). The heart of the induction is the zig-part, which employs a special technique to eliminate the dependence of the covariate estimates on all past coefficient estimates. The theoretical analysis of this dependence also indicates that, for the coefficient estimates to converge to the true value, the covariance estimates at the beginning of Phase 3 must be sufficiently close to the covariance matrix induced by the optimal policy. This motivates Phase 2 in our algorithm, which stabilizes the covariance estimates.

Finally, this technique establishes the rate of convergence of the estimates. These rates are used to analyze the regret of the algorithm. They are also used to establish the asymptotic distribution of the parameter estimators at the terminal period of the algorithm and, therefore, allow for the construction of confidence intervals and hypothesis testing.

Our paper contributes to the literature on dynamic learning. Within this literature, a key emphasis has been the trade-off between exploration and exploitation (March 1991). The workhorse for analyzing this trade-off is the multi-armed bandit models. Economists have extensively studied their theoretical and empirical implications (see Bergemann and Välimäki 2008 for a review). Recently, there is a growing sub-literature that studies algorithmic decision-making based on multi-armed bandit models (Perchet et al. 2016, Caria et al. 2020, Currie and MacLeod 2020, Kasy and Sautmann 2021, Kawaguchi 2021, Kawaguchi, Uetake, and Watanabe 2021)[2]. The closest one to our study is that of Li, Raymond, and Bergman (2020), which also examines the performances of UCB-type algorithms on the contextual bandit problem. We add to this sub-literature by studying the biases that arise in algorithmic decision-making when there is an endogeneity problem, and we also show how to correct for it.

Our paper also contributes to the vast literature of casual inference (see, for example, Imbens and Rubin 2015 for a review). Most of this literature does not examine the environment of online learning, where the actions influence and are influenced by data. Two notable exceptions are Nambiar, Simchi-Levi, and Wang (2019) and Li, Luo, and Zhang (2021). Nambiar, Simchi-Levi, and Wang (2019) consider a dynamic pricing problem with learning, where the endogeneity problem arises from mis-specification of demand functions. Li, Luo, and Zhang (2021) consider a Markov decision process setup with learning and show that online learning environments may exacerbate the estimation bias. Different from these papers, our focus is on the dynamic selection problem where current data, through affecting the choice of actions, affects how future data is generated. This dynamic selection problem also differs from selection problems in Heckman (1979, 1990), Rosenbaum and Rubin (1983), Altonji, Elder, and Taber (2005), and Oster (2019), which do not feature how past selection problems affect future data selection.

The remainder of the paper is organized in the following manner. Section 2 introduces the background of the contextual bandit problem. Section 3 discusses the case where there exists at least one endogenous variable in the contextual bandit problem, which leads to the dynamic

[2] While we focus on online learning environments, there is a rich line of work that studies offline learning of optimal policies *after* the experimental or observational data is collected (Kitagawa and Tetenov 2018, Narita, Yasui, and Yata 2019, Athey and Wager 2021, Kallus and Zhou 2021, Narita, Yasui, and Yata 2021, Shi et al. 2021, Zhan et al. 2021).

selection problem and self-fulfilling bias. Section 4 presents a set of regularity conditions for IVs and the contextual bandit setup as well as proposes a class of IV-based bandit algorithms. Section 5 establishes the main results including regret bounds and asymptotic distribution for the proposed algorithms. Section 6 demonstrates the performance of these algorithms in a series of simulations. Section 7 presents the conclusions. Technical proofs are included the online Supplemental Material.

2. Background

Consider an agent who makes a sequence of decisions over a time horizon of T periods. At each time $t = 1, \ldots, T$, the agent observes a vector of covariates $\mathbf{v}_t \in \mathbb{R}^p$, takes an action (i.e., pulls an arm) $a_t \in \mathcal{A} = \{1, \ldots, M\}$ with $M \geq 2$ being a fixed positive integer, and receives a *random* reward R_t:

$$R_t = \sum_{i=1}^{M} \mathbb{I}(a_t = i)\mu_i(\mathbf{v}_t) + \epsilon_t, \tag{1}$$

where $\mu_i : \mathbb{R}^p \mapsto \mathbb{R}$ is an unknown function, representing the expected reward associated with arm i for a given covariate, $i = 1, \ldots, M$, and ϵ_t is the noise having mean zero and standard deviation (SD) σ. We use $\mathcal{H}_t \coloneqq (\mathbf{v}_1, a_1, R_1, \ldots, \mathbf{v}_t, a_t, R_t)$ to denote the history up to the end of period t. The agent seeks a *nonanticipating* policy π that maps $(\mathcal{H}_{t-1}, \mathbf{v}_t)$ to a_t to maximize the expected cumulative reward over time. We write $R_t^\pi = R_t$ to emphasize its dependence on π.

This sequential decision-making problem is commonly known as the contextual bandit problem (Slivkins 2019, chapter 8). For this class of problems, a standard metric for measuring the performance of π is its *expected cumulative regret*, which is referred to just as "regret" hereafter. Specifically, we compare π with an oracle policy π^*, which knows the functions $\{\mu_i\}_{i=1}^M$ *a priori* and pulls arm $\pi^*(\mathbf{v}_t) \coloneqq \arg\max_{i=1,\ldots,M} \mu_i(\mathbf{v}_t)$ at each time t. Following policy π, the agent incurs regret

$$\mathsf{Regret}^\pi(T) \coloneqq \sum_{t=1}^{T} \mathbb{E}\big[R_t^{\pi^*} - R_t^\pi\big], \tag{2}$$

which measures the difference in expected cumulative reward between π^* and π. Thus, the agent aims to approach the oracle's performance by learning the reward functions gradually. In the sequel, we assume, for simplicity, that for each arm i, the reward function μ_i is linear in the covariate vector—that is, $\mu_i(\mathbf{v}) = \mathbf{v}^\intercal \boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i \in \mathbb{R}^p$ is a vector of unknown parameters. Then, the problem is reduced to learning these linear coefficients over time.

When the covariates are exogenous, a number of algorithms of Greedy or UCB types have been proposed to learn the coefficients. These algorithms differ in their finite-time and asymptotic properties, but they learn the true value of coefficients in the long run and, therefore, have a *sub-linear* regret—that is, $T^{-1}\mathsf{Regret}^\pi(T) \to 0$ as $T \to \infty$. This basically implies that at least the agent is taking the optimal action in most of the time periods as the horizon increases.

3. Self-fulfilling Bias: An Example

We now study a case in which the covariates are endogenous—that is, $\mathbb{E}[\epsilon_t|\mathbf{v}_t] \neq 0$. We show that the endogeneity of covariates leads to distortion in actions, which generates another type of endogeneity; that is, it creates "endogeneity spillover". The additional endogeneity generates a dynamic selection problem and leads to a new type of bias, which we refer to as self-fulfilling bias.

Consider the following special case of our setup. Assume that there exist two arms, and between the two arms $\mathcal{A} = \{1, 2\}$, one is the safe arm, having a *known* expected reward that is independent of the covariates, while the other is the risky arm, whose expected reward has an unknown linear dependence on the covariates. Further, assume that the covariates are one-dimensional. In this simple example, the reward functions are

$$\mu_1(v) \equiv c \quad \text{and} \quad \mu_2(v) = \alpha v, \quad \forall v \in \mathbb{R}.$$

It follows that, if α is known and positive, then the optimal policy for the agent is to pull arm 2 if $v > c/\alpha$, and pull arm 1 otherwise.

However, because α is unknown, the agent needs to estimate its value over time and makes his decisions accordingly. In contrast to the standard regression analysis in which data is taken as given, in the dynamic environment considered here, the data used for estimating α is only available when arm 2 is pulled (pulling arm 1 generates noisy observations of c, which are irrelevant to α). Now, suppose the agent uses ordinary least squares (OLS) to estimate α and the estimates converge to a limit, say $\widehat{\alpha}$. Then, in the long run,

$$\widehat{\alpha} = \frac{\text{Cov}[v_t, R_t | R_t \text{ is generated from arm 2}]}{\text{Var}[v_t | R_t \text{ is generated from arm 2}]}. \tag{3}$$

Note that given $\widehat{\alpha}$, the agent's policy in the long run is to pull arm 2 when $v > c/\widehat{\alpha}$. Thus, the conditioning event in equation (3) is identical to $\{v_t > c/\widehat{\alpha}\}$. Using the fact that $R_t = \alpha v_t + \epsilon_t$ for arm 2, we have

$$\widehat{\alpha} = \alpha + \frac{\text{Cov}[v_t, \epsilon_t | v_t > c/\widehat{\alpha}]}{\text{Var}[v_t | v_t > c/\widehat{\alpha}]}. \tag{4}$$

It is evident from equation (4) that $\widehat{\alpha}$ is a fixed point as it appears on both sides of the equation. It also shows that the bias of the estimate, $\widehat{\alpha} - \alpha$, has an expression that is similar to, but different from, that of the usual OLS bias, which is given by $\frac{\text{Cov}[v_t, \epsilon_t]}{\text{Var}[v_t]}$. We refer to the difference between them, $(\widehat{\alpha} - \alpha) - \frac{\text{Cov}[v_t, \epsilon_t]}{\text{Var}[v_t]}$, as the *self-fulfilling bias* to reflect that the limit policy of the agent is induced by his limit belief (i.e., the limit estimate $\widehat{\alpha}$), and the limit belief is confirmed by the data generated from the limit policy.

The self-fulfilling bias arises because there are two—rather than one—types of endogeneity problems. First, when correlation exists between the noise and the covariates, it directly creates a bias in the estimate in a manner that is similar to the usual OLS bias. Second, this bias affects the agent's actions, creating another source of endogeneity, which is essentially a type of bias caused by sample selection. Moreover, such sample selection is dynamic because when the policy that determines agent's actions, the estimate of α changes over time.

To illustrate the problem formally, we rewrite equation (1) as

$$R_t = c + \alpha v_t \mathbb{I}(a_t = 2) - c\,\mathbb{I}(a_t = 2) + \epsilon_t. \tag{5}$$

When ϵ_t and v_t are correlated, the only endogenous variable in this regression is $v_t\,\mathbb{I}(a_t = 2)$ if the arm a_t is selected independent of v_t. But because a_t depends on v_t—through the relation $\mathbb{I}(a_t = 2) = \mathbb{I}(v_t > c/\widehat{\alpha})$ in the long run—the correlation between ϵ_t and v_t leads to a correlation between $\mathbb{I}(a_t = 2)$ and ϵ_t. Therefore, the regression equation (5) contains two endogenous variables—$v_t\,\mathbb{I}(a_t = 2)$ and $\mathbb{I}(a_t = 2)$, reflecting the endogeneity spillover problem.

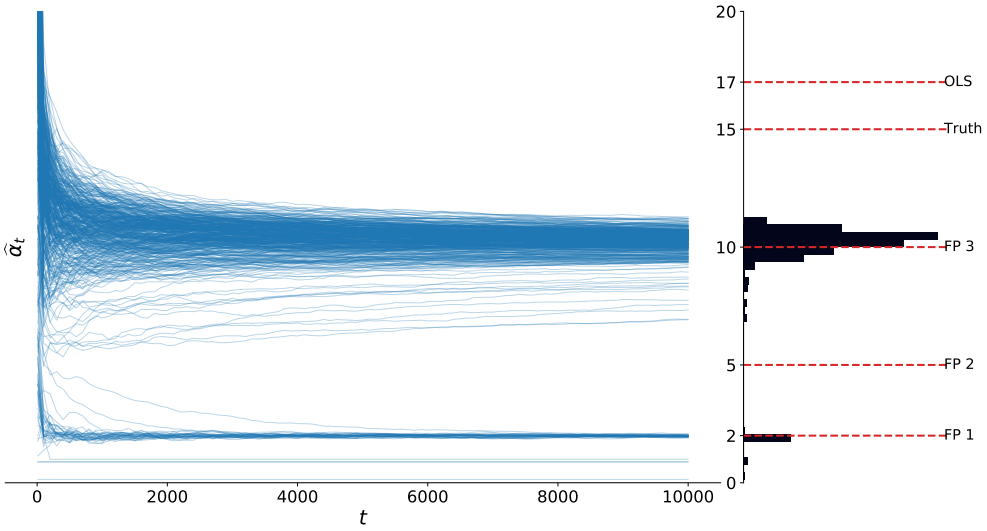
The additional endogenous variable complicates the dependence of the value of the limit coefficient estimate ($\widehat{\alpha}$) on the joint distribution of v_t and ϵ_t. In particular, if the term $\frac{\mathrm{Cov}[v_t,\epsilon_t|v_t>s]}{\mathrm{Var}[v_t|v_t>s]}$ is the same for all cutoff s—as is the case when v_t and ϵ_t have a joint normal distribution[3]—then $\widehat{\alpha}$ is identical to the OLS estimate. In general, however, the coefficient estimate can be either higher or lower than the OLS estimate, depending on the joint distribution of v_t and ϵ_t. When $\mathbb{E}[\epsilon_t|v_t]$ is a nonlinear function of v_t, there will be additional bias in general.

More importantly, the value of $\widehat{\alpha}$ (and hence the value of the self-fulling bias) may not be unique, which happens when equation (4) contains multiple solutions. Figure 1 illustrates one such case and shows that the agents' beliefs converge to two different values in the long run. With multiple self-fulfilling biases, this example shows that agents can adopt different policies in the long run, even if the underlying production environment is the same. It therefore provides one explanation for why seemingly similar enterprises may choose different practices and result in persistent differences in performance, which is a central question in organizational economics and strategy (see, for example, Gibbons and Henderson 2012 for a review).

Comment 1. There appear to be little restrictions on the relationship between the true value α and the belief that an agent has in the long run. Effectively, for a given α, the beliefs regarding α may converge in the long run to *any* number of *arbitrarily chosen* values in the long run. In other words, for a group of agents with the same initial beliefs, they can hold many very different beliefs in the long run. We discuss why this is the case in the Appendix. □

[3] If $\mathbb{E}[\epsilon_t|v_t]$ is a linear function of v_t, then $\frac{\mathrm{Cov}[v_t,\epsilon_t|v_t>s]}{\mathrm{Var}[v_t|v_t>s]}$ is a constant with respect to s. That (v_t, ϵ_t) has a joint normal distribution is a special case of this condition.

Figure 1 Multiplicity of the Self-fulfilling Bias



Note. Suppose that $c=1$, $\alpha=15$, v_t is uniformly distributed on $[0,1]$, and $\epsilon_t=\sum_{k=0}^{3}\beta_k v_t^k+\eta_t$, where β_k's are some constants and η_t is an independent standard normal random variable. Then, (i) the limit of the OLS estimate is 17 and (ii) equation (4) is reduced to a cubic equation in terms of $\widehat{\alpha}$, thereby yielding three distinct real fixed points (FPs)—2, 5, and 10—if β_k's are appropriately selected (see the Appendix for details). The left portion of this figure shows 500 sample paths of estimates of α following a greedy policy, while the right portion shows the histogram of the terminal values of these sample paths. The simulation results indicate that the greedy estimates mostly converge to either 2 or 10.

We illustrate the self-fulfilling bias using the simplest possible example. The underlying problem is more general because in any dynamic environment where the agent's actions depend on the covariates, endogeneity in the covariates will likely generate additional endogeneity with the actions. Note that when the coefficient estimate differs from the true parameter in the long run because of the self-fulling bias, the fraction of time a wrong decision is taken is positive in the limit. In other words, not only is the policy suboptimal in the long run, the number of periods in which a wrong action is taken is proportional to the time horizon T. Thus, the regret of the policy with self-fulfilling bias is in the order of $\mathcal{O}(T)$.

To correct for the self-fulfilling bias, it is tempting to have the agent to carry out the exploration in the conventional sense, i.e., to choose actions randomly sometimes. However, as long as the data from non-random actions is used, the agent's parameter estimate is biased because the endogeneity spillover problem renders $\mathbf{v}_t\,\mathbb{I}(a_t=2)$ an endogenous term in the regression equation (5). To correct for the endogeneity in $\mathbf{v}_t$, the solution is not more exploration in the conventional sense. Rather, it is *causal exploration*—that is, to use the IVs to "perturb" the data-generating process. We discuss how to correct for the self-fulfilling bias in the next section.

4. IV-Bandit Algorithms

In this section, we propose a class of algorithms that incorporates IVs into the online bandit algorithm. These algorithms update a policy function $\pi(\cdot|v)$ that approaches the optimal one gradually, while retaining a regret of the order of $\mathcal{O}(\log(T))$ or $\mathcal{O}(\log^2(T))$ in the long run.

Throughout the rest of the paper, we use $\|\mathbf{v}\|$ to denote the Euclidean norm if $\mathbf{v}$ is a vector in a finite dimensional Euclidean space. The following definitions are also needed.

DEFINITION 1. Let $\mathbf{M} \in \mathbb{R}^{k\times\ell}$ be a real matrix. If $k \geq \ell$, then a non-negative real number s is said to be a *singular value* of $\mathbf{M}$ if s^2 is an eigenvalue of $\mathbf{M}^\intercal\mathbf{M} \in \mathbb{R}^{\ell\times\ell}$. If $k < \ell$, then a non-negative real number s is said to be a singular value of $\mathbf{M}$ if s^2 is an eigenvalue of $\mathbf{M}\mathbf{M}^\intercal \in \mathbb{R}^{k\times k}$. Hence, $\mathbf{M} \in \mathbb{R}^{k\times\ell}$ has $\min\{k,\ell\}$ singular values in total. Let $\phi_{\max}(\mathbf{M})$ and $\phi_{\min}(\mathbf{M})$ denote the largest and smallest singular values of $\mathbf{M}$, respectively.

DEFINITION 2. The *spectral norm* of a matrix $\mathbf{M} \in \mathbb{R}^{k\times\ell}$ is defined as $\|\mathbf{M}\| \coloneqq \sup_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1} \|\mathbf{M}\mathbf{u}\|$, and its *max norm* is defined as $\|\mathbf{M}\|_{\max} \coloneqq \max_{1\leq i\leq k, 1\leq j\leq \ell} |M_{ij}|$, where M_{ij} denotes the (i,j)-th entry of $\mathbf{M}$.

DEFINITION 3. A random variable X is said to be *sub-Gaussian* with variance proxy ς^2 if $\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\lambda^2\varsigma^2/2}$ for all $\lambda \in \mathbb{R}$.

4.1. General Setup

Similar to (5), because $\sum_{i=1}^M \mathbb{I}(a_t = i) = 1$, we rewrite the reward function (1) as

$$R_t = \mathbf{v}_t^\intercal\boldsymbol{\alpha}_1 + \mathbb{I}(a_t = 2)\mathbf{v}_t^\intercal(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1) + \cdots + \mathbb{I}(a_t = M)\mathbf{v}_t^\intercal(\boldsymbol{\alpha}_M - \boldsymbol{\alpha}_1) + \epsilon_t. \tag{6}$$

With this formulation, we note that there are at most Mp endogenous variables in (6), where p is the dimensionality of $\mathbf{v}_t$. To see this, note that, first, there are at most p endogenous variables within $\mathbf{v}_t$. In addition, as our discussion following (5) illustrates, the optimal policy is

$$a_t^* \coloneqq \arg\max_{i=1,\ldots,M}\{\mathbf{v}_t^\intercal\boldsymbol{\alpha}_i\},$$

which is dependent on $\mathbf{v}_t$. As the action a_t converges to a_t^*, a_t will also be endogenous. This brings another p endogenous variables apart from $\mathbf{v}_t$ itself. To cope with these endogenous variables, we assume that there are at least Mp IVs.

ASSUMPTION 1. *Let $\{(\mathbf{v}_t, \mathbf{z}_t, \epsilon_t) : t = 1,\ldots,T\}$ be an i.i.d. sequence, where $\mathbf{z}_t \in \mathbb{R}^q$ is a vector of valid IVs with $q \geq Mp$.*

(i) ϵ_t is sub-Gaussian with variance proxy ς^2

(ii) $\mathbb{E}[\mathbf{z}_t\epsilon_t] = 0$ and $\mathrm{Var}(\epsilon_t) = \sigma^2 < \infty$.

(iii) There exist positive constants $\bar{v}$ and $\bar{z}$, such that $\|\mathbf{z}_t\|_\infty \le \bar{z}$ and $\|\mathbf{v}_t\|_\infty \le \bar{v}$ almost surely, where $\|\cdot\|_\infty$ denotes the L^∞ norm of a vector.

(iv) Let $\boldsymbol{\delta}_{i,j} \coloneqq \boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j$, $\boldsymbol{\Sigma}^*_{vz} \coloneqq \mathbb{E}[\mathbf{v}^*_t \mathbf{z}_t^\intercal]$, and $\boldsymbol{\Sigma}_{zz} \coloneqq \mathbb{E}[\mathbf{z}_t \mathbf{z}_t^\intercal]$, where

$$\mathbf{v}^*_t \coloneqq \begin{pmatrix} \mathbb{I}(a^*_t = 1)\mathbf{v}_t \\ \vdots \\ \mathbb{I}(a^*_t = M)\mathbf{v}_t \end{pmatrix} = \begin{pmatrix} \mathbb{I}(\min_{j\neq 1} \mathbf{v}_t^\intercal \boldsymbol{\delta}_{1j} > 0)\mathbf{v}_t \\ \vdots \\ \mathbb{I}(\min_{j\neq M} \mathbf{v}_t^\intercal \boldsymbol{\delta}_{Mj} > 0)\mathbf{v}_t \end{pmatrix} \in \mathbb{R}^{Mp}.$$

Then, $\boldsymbol{\Sigma}^*_{vz}$ is a full-rank matrix and $\boldsymbol{\Sigma}_{zz}$ is a positive definite matrix.

(v) There exists a constant $L > 0$, such that $\Pr(|\mathbf{v}_t^\intercal \boldsymbol{\delta}_{i,j}| \le c) \le Lc$ for all $c > 0$ and $1 \le i \neq j \le M$.

Condition (i) is common in the bandit literature and is key for establishing a $\log(T)$-like regret bound. For example, Goldenshluger and Zeevi (2013) assume normality on ϵ_t, and Zhong, Hong, and Liu (2021) assume boundedness, both of which are special cases of Condition (i). Bastani, Bayati, and Khosravi (2021) impose the same sub-Gaussian condition as we do. If Condition (i) is replaced with moment restrictions on ϵ_t, one might still obtain a sublinear regret, but not a $\log(T)$ regret. Conditions (ii) and (iii) are again common assumptions in the IV and contextual bandit literature, respectively.

Condition (iv) implies that $\boldsymbol{\Sigma}^*_{vz}\boldsymbol{\Sigma}^{-1}_{zz}(\boldsymbol{\Sigma}^*_{vz})^\intercal \in \mathbb{R}^{(Mp)\times(Mp)}$ is a positive definite matrix. This is a key condition for parameter identification. It requires that the optimal policy must be dependent on $\mathbf{v}_t$; more specifically, $\mathbb{I}(a^*_t = 1)\mathbf{v}_t, \ldots, \mathbb{I}(a^*_t = M)\mathbf{v}_t$ should not be co-linear when projected to the $\mathbf{z}_t$ space. Condition (iv) holds in general if $\boldsymbol{\delta}_{i,j} \neq \mathbf{0}$ for all $1 \le i \neq j \le M$.

Further, Condition (v) states that the density of $\mathbf{v}_t$ is Lipschitz continuous in the vicinity of the hyperplance $\{\mathbf{v} : \mathbf{v}^\intercal \boldsymbol{\delta}_{i,j} = 0\}$. This condition is often referred to as the "margin condition" in statistical learning literature (Tsybakov 2004). It is also a standard condition in the literature on contextual bandits. Moreover, Conditions (iv) and (v) together imply that $\Pr(\mathbf{v}_t^\intercal \boldsymbol{\delta}_{i,j} < 0) > 0$ and $\Pr(\mathbf{v}_t^\intercal \boldsymbol{\delta}_{i,j} > 0) > 0$, which is referred to as the "diversity condition" in Goldenshluger and Zeevi (2013).

COMMENT 2. The covariate vector $\mathbf{v}_t \in \mathbb{R}^p$ may be decomposed into two parts: $\mathbf{x}_t$ and $\mathbf{d}_t$, where $\mathbf{x}_t \in \mathbb{R}^\ell$ is exogenous, $\mathbf{d}_t \in \mathbb{R}^m$ is potentially endogenous, and $p = \ell + m$. Then, following the discussion below (6), there are $((M-1)\ell + Mm)$ number of endogenous variables in (6), with m of them from the term $\mathbf{v}_t^\intercal \boldsymbol{\alpha}_1$ and the other $(M-1)(l+m)$ of them from the term $\mathbb{I}(a_t = i)\mathbf{v}_t^\intercal(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_1)$, $i = 2, \ldots, M$. Thus, the minimal number of IVs that we need is $((M-1)\ell + Mm)$. However, for the convenience of exposition, we stipulate in Assumption 1 that there are a total of $q \ge Mp = M(\ell+m)$ IVs. □

COMMENT 3. To construct a sufficient number of IVs, following the decomposition of $\mathbf{v}_t$ in Comment 2, practitioners may begin with a set of IVs for $\mathbf{d}_t$, denoted as $\check{\mathbf{z}}_t \in \mathbb{R}^{\check{q}}$ with $\check{q} \geq m$. The exogenous variables $\mathbf{x}_t$ can serve as IVs for themselves. A natural means to construct additional IVs for $\mathbb{I}(a_t = i)\mathbf{v}_t$, $i = 2, \ldots, M$, in (6) is to use such variables as $\mathbb{I}(\mathbf{x}_t \geq \mathring{x})$, $\mathbb{I}(\check{\mathbf{z}}_t \geq \mathring{z})$, and their interaction terms—for example, $\mathbb{I}(\mathbf{x}_t \geq \mathring{x})\mathbf{x}_t$ and $\mathbb{I}(\check{\mathbf{z}}_t \geq \mathring{z})\check{\mathbf{z}}_t$ for some threshold values $\mathring{x}$ and $\mathring{z}$. See the simulation example in Section 6 for an illustration. Other nonlinear transformations of $\mathbf{x}_t$ and $\check{\mathbf{z}}_t$ may also work as IVs, if

$$\mathbb{E}[\epsilon_t | \mathbf{x}_t, \check{\mathbf{z}}_t] = 0. \tag{7}$$

This condition and the use of splines as IVs are widely employed in the conditional moment restriction literature (Chen and Pouzo 2012). This implies Condition (iii) in Assumption 1 and, thus, it is slightly stronger. For optimal selection of IVs under the conditional moment restriction (7), we refer to Donald, Imbens, and Newey (2009) and Belloni et al. (2012). □

4.2. Structure

In this subsection, we propose a class of IV-based algorithms for linear contextual bandits and name it *IV-Bandit*. This algorithm consists of three phases. In the first phase—which we refer to as *Coefficient Stabilization*—we seek a stable estimate of the coefficient vector $\boldsymbol{\alpha}$, thereby implying that the estimate should stay close to the true value with a high probability. For this purpose, our *behavior policy* is to select each arm randomly—irrespective of the observed covariates—with equal probability. At the end of Phase 1, we run *arm-specific-2SLS* to obtain an estimate of the coefficient vector α. Specifically, depending on the arm pulled in each period—whether $a_t = 1$, 2, etc.—we divide the time periods into M groups and then conduct 2SLS separately on the data (i.e., rewards, covariates, and IVs) in each group. Note that the estimate is unbiased because the actions are generated randomly, and the IVs are used to correct for the endogeneity in the covariates. But the absence of bias is achieved at the cost of a regret that is linear in the number of periods $\mathtt{T}_1$. To balance the regret and the purpose of coefficient stabilization, we choose the duration of Phase 1 to be $\mathtt{T}_1 = \mathcal{O}(\log(T))$.

In the second phase—which we refer to as *Covariance Stabilization*—we seek a stable estimate of the covariance matrix $\boldsymbol{\Omega}^* := \left(\boldsymbol{\Sigma}_{vz}^* \boldsymbol{\Sigma}_{zz}^{-1} (\boldsymbol{\Sigma}_{vz}^*)^\intercal\right)^{-1}$. In this phase, the behavioral policy takes the form of a greedy policy—that is, it uses the estimate of $\boldsymbol{\alpha}$ from Phase 1 to evaluate the expected payoff of each arm for a given value of the covariates, and then selects an arm that maximizes the expected payoff. During this phase, which lasts from period $\mathtt{T}_1 + 1$ to $\mathtt{T}_2$, no updates on the coefficients are performed. As explained below, fixing the coefficients helps the algorithm to obtain a

good estimate of $\mathbf{\Omega}^*$ at the beginning of Phase 3, which is crucial for the estimate of $\boldsymbol{\alpha}$ to converge to the true value. Similar to Phase 1, we set the duration of this phase to be $\mathtt{T}_2 - \mathtt{T}_1 = \mathcal{O}(\log(T))$ so that sufficient data is generated without causing excessive regret.

In the third phase—which we refer to as *Policy Improvement*—our algorithm allows for a class of possible behavioral policies, depending on the choice of a parameter θ. In the simplest form, which corresponds to the case of $\theta = 0$, the behavioral policy is again greedy: we select the arm that maximizes the estimated expected reward in each period. However, in contrast to Phase 2, the criterion for evaluating the expected reward is constantly updated. In each period, the expected reward for each arm is calculated from the most updated estimate of $\boldsymbol{\alpha}$. When $\theta > 0$, the behavioral policy in Phase 3 of our algorithm becomes a UCB type. UCB is a common strategy for solving contextual bandit problems (Chu et al. 2011, Dimakopoulou et al. 2019, Zhong, Hong, and Liu 2021). In contrast to the greedy policy, UCB takes into account the uncertainty regarding its expected reward when deciding an arm, thereby balancing between exploitation and exploration. Specifically, UCB selects the arm that maximizes a weighted sum of the estimated expected reward and the estimated SD. In our algorithm, the value of θ represents the relative significance of the SD, thereby reflecting the level of exploration. Again, both the expected reward and SD in each period are estimated via the joint-2SLS using all the available data from the beginning of Phase 2.

Two features of our algorithms, the use of joint-2SLS in Phase 3 and the existence of Phase 1 and 2, are crucial for generating unbiased estimation by dealing with the endogeneity spillover issue. In particular, one might consider applying the arm-specific-2SLS in Phase 3 (as done in Phase 1) to obtain estimates of $\boldsymbol{\alpha}$ and $\mathbf{\Omega}^*$. This approach leads to biased estimates because it copes with only the endogeneity in the covariate $\mathbf{v}_t$, but not the endogeneity in the action a_t, which—as explained in detail in Section 5—is propagated from that in $\mathbf{v}_t$ through the behavior policy that again depends on $\mathbf{v}_t$. To account for this problem, joint-2SLS conducts 2SLS with M regressors $\{\mathbb{I}(a_t = i)\mathbf{v}_t : i = 1, \ldots, M\}$ in addition to the IVs $\mathbf{z}_t$. This formulation clarifies that the minimal number of IVs should be Mp, rather than p, as the number of regressors is M times as many.

In addition, one might consider beginning the algorithm immediately with Phase 3—without Phases 1 and 2—using the joint-2SLS starting from the first time period to estimate both the coefficients $\boldsymbol{\alpha}$ and the covariance matrix $\mathbf{\Omega}^*$ and select the arms accordingly. However, doing so again leads to biased estimates in the long run. This problem also arises because of the interactions between the estimates and the chosen actions. Poor estimates of the coefficients and the covariance matrix in the past, through their effects on the chosen actions, have persistent effects in the future and lead to biased estimates in the long run. Phase 1 ensures that the initial coefficient estimate is

sufficiently accurate. Then, Phase 2 uses the greedy policy based on the initial coefficient estimate to generate "proxicies" of $\mathbf{v}_t^*$'s, which, by definition, are induced by the unknown optimal policy. This allows us to obtain a stable estimate in the first-stage regression of the joint-2SLS, which is critical for producing a sample path of $\widehat{\boldsymbol{\alpha}}_t$ that converges to the true value $\boldsymbol{\alpha}$. In Section 5, we discuss how our algorithm deals with this problem.

COMMENT 4. Because the reward associated with an arm is sampled only if that arm is selected, and the decision to select an arm depends on the estimate of $\boldsymbol{\alpha}$, each arm effectively faces a sample selection problem. Apart from our algorithm, one may apply the bias correction technique (Heckman 1979, 1990) to address the sample selection problem. This method often requires a normality assumption or some other parametric assumption on the distribution of the error terms. A less restrictive method is the semi-parametric maximum likelihood estimation (Cosslett 2004). Both of these methods usually require the error terms to be identically distributed, while the IV-Bandit algorithm may work in the presence of heteroskedasticity. Moreover, they are developed in a static environment. Extending them to an online learning context is beyond the scope of the present paper and we defer the investigation to future research. □

4.3. Details

To facilitate the subsequent presentation, we introduce the following notations. For each $i = 1, \ldots, M$ and t, let $\mathcal{T}_{i,t} \coloneqq \{s \le t | a_s = i\}$ be the set of time periods when arm i is pulled within the first t periods. Let $\mathbf{V}_{i,t} \in \mathbb{R}^{|\mathcal{T}_{i,t}| \times p}$ denote the matrix for data collected in time periods in $\mathcal{T}_{i,t}$, composed of rows $\mathbf{v}_s^\intercal$ for all $s \in \mathcal{T}_{i,t}$. Likewise, we denote by $\mathbf{R}_{i,t} \in \mathbb{R}^{|\mathcal{T}_{i,t}|}$ the vector composed of elements $R_{i,s}$ and by $\mathbf{Z}_{i,t} \in \mathbb{R}^{|\mathcal{T}_{i,t}| \times q}$ the vector composed of rows $\mathbf{z}_{i,s}^\intercal$, for all $s \in \mathcal{T}_{i,t}$. Let

$$\boldsymbol{\alpha} \coloneqq \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_M \end{pmatrix} \in \mathbb{R}^{Mp}, \quad \widehat{\boldsymbol{\alpha}}_t \coloneqq \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{1,t} \\ \vdots \\ \widehat{\boldsymbol{\alpha}}_{M,t} \end{pmatrix} \in \mathbb{R}^{Mp}, \quad \widetilde{\mathbf{v}}_t \coloneqq \begin{pmatrix} \mathbb{I}(a_t = 1)\mathbf{v}_t \\ \vdots \\ \mathbb{I}(a_t = M)\mathbf{v}_t \end{pmatrix} \in \mathbb{R}^{Mp}.$$

For any $t_1 < t_2$, let

$$\widetilde{\mathbf{V}}_{t_1:t_2} \coloneqq \begin{pmatrix} \widetilde{\mathbf{v}}_{t_1+1}^\intercal \\ \vdots \\ \widetilde{\mathbf{v}}_{t_2}^\intercal \end{pmatrix} \in \mathbb{R}^{(t_2-t_1)\times(Mp)}, \quad \mathbf{Z}_{t_1:t_2} \coloneqq \begin{pmatrix} \mathbf{z}_{t_1+1}^\intercal \\ \vdots \\ \mathbf{z}_{t_2}^\intercal \end{pmatrix} \in \mathbb{R}^{(t_2-t_1)\times q}, \quad \mathbf{R}_{t_1:t_2} \coloneqq \begin{pmatrix} R_{t_1+1} \\ \vdots \\ R_{t_2} \end{pmatrix} \in \mathbb{R}^{t_2-t_1}.$$

We define a projection operator $\mathcal{P}$, which takes effect on a matrix $\mathbf{Z}$, as $\mathcal{P}[\mathbf{Z}] = \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal$. The IV-Bandit algorithm is formally presented in Algorithm 1. For the ease of reference, we call the algorithm *IV-Greedy* when $\theta = 0$ and *IV-UCB* when $\theta > 0$.

Algorithm 1: IV-Bandit

Initialization : Set θ, $\mathtt{T}_1$, $\mathtt{T}_2$, and T

Coefficient Stabilization : **for** $t = 1, \ldots, \mathtt{T}_1$ **do**

  Observe $(\mathbf{v}_t, \mathbf{z}_t)$, pull arm $a_t = 1, \ldots, M$ purely randomly, and collect R_t.

end

for $i = 1, \ldots, M$ **do**

  Run arm-specific-2SLS on $(\mathbf{V}_{i,\mathtt{T}_1}, \mathbf{Z}_{i,\mathtt{T}_1}, \mathbf{R}_{i,\mathtt{T}_1})$ to estimate $\boldsymbol{\alpha}$:

$$\widehat{\boldsymbol{\alpha}}_{i,\mathtt{T}_1} \coloneqq \left(\mathbf{V}_{i,\mathtt{T}_1}^\intercal \mathcal{P}[\mathbf{Z}_{i,\mathtt{T}_1}] \mathbf{V}_{i,\mathtt{T}_1}\right)^{-1} \mathbf{V}_{i,\mathtt{T}_1}^\intercal \mathcal{P}[\mathbf{Z}_{i,\mathtt{T}_1}] \mathbf{R}_{i,\mathtt{T}_1}.$$

end

Covariance Stabilization : **for** $t = \mathtt{T}_1 + 1, \ldots, \mathtt{T}_2$ **do**

  Observe $(\mathbf{v}_t, \mathbf{z}_t)$, pull arm $a_t = \underset{i=1,\ldots,M}{\arg\max} \left\{ \mathbf{v}_t^\intercal \widehat{\boldsymbol{\alpha}}_{i,\mathtt{T}_1} \right\}$, and collect R_t.

end

Policy Improvement : **for** $t = \mathtt{T}_2 + 1, \ldots, T$ **do**

  if $t = \mathtt{T}_2 + 1$ **then**

    for $i = 1, \ldots, M$ **do**

      Set $\widehat{\boldsymbol{\alpha}}_{i,\mathtt{T}_2} \coloneqq \widehat{\boldsymbol{\alpha}}_{i,\mathtt{T}_1}$ and $\widehat{\boldsymbol{\Omega}}_{i,\mathtt{T}_2} \coloneqq \left(\mathbf{V}_{i,\mathtt{T}_1}^\intercal \mathcal{P}[\mathbf{Z}_{i,\mathtt{T}_1}] \mathbf{V}_{i,\mathtt{T}_1}\right)^{-1}$.

    end

    Set $\widehat{\sigma}_{\mathtt{T}_2} \coloneqq \left(\mathtt{T}_1^{-1} \sum_{i=1}^M \|\mathbf{R}_{i,\mathtt{T}_1} - \mathbf{V}_{i,\mathtt{T}_1} \widehat{\boldsymbol{\alpha}}_{i,\mathtt{T}_1}\|^2\right)^{\frac{1}{2}}$.

  end

  Observe $(\mathbf{v}_t, \mathbf{z}_t)$, pull arm $a_t = \underset{i=1,\ldots,M}{\arg\max} \left\{ \mathbf{v}_t^\intercal \widehat{\boldsymbol{\alpha}}_{i,t-1} + \widehat{\sigma}_{t-1} \sqrt{2\theta \log(t - \mathtt{T}_1) \cdot \mathbf{v}_t^\intercal \widehat{\boldsymbol{\Omega}}_{i,t-1} \mathbf{v}_t} \right\}$, and collect R_t.

  Run joint-2SLS on $(\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}, \mathbf{Z}_{\mathtt{T}_1:t}, \mathbf{R}_{\mathtt{T}_1:t})$ to update the estimates of $\boldsymbol{\Omega}^*$ and $\boldsymbol{\alpha}$:

$$\widehat{\boldsymbol{\Omega}}_t \coloneqq (\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}^\intercal \mathcal{P}[\mathbf{Z}_{\mathtt{T}_1:t}] \widetilde{\mathbf{V}}_{\mathtt{T}_1:t})^{-1} \quad \text{and} \quad \widehat{\boldsymbol{\alpha}}_t \coloneqq \widehat{\boldsymbol{\Omega}}_t \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}^\intercal \mathcal{P}[\mathbf{Z}_{\mathtt{T}_1:t}] \mathbf{R}_{\mathtt{T}_1:t}.$$

  Set $\widehat{\boldsymbol{\Omega}}_{i,t}$ as the i-th $p \times p$ diagonal block of $\widehat{\boldsymbol{\Omega}}_t \in \mathbb{R}^{(Mp)\times(Mp)}$, respectively.

  if $\theta > 0$ **then**

    Update the estimate of σ: $\widehat{\sigma}_t = (t - \mathtt{T}_1)^{-\frac{1}{2}} \left\| \mathbf{R}_{\mathtt{T}_1:t} - \widetilde{\mathbf{V}}_{\mathtt{T}_1:t} \widehat{\boldsymbol{\alpha}}_t \right\|$.

  end

end

COMMENT 5. Algorithm 1 may be enhanced by using $\breve{\boldsymbol{\alpha}}_t \coloneqq \frac{\mathtt{T}_1}{t} \widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1} + \left(1 - \frac{\mathtt{T}_1}{t}\right) \widehat{\boldsymbol{\alpha}}_t$, a weighted average between $\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}$ and $\widehat{\boldsymbol{\alpha}}_t$, as the estimate of $\boldsymbol{\alpha}$ in Phase 3. This may somewhat reduce the necessary duration of Phase 2, which nevertheless would still be of order $\mathcal{O}(\log(T))$. In other words, this variant would not affect the order of the regret bound. Due to the additional technical complexity this variant would introduce, we focus on the current formulation of Algorithm 1. □

5. Theoretical Results

In this section, we analyze the theoretical properties of the IV-Bandit algorithms. We show that IV-Greedy produces a regret of order $\mathcal{O}(\log(T))$, while IV-UCB produces a regret of order $\mathcal{O}(\log^2(T))$. We also show that these algorithms deliver consistent and asymptotically normal estimates of $\boldsymbol{\alpha}$.

5.1. IV-Greedy

Below, we first overview the regret analysis of the IV-Greedy algorithm, and then discuss a new technique—zig-zag induction—that we develop for analyzing algorithms associated with online learning environments

5.1.1. Overview of the Regret Analysis

THEOREM 1. *Let* $\mathtt{T}_1 = C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2 = (C_{\mathtt{T}_1} + C_{\mathtt{T}_2})\log(T)$ *for some sufficiently large constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$*. Then, under Assumption 1, the regret of the IV-Greedy Algorithm (i.e., Algorithm 1 with* $\theta = 0$*) is* $\mathsf{Regret}(T) = \mathcal{O}(\log(T))$ *for all* $T > \mathtt{T}_2$*.*

Note that in the presence of endogeneity, an algorithm that does not produce consistent estimates of the coefficient vector $\boldsymbol{\alpha}$ will generally suffer a linear regret (Nambiar, Simchi-Levi, and Wang 2019). Therefore, it is necessary to establish consistency of the coefficient estimate. This requires the analysis of $\Pr(\|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}\| < \eta)$ for small η over Phase 3 of the algorithm. In Phase 3, the 2SLS formula yields

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_t &= \boldsymbol{\alpha} + \widehat{\boldsymbol{\Omega}}_t \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}^{\intercal} \mathcal{P}[\mathbf{Z}_{\mathtt{T}_1:t}] \boldsymbol{\epsilon}_{\mathtt{T}_1:t} \\ &= \boldsymbol{\alpha} + (\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1} \widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}^{\intercal})^{-1} \widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1} \widehat{\mathbf{G}}_{\mathtt{T}_1:t}, \end{aligned} \tag{8}$$

where for any $t_1 < t_2$, $\boldsymbol{\epsilon}_{t_1:t_2} \coloneqq (\epsilon_{t_1+1}, \ldots, \epsilon_{t_2})^{\intercal} \in \mathbb{R}^{t_2 - t_1}$, and

$$\widehat{\boldsymbol{\Sigma}}_{vz,t_1:t_2} \coloneqq \frac{1}{t_2 - t_1} \sum_{s=t_1+1}^{t_2} \widetilde{\mathbf{v}}_s \mathbf{z}_s^{\intercal} = \frac{1}{t_2 - t_1} \widetilde{\mathbf{V}}_{t_1:t_2}^{\intercal} \mathbf{Z}_{t_1:t_2} \in \mathbb{R}^{(Mp)\times q},$$

$$\widehat{\boldsymbol{\Sigma}}_{zz,t_1:t_2} \coloneqq \frac{1}{t_2 - t_1} \sum_{s=t_1+1}^{t_2} \mathbf{z}_s \mathbf{z}_s^{\intercal} = \frac{1}{t_2 - t_1} \mathbf{Z}_{t_1:t_2}^{\intercal} \mathbf{Z}_{t_1:t_2} \in \mathbb{R}^{q\times q},$$

$$\widehat{\mathbf{G}}_{t_1:t_2} \coloneqq \frac{1}{t_2 - t_1} \sum_{s=t_1+1}^{t_2} \mathbf{z}_s \epsilon_s = \frac{1}{t_2 - t_1} \mathbf{Z}_{t_1:t_2}^{\intercal} \boldsymbol{\epsilon}_{t_1:t_2} \in \mathbb{R}^{q}.$$

Expression (8) shows that $\widehat{\boldsymbol{\alpha}}_t$ depend on $\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}$, $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$, and $\widehat{\mathbf{G}}_{\mathtt{T}_1:t}$. The calculations of $\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}$ and $\widehat{\mathbf{G}}_{\mathtt{T}_1:t}$ are standard. They are averages of samples of i.i.d. random variables and their statistical properties are standard. For expositional ease, we do not discuss these two terms below and relegate their analysis to the proof in the Supplemental Material. In the following discussion, we refer to $\widehat{\boldsymbol{\alpha}}_t$ as the coefficient estimate and to $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ as the covariance estimate.

However, the calculation of $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ is not standard because it depends on the past estimates $\{\widehat{\boldsymbol{\alpha}}_s : \mathtt{T}_1 \le s \le t-1\}$ through $\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}$—that is,

$$\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t} = \frac{1}{t - \mathtt{T}_1} \sum_{s=\mathtt{T}_1+1}^{t} \begin{pmatrix} \mathbb{I}(a_s = 1)\mathbf{v}_s \\ \vdots \\ \mathbb{I}(a_s = M)\mathbf{v}_s \end{pmatrix} \mathbf{z}_s^{\intercal}, \tag{9}$$

where a_s is selected by the greedy algorithm and depends on $\widehat{\boldsymbol{\alpha}}_s$.

Note that if the optimal arm a_s^* were known and selected in each period, we would simply calculate

$$\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t} \coloneqq \frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\begin{pmatrix}\mathbb{I}(a_s^*=1)\mathbf{v}_s\\ \vdots \\ \mathbb{I}(a_s^*=M)\mathbf{v}_s\end{pmatrix}\mathbf{z}_s^\intercal, \tag{10}$$

and use it in place of $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ in (8). The calculation of $\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}$ is again standard because it is the average of samples from i.i.d. random variables.

However, because the optimal arms are unknown, the analysis of $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ poses a challenge. First, since $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ depends on the choice of arm a_s for all s from $\mathtt{T}_1+1$ to t, any past coefficient estimate $\widehat{\boldsymbol{\alpha}}_s$, through its effect on the selected arm, has an effect on $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$. Second, these coefficient estimates are serially correlated because they are all calculated from the same data source. Lastly, not only does $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ depend on $\{\widehat{\boldsymbol{\alpha}}_s : \mathtt{T}_1 \le s \le t-1\}$, $\widehat{\boldsymbol{\alpha}}_t$ also depends on $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$. This two-way dependence implies that the coefficient estimates and the covariance estimates are entangled. The past coefficient estimate $\widehat{\boldsymbol{\alpha}}_s$ affects the choice of arm a_s, which affects $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ for all $t>s$. These covariance estimates, through expression (8), again affects $\widehat{\boldsymbol{\alpha}}_t$.

The two-way dependence, a common feature of online learning environments, makes it difficult for the coefficient estimates to converge to the true value in the long run. A poor coefficient estimate can lead to a wrong action, which contaminates the generated data (since the data associated with the right arm is no longer observed). The wrong data then worsens the future coefficient estimates and, thus, the future data-generating process. In summary, errors made in the past do not disappear; instead, they make future errors more likely.

To deal with this problem, we now describe an approach that untangles the dependence via a "zig-zag" induction. As an overview of its structure, assume that, in period $t-1$ in Phase 3, the coefficient estimate $\widehat{\boldsymbol{\alpha}}_s$ is sufficiently close to the true value for *all* s from $\mathtt{T}_2+1$ to $t-1$ with a sufficiently high probability. The goal is to show that this remains true for period t, and this takes two steps. First, we show that the covariance estimate $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ is sufficiently accurate in period t. This is the zig-step. Second, we use this result from the zig-step to show that the induction hypothesis holds at time t. The difficult part of the induction is the zig-step. As mentioned earlier, the covariance estimate in the zig-step depends on all the coefficient estimates in the past, which are serially correlated. Our technique finds a means to replace the dependence (of the covariance estimates) on the past coefficient estimates with the dependence on the realizations of past covariates, which are i.i.d. over time.

5.1.2. Details of the Zig-Zag Induction We now provide a detailed description of the zig-zag induction. Our description comprises three parts. First, we define the relevant objects for the induction. Second, we describe the zig-zag decomposition within the induction step. Third, we describe the observation that allows us to deal with the difficulties in the zig-step of the induction.

Part I. Let

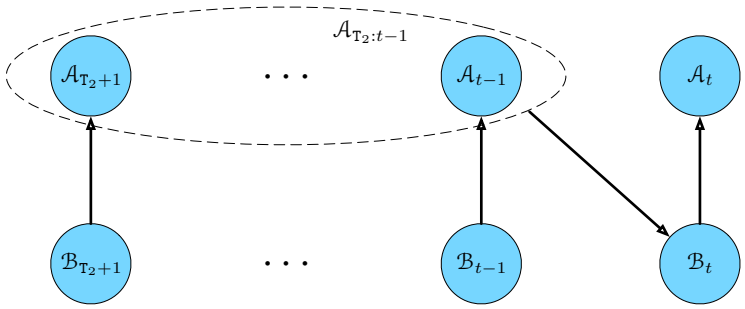
$$\mathcal{A}_t := \left\{ \|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}\| \le C_A \sqrt{\frac{\log(T)}{t - \mathtt{T}_1}} \right\} \quad \text{and} \quad \mathcal{B}_t := \left\{ \|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t} - \boldsymbol{\Sigma}^*_{vz}\| \le \eta \right\},$$

for some constant $C_A > 0$. Further, let $\mathcal{A}_{\mathtt{T}_2:t} := \cap_{s=\mathtt{T}_2+1}^{t} \mathcal{A}_s$, where we note that the intersection begins at the beginning of Phase 2. This is the event in which all the past coefficient estimates from Phase 2 are sufficiently accurate. The goal of the induction is to show that there exists a constant $K > 0$, such that

$$\Pr(\mathcal{A}_{\mathtt{T}_2:t}) \ge 1 - K(t - \mathtt{T}_2)T^{-2}, \quad t = \mathtt{T}_2 + 1, \ldots, T.$$

This is achieved via a two-step induction using $\mathcal{B}_t$ as a bridge—first from $\mathcal{A}_{\mathtt{T}_2:t-1}$ to $\mathcal{B}_t$ and then from $\mathcal{B}_t$ to $\mathcal{A}_t$ (see Figure 2).

Figure 2 The Zig-Zag Approach.



Note that we perform the induction on $\mathcal{A}_{\mathtt{T}_2:t}$ instead of $\mathcal{A}_t$. This is because the zig-zag induction step requires the analysis of the covariance estimate. As we emphasize above, the covariance estimate depends on the past coefficient estimates. Therefore, showing that the covariance estimate is sufficiently accurate requires event $\mathcal{A}_{\mathtt{T}_2:t}$ to occur—that is, all the past coefficient estimates from Phase 2 on to be sufficiently accurate.

Part II. We decompose the induction step into the zig-zag steps. In particular, suppose the induction claim on $\mathcal{A}_{\mathtt{T}_2:t-1}$ holds:

$$\Pr(\mathcal{A}_{\mathtt{T}_2:t-1}) \ge 1 - K(t - 1 - \mathtt{T}_2)T^{-2}.$$

Note that

$$\Pr(\mathcal{A}_{\mathtt{T}_2:t}) = \Pr(\mathcal{A}_t \cap \mathcal{A}_{\mathtt{T}_2:t-1}) = \Pr(\mathcal{A}_{\mathtt{T}_2:t-1}) - \Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1}).$$

Therefore, the induction step requires establishing an upper bound on $\Pr(\mathcal{A}_t^c \cap \mathcal{A}_{\mathtt{T}_2:t-1})$. The zig-zag approach decomposes $\mathcal{A}_t^c \cap \mathcal{A}_{\mathtt{T}_2:t-1}$ into two events depending on whether the covariance estimate in period t is sufficiently accurate (i.e., $\mathcal{B}_t$ or $\mathcal{B}_t^c$):

$$\begin{aligned}\Pr(\mathcal{A}_t^c \cap \mathcal{A}_{\mathtt{T}_2:t-1}) &= \Pr(\mathcal{A}_t^c \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_t) + \Pr(\mathcal{A}_t^c \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_t^c) \\ &\leq \Pr(\mathcal{A}_t^c \cap \mathcal{B}_t) + \Pr(\mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_t^c). \end{aligned} \tag{11}$$

The analysis of $\Pr(\mathcal{A}_t^c \cap \mathcal{B}_t)$, which corresponds to the second step of the induction (from $\mathcal{B}_t$ to $\mathcal{A}_t$ in Figure 2), is standard. When $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ is sufficiently accurate—that is, in the event of $\mathcal{B}_t$—the formula for $\widehat{\boldsymbol{\alpha}}_t$ in (8) can be used to show that that $\mathcal{A}_t^c$ occurs with a sufficiently small probability.

However, the analysis of $\Pr(\mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_t^c)$, which corresponds to the first step of the induction (from $\mathcal{A}_{\mathtt{T}_2:t-1}$ to $\mathcal{B}_t$ in Figure 2) is nonstandard. Recall that if the optimal arms were known and selected in each period, then it would suffice to analyze $\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}$ defined in (10) instead of analyzing $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$. The analysis of the former would be standard, since it would only involve realizations of the covariates and the IVs $\{(\mathbf{v}_s, \mathbf{z}_s) : s = \mathtt{T}_1 + 1, \ldots, t\}$, which are i.i.d. over time. In contrast, because the optimal arms are not known and the selected arm depends on the past coefficient estimates, the calculation of $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ depends on all the past coefficient estimates, which are not i.i.d.

Part III. Finally, we make the following key observation. When the induction assumption holds in period $t-1$ of Phase 3, we can show (with details in the proof) that there exists a constant $K' > 0$, such that

$$\{a_s \neq a_s^*\} \subseteq \cup_{1 \leq j \neq a_s^* \leq M} \left\{ |\mathbf{v}_s^\intercal \boldsymbol{\alpha}_j - \mathbf{v}_s^\intercal \boldsymbol{\alpha}_{a_s^*}| \leq K' \eta \right\}, \tag{12}$$

for all $s = \mathtt{T}_2 + 1, \ldots, t-1$. Therefore, this relation states that a wrong arm is selected only when the difference in the actual expected reward between this arm and some other arm is small. (We can also show, with a similar argument, that (12) holds all $s = \mathtt{T}_1 + 1, \ldots, \mathtt{T}_2$, i.e., Phase 2 of the algorithm, provided that the coefficient estimate at the end of Phase 1 is sufficiently accurate.)

The intuition for this observation is that, when the coefficient estimate is sufficiently accurate, the estimated expected reward of choosing each arm (calculated from the coefficient estimate) is also close to the actual expected reward; and thus the arm selected by the algorithm is optimal if the actual expected reward from this arm is sufficiently different than that from any other arm. In other words, when a wrong arm is selected, it must be that the difference in the actual expected reward between this arm and some other arm is small.

A consequence of this observation is that, instead of estimating the probability that a wrong arm is selected, which depends on the past coefficient estimates, it suffices to estimate the probability that the covariate vector $\mathbf{v}_s$ falls within a bound. Specifically, we can show that

$$\left\| \widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t} - \widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t} \right\| \leq \tilde{K} \frac{1}{t - \mathtt{T}_1} \sum_{s=\mathtt{T}_1+1}^{t} \mathbb{I}(a_s \neq a_s^*)$$

$$\leq \tilde{K}\frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\sum_{1\leq j\neq a_s^*\leq M}\mathbb{I}\left(|\mathbf{v}_s^\intercal\boldsymbol{\alpha}_j-\mathbf{v}_s^\intercal\boldsymbol{\alpha}_{a_s^*}|\leq K'\eta\right), \tag{13}$$

for some constant $\tilde{K}>0$. Because the covariate vectors are i.i.d. over time periods s, so are $\mathbb{I}\left(|\mathbf{v}_s^\intercal\boldsymbol{\alpha}_j-\mathbf{v}_s^\intercal\boldsymbol{\alpha}_i|\leq K'\eta\right)$. In other words, the right-hand-side of the inequality (13) is a sample average of i.i.d. random variables; and therefore, bounds on its tail probability can be established using standard concentration inequalities. This allows us to show that, with a sufficiently high probability, $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$ is close to $\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}$. Moreover, since it is standard to show that $\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}$ is close to the true value with a sufficiently high probability, we can therefore show that the probability of $\mathcal{B}_t^\mathsf{c}$ is small. This completes the analysis of $\Pr(\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_t^\mathsf{c})$ in the decomposition (11), and together with the analysis of $\Pr(\mathcal{A}_t^\mathsf{c}\cap\mathcal{B}_t)$ discussed earlier, we complete the induction step.

The induction shows that, with a sufficiently high probability, the coefficient estimates remains sufficiently accurate over the entire Phase 3 of the algorithm. Standard calculation (Bastani, Bayati, and Khosravi 2021) can then be applied to show that the regret in Phase 3 is $\mathcal{O}(\log(T))$. Since the durations of Phases 1 and 2 are both $\mathcal{O}(\log(T))$, this shows that the total regret is $\mathcal{O}(\log(T))$. Goldenshluger and Zeevi (2013) prove that in a standard linear contextual bandit problem—with no endogeneity involved—the best possible lower bound for any algorithm is $C\log(T)$ for some constant C. Since our setup allows endogeneity in the covariates and includes standard linear contextual bandits as special cases, $C\log(T)$ is also a lower bound on regret for our setup. Hence, Theorem 1 shows that IV-Greedy achieves the asymptotically minimal regret.

5.2. IV-UCB

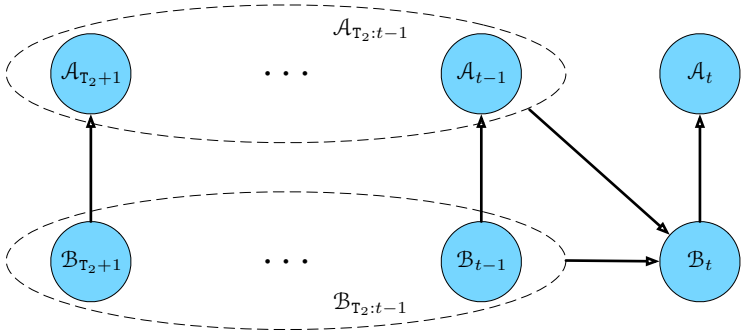
THEOREM 2. *Let* $\mathtt{T}_1=C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2=(C_{\mathtt{T}_1}+C_{\mathtt{T}_2})\log(T)$ *for some sufficiently large constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$*. Then, under Assumption 1, the regret of the IV-UCB algorithm (i.e., Algorithm 1 with* $\theta>0$*) is* $\mathsf{Regret}(T)=\mathcal{O}(\log^2(T))$ *for all* $T>\mathtt{T}_2$*.*

For the IV-UCB algorithm, in each period t of Phase 3, the selected arm a_t depends not only on the coefficient estimate $\widehat{\boldsymbol{\alpha}}_{t-1}$, as is in the case of IV-Greedy, but also on the covariance estimate $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t-1}$ (again, through the calculation of $\widehat{\boldsymbol{\Omega}}_{t-1}$). This arm selection criterion deepens the two-way dependence between the coefficient estimates and the covariance estimates. In particular, whereas the covariance estimates only affect the arm choices indirectly (through their effects on the coefficient estimates) in IV-Greedy, they now affect the arm choices directly. This implies that, to establish the convergence of the coefficient estimates, both the past coefficient estimates and the past covariance estimates must be well controlled.

Despite the additional complexity, the zig-zag induction approach can be adapted to untangle the two-way dependence in Phase 3 of the algorithm. Different from the approach for IV-Greedy,

the induction step from period $t-1$ to period t will now also require the occurrence of event $\mathcal{B}_{\mathtt{T}_2:t-1} := \cap_{\mathtt{T}_2+1}^{t-1}\mathcal{B}_s$ (i.e., all the past covariance estimates) to be sufficiently accurate, as opposed to only the occurrence of $\mathcal{B}_{t-1}$. Figure 3 illustrates the flow of the induction and indicates that in order to show the covariance is well estimated in period t, there is a requirement on both the coefficient and covariance estimates.

Figure 3 The Two-way Zig-Zag Approach.



In contrast to IV-Greedy, the regret of IV-UCB is $\mathcal{O}(\log^2(T))$. The additional factor of $\log(T)$ is incurred because—in addition to the dependence on the coefficient estimate—the direct dependence of the UCB criterion on the covariance estimates introduces an additional source of error in selecting the optimal arm. It takes more time to eliminate the additional error, thereby leading to a higher regret. Note that Zhong, Hong, and Liu (2021) also establish a $\mathcal{O}(\log^2(T))$ regret of a UCB-type algorithm for standard linear contextual bandits. However, the result that IV-UCB has a higher regret than IV-Greedy must be interpreted in an asymptotic sense as the time horizon increases to infinity. The former may still have a lower regret for a given finite time horizon.

5.3. Statistical Inference

The contextual bandit literature mostly focuses on regret analysis, whereas results on statistical inference are relatively scarce. We provide such results for the coefficient estimates.

THEOREM 3. *Let* $\mathtt{T}_1 = C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2 = (C_{\mathtt{T}_1} + C_{\mathtt{T}_2})\log(T)$ *for some sufficiently large constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$*. Then, under Assumption 1, the estimator* $\widehat{\boldsymbol{\alpha}}_T$ *of Algorithm 1 satisfies*

$$\sqrt{T-\mathtt{T}_1}(\widehat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}) \rightsquigarrow \mathcal{N}\Big(\mathbf{0}, \sigma^2\big(\boldsymbol{\Sigma}_{vz}^* \boldsymbol{\Sigma}_{zz}^{-1} (\boldsymbol{\Sigma}_{vz}^*)^\intercal\big)^{-1}\Big), \tag{14}$$

as $T \to \infty$*, where* $\boldsymbol{\Sigma}_{vz}^* = \mathbb{E}[\mathbf{v}_t^* \mathbf{z}_t^\intercal]$*,* $\boldsymbol{\Sigma}_{zz} = \mathbb{E}[\mathbf{z}_t \mathbf{z}_t^\intercal]$*, and* $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ *denotes the multivariate normal distribution with zero mean vector and covariance matrix* $\boldsymbol{\Sigma}$*.*

To construct a confidence interval for $\boldsymbol{\alpha}$, we need a consistent estimator of the covariance matrix $\sigma^2\big(\boldsymbol{\Sigma}_{vz}^*\boldsymbol{\Sigma}_{zz}^{-1}(\boldsymbol{\Sigma}_{vz}^*)^\intercal\big)^{-1}$ in (14). To this end, note that $\widehat{\sigma}_T^2$, $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T}$, and $\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:T}$ are consistent estimators of σ^2, $\boldsymbol{\Sigma}_{vz}^*$, and $\boldsymbol{\Sigma}_{zz}$, respectively. Moreover, by the definition of $\widehat{\boldsymbol{\Omega}}_T$ in Algorithm 1,

$$(T-\mathtt{T}_1)\widehat{\boldsymbol{\Omega}}_T=\big(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:T}^{-1}(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T})^\intercal\big)^{-1}.$$

Thus, we use $(T-\mathtt{T}_1)\widehat{\sigma}_T^2\widehat{\boldsymbol{\Omega}}_T$ to consistently estimate $\sigma^2\big(\boldsymbol{\Sigma}_{vz}^*\boldsymbol{\Sigma}_{zz}^{-1}(\boldsymbol{\Sigma}_{vz}^*)^\intercal\big)^{-1}$. Then, standard calculation leads to the following result, which facilitates inference on $\boldsymbol{\alpha}$.

COROLLARY 1. *Let* $\mathtt{T}_1=C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2=(C_{\mathtt{T}_1}+C_{\mathtt{T}_2})\log(T)$ *for some sufficiently large constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$*. Then, under Assumption 1, the estimator* $\widehat{\boldsymbol{\alpha}}_T$ *of Algorithm 1 satisfies*

$$(T-\mathtt{T}_1)^{\frac{1}{2}}\big(\widehat{\sigma}_T^2\widehat{\boldsymbol{\Omega}}_T\big)^{-\frac{1}{2}}(\widehat{\boldsymbol{\alpha}}_T-\boldsymbol{\alpha})\rightsquigarrow\mathcal{N}(0,\mathbf{I}_{Mp}),$$
$$(T-\mathtt{T}_1)(\widehat{\boldsymbol{\alpha}}_T-\boldsymbol{\alpha})^\intercal\big(\widehat{\sigma}_T^2\widehat{\boldsymbol{\Omega}}_T\big)^{-1}(\widehat{\boldsymbol{\alpha}}_T-\boldsymbol{\alpha})\rightsquigarrow\chi^2_{Mp}.$$

as $T\to\infty$*, where* $\widehat{\sigma}_T$ *and* $\widehat{\boldsymbol{\Omega}}_T$ *are defined in Algorithm 1,* $\mathbf{I}_{Mp}$ *is the identity matrix of size* $(Mp)\times(Mp)$*, and* χ^2_{Mp} *denotes the chi-square distribution with* Mp *degrees of freedom.*

6. Simulation Experiments

In this section, we present a simulation study to demonstrate the performances of IV-Greedy and IV-UCB and compare them with the following alternatives. These alternatives—similar to both IV-Greedy and IV-UCB—all have a randomization phase of duration $\mathtt{T}_1$, in which each arm is selected uniformly at random.

(i) Naïve-IV-UCB. In each period $t=\mathtt{T}_1+1,\ldots,T$, this algorithm performs arm-specific 2SLS to obtain an estimate of $\boldsymbol{\alpha}$ and selects an arm following the same UCB criterion as the IV-UCB algorithm, except that $\widehat{\sigma}_t$ and $\widehat{\boldsymbol{\Omega}}_{i,t}$ are defined and computed from the arm-specific 2SLS.

(ii) OLS-UCB. In each period $t=\mathtt{T}_1+1,\ldots,T$, this algorithm performs arm-specific OLS to obtain an estimate of $\boldsymbol{\alpha}$ and selects an arm following the same UCB criterion as the IV-UCB algorithm, except that $\widehat{\sigma}_t$ and $\widehat{\boldsymbol{\Omega}}_{i,t}$ are defined and computed from the arm-specific OLS.

(iii) Randomize-then-commit (RTC). After the first $\mathtt{T}_1$ periods, this algorithm performs arm-specific 2SLS to obtain an estimate of $\boldsymbol{\alpha}$ and commit to it. In each period $t=\mathtt{T}_1+1,\ldots,T$, the algorithm continues to use this estimate without updates to select an arm in a greedy manner.

Note that IV-UCB, Naïve-IV-UCB, and OLS-UCB all involve a parameter θ in their respective UCB criterion, and we set $\theta=0.5$ for them all. For all the five algorithms, we set $\mathtt{T}_1=50$ and the total run length $T=20000$. In addition, we set $\mathtt{T}_2=100$ for both IV-Greedy and IV-UCB.

We evaluate these algorithms via regret and number of wrong arms pulled, both of which are common metrics for online learning algorithms. We also perform statistical inference on the coefficients in the reward function upon termination of the algorithms.

To this end, we consider a two-armed linear contextual bandit problem with three-dimensional covariates, i.e., $M = 2$. Let the covariate vector $\mathbf{v}_t = (1, x_t, d_t)^\intercal$, where $x_t \in \mathbb{R}$ is an exogenous variable and $d_t \in \mathbb{R}$ is an endogenous variable. The endogeneity is introduced in the following manner. Suppose the random reward at time t is

$$R_t = \sum_{i=1}^{2} \mu_i(\mathbf{v}_t)\, \mathbb{I}(a_t = i) + \epsilon_t,$$

where $\mu_i(\mathbf{v}_t) = \beta_{i,0} + \beta_{i,1} x_t + \gamma_i d_t$, for some parameters $\beta_{i,0}$, $\beta_{i,1}$, and γ_i. Suppose also

$$d_t = \sqrt{x_t} + \rho_{\check{z}} \check{z}_t + \rho_\eta \eta_t \quad \text{and} \quad \epsilon_t = \tilde{\epsilon}_t + 2\eta_t,$$

where $\rho_{\check{z}}$ and ρ_η are some positive constants, and x_t, $\check{z}_t$, η_t, and $\tilde{\epsilon}_t$ are independent random variables with the following distributions:

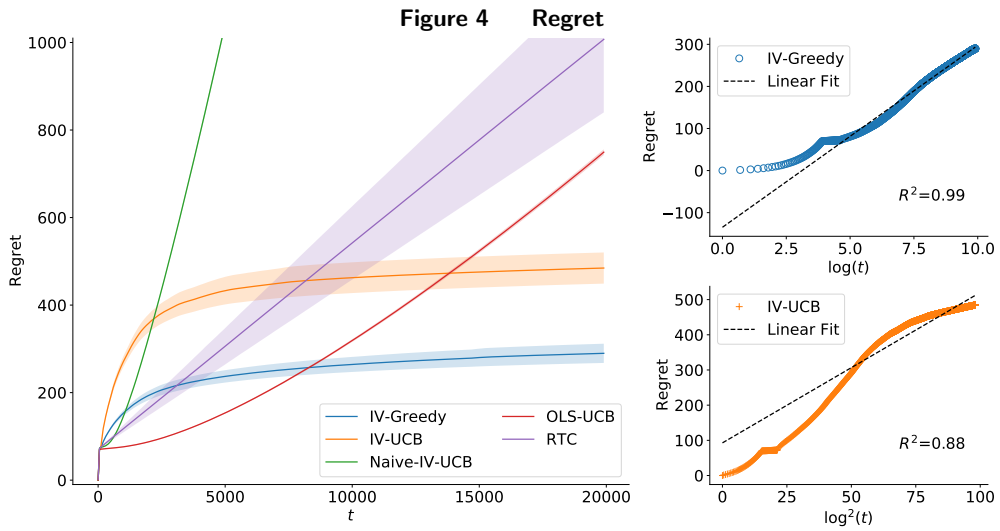
$$x_t \sim \mathsf{TruncNorm}(\mu_x, \sigma_x^2, l_x, u_x), \quad \check{z}_t \sim \mathsf{TruncNorm}(\mu_{\check{z}}, \sigma_{\check{z}}^2, l_{\check{z}}, u_{\check{z}}),$$
$$\eta_t \sim \mathsf{TruncNorm}(\mu_\eta, \sigma_\eta^2, l_\eta, u_\eta), \quad \tilde{\epsilon}_t \sim \mathcal{N}(0, \sigma_{\tilde{\epsilon}}^2).$$

Here, we use $\mathsf{TruncNorm}(\mu, \sigma^2, l, u)$ to denote the truncated normal distribution that is derived from bounding the $\mathcal{N}(\mu, \sigma^2)$ distribution on the interval (l, u). The parameters are specified as follows:

$$(\beta_{1,0}, \beta_{1,1}, \gamma_1) = (1, 4, 4), \quad (\beta_{2,0}, \beta_{2,1}, \gamma_2) = (8, 2, 2),$$
$$\rho_{\check{z}} = 0.5, \quad \rho_\eta = 1.5, \quad \sigma_{\tilde{\epsilon}}^2 = 0.25,$$
$$(\mu_x, \sigma_x^2, l_x, u_x) = (0, 1, 0, 10), \quad (\mu_{\check{z}}, \sigma_{\check{z}}^2, l_{\check{z}}, u_{\check{z}}) = (0, 4, 0, 10),$$
$$(\mu_\eta, \sigma_\eta^2, l_\eta, u_\eta) = (0, 0.25, -5, 5).$$

Note that ϵ_t and d_t are correlated through η_t and, thus, d_t becomes endogenous. Moreover, $\check{z}_t$ is correlated with d_t but is uncorrelated with ϵ_t, so $\check{z}_t$ serves as an IV for d_t. However, for IV-Greedy and IV-UCB to work, the number of IVs must be at least twice as many as the number of endogenous variables in order to account for the endogeneity spillover issue; see Comment 2. As discussed in Comment 3, we create additional IVs using $\mathbb{I}(x_t \geq \mathring{x})$, $\mathbb{I}(\check{z}_t \geq \mathring{z})$, and the interaction terms between them and x_t or $\check{z}_t$, for some threshold values $\mathring{x}$ and $\mathring{z}$. Specifically, we use the following list of IVs:

$$\big\{1,\, x_t,\, \check{z}_t,\, \mathbb{I}(x_t \geq 1),\, \mathbb{I}(x_t \geq 1)\check{z}_t,\, \mathbb{I}(x_t \geq 1.5),\, \mathbb{I}(x_t \geq 1.5)\check{z}_t,\, \mathbb{I}(\check{z}_t \geq 2),\, \mathbb{I}(\check{z}_t \geq 2)\check{z}_t\big\}.$$



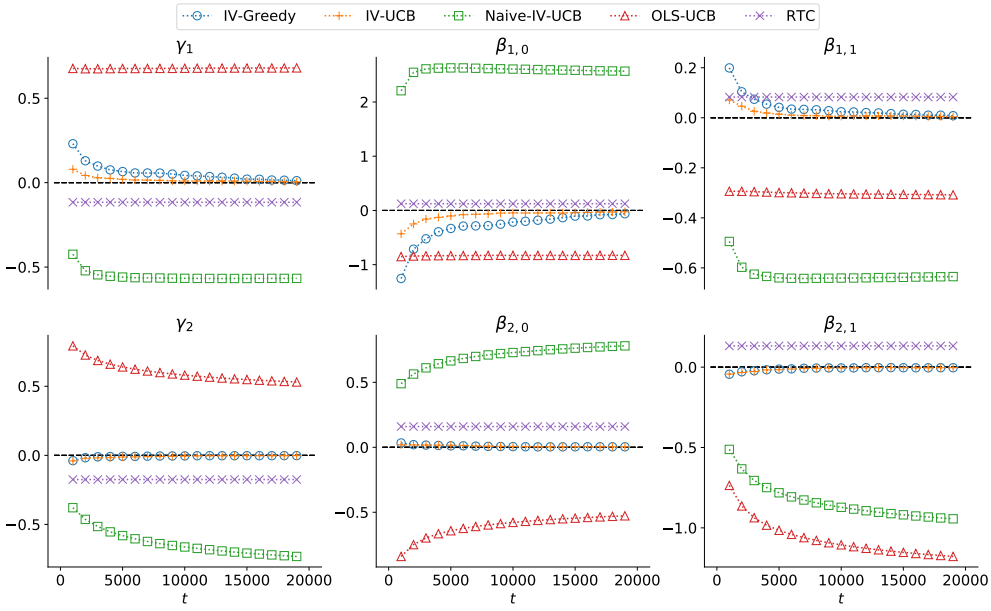
Note. The shaded areas represent 95% confidence bands calculated from 1000 replications.

Further, we run the five algorithms (IV-Greedy, IV-UCB, Naïve-IV-UCB, OLS-UCB, and RTC) 1000 times each and present the results in Figures 4 and 5 as well as Table 1.

Figure 4 presents the regret of each algorithm. First, in contrast to the others, both IV-Greedy and IV-UCB achieve a regret with a diminishing growth rate after a sufficient number of iterations (e.g., $T = 5000$). The confidence bands further show that the performances of these two algorithms are stable in the learning process. Second, to verify our theoretical results in Theorems 1 and 2, we conduct a linear regression analysis of the regrets of IV-Greedy and IV-UCB on $\log(t)$ and $\log^2(t)$, respectively. The value of R^2 is 0.99 for IV-Greedy and 0.88 for IV-UCB, thereby providing strong empirical evidence for our theory. Third, although the theoretical analysis requires that Phases 1 and 2 of IV-Greedy and IV-UCB should be "sufficiently long," the simulation results indicate that they can be set to be moderate values ($\mathtt{T}_1 = 50$ and $\mathtt{T}_2 = 100$) in practice.

Figure 5 presents the biases of the coefficient estimates. For the estimates produced by either IV-Greedy or IV-UCB, the bias vanishes as the number of iterations increases, while the other algorithms are generally biased. In particular, the fact that Naïve-IV-UCB produces inconsistent estimates of the coefficients demonstrates the existence of endogeneity spillover in this online learning example (the endogeneity in the covariates is addressed by the use of IVs in the arm-specific 2SLS, but the endogeneity that spills over to the actions remains). Recall that RTC first performs arm-specific 2SLS to data from randomization and then commits to the coefficient estimates. This results in a dilemma: on the one hand, if the randomization period is long, then the regret is large, because the randomization produces a linear regret; on the other hand, if the randomization period

Figure 5 Biases in Coefficient Estimates.



Note. For each parameter α, the bias $\widehat{\alpha}_t - \alpha$ is calculated for 1000 replications to obtain an average.

is short, then the coefficient estimates are inaccurate,[4] which also leads to a large regret in the long run. Thus, the regret of RTC is, at best, a fractal polynomial of T, which is significantly worse than $\log(T)$.

In Table 1, we report main statistics for the coefficient estimates after 20000 iterations of the five algorithms. For all the coefficients, both IV-Greedy and IV-UCB result in almost zero bias and the corresponding confidence intervals achieve proper coverage of the true values (the coverage is close to 95% for 95% confidence intervals). Note that Naïve-IV-UCB produces both substantially biased estimates and confidence intervals with zero coverage, which, again, demonstrates the endogeneity spillover issue. Note also that the quality of the inference results produced by RTC is fair, but the SD is much larger than both IV-Greedy and IV-UCB.

7. Conclusions

We study the dynamic selection problems that arise in algorithmic decision-making when the data has endogeneity problems. In a contextual bandit model, endogeneity in data influences the arms pulled, causing a non-random sampling of data that results in self-fulfilling bias. We correct for the bias by incorporating IVs to existing online learning algorithms. The new algorithms lead to the true parameter values and meanwhile attain low (logarithmic-like) regret levels. We also prove

[4] In general, the 2SLS estimates have a second-order bias if the sample size is small or the IVs are weak (Andrews, Stock, and Sun 2019).

Table 1 Statistics at Time $T = 20000$.

	γ_1			$\beta_{1,0}$			$\beta_{1,1}$		
	Bias	SD	Coverage	Bias	SD	Coverage	Bias	SD	Coverage
IV-Greedy	0.009	0.112	0.961	-0.049	0.527	0.960	0.007	0.092	0.960
IV-UCB	0.004	0.079	0.958	-0.020	0.352	0.960	0.003	0.080	0.959
Naïve-IV-UCB	-0.566	0.028	0.000	2.565	0.075	0.000	-0.634	0.025	0.000
OLS-UCB	0.680	0.016	0.000	-0.828	0.059	0.000	-0.308	0.021	0.000
RTC	-0.116	1.177	0.913	0.124	1.218	0.921	0.083	1.002	0.937
	γ_2			$\beta_{2,0}$			$\beta_{2,1}$		
	Bias	SD	Coverage	Bias	SD	Coverage	Bias	SD	Coverage
IV-Greedy	-0.002	0.032	0.955	0.004	0.041	0.955	-0.004	0.037	0.959
IV-UCB	-0.002	0.034	0.954	0.004	0.043	0.956	-0.004	0.039	0.959
Naïve-IV-UCB	-0.736	0.064	0.000	0.784	0.057	0.000	-0.949	0.051	0.000
OLS-UCB	0.527	0.030	0.000	-0.524	0.029	0.000	-1.183	0.050	0.000
RTC	-0.174	1.543	0.928	0.160	1.757	0.933	0.132	1.299	0.942

Note. The coverage is calculated for the 95% confidence interval based on 1000 replications.

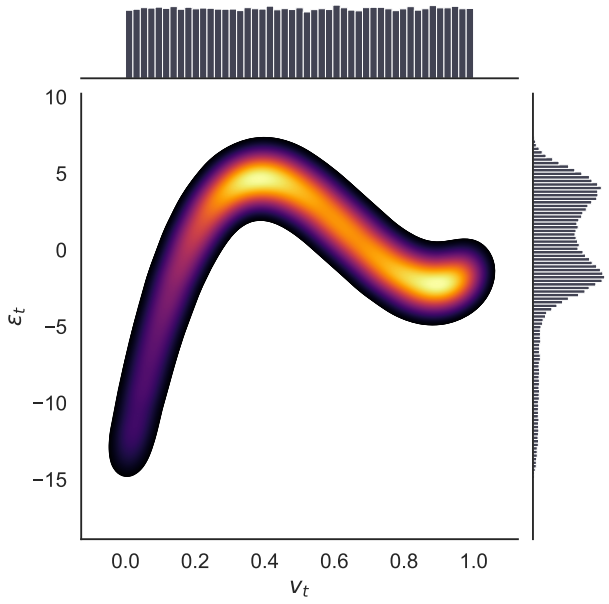
a central limit theorem for statistical inference of the parameters of interest. To establish these properties, we develop a general technique that untangles the interdependence between data and actions.

The use of IVs corrects the bias because they perturb the data-generating process. Therefore, our approach features "ex-ante randomization", which contrasts with the "ex-post randomization"—that is, randomizing the choices given to the users once they arrive. The ex-post randomization, typically under the name of exploration, is the focus of most research. However, in the presence of endogenous data, our analysis highlights the importance of ex-ante randomization. Understanding how best to carry out ex-ante randomization and how these two types of randomization interact offer fruitful venues for future research.

Appendix. Example of Multiplicity of the Self-fulfilling Bias

Consider the example in Section 3: the expected reward for the two arms are $\mu_1(v) \equiv c$ and $\mu_2(v) = \alpha v$. Let v_t be a uniform random variable on $[0,1]$, η_t be an independent standard normal random variable, and $\epsilon_t = \sum_{k=0}^{3} \beta_k v_t^k + \eta_t$ for some constants $\beta_k \in \mathbb{R}$, $k = 0,1,2,3$. See Figure 6 for the joint distribution of (v_t, ϵ_t).

Figure 6 Joint Distribution of (v_t, ϵ_t)



Note. $c = 1$, $\alpha = 15$, and $\{\beta_k : k = 0, 1, 2, 3\}$ are determined by equations (18) and (19) with $(r_1, r_2, r_3) = (2, 5, 10)$.

Direct calculations reveals that for any $b \in [0, 1]$,

$$\begin{aligned}
\mathrm{Var}[v_t | v_t > b] &= \frac{1}{12}(1-b)^2, \\
\mathrm{Cov}[v_t, v_t^2 | v_t > b] &= \frac{1}{12}(1-b)^2(1+b), \\
\mathrm{Cov}[v_t, v_t^3 | v_t > b] &= \frac{1}{40}(1-b)^2(3b^2+4b+3).
\end{aligned}$$

Consequently,

$$\frac{\mathrm{Cov}[v_t, \epsilon_t | v_t > b]}{\mathrm{Var}[v_t | v_t > b]} = \frac{\sum_{k=1}^3 \beta_k \mathrm{Cov}[v_t, v_t^k | v_t > b]}{\mathrm{Var}[v_t | v_t > b]} = \beta_1 + \beta_2(1+b) + \frac{3}{10}\beta_3(3b^2+4b+3). \tag{15}$$

By setting $b = 0$ in equation (15), we derive the limit of the OLS estimate of α in the following manner:

$$\widehat{\alpha}_{\mathrm{OLS}} \coloneqq \alpha + \frac{\mathrm{Cov}[v_t, \epsilon_t]}{\mathrm{Var}[v_t]} = \alpha + \frac{\mathrm{Cov}[v_t, \epsilon_t | v_t > 0]}{\mathrm{Var}[v_t | v_t > 0]} = \alpha + \beta_1 + \beta_2 + \frac{9}{10}\beta_3.$$

Moreover, applying equation (15) to equation (4) with $b = c/\widehat{\alpha}$ yields that

$$\widehat{\alpha}^3 - \left(\alpha + \beta_1 + \beta_2 + \frac{9}{10}\beta_3\right)\widehat{\alpha}^2 - c\left(\beta_2 + \frac{6}{5}\beta_3\right)\widehat{\alpha} - \frac{9}{10}c^2\beta_3 = 0, \tag{16}$$

which is a cubic equation in terms of $\widehat{\alpha}$.

Assume that equation (16) has three real roots: r_k, $k = 1, 2, 3$. Then, $\prod_{k=1}^k (\widehat{\alpha} - r_k) = 0$—that is,

$$\widehat{\alpha}^3 - (r_1 + r_2 + r_3)\widehat{\alpha}^2 + (r_1 r_2 + r_1 r_3 + r_2 r_3)\widehat{\alpha} - r_1 r_2 r_3 = 0. \tag{17}$$

By matching the coefficients of equation (16) and equation (17), we have

$$\beta_3 = \frac{10 r_1 r_2 r_3}{9c^2}, \quad \beta_2 = \frac{-(r_1 r_2 + r_1 r_3 + r_2 r_3)}{c} - \frac{6}{5}\beta_3, \quad \text{and} \quad \beta_1 = r_1 + r_2 + r_3 - \left(\alpha + \beta_2 + \frac{9}{10}\beta_3\right). \tag{18}$$

Further, we may set β_0 such that $\mathbb{E}[\epsilon_t] = \sum_{k=0}^{3} \beta_k v_t^k = 0$—that is,

$$\beta_0 = -\sum_{k=1}^{3} \beta_k \, \mathbb{E}[v_t^k] = -\sum_{k=1}^{3} \frac{1}{k+1} \beta_k. \tag{19}$$

Therefore, for any arbitrary (r_1, r_2, r_3) such that $c/r_k \in (0,1)$, $k = 1, 2, 3$, there exist $(\beta_0, \beta_1, \beta_2, \beta_3)$—given by equations (18) and (19)—such that r_1, r_2, and r_3 are three real fixed points of equation (4). Note that our method of constructing multiple self-fulling biases can be generalized. For an arbitrary number of self-fulling biases with arbitrary values, we can again let the covariates be uniformly distributed and construct the noise as a higher-order polynomial of the covariates. The coefficients of the polynomial can again be solved via linear equations.

References

Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. *Proceedings of the 30th Annual International Conference on Machine Learning*, 1220–1228.

Altonji JG, Elder TE, Taber CR (2005) Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113:151–184.

Andrews I, Stock J, Sun L (2019) Weak instruments in IV regression: Theory and practice. *Annual Review of Economics* 11:727–753.

Athey S, Imbens G (2017) The econometrics of randomized experiments. Banerjee AV, Duflo E, eds., *Handbook of Field Experiments*, volume 1, 73–140 (North-Holland).

Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.

Azevedo EM, Deng A, Montiel Olea JL, Rao J, Weyl EG (2020) A/B testing with fat tails. *Journal of Political Economy* 128(12):4614–4672.

Bastani H, Bayati M, Khosravi K (2021) Mostly exploration-free algorithms for contextual bandits. *Management Science* 67(3):1329–1349.

Belloni A, Chen D, Chernozhukov V, Hansen C (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6):2369–2429.

Bergemann D, Välimäki J (2008) Bandit problems. Durlauf SN, Blume LE, eds., *The New Palgrave Dictionary of Economics* (Basingstoke, UK: Palgrave Macmillan).

Blake T, Nosko C, Tadelis S (2015) Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83(1):155–174.

Blattner L, Nelson S, Spiess J (2021) Unpacking the black box: Regulating algorithmic decisions. Preprint available at *arXiv:2110.03443*.

Caria S, Gordon G, Kasy M, Quinn S, Shami S, Teytelboym A (2020) An adaptive targeted field experiment: Job search assistance for refugees in Jordan. Preprint available at *SSRN:3689456*.

Chen X, Pouzo D (2012) Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* 80(1):277–321.

Chu W, Li L, Reyzin L, Schapire R (2011) Contextual bandits with linear payoff functions. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 208–214.

Cosslett SR (2004) Efficient semiparametric estimation of censored and truncated regressions via a smoothed self-consistency equation. *Econometrica* 72(4):1277–1293.

Currie JM, MacLeod WB (2020) Understanding doctor decision making: The case of depression treatment. *Econometrica* 88(3):847–878.

Dimakopoulou M, Zhou Z, Athey S, Imbens G (2019) Balanced linear contextual bandits. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 3445–3453.

Donald SG, Imbens GW, Newey WK (2009) Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics* 152(1):28–36.

Gibbons R, Henderson R (2012) What do managers do?: Exploring persistent performance differences among seemingly similar enterprises. Gibbons R, Roberts J, eds., *The Handbook of Organizational Economics*, 680–731 (Princeton University Press).

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

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Heckman JJ (1990) Varieties of selection bias. *American Economic Review: Papers and Proceedings* 80(2):313–318.

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press).

Kallus N, Zhou A (2021) Minimax-optimal policy learning under unobserved confounding. *Management Science* 67(5):2870–2890.

Kasy M, Sautmann A (2021) Adaptive treatment assignment in experiments for policy choice. *Econometrica* 89(1):113–132.

Kawaguchi K (2021) When will workers follow an algorithm? A field experiment with a retail business. *Management Science* 67(3):1670–1695.

Kawaguchi K, Uetake K, Watanabe Y (2021) Designing context-based marketing: Product recommendations under time pressure. *Management Science*, forthcoming.

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Kohavi R, Tang D, Xu Y (2020) *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge University Press).

Koning R, Hasan S, Chatterji A (2019) Experimentation and startup performance: Evidence from A/B testing. Preprint available at *SSRN:3440291*.

Li D, Raymond L, Bergman P (2020) Hiring as exploration. Preprint available at *SSRN:3630630*.

Li J, Luo Y, Zhang X (2021) Causal reinforcement learning: An instrumental variable approach. Preprint available at *SSRN:3792824*.

March JG (1991) Exploration and exploitation in organizational learning. *Organization Science* 2(1):71–87.

Nambiar M, Simchi-Levi D, Wang H (2019) Dynamic learning and pricing with model misspecification. *Management Science* 65(11):4980–5000.

Narita Y, Yasui S, Yata K (2019) Efficient counterfactual learning from bandit feedback. *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, 4634–4641 (AAAI Press).

Narita Y, Yasui S, Yata K (2021) Debiased off-policy evaluation for recommendation systems. *Fifteenth ACM Conference on Recommender Systems (RecSys '21)*, 372–379.

Oster E (2019) Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics* 37(2):187–204.

Perchet V, Rigollet P, Chassang S, Snowberg E (2016) Batched bandit problems. *Annals of Statistics* 44(2):660–681.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Shi C, Wan R, Chernozhukov V, Song R (2021) Deeply-debiased off-policy interval estimation. *Proceedings of the 38th International Conference on Machine Learning*, 9580–9591.

Slivkins A (2019) Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* 12(1-2):1–286.

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

Zhan R, Hadad V, Hirshberg DA, Athey S (2021) Off-policy evaluation via adaptive weighting with data from contextual bandits. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2125–2135.

Zhong Y, Hong LJ, Liu G (2021) Earning and learning with varying cost. *Production and Operations Management*, forthcoming.

Supplemental Material

EC.1. Concentration Inequalities

LEMMA EC.1. *For any matrix* $\mathbf{M} \in \mathbb{R}^{k\times\ell}$,

$$\phi_{\max}(\mathbf{M}) = \sup_{\mathbf{u}\in\mathbb{R}^{\ell}:\|\mathbf{u}\|=1} \|\mathbf{M}\mathbf{u}\| \quad \textit{and} \quad \phi_{\min}(\mathbf{M}) = \inf_{\mathbf{u}\in\mathbb{R}^{\ell}:\|\mathbf{u}\|=1} \|\mathbf{M}\mathbf{u}\|,$$

where $\|\mathbf{u}\|$ *denotes the Euclidean norm of* $\mathbf{u}$.

Proof. See page 78 of Golub and Van Loan (2013). □

LEMMA EC.2. *For any matrices* $\mathbf{M} \in \mathbb{R}^{k\times\ell}$ *and* $\mathbf{N} \in \mathbb{R}^{\ell\times m}$, *we have*

$$\phi_{\max}(\mathbf{M}\mathbf{N}) \le \phi_{\max}(\mathbf{M})\phi_{\max}(\mathbf{N}) \quad \textit{and} \quad \phi_{\min}(\mathbf{M}\mathbf{N}) \ge \phi_{\min}(\mathbf{M})\phi_{\min}(\mathbf{N}). \tag{EC.1.1}$$

If $\mathbf{M}$ *is an invertible square matrix, then*

$$\phi_{\max}(\mathbf{M}^{-1}) = (\phi_{\min}(\mathbf{M}))^{-1} \quad \textit{and} \quad \phi_{\min}(\mathbf{M}^{-1}) = (\phi_{\max}(\mathbf{M}))^{-1}. \tag{EC.1.2}$$

Proof. The inequalities in (EC.1.1) follow from Theorem 3 of Wang and Xi (1997), and the identities in (EC.1.2) follow from the definition of singular values. □

LEMMA EC.3. *For any matrix* $\mathbf{M} \in \mathbb{R}^{k\times\ell}$, $\|\mathbf{M}\| \le \sqrt{k\ell}\|\mathbf{M}\|_{\max}$.

Proof. See page 72 of Golub and Van Loan (2013). □

LEMMA EC.4. *Let* $\{\mathbf{M}_i\}_{i=1}^N$ *be a sequence of zero-mean independent* k*-by-*ℓ *random matrices. For each* $i = 1,\dots,N$, *assume that* $\|\mathbf{M}_i\| \le b$ *almost surely for some constant* $b > 0$. *Then, for all* $\tau \ge 0$,

$$\Pr\left(\left\|\frac{1}{N}\sum_{i=1}^N \mathbf{M}_i\right\| \ge \tau\right) \le 2(k+\ell)\exp\left(\frac{-N\tau^2}{2b^2}\right),$$

Moreover, if $k = \ell$, *then the leading constant* $2(k+\ell)$ *on the right-hand-side of the above inequalities can be reduced to* $2k$.

Proof. This is Hoeffding's inequality for bounded random matrices. See pages 174–175 of Wainwright (2019). □

COROLLARY EC.1. *Let* $\{\mathbf{M}_i\}_{i=1}^N$ *be a sequence of independent* k*-by-*ℓ *random matrices. For each* $i = 1,\dots,N$, *assume that* $\|\mathbf{M}_i\| \le b$ *almost surely for some constant* $b > 0$. *Then, for all* $\tau \ge 0$,

$$\Pr\left(\phi_{\max}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{M}_i\right) \ge \phi_{\max}(\mathbb{E}[\mathbf{M}_1]) + \tau\right) \le 2(k+\ell)\exp\left(\frac{-N\tau^2}{8b^2}\right), \tag{EC.1.3}$$

$$\Pr\left(\phi_{\min}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i\right)\le\phi_{\min}(\mathbb{E}[\mathbf{M}_1])-\tau\right)\le 2(k+\ell)\exp\left(\frac{-N\tau^2}{8b^2}\right). \tag{EC.1.4}$$

Moreover, if $k=\ell$, then the leading constant $2(k+\ell)$ on the right-hand-side of the above inequalities can be reduced to $2k$.

Proof. Let $\widetilde{\mathbf{M}}_i=\mathbf{M}_i-\mathbb{E}[\mathbf{M}_i]$. Then, $\|\widetilde{\mathbf{M}}_i\|\le\|\mathbf{M}_i\|+\|\mathbb{E}[\mathbf{M}_i]\|\le 2b$. By Lemmas EC.1 and EC.4,

$$\phi_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{M}}_i\right)=\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{M}}_i\right\|\le\tau \tag{EC.1.5}$$

holds with probability at least $1-2(k+\ell)\exp\left(-\frac{N\tau^2}{8b^2}\right)$.

Given the event (EC.1.5) holds, it follows from Lemma EC.1 that

$$\begin{aligned}\phi_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i\right)&=\sup_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i\mathbf{u}\right\|\\&\le\sup_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1}\|\mathbb{E}[\mathbf{M}_1]\mathbf{u}\|+\sup_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1}\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{M}}_i\mathbf{u}\right\|\\&\le\phi_{\max}(\mathbb{E}[\mathbf{M}_1])+\tau\end{aligned}$$

holds with probability at least $1-2(k+\ell)\exp\left(-\frac{N\tau^2}{8b^2}\right)$, proving (EC.1.3). Likewise, given the event (EC.1.5) holds,

$$\begin{aligned}\phi_{\min}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i\right)&=\inf_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i\mathbf{u}\right\|\\&\ge\inf_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1}\|\mathbb{E}[\mathbf{M}_1]\mathbf{u}\|-\sup_{\mathbf{u}\in\mathbb{R}^\ell:\|\mathbf{u}\|=1}\left\|\frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{M}}_i\mathbf{u}\right\|\\&\ge\phi_{\min}(\mathbb{E}[\mathbf{M}_1])-\tau,\end{aligned}$$

which proves (EC.1.4). □

LEMMA EC.5. *Let $\{X_i\}_{i=1}^N$ be a sequence of i.i.d. zero-mean sub-Gaussian random variables with variance proxy ς^2. Then, for all $\tau\ge 0$,*

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N}X_i\right|\ge\tau\right)\le 2\exp\left(-\frac{N\tau^2}{2\varsigma^2}\right).$$

Proof. This is Hoeffding's inequality for sub-Gaussian random variables. See page 24 of Wainwright (2019). □

COROLLARY EC.2. *Let $\mathbf{g}\in\mathbb{R}^k$ be a vector of zero-mean sub-Gaussian random variables with variance proxy ς^2. Let $\{\mathbf{g}_i\}_{i=1}^N$ be a sequence of i.i.d. random vectors having the same distribution as $\mathbf{g}$. Then, for all $\tau\ge 0$,*

$$\Pr\left(\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i\right\|\ge\tau\right)\le 2k\exp\left(-\frac{N\tau^2}{2k^2\varsigma^2}\right).$$

Proof. Let g_{ij} denote the j-th entry of $\mathbf{g}_i$, for all $i=1,\dots,N$ and $j=1,\dots,k$. Then, for each j, $\{g_{ij}\}_{i=1}^N$ is a sequence of i.i.d. zero-mean sub-Gaussian random variables with variance proxy ς^2. Note that if $|\frac{1}{N}\sum_{i=1}^N g_{ij}| \le \frac{\tau}{k}$ for all $j=1,\dots,k$, then $\|\frac{1}{N}\sum_{i=1}^N \mathbf{g}_i\| \le \tau$. Hence,

$$\begin{aligned}\Pr\left(\left\|\frac{1}{N}\sum_{i=1}^N \mathbf{g}_i\right\| \ge \tau\right) &\le \Pr\left(\left|\frac{1}{N}\sum_{i=1}^N g_{ij}\right| \ge \frac{\tau}{k} \text{ for some } j=1,\dots,k\right)\\ &\le \sum_{j=1}^k \Pr\left(\left|\frac{1}{N}\sum_{i=1}^N g_{ij}\right| \ge \frac{\tau}{k}\right) \le 2k\exp\left(-\frac{N\tau^2}{2k^2\varsigma^2}\right),\end{aligned}$$

where the last inequality follows from Lemma EC.5. □

Definition EC.1 (Sub-Exponential Random Variables). A random variable X is said to be *sub-exponential* if its *sub-exponential norm*, defined as $\|X\|_{\psi_1} \coloneqq \inf\{w>0 : \mathbb{E}[e^{|X|/w}] \le 2\}$, is finite.

Lemma EC.6. *Let $\{X_i\}_{i=1}^N$ be a sequence of i.i.d. zero-mean sub-exponential random variables. Then, for all $\tau \ge 0$,*

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \ge \tau\right) \le 2\exp\left(-N\mathsf{C}_{\mathsf{Bern}}\cdot\min\left\{\frac{\tau^2}{K^2}, \frac{\tau}{K}\right\}\right),$$

where $\mathsf{C}_{\mathsf{Bern}}$ is an absolute constant and $K = \|X_1\|_{\psi_1}$.

Proof. This is Bernstein's inequality for sub-exponential random variables. See Theorem 2.8.2 of Vershynin (2018). □

EC.2. Coefficient Stabilization Phase: $t = 1,\dots,\mathtt{T}_1$

Throughout the rest of the Supplmental Material, we let $\mathtt{T}_1 = C_{\mathtt{T}_1}\log(T)$ and $\mathtt{T}_2 = (C_{\mathtt{T}_1} + C_{\mathtt{T}_2})\log(T)$ for some constant $C_{\mathtt{T}_2}$ for some constants $C_{\mathtt{T}_1}$ and $C_{\mathtt{T}_2}$; we also suppose that Assumption 1 holds and the data are generated by Algorithm 1.

We first define a set of notations before the analysis. For any t, let

$$\widetilde{\mathbf{v}}_t \coloneqq \begin{pmatrix}\mathbb{I}(a_t=1)\mathbf{v}_t\\ \vdots \\ \mathbb{I}(a_t=M)\mathbf{v}_t\end{pmatrix} \in \mathbb{R}^{Mp}, \quad \widetilde{\mathbf{z}}_t \coloneqq \begin{pmatrix}\mathbb{I}(a_t=1)\mathbf{z}_t\\ \vdots \\ \mathbb{I}(a_t=M)\mathbf{z}_t\end{pmatrix} \in \mathbb{R}^{Mq},$$

$$\widetilde{\mathbf{V}}_t \coloneqq \begin{pmatrix}\widetilde{\mathbf{v}}_1^\intercal\\ \vdots \\ \widetilde{\mathbf{v}}_t^\intercal\end{pmatrix} \in \mathbb{R}^{t\times(Mp)}, \quad \widetilde{\mathbf{Z}}_t \coloneqq \begin{pmatrix}\widetilde{\mathbf{z}}_1^\intercal\\ \vdots \\ \widetilde{\mathbf{z}}_t^\intercal\end{pmatrix} \in \mathbb{R}^{t\times(Mq)}, \quad \boldsymbol{\epsilon}_t \coloneqq \begin{pmatrix}\epsilon_1\\ \vdots \\ \epsilon_t\end{pmatrix} \in \mathbb{R}^t,$$

$$\widetilde{\boldsymbol{\Sigma}}_{vz,t} \coloneqq \frac{1}{t}\sum_{s=1}^t \widetilde{\mathbf{v}}_s\widetilde{\mathbf{z}}_s^\intercal = \frac{1}{t}\widetilde{\mathbf{V}}_t^\intercal\widetilde{\mathbf{Z}}_t \in \mathbb{R}^{(Mp)\times(Mq)},$$

$$\widetilde{\mathbf{\Sigma}}_{zz,t} \coloneqq \frac{1}{t}\sum_{s=1}^{t}\widetilde{\mathbf{z}}_s\widetilde{\mathbf{z}}_s^\intercal = \frac{1}{t}\widetilde{\mathbf{Z}}_t^\intercal\widetilde{\mathbf{Z}}_t \in \mathbb{R}^{(Mq)\times(Mq)},$$

$$\widetilde{\mathbf{G}}_t \coloneqq \frac{1}{t}\sum_{s=1}^{t}\widetilde{\mathbf{z}}_s\epsilon_s = \frac{1}{t}\widetilde{\mathbf{Z}}_t^\intercal\boldsymbol{\epsilon}_t \in \mathbb{R}^{2q},$$

$$\mathbf{\Sigma}_{vz} \coloneqq \mathbb{E}[\mathbf{v}_t\mathbf{z}_t^\intercal] \in \mathbb{R}^{p\times q}, \quad \mathbf{\Sigma}_{zz} \coloneqq \mathbb{E}[\mathbf{z}_t\mathbf{z}_t^\intercal] \in \mathbb{R}^{q\times q},$$

$$\breve{\mathbf{\Sigma}}_{vz} \coloneqq \mathbb{E}[\widetilde{\mathbf{v}}_t\widetilde{\mathbf{z}}_t^\intercal] \in \mathbb{R}^{(Mp)\times(Mq)}, \quad \breve{\mathbf{\Sigma}}_{zz} \coloneqq \mathbb{E}[\widetilde{\mathbf{z}}_t\widetilde{\mathbf{z}}_t^\intercal] \in \mathbb{R}^{(Mq)\times(Mq)}.$$

LEMMA EC.7. *Both $\breve{\mathbf{\Sigma}}_{zz}$ and $\breve{\mathbf{\Sigma}}_{vz}$ have full rank.*

Proof. By condition (iii) of Assumption 1, $\mathbf{\Sigma}_{vz}^* = \mathbb{E}[\mathbf{v}_t^*\mathbf{z}_t^\intercal]$ is full rank and $\mathbf{\Sigma}_{zz} = \mathbb{E}[\mathbf{z}_t\mathbf{z}_t^\intercal]$ is positive definite. By definition, for $t = 1, 2, \ldots, \mathtt{T}_1$, a_t is randomly selected such that $\Pr(a_t = i) = \frac{1}{M}$ for all $i = 1, \ldots, M$. Thus,

$$\breve{\mathbf{\Sigma}}_{vz} = \frac{1}{M}\begin{pmatrix} \mathbf{\Sigma}_{zz} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Sigma}_{zz} \end{pmatrix},$$

and thus $\breve{\mathbf{\Sigma}}_{vz}$ is full rank and positive definite.

Because $\sum_{i=1}^{M}\mathbb{I}(a_t^* = i) = 1$ and $\mathbf{\Sigma}_{vz}^*$ is a full-rank matrix of size $(Mp)\times q$, $\mathbf{\Sigma}_{vz} = \begin{pmatrix}\mathbf{I}_p & \cdots & \mathbf{I}_p\end{pmatrix}\mathbf{\Sigma}_{vz}^*$ is a full-rank matrix of size $p\times q$, where $\mathbf{I}_p$ is the identity matrix of size $p\times p$. As a result,

$$\breve{\mathbf{\Sigma}}_{vz} = \mathbb{E}[\widetilde{\mathbf{v}}_t\widetilde{\mathbf{z}}_t^\intercal] = \begin{pmatrix} \mathbb{E}[\mathbb{I}(a_t = 1)\mathbf{v}_t\mathbf{z}_t^\intercal] & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbb{E}[\mathbb{I}(a_t = 2)\mathbf{v}_t\mathbf{z}_t^\intercal] \end{pmatrix} = \frac{1}{M}\begin{pmatrix} \mathbf{\Sigma}_{vz} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Sigma}_{vz} \end{pmatrix}.$$

Thus, $\breve{\mathbf{\Sigma}}_{vz}$ is a full-rank matrix of size $(Mp)\times(Mq)$. □

LEMMA EC.8. *Define*

$$\breve{\kappa} \coloneqq \frac{2\phi_{\max}(\breve{\mathbf{\Sigma}}_{vz})\phi_{\max}(\breve{\mathbf{\Sigma}}_{zz})}{(\phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}))^2\phi_{\min}(\breve{\mathbf{\Sigma}}_{zz})} \tag{EC.2.1}$$

and

$$\mathcal{A}_{\mathtt{T}_1} \coloneqq \left\{\left\|\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1} - \boldsymbol{\alpha}\right\| \le 2Mq\bar{z}\varsigma\breve{\kappa}C_{\mathtt{T}_1}^{-\frac{1}{2}}\right\}. \tag{EC.2.2}$$

Suppose

$$C_{\mathtt{T}_1} \ge \left(\frac{40M\max\{\sqrt{pq}\bar{v}\bar{z},\, q\bar{z}^2\}}{\min\{\phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}), \phi_{\min}(\breve{\mathbf{\Sigma}}_{zz})\}}\right)^2. \tag{EC.2.3}$$

Then, $\Pr(\mathcal{A}_{\mathtt{T}_1}) \ge 1 - 2M(2p+5q)T^{-2}$ *for all* $T \ge \mathtt{T}_1$.

Proof. It follows from Lemma EC.7 that $0 < \phi_{\min}(\breve{\mathbf{\Sigma}}_{zz}) \le \phi_{\max}(\breve{\mathbf{\Sigma}}_{zz}) < \infty$ and $0 < \phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}) \le \phi_{\max}(\breve{\mathbf{\Sigma}}_{vz}) < \infty$. Hence, $\breve{\kappa}$ is a positive finite constant.

Because in the coefficient stabilization phase a_t is randomly selected for all $t=1,\ldots,\mathtt{T}_1$, we may rewrite the arm-specific 2SLS for the M arms together equivalently as the joint-2SLS using the notations $\widetilde{\mathbf{V}}_{\mathtt{T}_1}$ and $\widetilde{\mathbf{Z}}_{\mathtt{T}_1}$. Hence,

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}-\boldsymbol{\alpha}&=\left(\widetilde{\mathbf{V}}_{\mathtt{T}_1}^\intercal\mathcal{P}[\widetilde{\mathbf{Z}}_{\mathtt{T}_1}]\widetilde{\mathbf{V}}_{\mathtt{T}_1}\right)^{-1}\widetilde{\mathbf{V}}_{\mathtt{T}_1}^\intercal\mathcal{P}[\widetilde{\mathbf{Z}}_{\mathtt{T}_1}]\boldsymbol{\epsilon}_{\mathtt{T}_1}\\&=\left(\widetilde{\mathbf{V}}_{\mathtt{T}_1}^\intercal\widetilde{\mathbf{Z}}_{\mathtt{T}_1}(\widetilde{\mathbf{Z}}_{\mathtt{T}_1}^\intercal\widetilde{\mathbf{Z}}_{\mathtt{T}_1})^{-1}\widetilde{\mathbf{Z}}_{\mathtt{T}_1}^\intercal\widetilde{\mathbf{V}}_{\mathtt{T}_1}\right)^{-1}\widetilde{\mathbf{V}}_{\mathtt{T}_1}^\intercal\widetilde{\mathbf{Z}}_{\mathtt{T}_1}(\widetilde{\mathbf{Z}}_{\mathtt{T}_1}^\intercal\widetilde{\mathbf{Z}}_{\mathtt{T}_1})^{-1}\widetilde{\mathbf{Z}}_{\mathtt{T}_1}^\intercal\boldsymbol{\epsilon}_{\mathtt{T}_1}\\&=\left(\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1}^{-1}\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}^\intercal\right)^{-1}\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1}^{-1}\widetilde{\mathbf{G}}_{\mathtt{T}_1},\end{aligned}$$

where the projection operator $\mathcal{P}$ maps a matrix $\mathbf{Z}$ to $\mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal$. Therefore,

$$\|\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}-\boldsymbol{\alpha}\|\le\left\|\left(\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1}^{-1}\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}^\intercal\right)^{-1}\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1}^{-1}\right\|\cdot\left\|\widetilde{\mathbf{G}}_{\mathtt{T}_1}\right\| \tag{EC.2.4}$$

In what follows, we analyze the two norms on the right-hand-side of (EC.2.4) separately.

Because the samples are i.i.d.,

$$\mathbb{E}[\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1}]=\mathbb{E}[\widetilde{\mathbf{v}}_s\widetilde{\mathbf{z}}_s^\intercal]=\breve{\boldsymbol{\Sigma}}_{vz}\quad\text{and}\quad\mathbb{E}[\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1}]=\mathbb{E}[\widetilde{\mathbf{z}}_s\widetilde{\mathbf{z}}_s^\intercal]=\breve{\boldsymbol{\Sigma}}_{zz}.$$

Moreover, because $\widetilde{\mathbf{v}}_s\widetilde{\mathbf{z}}_s^\intercal$ is of size (Mp)-by-(Mq), we know by Lemma EC.3 that

$$\|\widetilde{\mathbf{v}}_s\widetilde{\mathbf{z}}_s^\intercal\|\le\sqrt{(Mp)\times(Mq)}\|\widetilde{\mathbf{v}}_s\widetilde{\mathbf{z}}_s^\intercal\|_{\max}\le M\sqrt{pq}\bar{v}\bar{z}.$$

Likewise, $\|\widetilde{\mathbf{z}}_s\widetilde{\mathbf{z}}_s^\intercal\|\le Mq\bar{z}^2$. Hence, if we define the following events:

$$\begin{aligned}\mathcal{S}_{vz,\min}&\coloneqq\left\{\phi_{\min}(\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1})\le\phi_{\min}(\breve{\boldsymbol{\Sigma}}_{vz})-\tau\right\},\\\mathcal{S}_{vz,\max}&\coloneqq\left\{\phi_{\max}(\widetilde{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1})\ge\phi_{\max}(\breve{\boldsymbol{\Sigma}}_{vz})+\tau\right\},\\\mathcal{S}_{zz,\min}&\coloneqq\left\{\phi_{\min}(\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1})\le\phi_{\min}(\breve{\boldsymbol{\Sigma}}_{zz})-\tau\right\},\\\mathcal{S}_{zz,\max}&\coloneqq\left\{\phi_{\max}(\widetilde{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1})\ge\phi_{\max}(\breve{\boldsymbol{\Sigma}}_{zz})+\tau\right\},\end{aligned}$$

then we can apply Corollary EC.1 to conclude that for all $\tau\ge0$,

$$\begin{aligned}\Pr(\mathcal{S}_{vz,\min})&\le2M(p+q)\exp\left(-\frac{\tau^2\mathtt{T}_1}{8M^2pq\bar{v}^2\bar{z}^2}\right),\\\Pr(\mathcal{S}_{vz,\max})&\le2M(p+q)\exp\left(-\frac{\tau^2\mathtt{T}_1}{8M^2pq\bar{v}^2\bar{z}^2}\right),\\\Pr(\mathcal{S}_{zz,\min})&\le2Mq\exp\left(-\frac{\tau^2\mathtt{T}_1}{8M^2q^2\bar{z}^4}\right),\\\Pr(\mathcal{S}_{zz,\max})&\le2Mq\exp\left(-\frac{\tau^2\mathtt{T}_1}{8M^2q^2\bar{z}^4}\right).\end{aligned}$$

Setting $\tau = \zeta_1 \coloneqq 4M \max\{\sqrt{pq}\bar{v}\bar{z},\, q\bar{z}^2\} C_{\mathtt{T}_1}^{-\frac{1}{2}}$ leads to

$$\exp\left(-\frac{\zeta_1^2 \mathtt{T}_1}{8M^2 pq\bar{v}^2\bar{z}^2}\right) \le T^{-2} \quad \text{and} \quad \exp\left(-\frac{\zeta_1^2 \mathtt{T}_1}{8M^2 q^2\bar{z}^4}\right) \le T^{-2}.$$

Moreover, by the condition (EC.2.3), we have

$$\zeta_1 = 4M \max\{\sqrt{pq}\bar{v}\bar{z}, q\bar{z}^2\} C_{\mathtt{T}_1}^{-\frac{1}{2}} \le \frac{1}{10} \min\left\{\phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}), \phi_{\min}(\breve{\mathbf{\Sigma}}_{zz})\right\}. \tag{EC.2.5}$$

Therefore, upon defining the joint event $\mathcal{S}_{\phi,\mathtt{T}_1} \coloneqq \mathcal{S}^{\mathsf{c}}_{vz,\min} \cap \mathcal{S}^{\mathsf{c}}_{vz,\max} \cap \mathcal{S}^{\mathsf{c}}_{zz,\min} \cap \mathcal{S}^{\mathsf{c}}_{zz,\max}$, it follows from the Bonferroni inequality that

$$\begin{aligned}\Pr(\mathcal{S}_{\phi,\mathtt{T}_1}) &\ge 1 - \left[\Pr(\mathcal{S}_{vz,\min}) + \Pr(\mathcal{S}_{vz,\max}) + \Pr(\mathcal{S}_{zz,\min}) + \Pr(\mathcal{S}_{zz,\max})\right] \\ &\ge 1 - 4M(p+2q)T^{-2}. \end{aligned} \tag{EC.2.6}$$

Note that conditional on the event $\mathcal{S}_{\phi,\mathtt{T}_1}$ with $\tau = \zeta_1$, we have

$$\begin{aligned}&\phi_{\min}(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1}) > \phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}) - \zeta_1 > 0, \quad \phi_{\max}(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1}) < \phi_{\max}(\breve{\mathbf{\Sigma}}_{vz}) + \zeta_1, \\ &\phi_{\min}(\widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}) > \phi_{\min}(\breve{\mathbf{\Sigma}}_{zz}) - \zeta_1 > 0, \quad \phi_{\max}(\widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}) < \phi_{\max}(\breve{\mathbf{\Sigma}}_{zz}) + \zeta_1. \end{aligned} \tag{EC.2.7}$$

This implies that

$$\begin{aligned}&\left\| \left(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1} \widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}^{-1} \widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1}^{\intercal}\right)^{-1} \widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1} \widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}^{-1} \right\| \\ &= \phi_{\max}\left(\left(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1} \widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}^{-1} \widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1}^{\intercal}\right)^{-1} \widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1} \widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}^{-1}\right) \\ &\le \frac{\phi_{\max}(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1})\phi_{\max}(\widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1})}{(\phi_{\min}(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1}))^2 \phi_{\min}(\widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1})} && \text{(EC.2.8)} \\ &\le \frac{(\phi_{\max}(\breve{\mathbf{\Sigma}}_{vz}) + \zeta_1)(\phi_{\max}(\breve{\mathbf{\Sigma}}_{zz}) + \zeta_1)}{(\phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}) - \zeta_1)^2 (\phi_{\min}(\breve{\mathbf{\Sigma}}_{zz}) - \zeta_1)} && \text{(EC.2.9)} \\ &\le \left(\frac{11}{10}\right)^2 \left(\frac{10}{9}\right)^3 \frac{\phi_{\max}(\breve{\mathbf{\Sigma}}_{vz})\phi_{\max}(\breve{\mathbf{\Sigma}}_{zz})}{(\phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}))^2 \phi_{\min}(\breve{\mathbf{\Sigma}}_{zz})} && \text{(EC.2.10)} \\ &\le \breve{\kappa}, && \text{(EC.2.11)} \end{aligned}$$

where (EC.2.8) follows from Lemma EC.2, (EC.2.9) from (EC.2.7), (EC.2.10) from (EC.2.5), and (EC.2.11) from (EC.2.1). Therefore,

$$\mathcal{S}_{\phi,\mathtt{T}_1} \subseteq \left\{ \left\| \left(\widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1} \widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}^{-1} \widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1}^{\intercal}\right)^{-1} \widetilde{\mathbf{\Sigma}}_{vz,\mathtt{T}_1} \widetilde{\mathbf{\Sigma}}_{zz,\mathtt{T}_1}^{-1} \right\| \le \breve{\kappa} \right\}. \tag{EC.2.12}$$

Note that $\mathbb{E}[\widetilde{\mathbf{G}}_{\mathtt{T}_1}] = \mathbb{E}[\widetilde{\mathbf{z}}_s \epsilon_s] = \mathbf{0}$. Moreover, note that $\widetilde{\mathbf{z}}_s \epsilon_s \in \mathbb{R}^{Mq}$ is a vector of zero-mean sub-Gaussian random variables with variance proxy $\bar{z}^2\varsigma^2$. Hence, applying Corollary EC.2 to $\widetilde{\mathbf{G}}_{\mathtt{T}_1}$ yields that for all $\tau \ge 0$,

$$\Pr\left(\left\|\widetilde{\mathbf{G}}_{\mathtt{T}_1}\right\| \ge \tau\right) \le 2Mq \exp\left(-\frac{\tau^2 \mathtt{T}_1}{2M^2 q^2 \bar{z}^2 \varsigma^2}\right).$$

Thus, setting $\tau = \zeta_2 \coloneqq 2Mq\bar{z}\varsigma C_{\mathtt{T}_1}^{-\frac{1}{2}}$ yields

$$\Pr\Big(\big\|\widetilde{\mathbf{G}}_{\mathtt{T}_1}\big\| > \zeta_2\Big) \le 2Mq\exp\big(-2\mathtt{T}_1 C_{\mathtt{T}_1}^{-1}\big) = 2Mq\exp(-2\log(T)) = 2MqT^{-2}. \tag{EC.2.13}$$

It follows from (EC.2.4) and (EC.2.12) that

$$\mathcal{S}_{\phi,\mathtt{T}_1} \cap \big\{\big\|\widetilde{\mathbf{G}}_{\mathtt{T}_1}\big\| \le \zeta_2\big\} \subseteq \big\{\|\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1} - \boldsymbol{\alpha}\| \le \zeta_2\breve{\kappa}\big\} = \mathcal{A}_{\mathtt{T}_1}.$$

Hence,

$$\Pr(\mathcal{A}_{\mathtt{T}_1}) \ge \Pr\Big(\mathcal{S}_{\phi,\mathtt{T}_1} \cap \big\{\|\widetilde{\mathbf{G}}_{\mathtt{T}_1}\| \le \zeta_2\big\}\Big) \ge 1 - 2M(2p+5q)T^{-2},$$

where the second inequality follows from (EC.2.6) and (EC.2.13). □

EC.3. Covariance Stabilization Phase and Beyond: $t > \mathtt{T}_1$

For any t, let

$$\widetilde{\mathbf{v}}_t \coloneqq \begin{pmatrix} \mathbb{I}(a_t = 1)\mathbf{v}_t \\ \vdots \\ \mathbb{I}(a_t = M)\mathbf{v}_t \end{pmatrix} \in \mathbb{R}^{Mp}, \quad \mathbf{v}_t^* \coloneqq \begin{pmatrix} \mathbb{I}(a_t^* = 1)\mathbf{v}_t \\ \vdots \\ \mathbb{I}(a_t^* = M)\mathbf{v}_t \end{pmatrix} \in \mathbb{R}^{Mp}.$$

For any $t_1 < t_2$, let

$$\widetilde{\mathbf{V}}_{t_1:t_2} \coloneqq \begin{pmatrix} \widetilde{\mathbf{v}}_{t_1+1}^\intercal \\ \vdots \\ \widetilde{\mathbf{v}}_{t_2}^\intercal \end{pmatrix} \in \mathbb{R}^{(t_2-t_1)\times(Mp)}, \quad \widetilde{\mathbf{V}}_{t_1:t_2}^* \coloneqq \begin{pmatrix} (\mathbf{v}_{t_1+1}^*)^\intercal \\ \vdots \\ (\mathbf{v}_{t_2}^*)^\intercal \end{pmatrix} \in \mathbb{R}^{(t_2-t_1)\times(Mp)},$$

$$\mathbf{Z}_{t_1:t_2} \coloneqq \begin{pmatrix} \mathbf{z}_{t_1+1}^\intercal \\ \vdots \\ \mathbf{z}_{t_2}^\intercal \end{pmatrix} \in \mathbb{R}^{(t_2-t_1)\times q}, \quad \mathbf{R}_{t_1:t_2} \coloneqq \begin{pmatrix} R_{t_1+1} \\ \vdots \\ R_{t_2} \end{pmatrix} \in \mathbb{R}^{t_2-t_1}, \quad \boldsymbol{\epsilon}_{t_1:t_2} \coloneqq \begin{pmatrix} \epsilon_{t_1+1} \\ \vdots \\ \epsilon_{t_2} \end{pmatrix} \in \mathbb{R}^{t_2-t_1},$$

$$\widehat{\boldsymbol{\Sigma}}_{vz,t_1:t_2} \coloneqq \frac{1}{t_2-t_1}\sum_{s=t_1+1}^{t_2} \widetilde{\mathbf{v}}_s\mathbf{z}_s^\intercal = \frac{1}{t_2-t_1}\widetilde{\mathbf{V}}_{t_1:t_2}^\intercal\mathbf{Z}_{t_1:t_2} \in \mathbb{R}^{(Mp)\times q},$$

$$\widehat{\boldsymbol{\Sigma}}_{vz,t_1:t_2}^* \coloneqq \frac{1}{t_2-t_1}\sum_{s=t_1+1}^{t_2} \mathbf{v}_t^*\mathbf{z}_t^\intercal = \frac{1}{t_2-t_2}(\mathbf{V}_{t_1:t_2}^*)^\intercal\mathbf{Z}_{t_1:t_2} \in \mathbb{R}^{(Mp)\times q},$$

$$\widehat{\boldsymbol{\Sigma}}_{zz,t_1:t_2} \coloneqq \frac{1}{t_2-t_1}\sum_{s=t_1+1}^{t_2} \mathbf{z}_s\mathbf{z}_s^\intercal = \frac{1}{t_2-t_1}\mathbf{Z}_{t_1:t_2}^\intercal\mathbf{Z}_{t_1:t_2} \in \mathbb{R}^{q\times q},$$

$$\widehat{\mathbf{G}}_{t_1:t_2} \coloneqq \frac{1}{t_2-t_1}\sum_{s=t_1+1}^{t_2} \mathbf{z}_s\epsilon_s = \frac{1}{t_2-t_1}\mathbf{Z}_{t_1:t_2}^\intercal\boldsymbol{\epsilon}_{t_1:t_2} \in \mathbb{R}^q,$$

$$\boldsymbol{\Sigma}_{vz}^* \coloneqq \mathbb{E}[\mathbf{v}_t^*\mathbf{z}_t^\intercal] \in \mathbb{R}^{(Mp)\times q}, \quad \boldsymbol{\Sigma}_{zz} \coloneqq \mathbb{E}[\mathbf{z}_t\mathbf{z}_t^\intercal] \in \mathbb{R}^{q\times q}.$$

Note that for $t > \mathtt{T}_1$, the 2SLS estimator $\widehat{\boldsymbol{\alpha}}_t$ in Algorithm 1 satisfies

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha} &= \big(\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}^\intercal \mathcal{P}[\mathbf{Z}_{\mathtt{T}_1:t}]\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\big)^{-1}\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}^\intercal \mathcal{P}[\mathbf{Z}_{\mathtt{T}_1:t}]\boldsymbol{\epsilon}_{\mathtt{T}_1:t} \\ &= \Big(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}^\intercal\Big)^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}. \end{aligned} \tag{EC.3.1}$$

Moreover, recall that in Algorithm 1 we use the following estimator for the variance of noise terms:

$$\widehat{\sigma}_t \coloneqq (t-\mathtt{T}_1)^{-\frac{1}{2}}\big\|\mathbf{R}_{\mathtt{T}_1:t} - \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\widehat{\boldsymbol{\alpha}}_t\big\| \in \mathbb{R}_+. \tag{EC.3.2}$$

In order to establish an probabilistic bound on $\|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}\|$, we establish below bounds on a set of events that characterize how accurate the estimators such as $\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}$, $\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}$, and $\widehat{\sigma}_t^2$ are relative to their corresponding true values.

Specifically, we fix arbitrary constants $C_{\mathcal{A}}, \eta, \Delta > 0$, and define for any $t > \mathtt{T}_1$ the following events:

$$\begin{aligned}\mathcal{A}_t &\coloneqq \left\{\|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}\| \le C_{\mathcal{A}}\sqrt{\frac{\log(T)}{t-\mathtt{T}_1}}\right\}, \quad \mathcal{B}_t \coloneqq \big\{\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t} - \boldsymbol{\Sigma}_{vz}^*\| \le \eta\big\}, \\ \mathcal{C}_t &\coloneqq \big\{\|\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t} - \boldsymbol{\Sigma}_{zz}\| \le \eta\big\}, \quad \mathcal{D}_t \coloneqq \big\{|\widehat{\sigma}_t^2 - \sigma^2| \le \Delta\big\}.\end{aligned}$$

In Lemmas EC.9—EC.13 below, we will establish probabilistic bounds of these events and their various intersections.

Lemma EC.9. *Define*

$$\bar{\kappa} \coloneqq \frac{2\phi_{\max}(\boldsymbol{\Sigma}_{vz}^*)\phi_{\max}(\boldsymbol{\Sigma}_{zz})}{(\phi_{\min}(\boldsymbol{\Sigma}_{vz}^*))^2\phi_{\min}(\boldsymbol{\Sigma}_{zz})}. \tag{EC.3.3}$$

Suppose that the constants $C_{\mathcal{A}}$ and η in the definitions of $\mathcal{A}_t$, $\mathcal{B}_t$ and $\mathcal{C}_t$ satisfy

$$C_{\mathcal{A}} \ge 2q\bar{z}\varsigma\bar{\kappa}, \tag{EC.3.4}$$

$$\eta \le \frac{1}{10}\min\Big\{\phi_{\min}(\boldsymbol{\Sigma}_{vz}^*),\, \phi_{\min}(\boldsymbol{\Sigma}_{zz})\Big\}. \tag{EC.3.5}$$

Then, for all $t = \mathtt{T}_1 + 1, \ldots, T$,

$$\Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t) \le 2qT^{-2}.$$

Proof. It follows from Assumption 1 that $0 < \phi_{\min}(\boldsymbol{\Sigma}_{zz}) \le \phi_{\max}(\boldsymbol{\Sigma}_{zz}) < \infty$ and $0 < \phi_{\min}(\boldsymbol{\Sigma}_{vz}^*) \le \phi_{\max}(\boldsymbol{\Sigma}_{vz}^*) < \infty$. Hence, $\bar{\kappa}$ is a positive finite constant.

We apply an analysis that is similar to that used in the proof of Lemma EC.8. Specifically, it follows from (EC.3.1) that for all $t > \mathtt{T}_1$,

$$\|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}\| \le \left\|\Big(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}^\intercal\Big)^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\right\| \cdot \big\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\big\|. \tag{EC.3.6}$$

We analyze below the two norms on the right-hand-side of (EC.3.6) separately.

Note that conditional on the event $\mathcal{B}_t \cap \mathcal{C}_t$ with $\eta < \min\{\phi_{\min}(\mathbf{\Sigma}^*_{vz}), \phi_{\min}(\mathbf{\Sigma}_{zz})\}$, which is implied by condition (EC.3.5), we have

$$\begin{aligned} &\phi_{\min}(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}) > \phi_{\min}(\mathbf{\Sigma}^*_{vz}) - \eta > 0, \quad \phi_{\max}(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}) < \phi_{\max}(\mathbf{\Sigma}^*_{vz}) + \eta, \\ &\phi_{\min}(\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t}) > \phi_{\min}(\mathbf{\Sigma}_{zz}) - \eta > 0, \quad \phi_{\max}(\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t}) < \phi_{\max}(\mathbf{\Sigma}_{zz}) + \eta. \end{aligned} \tag{EC.3.7}$$

This implies that

$$\begin{aligned} \left\| \left(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{-1}_{zz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{\intercal}_{vz,\mathtt{T}_1:t} \right)^{-1} \widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{-1}_{zz,\mathtt{T}_1:t} \right\| &= \phi_{\max} \left(\left(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{-1}_{zz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{\intercal}_{vz,\mathtt{T}_1:t} \right)^{-1} \widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{-1}_{zz,\mathtt{T}_1:t} \right) \\ &\le \frac{\phi_{\max}(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}) \phi_{\max}(\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t})}{(\phi_{\min}(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}))^2 \phi_{\min}(\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t})} \qquad \text{(EC.3.8)} \\ &\le \frac{(\phi_{\max}(\mathbf{\Sigma}^*_{vz}) + \eta)(\phi_{\max}(\mathbf{\Sigma}_{zz}) + \eta)}{(\phi_{\min}(\mathbf{\Sigma}^*_{vz}) - \eta)^2 (\phi_{\min}(\mathbf{\Sigma}_{zz}) - \eta)} \qquad \text{(EC.3.9)} \\ &\le \left(\frac{11}{10} \right)^2 \left(\frac{10}{9} \right)^3 \frac{\phi_{\max}(\mathbf{\Sigma}^*_{vz}) \phi_{\max}(\mathbf{\Sigma}_{zz})}{(\phi_{\min}(\mathbf{\Sigma}^*_{vz}))^2 \phi_{\min}(\mathbf{\Sigma}_{zz})} \qquad \text{(EC.3.10)} \\ &\le \bar{\kappa}, \qquad \text{(EC.3.11)} \end{aligned}$$

where (EC.3.8) follows from Lemma EC.2, (EC.3.9) from (EC.3.7), (EC.3.10) from (EC.3.5), and (EC.3.11) from (EC.3.3). Therefore,

$$\mathcal{B}_t \cap \mathcal{C}_t \subseteq \left\{ \left\| \left(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{-1}_{zz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{\intercal}_{vz,\mathtt{T}_1:t} \right)^{-1} \widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t} \widehat{\mathbf{\Sigma}}^{-1}_{zz,\mathtt{T}_1:t} \right\| \le \bar{\kappa} \right\}. \tag{EC.3.12}$$

Next, note that $\mathbb{E}[\widehat{\mathbf{G}}_{\mathtt{T}_1:t}] = \mathbb{E}[\mathbf{z}_s \epsilon_s] = \mathbf{0}$. Moreover, it is straightforward to see that $\mathbf{z}_s \epsilon_s \in \mathbb{R}^q$ is a vector of zero-mean sub-Gaussian random variables with variance proxy $\bar{z}^2 \varsigma^2$. Hence, applying Corollary EC.2 to $\widehat{\mathbf{G}}_{\mathtt{T}_1:t}$ yields that for all $\tau \ge 0$,

$$\Pr\left(\left\| \widehat{\mathbf{G}}_{\mathtt{T}_1:t} \right\| \ge \tau \right) \le 2q \exp\left(-\frac{\tau^2 (t - \mathtt{T}_1)}{2q^2 \bar{z}^2 \varsigma^2} \right). \tag{EC.3.13}$$

Thus, setting $\tau = \zeta_3 := 2q\bar{z}\varsigma \sqrt{\frac{\log(T)}{t - \mathtt{T}_1}}$ yields

$$\Pr\left(\| \widehat{\mathbf{G}}_{\mathtt{T}_1:t} \| > \zeta_3 \right) \le 2q \exp(-2\log(T)) = 2qT^{-2}. \tag{EC.3.14}$$

It follows from (EC.3.6) and (EC.3.12) that

$$\left\{ \left\| \widehat{\mathbf{G}}_{\mathtt{T}_1:t} \right\| \le \zeta_3 \right\} \cap \mathcal{B}_t \cap \mathcal{C}_t \subseteq \left\{ \| \widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha} \| \le \zeta_3 \bar{\kappa} \right\} \subseteq \left\{ \| \widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha} \| \le C_{\mathcal{A}} \sqrt{\frac{\log(T)}{t - \mathtt{T}_1}} \right\} = \mathcal{A}_t,$$

where the second step follows from the assumption on $C_{\mathcal{A}}$ in (EC.3.4) as well as the definitions of ζ_3 and $\bar{\kappa}$. Hence,

$$\left\{ \left\| \widehat{\mathbf{G}}_{\mathtt{T}_1:t} \right\| \le \zeta_3 \right\} \cap \mathcal{B}_t \cap \mathcal{C}_t \subseteq \mathcal{A}_t \cap \mathcal{B}_t \cap \mathcal{C}_t,$$

and thus

$$\left\{\left\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\right\| > \zeta_3\right\} \cap \mathcal{B}_t \cap \mathcal{C}_t \supseteq \mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t.$$

It follows that

$$\Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t) \le \Pr\left(\left\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\right\| > \zeta_3\right) \le 2qT^{-2},$$

where the second inequality follows from (EC.3.14). □

We now define for any $t > \mathtt{T}_2$ the following joint events:

$$\mathcal{A}_{\mathtt{T}_2:t} = \bigcap_{s=\mathtt{T}_2+1}^{t} \mathcal{A}_s, \quad \mathcal{B}_{\mathtt{T}_2:t} = \bigcap_{s=\mathtt{T}_2+1}^{t} \mathcal{B}_s, \quad \mathcal{C}_{\mathtt{T}_2:t} = \bigcap_{s=\mathtt{T}_2+1}^{t} \mathcal{C}_s, \quad \mathcal{D}_{\mathtt{T}_2:t} := \bigcap_{s=\mathtt{T}_2+1}^{t} \mathcal{D}_s.$$

LEMMA EC.10. *Suppose $C_{\mathtt{T}_2} \ge 6q^2\bar{z}^4\eta^{-2}$. Then, for all $t = \mathtt{T}_2+1, \ldots, T$,*

$$\Pr(\mathcal{C}_t) \ge 1 - 2qT^{-3} \quad \textit{and} \quad \Pr(\mathcal{C}_{\mathtt{T}_2:t}) \ge 1 - 2q(t-\mathtt{T}_2)T^{-3}.$$

Proof. Note that by the definition of $\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t}$,

$$\mathbb{E}[\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t}] = \frac{1}{\mathtt{T}_1 - t} \sum_{s=\mathtt{T}_1+1}^{t} \mathbb{E}[\mathbf{z}_s\mathbf{z}_s^\intercal] = \mathbf{\Sigma}_{zz}.$$

In addition, $\|\mathbf{z}_s\mathbf{z}_s^\intercal\| \le q\bar{z}^2$. Hence, it follows from Lemma EC.4 that for all $t \ge \mathtt{T}_2$,

$$\begin{aligned}
\Pr(\mathcal{C}_t^{\mathsf{c}}) = \Pr\left(\|\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:t} - \mathbf{\Sigma}_{zz}\| > \eta\right) &\le 2q\exp\left(-\frac{\eta^2(t-\mathtt{T}_1)}{2q^2\bar{z}^4}\right) \\
&\le 2q\exp\left(-\frac{\eta^2(\mathtt{T}_2-\mathtt{T}_1)}{2q^2\bar{z}^4}\right) = 2q\exp\left(-\frac{\eta^2 C_{\mathtt{T}_2}\log(T)}{2q^2\bar{z}^4}\right) \\
&\le 2q\exp(-3\log(T)) = 2qT^{-3},
\end{aligned}$$

where the last inequality holds because $C_{\mathtt{T}_2} \ge 6q^2\bar{z}^4\eta^{-2}$. Therefore, $\Pr(\mathcal{C}_t) \ge 1 - 2qT^{-3}$; further, by the Bonferroni inequality,

$$\Pr(\mathcal{C}_{\mathtt{T}_2:t}) \ge 1 - \sum_{s=\mathtt{T}_2+1}^{t} \Pr(\mathcal{C}_s^{\mathsf{c}}) \ge 1 - 2q(t-\mathtt{T}_2)T^{-3}. \qquad \square$$

LEMMA EC.11. *Suppose*

$$C_{\mathtt{T}_2} \ge \max\left\{\frac{3}{\mathsf{C}_{\mathsf{Bern}}\min\left\{\frac{\Delta^2}{4K_\epsilon^2}, \frac{\Delta}{2K_\epsilon}\right\}}, \frac{6Mpq^2\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2}{\min\left\{\frac{\Delta^2}{128\sigma^2}, \frac{\Delta}{64}\right\}}\right\}, \tag{EC.3.15}$$

where $\mathsf{C}_{\mathsf{Bern}}$ is the constant in Lemma EC.6 and $K_\epsilon = \|\epsilon_1^2 - \sigma^2\|_{\psi_1} < \infty$. Suppose that the constant η in the definitions of $\mathcal{B}_t$ and $\mathcal{C}_t$ satisfies

$$\eta \le \frac{1}{10}\min\left\{\phi_{\min}(\mathbf{\Sigma}_{vz}^*),\ \phi_{\min}(\mathbf{\Sigma}_{zz})\right\}. \tag{EC.3.16}$$

Then, for all $t = \mathtt{T}_2+1, \ldots, T$,

$$\Pr(\mathcal{B}_t \cap \mathcal{C}_t \cap \mathcal{D}_t^{\mathsf{c}}) \le (2q+2)T^{-3}.$$

Proof. Denote the estimated residual by $\widehat{\epsilon}_t \coloneqq R_t - \widetilde{\mathbf{v}}_t^\intercal \widehat{\boldsymbol{\alpha}}_t$, and define for all $t > \mathtt{T}_1$,

$$\begin{aligned}\widehat{\boldsymbol{\epsilon}}_{\mathtt{T}_1:t} &\coloneqq \begin{pmatrix} \widehat{\epsilon}_{\mathtt{T}_1} \\ \vdots \\ \widehat{\epsilon}_t \end{pmatrix} = \mathbf{R}_{\mathtt{T}_1:t} - \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\widehat{\boldsymbol{\alpha}}_t = \boldsymbol{\epsilon}_{\mathtt{T}_1:t} + \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\boldsymbol{\alpha} - \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\widehat{\boldsymbol{\alpha}}_t \\ &= \boldsymbol{\epsilon}_{\mathtt{T}_1:t} - \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}, \end{aligned} \tag{EC.3.17}$$

where $\mathbf{M}_{\mathtt{T}_1:t} \coloneqq \left(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}^\intercal\right)^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}$, and the last equality follows from (EC.3.1). Then, it follows from (EC.3.2) that

$$\begin{aligned}\widehat{\sigma}_t^2 &= \frac{1}{t-\mathtt{T}_1}\|\widehat{\boldsymbol{\epsilon}}_{\mathtt{T}_1:t}\|^2 = \frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t} - \widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2 \\ &= \frac{1}{t-\mathtt{T}_1}\Big(\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2 - 2\boldsymbol{\epsilon}_{\mathtt{T}_1:t}^\intercal\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t} + \|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2\Big) \\ &\le \frac{1}{t-\mathtt{T}_1}\Big(\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2 + 2\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\| + \|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2\Big). \end{aligned} \tag{EC.3.18}$$

Therefore, to examine the event $\mathcal{D}_t^{\mathsf{c}} = \{|\widehat{\sigma}_t^2 - \sigma^2| \le \Delta\}$, it suffices to analyze the following two events separately: (i) $\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2$ is in the proximity of σ^2, and (ii) $\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2$ is small.

We first analyze the former. Define $\mathcal{E}_t \coloneqq \left\{\left|\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2 - \sigma^2\right| \le \frac{\Delta}{2}\right\}$. Note that

$$\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2 - \sigma^2 = \frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}(\epsilon_s^2 - \sigma^2),$$

and $\mathbb{E}[\epsilon_s^2] = \sigma^2$. Moreover, since ϵ_s is sub-Gaussian, it follows from Lemma 2.7.6 of Vershynin (2018) that $\epsilon_s^2 - \sigma^2$ is sub-exponential. Hence, by Lemma EC.6,

$$\begin{aligned}\Pr(\mathcal{E}_t^{\mathsf{c}}) &\le 2\exp\left(-(t-\mathtt{T}_1)\mathsf{C}_{\mathsf{Bern}}\min\left\{\frac{\Delta^2}{4K_\epsilon^2}, \frac{\Delta}{2K_\epsilon}\right\}\right) \\ &\le 2\exp\left(-C_{\mathtt{T}_2}\log(T)\mathsf{C}_{\mathsf{Bern}}\min\left\{\frac{\Delta^2}{4K_\epsilon^2}, \frac{\Delta}{2K_\epsilon}\right\}\right) \\ &\le 2\exp(-3\log(T)) = 2T^{-3}, \end{aligned} \tag{EC.3.19}$$

where the last inequality holds because of the condition (EC.3.15).

Now, we analyze the term $\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2$. Note that by Lemma EC.3,

$$\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\|^2 \le \frac{1}{t-\mathtt{T}_1}\left(\sqrt{Mp(t-\mathtt{T}_1)}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\|_{\max}\right)^2 \le Mp\bar{v}^2. \tag{EC.3.20}$$

It follows from (EC.3.11) that conditional on the event $\mathcal{B}_t \cap \mathcal{C}_t$ with η satisfying (EC.3.16),

$$\|\mathbf{M}_{\mathtt{T}_1:t}\|^2 = \left\|\left(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}^\intercal\right)^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:t}^{-1}\right\|^2 \le \bar{\kappa}^2. \tag{EC.3.21}$$

Note that

$$\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2 \le \frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\|^2\|\mathbf{M}_{\mathtt{T}_1:t}\|^2\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2.$$

It follows from (EC.3.20) and (EC.3.21) that conditional on the event $\mathcal{B}_t\cap\mathcal{C}_t\cap\{\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|\le\zeta_4\}$ with $\zeta_4 \coloneqq \sqrt{6}q\bar{z}\varsigma\sqrt{\frac{\log(T)}{t-\mathtt{T}_1}}$, we have

$$\begin{aligned}\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2 &\le Mp\bar{v}^2\zeta_4^2\bar{\kappa}^2 = 6Mpq^2\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2\frac{\log(T)}{t-\mathtt{T}_1}\\ &\le \frac{6Mpq^2\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2}{C_{\mathtt{T}_2}} \le \zeta_5 \end{aligned} \tag{EC.3.22}$$

where $\zeta_5 \coloneqq \min\{\frac{\Delta^2}{128\sigma^2},\frac{\Delta}{64}\}$. Note that the second inequality holds because $t-\mathtt{T}_1\ge\mathtt{T}_2-\mathtt{T}_1\ge C_{\mathtt{T}_2}\log(T)$, and the last inequality because of the condition (EC.3.15). Consequently,

$$\{\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|\le\zeta_4\}\cap\mathcal{B}_t\cap\mathcal{C}_t\subseteq\left\{\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2\le\zeta_5\right\}\coloneqq\mathcal{F}_t. \tag{EC.3.23}$$

Note that conditional on the event $\mathcal{E}_t\cap\mathcal{F}_t$, we have

$$\begin{aligned}|\widehat{\sigma}_t^2-\sigma^2| &\le \left|\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2-\sigma^2\right|+\left|\widehat{\sigma}_t^2-\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2\right|\\ &\le \left|\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2-\sigma^2\right|+2\left(\frac{1}{t-\mathtt{T}_1}\|\boldsymbol{\epsilon}_{\mathtt{T}_1:t}\|^2\cdot\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2\right)^{1/2}\\ &\quad+\frac{1}{t-\mathtt{T}_1}\|\widetilde{\mathbf{V}}_{\mathtt{T}_1:t}\mathbf{M}_{\mathtt{T}_1:t}\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|^2\\ &\le \frac{\Delta}{2}+2\sqrt{\left(\sigma^2+\frac{\Delta}{2}\right)\zeta_5}+\zeta_5, \end{aligned} \tag{EC.3.24}$$

where the second inequality follows from (EC.3.18). Further, by the definition of ζ_5, we have

$$\left(\sigma^2+\frac{\Delta}{2}\right)\zeta_5\le\sigma^2\cdot\frac{\Delta^2}{128\sigma^2}+\frac{\Delta}{2}\cdot\frac{\Delta}{64}=\frac{\Delta^2}{64}.$$

It then follows from (EC.3.24) that $|\widehat{\sigma}_t^2-\sigma^2|\le\frac{\Delta}{2}+\frac{\Delta}{4}+\frac{\Delta}{64}<\Delta$. Hence, $\mathcal{E}_t\cap\mathcal{F}_t\subseteq\mathcal{D}_t$, which in conjunction with (EC.3.23) implies that

$$\mathcal{E}_t\cap\{\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|\le\zeta_4\}\cap\mathcal{B}_t\cap\mathcal{C}_t\subseteq\mathcal{E}_t\cap\mathcal{F}_t\subseteq\mathcal{D}_t,$$

and thus $\mathcal{E}_t\cap\{\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|\le\zeta_4\}\cap\mathcal{B}_t\cap\mathcal{C}_t\subseteq\mathcal{B}_t\cap\mathcal{C}_t\cap\mathcal{D}_t$. Hence,

$$\mathcal{B}_t\cap\mathcal{C}_t\cap\mathcal{D}_t^{\mathsf{c}}\subseteq\left(\mathcal{E}_t\cap\{\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|\le\zeta_4\}\right)^{\mathsf{c}}\cap\mathcal{B}_t\cap\mathcal{C}_t\subseteq\mathcal{E}_t^{\mathsf{c}}\cup\{\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|>\zeta_4\}. \tag{EC.3.25}$$

Note that by (EC.3.13), we have

$$\Pr\left(\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|>\zeta_4\right)\le 2q\exp\left(-\frac{\zeta_4^2(t-\mathtt{T}_1)}{2q^2\bar{z}^2\varsigma^2}\right)\le 2qT^{-3}. \tag{EC.3.26}$$

Combining (EC.3.19), (EC.3.25), and (EC.3.26) yields

$$\Pr(\mathcal{B}_t\cap\mathcal{C}_t\cap\mathcal{D}_t^{\mathsf{c}})\le\Pr(\mathcal{E}_t^{\mathsf{c}})+\Pr\left(\|\widehat{\mathbf{G}}_{\mathtt{T}_1:t}\|>\zeta_4\right)\le(2q+2)T^{-3}. \quad \square$$

LEMMA EC.12. *Define*

$$\mathring{\kappa} \coloneqq \frac{2\phi_{\max}(\mathbf{\Sigma}_{zz})}{(\phi_{\min}(\mathbf{\Sigma}^*_{vz}))^2} \tag{EC.3.27}$$

Suppose

$$C_{\mathtt{T}_1} \geq 1024M^4p^2q^3\bar{v}^4\bar{z}^4\varsigma^2\breve{\kappa}^2L^2\eta^{-2}, \tag{EC.3.28}$$

$$C_{\mathtt{T}_2} \geq \eta^{-2}\max\left\{128(M-1)^2pq\bar{v}^2\bar{z}^2,\, 256M^2L^2p^2q\bar{v}^4\bar{z}^2\left(C_{\mathcal{A}}+\sqrt{\theta\mathring{\kappa}(\sigma^2+\Delta)}\right)^2\right\}. \tag{EC.3.29}$$

Suppose that the constant η in the definitions of $\mathcal{B}_t$ and $\mathcal{C}_t$ satisfies (EC.3.16)*. Then, for all $t=\mathtt{T}_2+1,\ldots,T$,*

$$\Pr(\mathcal{B}_t^{\mathsf{c}}\cap\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_{\mathtt{T}_2:t-1}\cap\mathcal{C}_{\mathtt{T}_2:t-1}\cap\mathcal{D}_{\mathtt{T}_2:t-1}) \leq 2(Mp+q+1)T^{-2}.$$

Proof. A subtlety in analyzing the event $\mathcal{B}_t^c=\{\|\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}-\mathbf{\Sigma}^*_{vz}\|>\eta\}$ for $t>\mathtt{T}_1$ is that $\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}$ is not an unbiased estimator of $\mathbf{\Sigma}^*_{vz}$—that is, $\mathbb{E}[\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}]\neq\mathbf{\Sigma}^*_{vz}$. This is because $\mathbf{\Sigma}^*_{vz}$ is defined with respect to the (unknown) optimal policy (i.e., the optimal arm a_s^* is taken for all $s=\mathtt{T}_1+1,\ldots,t$); whereas $\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}$ is defined with respect to the actual policy that is used to generate the data. To address the issue, we define

$$\widehat{\mathbf{\Sigma}}^*_{vz,\mathtt{T}_1:t} \coloneqq \frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\mathbf{v}_t^*\mathbf{z}_t^{\intercal}\in\mathbb{R}^{(Mp)\times q},$$

and analyze separately the two components of the following decomposition:

$$\|\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}-\mathbf{\Sigma}^*_{vz}\| \leq \underbrace{\|\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}-\widehat{\mathbf{\Sigma}}^*_{vz,\mathtt{T}_1:t}\|}_{I_1}+\underbrace{\|\widehat{\mathbf{\Sigma}}^*_{vz,\mathtt{T}_1:t}-\mathbf{\Sigma}^*_{vz}\|}_{I_2}. \tag{EC.3.30}$$

For I_2, note that $\mathbb{E}[\mathbf{v}_t^*\mathbf{z}_t^{\intercal}]=\mathbf{\Sigma}^*_{vz}$ and that, by Lemma EC.3,

$$\|\mathbf{v}_t^*\mathbf{z}_t^{\intercal}-\mathbf{\Sigma}^*_{vz}\| \leq \|\mathbf{v}_t^*\mathbf{z}_t^{\intercal}\|+\|\mathbf{\Sigma}^*_{vz}\| \leq 2\sqrt{Mpq}\bar{v}\bar{z}.$$

It then follows from Lemma EC.4 that for all $t\geq\mathtt{T}_2$,

$$\begin{aligned}\Pr\left(\|\widehat{\mathbf{\Sigma}}^*_{vz,\mathtt{T}_1:t}-\mathbf{\Sigma}^*_{vz}\|>\frac{\eta}{2}\right) &\leq 2(Mp+q)\exp\left(-\frac{(t-\mathtt{T}_1)\eta^2/4}{2(2\sqrt{Mpq}\bar{v}\bar{z})^2}\right)\\ &\leq 2(Mp+q)\exp\left(-\frac{\eta^2C_{\mathtt{T}_2}\log(T)}{32Mpq\bar{v}^2\bar{z}^2}\right)\\ &\leq 2(Mp+q)\exp(-2\log(T))=2(Mp+q)T^{-2},\end{aligned} \tag{EC.3.31}$$

where the second inequality holds because $t-\mathtt{T}_1\geq\mathtt{T}_2-\mathtt{T}_1=C_{\mathtt{T}_2}\log(T)$, and the third holds because of the condition (EC.3.29).

We now consider I_1. Note that

$$\left\|\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}-\widehat{\mathbf{\Sigma}}^*_{vz,\mathtt{T}_1:t}\right\|=\left\|\frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\begin{pmatrix}(\mathbb{I}(a_s=1)-\mathbb{I}(a_s^*=1))\mathbf{v}_s\\ \vdots\\ (\mathbb{I}(a_s=M)-\mathbb{I}(a_s^*=M))\mathbf{v}_s\end{pmatrix}\mathbf{z}_s^\intercal\right\|$$
$$\le\frac{2}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\mathbb{I}(a_s\neq a_s^*)\|\mathbf{v}_s\mathbf{z}_s^\intercal\|\le\frac{2\sqrt{pq}\bar{v}\bar{z}}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\mathbb{I}(a_s\neq a_s^*),\tag{EC.3.32}$$

where the first inequality holds because there are at most non-zero entries in the vector $(\mathbb{I}(a_s=1)-\mathbb{I}(a_s^*=1),\ldots,\mathbb{I}(a_s=M)-\mathbb{I}(a_s^*=M))^\intercal$; the last inequality follows from Lemma EC.3. We show next that conditional on some other events,

$$\{a_s=j\}\subseteq\left\{|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}}\right\},\quad\forall s\ge\mathtt{T}_1+1,\,\forall j\neq a_s^*,$$

where $\boldsymbol{\delta}_{i,j}=\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j$ and L is the constant defined in the margin condition of Assumption 1. We consider two cases: (i) $s=\mathtt{T}_1+1,\ldots,\mathtt{T}_2$ and (ii) $s=\mathtt{T}_2+1,\ldots,T$. Let $\widehat{\boldsymbol{\delta}}_{i,j,s}\coloneqq\widehat{\boldsymbol{\alpha}}_{i,s}-\widehat{\boldsymbol{\alpha}}_{j,s}$ for each s.

Case (i): $s=\mathtt{T}_1+1,\ldots,\mathtt{T}_2$**.** Recall that for $s=\mathtt{T}_1+1,\ldots,\mathtt{T}_2$, the selected arm is $a_s=\arg\max_{i=1,\ldots,M}\{\mathbf{v}_s^\intercal\widehat{\boldsymbol{\alpha}}_{i,\mathtt{T}_1}\}$. Hence, for any $j\neq a_s^*$, if $a_s=j$, then $\mathbf{v}_s^\intercal\widehat{\boldsymbol{\delta}}_{j,a_s^*,\mathtt{T}_1}=\mathbf{v}_s^\intercal\widehat{\boldsymbol{\alpha}}_{j,\mathtt{T}_1}-\mathbf{v}_s^\intercal\widehat{\boldsymbol{\alpha}}_{a_s^*,\mathtt{T}_1}>0$. Thus,

$$\begin{aligned}0\ge\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}>\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}-\mathbf{v}_s^\intercal\widehat{\boldsymbol{\delta}}_{j,a_s^*,\mathtt{T}_1}&=\mathbf{v}_s^\intercal[(\boldsymbol{\alpha}_j-\widehat{\boldsymbol{\alpha}}_{j,\mathtt{T}_1})-(\boldsymbol{\alpha}_{a_s^*}-\widehat{\boldsymbol{\alpha}}_{a_s^*,\mathtt{T}_1})]\\&\ge-2\|\mathbf{v}_s\|\|\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}\|\ge-2\sqrt{p}\bar{v}\|\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}\|.\end{aligned}$$

Further, conditional on $\mathcal{A}_{\mathtt{T}_1}=\left\{\|\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}-\boldsymbol{\alpha}\|\le 2Mq\bar{z}\varsigma\breve{\kappa}C_{\mathtt{T}_1}^{-\frac{1}{2}}\right\}$ defined in (EC.2.2), we have

$$|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le 2\sqrt{p}\bar{v}\|\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}_{\mathtt{T}_1}\|\le 4M\sqrt{p}q\bar{v}\bar{z}\varsigma\breve{\kappa}C_{\mathtt{T}_1}^{-\frac{1}{2}}\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}},$$

where the last inequality follows from the condition (EC.3.28). Hence,

$$\mathcal{A}_{\mathtt{T}_1}\cap\{a_s=j\}\subseteq\left\{|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}}\right\},\quad\forall s=\mathtt{T}_1+1,\ldots,\mathtt{T}_2,\,\forall j\neq a_s^*.\tag{EC.3.33}$$

Case (ii): $s=\mathtt{T}_2+1,\ldots,T$**.** Recall that for $s\ge\mathtt{T}_2+1$, the arm selected at time s is

$$a_s=\arg\max_{i=1,\ldots,M}\left\{\mathbf{v}_s^\intercal\widehat{\boldsymbol{\alpha}}_{i,s-1}+\widehat{\sigma}_{s-1}\sqrt{2\theta\log(s-\mathtt{T}_1)\cdot\mathbf{v}_s^\intercal\widehat{\mathbf{\Omega}}_{i,s-1}\mathbf{v}_s}\right\}$$

where $\widehat{\mathbf{\Omega}}_{i,s-1}\in\mathbb{R}^{p\times p}$ is the i-th diagonal $p\times p$ block of $\widehat{\mathbf{\Omega}}_{s-1}\in\mathbb{R}^{(Mp)\times(Mp)}$ defined as

$$\begin{aligned}\widehat{\mathbf{\Omega}}_{s-1}&=\left(\widetilde{\mathbf{V}}_{\mathtt{T}_1:s-1}^\intercal\mathcal{P}[\mathbf{Z}_{\mathtt{T}_1:s-1}]\widetilde{\mathbf{V}}_{\mathtt{T}_1:s-1}\right)^{-1}\\&=\frac{1}{s-1-\mathtt{T}_1}\left(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:s-1}\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:s-1}^{-1}\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:s-1}^\intercal\right)^{-1}.\end{aligned}$$

Thus, for any $j \neq a_s^*$, if $a_s = j$, then

$$\mathbf{v}_s^\intercal \widehat{\boldsymbol{\delta}}_{j,a_s^*,s-1} + \widehat{\sigma}_{s-1}\sqrt{2\theta\log(s-\mathtt{T}_1)}\left(\sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{j,s-1}\mathbf{v}_s} - \sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\right) > 0.$$

Since $\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*} \leq 0$, we have

$$\begin{aligned} 0 \geq \mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*} > & \underbrace{\mathbf{v}_s^\intercal(\boldsymbol{\delta}_{j,a_s^*} - \widehat{\boldsymbol{\delta}}_{j,a_s^*,s-1})}_{W_{1,j}} \\ & \underbrace{- \widehat{\sigma}_{s-1}\sqrt{2\theta\log(s-\mathtt{T}_1)}\left(\sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{j,s-1}\mathbf{v}_s} - \sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\right)}_{W_{2,j}}. \end{aligned} \tag{EC.3.34}$$

For $W_{1,j}$, note that conditional on $\mathcal{A}_{s-1} = \left\{\|\widehat{\boldsymbol{\alpha}}_{s-1} - \boldsymbol{\alpha}\| \leq C_{\mathcal{A}}\sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}}\right\}$, we have

$$\begin{aligned} |W_{1,j}| = |\mathbf{v}_s^\intercal[(\boldsymbol{\alpha}_j - \widehat{\boldsymbol{\alpha}}_{j,s-1}) - (\boldsymbol{\alpha}_{a_s^*} - \widehat{\boldsymbol{\alpha}}_{a_s^*,s-1})]| &\leq 2\|\mathbf{v}_s\|\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}_{s-1}\| \\ &\leq 2\sqrt{p}\bar{v}C_{\mathcal{A}}\sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}} \qquad \text{(EC.3.35)} \\ &\leq 2\sqrt{p}\bar{v}C_{\mathcal{A}}C_{\mathtt{T}_2}^{-\frac{1}{2}}. \qquad \text{(EC.3.36)} \end{aligned}$$

For $W_{2,j}$, note that

$$\begin{aligned} & \left(\sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{j,s-1}\mathbf{v}_s} - \sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\right)^2 \\ \leq & \left|\sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{j,s-1}\mathbf{v}_s} - \sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\right| \cdot \left|\sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{j,s-1}\mathbf{v}_s} + \sqrt{\mathbf{v}_s^\intercal \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\right| \\ \leq & |\mathbf{v}_s^\intercal(\widehat{\boldsymbol{\Omega}}_{j,s-1} - \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1})\mathbf{v}_s| \\ \leq & \|\mathbf{v}_s\|^2\|\widehat{\boldsymbol{\Omega}}_{j,s-1} - \widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\| \leq \|\mathbf{v}_s\|^2(\|\widehat{\boldsymbol{\Omega}}_{j,s-1}\| + \|\widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\|) \\ \leq & 2\|\mathbf{v}_s\|^2\|\widehat{\boldsymbol{\Omega}}_{s-1}\| \leq 2p\bar{v}^2\|\widehat{\boldsymbol{\Omega}}_{s-1}\|, \end{aligned} \tag{EC.3.37}$$

where the second to the last inequality holds because $\widehat{\boldsymbol{\Omega}}_{i,s-1}$ is a submatrix of $\widehat{\boldsymbol{\Omega}}_{s-1}$, so the norm of the former is upper bounded by that of the latter; see Golub and Van Loan (2013, page 72). Further, conditional on the event $\mathcal{B}_{s-1} \cap \mathcal{C}_{s-1} = \{\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1} - \boldsymbol{\Sigma}_{vz}^*\| \leq \eta\} \cap \{\|\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1} - \boldsymbol{\Sigma}_{zz}\| \leq \eta\}$, we have

$$\begin{aligned} (s-1-\mathtt{T}_1)\|\widehat{\boldsymbol{\Omega}}_{s-1}\| &= \phi_{\max}((s-1-\mathtt{T}_1)\widehat{\boldsymbol{\Omega}}_{s-1}) \\ &= \phi_{\max}\left(\left(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}^\intercal\right)^{-1}\right) \\ &\leq \frac{\phi_{\max}(\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1})}{(\phi_{\min}(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}))^2} \leq \frac{\phi_{\max}(\boldsymbol{\Sigma}_{zz}) + \eta}{(\phi_{\min}(\boldsymbol{\Sigma}_{vz}^*) - \eta)^2} \\ &\leq \left(\frac{11}{10}\right)\left(\frac{10}{9}\right)^2 \frac{\phi_{\max}(\boldsymbol{\Sigma}_{zz})}{(\phi_{\min}(\boldsymbol{\Sigma}_{vz}^*))^2} \leq \mathring{\kappa}, \end{aligned} \tag{EC.3.38}$$

where the first inequality follows from Lemma EC.2, the third from the condition (EC.3.16), and the last from the definition of $\mathring{\kappa}$ in (EC.3.27). Plugging (EC.3.38) in (EC.3.37) yields

$$\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1}\subseteq\left\{\sqrt{\mathbf{v}_s^\intercal\widehat{\mathbf{\Omega}}_{j,s-1}\mathbf{v}_s}-\sqrt{\mathbf{v}_s^\intercal\widehat{\mathbf{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\le\sqrt{\frac{2p\bar{v}^2\mathring{\kappa}}{s-1-\mathtt{T}_1}}\right\} \tag{EC.3.39}$$

In addition, note that

$$\mathcal{D}_{s-1}=\{|\widehat{\sigma}_{s-1}^2-\sigma^2|\le\Delta\}\subseteq\{|\widehat{\sigma}_{s-1}^2|\le\sigma^2+\Delta\}. \tag{EC.3.40}$$

Plugging (EC.3.39) and (EC.3.40) into the definition of $W_{2,j}$ in (EC.3.34), we have that conditional on $\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1}\cap\mathcal{D}_{s-1}$,

$$\begin{aligned}|W_{2,j}|&\le\sqrt{\sigma^2+\Delta}\cdot\sqrt{2\theta\log(s-\mathtt{T}_1)}\cdot\sqrt{\frac{2p\bar{v}^2\mathring{\kappa}}{s-1-\mathtt{T}_1}}\\&\le 2\sqrt{\theta p\bar{v}^2\mathring{\kappa}(\sigma^2+\Delta)}\sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}} &\text{(EC.3.41)}\\&\le 2\sqrt{\theta p\bar{v}^2\mathring{\kappa}(\sigma^2+\Delta)}C_{\mathtt{T}_2}^{-\frac{1}{2}}, &\text{(EC.3.42)}\end{aligned}$$

where the last inequality holds because $s-1-\mathtt{T}_1\ge\mathtt{T}_2-\mathtt{T}_1=C_{\mathtt{T}_2}\log(T)$.

Plugging (EC.3.36) and (EC.3.42) into (EC.3.34) yields that conditional on $\mathcal{A}_{s-1}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1}\cap\mathcal{D}_{s-1}$,

$$|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le|W_{1,j}|+|W_{2,j}|\le 2\sqrt{p}\bar{v}\left(C_{\mathcal{A}}+\sqrt{\theta\mathring{\kappa}(\sigma^2+\Delta)}\right)C_{\mathtt{T}_2}^{-\frac{1}{2}}\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}},$$

where the second inequality holds because of the condition (EC.3.29). Therefore,

$$\begin{aligned}&\mathcal{A}_{s-1}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1}\cap\mathcal{D}_{s-1}\cap\{a_s=j\}\\\subseteq&\left\{|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}}\right\},\quad\forall s=\mathtt{T}_2+1,\dots,T,\,\forall j\ne a_s^*.\end{aligned} \tag{EC.3.43}$$

Consequently, combining (EC.3.33) and (EC.3.43) yields that

$$\begin{aligned}&\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{s-1}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1}\cap\mathcal{D}_{s-1}\cap\{a_s=j\}\\\subseteq&\left\{|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}}\right\},\quad\forall s=\mathtt{T}_1+1,\dots,T,\,\forall j\ne a_s^*.\end{aligned}$$

Now, let us return the task of bounding I_1, which is defined in (EC.3.30). By (EC.3.32), conditional on $\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_{\mathtt{T}_2:t-1}\cap\mathcal{C}_{\mathtt{T}_2:t-1}\cap\mathcal{D}_{\mathtt{T}_2:t-1}$, we have

$$\|\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:t}-\widehat{\mathbf{\Sigma}}^*_{vz,\mathtt{T}_1:t}\|\le\frac{2\sqrt{pq}\bar{v}\bar{z}}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\mathbb{I}(a_s\ne a_s^*)\le\frac{2\sqrt{pq}\bar{v}\bar{z}}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\sum_{j\ne a_s^*}\xi_{j,s},$$

for all $t=\mathtt{T}_2+1,\ldots,T$, where $\xi_{j,s}\coloneqq\mathbb{I}\big(|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le\frac{\eta}{8ML\sqrt{pq\bar{v}\bar{z}}}\big)$. Hence,

$$\begin{aligned}
&\Pr\Big(\Big\{\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}-\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}\|\ge\frac{\eta}{2}\Big\}\cap\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_{\mathtt{T}_2:t-1}\cap\mathcal{C}_{\mathtt{T}_2:t-1}\cap\mathcal{D}_{\mathtt{T}_2:t-1}\Big)\\
&\le\Pr\bigg(\frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\sum_{j\neq a_s^*}\xi_{j,s}\ge\frac{\eta}{4\sqrt{pq}\bar{v}\bar{z}}\bigg)\\
&\le\Pr\bigg(\bigg|\frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\sum_{j\neq a_s^*}(\xi_{j,s}-\mathbb{E}[\xi_{j,s}])\bigg|\ge\frac{\eta}{4\sqrt{pq}\bar{v}\bar{z}}-\sum_{j\neq a_s^*}\mathbb{E}[\xi_{j,s}]\bigg)\\
&\le\Pr\bigg(\bigg|\frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}\sum_{j\neq a_s^*}(\xi_{j,s}-\mathbb{E}[\xi_{j,s}])\bigg|\ge\frac{\eta}{8\sqrt{pq}\bar{v}\bar{z}}\bigg)
\end{aligned}\tag{EC.3.44}$$

where the last step holds because

$$\sum_{j\neq a_s^*}\mathbb{E}[\xi_{j,s}]=\sum_{j\neq a_s^*}\Pr\bigg(|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\le\frac{\eta}{8ML\sqrt{pq}\bar{v}\bar{z}}\bigg)\le\sum_{j\neq a_s^*}\frac{\eta}{8M\sqrt{pq}\bar{v}\bar{z}}\le\frac{\eta}{8\sqrt{pq}\bar{v}\bar{z}},$$

by the margin condition in Assumption 1.

Note that $\{\sum_{j\neq a_s^*}\xi_{j,s}:s=\mathtt{T}_1+1,\ldots,T\}$ are i.i.d. bounded random variables taking values on $[0,M-1]$. Thus, they are sub-Gaussian with variance proxy $\frac{M-1}{4}$; see, e.g., Wainwright (2019, page 24). Then, setting $\zeta_6\coloneqq\frac{\eta}{8\sqrt{pq}\bar{v}\bar{z}}$ and applying Lemma EC.5 leads to

$$\begin{aligned}
\Pr\bigg(\bigg|\frac{1}{t-\mathtt{T}_1}\sum_{s=\mathtt{T}_1+1}^{t}(\xi_s-\mathbb{E}[\xi_s])\bigg|\ge\zeta_6\bigg)&\le2\exp\bigg(\frac{-2\zeta_6^2(t-\mathtt{T}_1)}{(M-1)^2}\bigg)\\
&\le2\exp\bigg(\frac{-\eta^2C_{\mathtt{T}_2}\log(T)}{32(M-1)^2pq\bar{v}^2\bar{z}^2}\bigg)\le2T^{-4},
\end{aligned}\tag{EC.3.45}$$

for all $t=\mathtt{T}_2+1,\ldots,T$, where the last inequality follows from the condition (EC.3.29). It then follows from (EC.3.44) that

$$\begin{aligned}
&\Pr\Big(\Big\{\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}-\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}\|\ge\frac{\eta}{2}\Big\}\cap\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_{\mathtt{T}_2:t-1}\cap\mathcal{C}_{\mathtt{T}_2:t-1}\cap\mathcal{D}_{\mathtt{T}_2:t-1}\Big)\\
&\le2T^{-4},\quad\forall t=\mathtt{T}_2+1,\ldots,T.
\end{aligned}\tag{EC.3.46}$$

Combining (EC.3.30), (EC.3.31), and (EC.3.46) results in

$$\begin{aligned}
&\Pr\Big(\Big\{\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}-\boldsymbol{\Sigma}^*_{vz}\|\ge\eta\Big\}\cap\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_{\mathtt{T}_2:t-1}\cap\mathcal{C}_{\mathtt{T}_2:t-1}\cap\mathcal{D}_{\mathtt{T}_2:t-1}\Big)\\
&\le\Pr\Big(\Big\{\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:t}-\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}\|\ge\frac{\eta}{2}\Big\}\cap\mathcal{A}_{\mathtt{T}_1}\cap\mathcal{A}_{\mathtt{T}_2:t-1}\cap\mathcal{B}_{\mathtt{T}_2:t-1}\cap\mathcal{C}_{\mathtt{T}_2:t-1}\cap\mathcal{D}_{\mathtt{T}_2:t-1}\Big)\\
&\quad+\Pr\Big(\|\widehat{\boldsymbol{\Sigma}}^*_{vz,\mathtt{T}_1:t}-\boldsymbol{\Sigma}^*_{vz}\|>\frac{\eta}{2}\Big)\\
&\le2(Mp+q)T^{-2}+2T^{-4}\le2(Mp+q+1)T^{-2},
\end{aligned}$$

for all $t=\mathtt{T}_2+1,\ldots,T$, which concludes the proof. □

We now summarize the conditions of Lemmas EC.8–EC.12 as follows.

ASSUMPTION EC.1. *The constants $C_{\mathtt{T}_1}$ and $C_{\mathtt{T}_2}$ satisfy*

$$C_{\mathtt{T}_1} \geq \max\left\{ \left(\frac{40M\max\{\sqrt{pq}\bar{v}\bar{z},\, q\bar{z}^2\}}{\min\{\phi_{\min}(\breve{\mathbf{\Sigma}}_{vz}), \phi_{\min}(\breve{\mathbf{\Sigma}}_{zz})\}} \right)^2,\ 1024M^4p^2q^3\bar{v}^4\bar{z}^4\varsigma^2\breve{\kappa}^2L^2\eta^{-2} \right\},$$

$$C_{\mathtt{T}_2} \geq \max\left\{ \frac{3}{\mathsf{C}_{\mathsf{Bern}}\min\left\{\frac{\Delta^2}{4K_\epsilon^2}, \frac{\Delta}{2K_\epsilon}\right\}},\ \frac{6Mpq^2\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2}{\min\left\{\frac{\Delta^2}{128\sigma^2}, \frac{\Delta}{64}\right\}},\ 6q^2\bar{z}^4\eta^{-2},\right.$$
$$\left. 128(M-1)^2pq\bar{v}^2\bar{z}^2\eta^{-2},\ 256M^2L^2p^2q\bar{v}^4\bar{z}^2\left(C_{\mathcal{A}} + \sqrt{\theta\mathring{\kappa}(\sigma^2+\Delta)}\right)^2\eta^{-2} \right\},$$

where $\breve{\kappa}$, $\bar{\kappa}$, and $\mathring{\kappa}$ are defined in (EC.2.1), (EC.3.3), *and* (EC.3.27), *respectively; $\mathsf{C}_{\mathsf{Bern}}$ and K_ϵ are constants defined in Lemma EC.11, $C_{\mathcal{A}} \geq 2q\bar{z}\varsigma\bar{\kappa}$, and $\eta \leq \frac{1}{10}\min\{\phi_{\min}(\mathbf{\Sigma}^*_{vz}),\, \phi_{\min}(\mathbf{\Sigma}_{zz})\}$.*

LEMMA EC.13. *Suppose $C_{\mathtt{T}_1}$ and $C_{\mathtt{T}_2}$ satisfy Assumption EC.1. Then, for all $t = \mathtt{T}_2+1, \ldots, T$,*

$$\Pr(\mathcal{A}_{\mathtt{T}_2:t}) \geq 1 - [6Mp + (10M+10)q + 3]T^{-1},$$
$$\Pr(\mathcal{B}_{\mathtt{T}_2:t}) \geq 1 - [6Mp + (10M+10)q + 3]T^{-1}.$$

Proof. We first summarize the results from Lemmas EC.8–EC.12: for all $t = \mathtt{T}_2+1, \ldots, T$,

$$\Pr(\mathcal{A}_{\mathtt{T}_1}) \geq 1 - 2M(2p+5q)T^{-2}, \tag{EC.3.47}$$
$$\Pr(A_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t) \leq 2qT^{-2}, \tag{EC.3.48}$$
$$\Pr(\mathcal{C}_t) \geq 1 - 2qT^{-3}, \tag{EC.3.49}$$
$$\Pr(\mathcal{C}_{\mathtt{T}_2:t}) \geq 1 - 2q(t-\mathtt{T}_2)T^{-3} \geq 1 - 2qT^{-2} \tag{EC.3.50}$$
$$\Pr(\mathcal{B}_t \cap \mathcal{C}_t \cap \mathcal{D}_t^{\mathsf{c}}) \leq (2q+2)T^{-3}, \tag{EC.3.51}$$
$$\Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_1} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1} \cap \mathcal{D}_{\mathtt{T}_2:t-1}) \leq 2(Mp+q+1)T^{-2}. \tag{EC.3.52}$$

By (EC.3.48) and (EC.3.49), we have

$$\begin{aligned}\Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t) &= \Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t) + \Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t^{\mathsf{c}}) \\ &\leq \Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t \cap \mathcal{C}_t) + \Pr(\mathcal{C}_t^{\mathsf{c}}) \leq 2q(T^{-2}+T^{-3}) \leq 4qT^{-2}. \end{aligned} \tag{EC.3.53}$$

By applying (EC.3.51) for $s = \mathtt{T}_2+1, \ldots, t$, we have

$$\begin{aligned}&\Pr\big(\mathcal{B}_{\mathtt{T}_2:t} \cap \mathcal{C}_{\mathtt{T}_2:t} \cap \mathcal{D}_{\mathtt{T}_2:t}^{\mathsf{c}}\big) = \Pr\big(\mathcal{B}_{\mathtt{T}_2:t} \cap \mathcal{C}_{\mathtt{T}_2:t} \cap (\cup_{s=\mathtt{T}_2+1}^t \mathcal{D}_s^{\mathsf{c}})\big) \\ \leq &\sum_{s=\mathtt{T}_2+1}^t \Pr(\mathcal{B}_{\mathtt{T}_2:t} \cap \mathcal{C}_{\mathtt{T}_2:t} \cap \mathcal{D}_s^{\mathsf{c}}) \leq \sum_{s=\mathtt{T}_2+1}^t \Pr(\mathcal{B}_s \cap \mathcal{C}_s \cap \mathcal{D}_s^{\mathsf{c}})\end{aligned}$$

$$\leq (t-\mathtt{T}_2)(2q+2)T^{-3} \leq (2q+2)T^{-2}. \tag{EC.3.54}$$

By (EC.3.47) and (EC.3.52), we have

$$\begin{aligned}
&\Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1} \cap \mathcal{D}_{\mathtt{T}_2:t-1}) \\
\leq\ &\Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_1} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1} \cap \mathcal{D}_{\mathtt{T}_2:t-1}) + \Pr\big(\mathcal{A}_{\mathtt{T}_1}^{\mathsf{c}}\big) \\
\leq\ &2(Mp+q+1)T^{-2} + 2M(2p+5q)T^{-2} = [6Mp+(10M+2)q+1]T^{-2}.
\end{aligned} \tag{EC.3.55}$$

Define $\mathcal{G}_t \coloneqq \mathcal{A}_{\mathtt{T}_2:t} \cap \mathcal{B}_{\mathtt{T}_2:t}$ for all $t = T_{\mathtt{T}_2}+1, \ldots, T$. Let $\Pr(\mathcal{G}_{\mathtt{T}_2}) = 1$ for convenience. Moreover, note that

$$\begin{aligned}
&\Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{G}_{t-1}) = \Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1}) \\
=\ &\Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1}) + \Pr\big(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1}^{\mathsf{c}}\big) \\
\leq\ &\Pr(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1} \cap \mathcal{D}_{\mathtt{T}_2:t-1}) \\
&+ \Pr\big(\mathcal{B}_t^{\mathsf{c}} \cap \mathcal{A}_{\mathtt{T}_2:t-1} \cap \mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1} \cap \mathcal{D}_{\mathtt{T}_2:t-1}^{\mathsf{c}}\big) + \Pr\big(\mathcal{C}_{\mathtt{T}_2:t-1}^{\mathsf{c}}\big) \\
\leq\ &[6Mp+(10M+2)q+1]T^{-2} + \Pr\big(\mathcal{B}_{\mathtt{T}_2:t-1} \cap \mathcal{C}_{\mathtt{T}_2:t-1} \cap \mathcal{D}_{\mathtt{T}_2:t-1}^{\mathsf{c}}\big) + 2qT^{-2} \\
\leq\ &[6Mp+(10M+6)q+3]T^{-2},
\end{aligned} \tag{EC.3.56}$$

where the second inequality follows from (EC.3.50) and (EC.3.55), and the third from (EC.3.54). Therefore,

$$\begin{aligned}
\Pr(\mathcal{G}_t) &= \Pr(\mathcal{G}_{t-1} \cap \mathcal{A}_t \cap \mathcal{B}_t) = \Pr(\mathcal{G}_{t-1}) - \Pr(\mathcal{G}_{t-1} \cap (\mathcal{A}_t^{\mathsf{c}} \cup \mathcal{B}_t^{\mathsf{c}})) \\
&= \Pr(\mathcal{G}_{t-1}) - \Pr(\mathcal{G}_{t-1} \cap \mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t) - \Pr(\mathcal{G}_{t-1} \cap \mathcal{B}_t^{\mathsf{c}}) \\
&\geq \Pr(\mathcal{G}_{t-1}) - \Pr(\mathcal{A}_t^{\mathsf{c}} \cap \mathcal{B}_t) - \Pr(\mathcal{G}_{t-1} \cap \mathcal{B}_t^{\mathsf{c}}) \\
&\geq \Pr(\mathcal{G}_{t-1}) - [10Mp+(10M+10)q+3]T^{-2},
\end{aligned} \tag{EC.3.57}$$

where the last inequality follows from (EC.3.53) and (EC.3.56).

Note that $\Pr(\mathcal{G}_{\mathtt{T}_2}) = 1$, applying (EC.3.57) iteratively, we have that for all $t > \mathtt{T}_2$,

$$\Pr(\mathcal{G}_t) \geq 1 - (t-\mathtt{T}_2)[10Mp+(10M+10)q+3]T^{-2} \geq 1 - [10Mp+(10M+10)q+3]T^{-1}.$$

The proof is concluded by noting that $\Pr(\mathcal{A}_{\mathtt{T}_2:t}) \geq \Pr(\mathcal{G}_t)$ and $\Pr(\mathcal{B}_{\mathtt{T}_2:t}) \geq \Pr(\mathcal{G}_t)$. $\square$

EC.4. Regret Analysis for IV-Greedy

We state a more specific version of Theorem 1 as follows.

THEOREM EC.1. *Let* $\mathtt{T}_1 = C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2 = (C_{\mathtt{T}_1} + C_{\mathtt{T}_2})\log(T)$ *for some constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$ *that satisfy Assumption EC.1. Let* $\bar{\delta} \coloneqq \max_{1\le i\ne j\le M}\|\boldsymbol{\delta}_{i,j}\|$, *where* $\boldsymbol{\delta}_{i,j} \coloneqq \boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j$. *Then, under Assumption 1, the regret of IV-Greedy satisfies that, for all* $T > \mathtt{T}_2$,

$$\begin{aligned}\mathsf{Regret}(T) \le{}& [\sqrt{p}\bar{v}\bar{\delta}(C_{\mathtt{T}_1} + C_{\mathtt{T}_2}) + 16(M-1)Lpq^3\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2]\log(T)\\ &+ \sqrt{p}\bar{v}\bar{\delta}(M-1)[12Mp + (20M+24)q + 7].\end{aligned}$$

Proof of Theorem EC.1. Note that for all $s \ge 1$,

$$R_s^{\pi^*} - R_s^{\pi} = \mathbf{v}_s^\intercal\boldsymbol{\alpha}_{a_s^*} - \mathbf{v}_s^\intercal\boldsymbol{\alpha}_{a_s} = \sum_{j\ne a_s^*}\mathbb{I}(a_s = j)|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|. \tag{EC.4.1}$$

We first focus on the case that $s = \mathtt{T}_2 + 2, \dots, T$. By applying (EC.3.34) in the online Supplemental Material with $\theta = 0$ (which corresponds to IV-Greedy), we know that for any $j \ne a_s^*$, conditional on the event $\{a_s = j\}$,

$$\begin{aligned}|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}| \le |\mathbf{v}_s^\intercal(\boldsymbol{\delta}_{j,a_s^*} - \widehat{\boldsymbol{\delta}}_{j,a_s^*,s-1})| &= |\mathbf{v}_s^\intercal[(\boldsymbol{\alpha}_j - \widehat{\boldsymbol{\alpha}}_{j,s-1}) - (\boldsymbol{\alpha}_{a_s^*} - \widehat{\boldsymbol{\alpha}}_{a_s^*,s-1})]|\\ &\le 2\|\mathbf{v}_s\|\|\widehat{\boldsymbol{\alpha}}_{s-1} - \boldsymbol{\alpha}\| \le 2\sqrt{p}\bar{v}\|\widehat{\boldsymbol{\alpha}}_{s-1} - \boldsymbol{\alpha}\|.\end{aligned} \tag{EC.4.2}$$

Further, recall from (EC.3.1) that

$$\widehat{\boldsymbol{\alpha}}_{s-1} - \boldsymbol{\alpha} = \Big(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}^\intercal\Big)^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1}^{-1}\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1},$$

and from (EC.3.12) in the online Supplemental Material that

$$\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1} \subseteq \Big\{\Big\|\Big(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1}^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}^\intercal\Big)^{-1}\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:s-1}\widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:s-1}^{-1}\Big\| \le \bar{\kappa}\Big\}.$$

It then follows from (EC.4.2) that for any $j \ne a_s^*$,

$$\{a_s = j\}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1} \subseteq \Big\{|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}| \le 2\sqrt{p}\bar{v}\bar{\kappa}\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big\}. \tag{EC.4.3}$$

Hence, for any $j \ne a_s^*$,

$$\begin{aligned}&\mathbb{E}[\mathbb{I}(a_s = j)|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|]\\ &= \mathbb{E}[(\mathbb{I}(\{a_s = j\}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1})|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|] + \mathbb{E}[\mathbb{I}(\{a_s = j\}\cap(\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1})^{\mathsf{c}})|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|]\\ &\le \mathbb{E}[(\mathbb{I}(\{a_s = j\}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1})|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|] + \mathbb{E}[\mathbb{I}((\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1})^{\mathsf{c}})|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|]\\ &\le \mathbb{E}\Big[\mathbb{I}\Big(|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}| \le 2\sqrt{p}\bar{v}\bar{\kappa}\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big)|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\Big] + 2\sqrt{p}\bar{v}\|\boldsymbol{\delta}_{j,a_s^*}\|\Pr\big(\mathcal{B}_{s-1}^{\mathsf{c}}\cup\mathcal{C}_{s-1}^{\mathsf{c}}\big)\end{aligned}$$

$$\leq 2\sqrt{p}\bar{v}\bar{\kappa}\underbrace{\mathbb{E}\Big[\mathbb{I}\Big(|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\leq 2\sqrt{p}\bar{v}\bar{\kappa}\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big)\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big]}_{U_{1,j}}+2\sqrt{p}\bar{v}\bar{\delta}\underbrace{\Pr\big(\mathcal{B}_{s-1}^{\mathsf{c}}\cup\mathcal{C}_{s-1}^{\mathsf{c}}\big)}_{U_2},$$

where the second inequality follows from (EC.4.3), and that $|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\leq\|\mathbf{v}_s\|\|\boldsymbol{\delta}_{j,a_s^*}\|\leq 2\sqrt{p}\bar{v}\bar{\delta}$. By applying (EC.4.1), we have

$$\mathbb{E}[R_s^{\pi^*}-R_s^{\pi}]=\sum_{j\neq a_s^*}\mathbb{E}[\mathbb{I}(a_s=j)|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|]\leq\sum_{j\neq a_s^*}(U_{1,j}+U_2). \tag{EC.4.4}$$

For U_2, we may apply Lemmas EC.10 and EC.13 to conclude that

$$\begin{aligned}U_2&\leq\Pr\big(\mathcal{B}_{s-1}^{c}\big)+\Pr\big(\mathcal{C}_{s-1}^{\mathsf{c}}\big)\leq\Pr\big(\mathcal{B}_{\mathtt{T}_2:s-1}^{\mathsf{c}}\big)+\Pr\big(\mathcal{C}_{s-1}^{\mathsf{c}}\big)\\&\leq[6Mp+(10M+10)q+3]T^{-1}+2qT^{-3}\leq[6Mp+(10M+12)q+3]T^{-1}.\end{aligned} \tag{EC.4.5}$$

For $U_{1,j}$, we have

$$\begin{aligned}U_{1,j}&=\mathbb{E}\bigg[\mathbb{E}\Big[\mathbb{I}\Big(|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\leq 2\sqrt{p}\bar{v}\bar{\kappa}\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big)\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big]\,\Big|\,\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\bigg]\\&\leq\mathbb{E}\bigg[\Pr\Big(|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}|\leq 2\sqrt{p}\bar{v}\bar{\kappa}\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\,\Big|\,\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\Big)\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\bigg]\\&\leq\mathbb{E}\Big[2L\sqrt{p}\bar{v}\bar{\kappa}\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|^2\Big],\end{aligned} \tag{EC.4.6}$$

where the second inequality follows from the margin condition of Assumption 1,

Let $\mathsf{p}(w)$ denote the probability density function of $\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|$. Then,

$$\begin{aligned}\mathbb{E}\Big[\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|^2\Big]&=\int_0^\infty w^2\mathsf{p}(w)\,\mathrm{d}w=\int_0^\infty\mathsf{p}(w)\int_0^w 2\tau\,\mathrm{d}\tau\,\mathrm{d}w\\&=\int_0^\infty 2\tau\int_\tau^\infty\mathsf{p}(w)\,\mathrm{d}w\,\mathrm{d}\tau=\int_0^\infty 2\tau\Pr\Big(\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\geq\tau\Big)\,\mathrm{d}\tau\,,\end{aligned} \tag{EC.4.7}$$

where the third equality follows from Fubini's theorem. We then apply (EC.3.13):

$$\int_0^\infty 2\tau\Pr\Big(\|\widehat{\mathbf{G}}_{\mathtt{T}_1:s-1}\|\geq\tau\Big)\,\mathrm{d}\tau\leq\int_0^\infty 4q\tau\exp\bigg(-\frac{\tau^2(s-\mathtt{T}_2-1)}{2q^2\bar{z}^2\varsigma^2}\bigg)\,\mathrm{d}\tau=\frac{4q^3\bar{z}^2\varsigma^2}{s-\mathtt{T}_2-1}. \tag{EC.4.8}$$

Combining (EC.4.6), (EC.4.7), and (EC.4.8), we have that for any $j\neq a_s^*$,

$$U_{1,j}\leq\frac{8L\sqrt{p}q^3\bar{v}\bar{z}^2\varsigma^2\bar{\kappa}}{s-\mathtt{T}_2-1}. \tag{EC.4.9}$$

Plugging in (EC.4.5) and (EC.4.9) into (EC.4.4), we have

$$\mathbb{E}[R_s^{\pi^*}-R_s^{\pi}]\leq(M-1)\bigg[\frac{16Lpq^3\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2}{s-\mathtt{T}_2-1}+2\sqrt{p}\bar{v}\bar{\delta}[6Mp+(10M+12)q+3]T^{-1}\bigg]. \tag{EC.4.10}$$

Note that $\sum_{s=\mathtt{T}_2+2}^{T}\frac{1}{s-\mathtt{T}_2-1}\leq\int_1^T\frac{\mathrm{d}u}{u}=\log(T)$. Then, summing (EC.4.10) over $s=\mathtt{T}_2+2,\ldots,T$ yields

$$\sum_{s=\mathtt{T}_2+2}^{T}\mathbb{E}[R_s^{\pi^*}-R_s^{\pi}]$$

$$\leq (M-1)\big[16Lpq^3\bar{v}^2\bar{z}^2\varsigma^2\bar{\kappa}^2\log(T)+2\sqrt{p}\bar{v}\bar{\delta}[6Mp+(10M+12)q+3]\big]. \tag{EC.4.11}$$

Moreover, for the first $\mathtt{T}_2+1$ periods, we can simply take advantage of the fact that each reward is bounded and $\mathtt{T}_2$ is of order $\mathcal{O}(\log(T))$. Specifically, by (EC.4.1),

$$\begin{aligned}\sum_{s=1}^{\mathtt{T}_2+1}\mathbb{E}[R_s^{\pi^*}-R_s^{\pi}] &= \sum_{s=1}^{\mathtt{T}_2+1}\mathbb{E}[\mathbb{I}(a_s\neq a_s^*)|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{a_s,a_s^*}|] \leq \sum_{s=1}^{\mathtt{T}_2+1}\mathbb{E}[\|\mathbf{v}_s\|\|\boldsymbol{\delta}_{a_s,a_s^*}\|]\\ &\leq \sqrt{p}\bar{v}\bar{\delta}(\mathtt{T}_2+1)=\sqrt{p}\bar{v}\bar{\delta}\big[(C_{\mathtt{T}_1}+C_{\mathtt{T}_2})\log(T)+1\big].\end{aligned} \tag{EC.4.12}$$

Combining (EC.4.11) and (EC.4.12) completes the proof. □

EC.5. Regret Analysis for IV-UCB

We state a more specific version of Theorem 2 as follows.

THEOREM EC.2. *Let* $\mathtt{T}_1=C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2=(C_{\mathtt{T}_1}+C_{\mathtt{T}_2})\log(T)$ *for some constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$ *that satisfy Assumption EC.1. Let* $\bar{\delta}:=\max_{1\leq i\neq j\leq M}\|\boldsymbol{\delta}_{i,j}\|$, *where* $\boldsymbol{\delta}_{i,j}:=\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j$. *Then, under Assumption 1, the regret of IV-UCB satisfies that, for all* $T>\mathtt{T}_2$,

$$\begin{aligned}\mathsf{Regret}(T) \leq\ & 4Lp\bar{v}^2(M-1)\left(C_{\mathcal{A}}+\sqrt{\theta\mathring{\kappa}(\sigma^2+\Delta)}\right)^2(\log(T))^2+\sqrt{p}\bar{v}\bar{\delta}(C_{\mathtt{T}_1}+C_{\mathtt{T}_2})\log(T)\\ &+\sqrt{p}\bar{v}\bar{\delta}\big[(M-1)[24Mp+(40M+48)q+16]+1\big].\end{aligned}$$

Proof of Theorem EC.2. We first focus on the case that $s=\mathtt{T}_2+2,\ldots,T$. By applying (EC.3.34) with $\theta>0$ (which corresponds to the IV-UCB algorithm), we know that for any $j\neq a_s^*$, conditional on the event $\{a_s=j\}$,

$$\begin{aligned}\|\mathbf{v}_s^\intercal\boldsymbol{\delta}_{j,a_s^*}\| \leq\ & |\underbrace{\mathbf{v}_s^\intercal(\boldsymbol{\delta}_{j,a_s^*}-\widehat{\boldsymbol{\delta}}_{j,a_s^*,s-1})}_{W_{1,j}}|\\ &+\underbrace{\widehat{\sigma}_{s-1}\sqrt{2\theta\log(s-\mathtt{T}_1)}\left(\sqrt{\mathbf{v}_s^\intercal\widehat{\boldsymbol{\Omega}}_{j,s-1}\mathbf{v}_s}-\sqrt{\mathbf{v}_s^\intercal\widehat{\boldsymbol{\Omega}}_{a_s^*,s-1}\mathbf{v}_s}\right)}_{W_{2,j}}.\end{aligned} \tag{EC.5.1}$$

Further, recall from (EC.3.35) that

$$\{a_s=j\}\cap\mathcal{A}_{s-1}\subseteq\left\{|W_{1,j}|\leq 2\sqrt{p}\bar{v}C_{\mathcal{A}}\sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}}\right\}; \tag{EC.5.2}$$

recall also from (EC.3.41) that

$$\{a_s=j\}\cap\mathcal{B}_{s-1}\cap\mathcal{C}_{s-1}\cap\mathcal{D}_{s-1}\subseteq\left\{|W_{2,j}|\leq 2\sqrt{\theta p\bar{v}^2\mathring{\kappa}(\sigma^2+\Delta)}\sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}}\right\}. \tag{EC.5.3}$$

It follows from (EC.5.1), (EC.5.2), and (EC.5.3) that for any $j \neq a_s^*$,

$$\begin{aligned} &\{a_s = j\} \cap \mathcal{A}_{s-1} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1} \cap \mathcal{D}_{s-1} \\ \subseteq &\left\{ |\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}| \leq \ddot{\kappa} \sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}} \right\} \coloneqq \mathcal{H}_{j,s}, \end{aligned} \tag{EC.5.4}$$

where $\ddot{\kappa} \coloneqq 2\sqrt{p}\bar{v}\left(C_{\mathcal{A}} + \sqrt{\theta \mathring{\kappa}(\sigma^2 + \Delta)}\right)$. By the margin condition in Assumption 1,

$$\Pr(\mathcal{H}_{j,s}) \leq L\ddot{\kappa} \sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}}. \tag{EC.5.5}$$

Hence,

$$\begin{aligned} &\mathbb{E}[\mathbb{I}(a_s = j)|\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}|] &\text{(EC.5.6)} \\ =&\, \mathbb{E}[(\mathbb{I}(\{a_s = j\} \cap \mathcal{A}_{s-1} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1} \cap \mathcal{D}_{s-1})|\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}|] \\ &+ \mathbb{E}[\mathbb{I}(\{a_s = j\} \cap (\mathcal{A}_{s-1} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1} \cap \mathcal{D}_{s-1})^{\mathsf{c}})|\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}|] \\ \leq&\, \mathbb{E}[(\mathbb{I}(\mathcal{H}_{j,s})|\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}|] + 2\sqrt{p}\bar{v}\bar{\delta} \Pr((\mathcal{A}_{s-1} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1} \cap \mathcal{D}_{s-1})^{\mathsf{c}}), &\text{(EC.5.7)} \end{aligned}$$

where the inequality follows from (EC.5.4).

For the first term on the right-hand-side of (EC.5.7), we have

$$\mathbb{E}[(\mathbb{I}(\mathcal{H}_{j,s})|\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}|] \leq \ddot{\kappa} \sqrt{\frac{\log(T)}{s-1-\mathtt{T}_1}} \cdot \Pr(\mathcal{H}_{j,s}) \leq \frac{L\ddot{\kappa}^2 \log(T)}{s-1-\mathtt{T}_1}, \tag{EC.5.8}$$

where the two inequalities follow from (EC.5.4) and (EC.5.5), respectively.

For the second term on the right-hand-side of (EC.5.7), we let $\mathcal{I}_{s-1} \coloneqq \mathcal{A}_{s-1} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1}$ and note that

$$\begin{aligned} (\mathcal{I}_{s-1} \cap \mathcal{D}_{s-1})^{\mathsf{c}} &= \mathcal{I}_{s-1}^{\mathsf{c}} \cup \mathcal{D}_{s-1}^{\mathsf{c}} \\ &= \mathcal{I}_{s-1}^{\mathsf{c}} \cup \left[(\mathcal{D}_{s-1}^{\mathsf{c}} \cap \mathcal{I}_{s-1}^{\mathsf{c}}) \cup (\mathcal{D}_{s-1}^{\mathsf{c}} \cap \mathcal{I}_{s-1}) \right] \\ &= \mathcal{I}_{s-1}^{\mathsf{c}} \cup (\mathcal{D}_{s-1}^{\mathsf{c}} \cap \mathcal{I}_{s-1}) \\ &\subseteq \mathcal{A}_{s-1}^{\mathsf{c}} \cup \mathcal{B}_{s-1}^{\mathsf{c}} \cup \mathcal{C}_{s-1}^{\mathsf{c}} \cup (\mathcal{D}_{s-1}^{\mathsf{c}} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1}). \end{aligned}$$

Thus,

$$\begin{aligned} &\Pr((\mathcal{A}_{s-1} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1} \cap \mathcal{D}_{s-1})^{\mathsf{c}}) \\ \leq&\, \Pr\big(\mathcal{A}_{s-1}^{\mathsf{c}}\big) + \Pr\big(\mathcal{B}_{s-1}^{\mathsf{c}}\big) + \Pr\big(\mathcal{C}_{s-1}^{\mathsf{c}}\big) + \Pr\big(\mathcal{D}_{s-1}^{\mathsf{c}} \cap \mathcal{B}_{s-1} \cap \mathcal{C}_{s-1}\big) \\ \leq&\, [6Mp + (10M+10)q + 3]T^{-1} + [6Mp + (10M+10)q + 3])T^{-1} \end{aligned}$$

$$+ 2qT^{-3} + (2q+2)T^{-3}$$
$$\le [12Mp + (20M+24)q + 8]T^{-1}, \tag{EC.5.9}$$

where the second inequality follows from Lemmas EC.10, EC.11, and EC.13.

Plugging (EC.5.8) and (EC.5.9) into (EC.5.7) and applying (EC.4.1), we have

$$\mathbb{E}[R_s^{\pi^*} - R_s^{\pi}] = \sum_{j \neq a_s^*} \mathbb{E}[\mathbb{I}(a_s = j)|\mathbf{v}_s^\intercal \boldsymbol{\delta}_{j,a_s^*}|]$$
$$\le (M-1)\left(\frac{L\ddot{\kappa}^2 \log(T)}{s-1-\mathtt{T}_1} + 2\sqrt{p}\bar{v}\bar{\delta}[12Mp + (20M+24)q + 8]T^{-1}\right). \tag{EC.5.10}$$

Note that $\sum_{s=\mathtt{T}_2+2}^{T} \frac{1}{s-\mathtt{T}_2-1} \le \int_1^T \frac{\mathrm{d}u}{u} = \log(T)$. Then, summing (EC.5.10) over $s = \mathtt{T}_2+2, \ldots, T$, we have

$$\sum_{s=\mathtt{T}_2+2}^{T} \mathbb{E}[R_s^{\pi^*} - R_s^{\pi}] \le (M-1)\left[L\ddot{\kappa}^2(\log(T))^2 + 2\sqrt{p}\bar{v}\bar{\delta}[12Mp + (20M+24)q + 8]\right].$$

The regret for the first $\mathtt{T}_2 + 1$ periods is calculated in (EC.4.12). Hence, the proof is completed by plugging the definition of $\ddot{\kappa}$ into the above inequality. □

EC.6. Proofs for Statistical Inference

We state a more specific version of Theorem 3 as follows.

THEOREM EC.3. *Let* $\mathtt{T}_1 = C_{\mathtt{T}_1}\log(T)$ *and* $\mathtt{T}_2 = (C_{\mathtt{T}_1} + C_{\mathtt{T}_2})\log(T)$ *for some constants* $C_{\mathtt{T}_1}$ *and* $C_{\mathtt{T}_2}$ *that satisfy Assumption EC.1. Then, under Assumption 1, the estimator* $\widehat{\boldsymbol{\alpha}}_T$ *of Algorithm 1 satisfies*

$$\sqrt{T-\mathtt{T}_1}(\widehat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}) \rightsquigarrow \mathcal{N}\Big(\mathbf{0}, \sigma^2\big(\boldsymbol{\Sigma}_{vz}^* \boldsymbol{\Sigma}_{zz}^{-1} (\boldsymbol{\Sigma}_{vz}^*)^\intercal\big)^{-1}\Big), \tag{EC.6.1}$$

as $T \to \infty$, *where* $\boldsymbol{\Sigma}_{vz}^* = \mathbb{E}[\mathbf{v}_t^* \mathbf{z}_t^\intercal]$ *and* $\boldsymbol{\Sigma}_{zz} = \mathbb{E}[\mathbf{z}_t \mathbf{z}_t^\intercal]$.

Proof of Theorem EC.3. By (EC.3.1), for all $T > \mathtt{T}_1$,

$$\widehat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha} = \Big(\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T} \widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:T}^{-1} \widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T}^\intercal\Big)^{-1} \widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T} \widehat{\boldsymbol{\Sigma}}_{zz,\mathtt{T}_1:T}^{-1} \widehat{\mathbf{G}}_{\mathtt{T}_1:T}.$$

It follows from the definition of $\widehat{\mathbf{G}}_{\mathtt{T}_1:T}$ and the central limit theorem that as $T \to \infty$,

$$\sqrt{T-\mathtt{T}_1}\widehat{\mathbf{G}}_{\mathtt{T}_1:T} = \frac{1}{\sqrt{T-\mathtt{T}_1}} \sum_{s=\mathtt{T}_1+1}^{T} \mathbf{z}_s \epsilon_s \rightsquigarrow \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{zz}).$$

By Lemmas EC.10 and EC.13, for arbitrarily small $\tau > 0$, for T large enough (dependent on τ), we have that:

$$\Pr\Big(\|\widehat{\boldsymbol{\Sigma}}_{vz,\mathtt{T}_1:T} - \boldsymbol{\Sigma}_{vz}^*\| \le \tau\Big) \ge 1 - [6Mp + (10M+10)q + 3]T^{-1},$$

$$\Pr\Big(\|\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:T}-\mathbf{\Sigma}_{zz}\|\le\tau\Big)\ge 1-2qT^{-3}.$$

Therefore, as $T\to\infty$,

$$\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:T}-\mathbf{\Sigma}^*_{vz}\to_p\mathbf{0}\quad\text{and}\quad\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:T}-\mathbf{\Sigma}_{zz}\to_p\mathbf{0}.$$

It then follows from the continuous mapping theorem that, as $T\to\infty$,

$$\begin{aligned}\sqrt{T-\mathtt{T}_1}(\widehat{\boldsymbol{\alpha}}_T-\boldsymbol{\alpha})&\rightsquigarrow\big(\mathbf{\Sigma}^*_{vz}\mathbf{\Sigma}_{zz}^{-1}(\mathbf{\Sigma}^*_{vz})^\intercal\big)^{-1}\mathbf{\Sigma}^*_{vz}\mathbf{\Sigma}_{zz}^{-1}\cdot\mathcal{N}(\mathbf{0},\sigma^2\mathbf{\Sigma}_{zz})\\&\overset{d}{=}\mathcal{N}\Big(\mathbf{0},\sigma^2\big(\mathbf{\Sigma}^*_{vz}\mathbf{\Sigma}_{zz}^{-1}(\mathbf{\Sigma}^*_{vz})^\intercal\big)^{-1}\Big).\quad\square\end{aligned}$$

Proof of Corollary 1. By Lemmas EC.10 and EC.13, for arbitrarily small $\tau>0$, for T large enough (dependent on τ), we have that:

$$\begin{aligned}&\Pr\Big(\|\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:T}-\mathbf{\Sigma}^*_{vz}\|\le\tau\Big)\ge 1-[6Mp+(10M+10)q+3]T^{-1},\\&\Pr\Big(\|\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:T}-\mathbf{\Sigma}_{zz}\|\le\tau\Big)\ge 1-2qT^{-3}.\end{aligned}$$

Therefore, as $T\to\infty$,

$$\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:T}-\mathbf{\Sigma}^*_{vz}\to_p\mathbf{0}\quad\text{and}\quad\widehat{\mathbf{\Sigma}}_{zz,\mathtt{T}_1:T}-\mathbf{\Sigma}_{zz}\to_p\mathbf{0}.$$

It implies that

$$\widehat{\mathbf{\Omega}}_T=\Big(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:T}\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:T}^{-1}(\widehat{\mathbf{\Sigma}}_{vz,\mathtt{T}_1:T})^\intercal\Big)^{-1}\to_p\mathbf{\Omega}^*.$$

Similarly, by Lemma EC.11, for arbitrarily small η and large enough T (dependent on η), we have that

$$\Pr(\mathcal{B}_T\cap\mathcal{C}_T\cap\mathcal{D}_T^\mathsf{c})\le(2q+2)T^{-3}.$$

By Lemma EC.10 and Lemma EC.13, we know that for T large enough,

$$\begin{aligned}\Pr(\mathcal{B}_T\cap\mathcal{C}_T)&\ge 1-\Pr(\mathcal{B}_T^\mathsf{c})-\Pr(\mathcal{C}_T^\mathsf{c})\\&\ge 1-[6Mp+(10M+10)q+3]T^{-1}-2qT^{-3}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Pr(\mathcal{D}_T^\mathsf{c})&=\Pr(\mathcal{B}_T\cap\mathcal{C}_T\cap\mathcal{D}_T^\mathsf{c})+\Pr((\mathcal{B}_T\cap\mathcal{C}_T)^\mathsf{c}\cap\mathcal{D}_T^\mathsf{c})\\&\le(2q+2)T^{-3}+[6Mp+(10M+10)q+3]T^{-1}-2qT^{-3}.\end{aligned}$$

Hence, $\Pr(\mathcal{D}_t)\to 1$ as $T\to\infty$. Since η can be arbitrarily small, it implies that $\widehat{\sigma}_T^2\to_p\sigma^2$. Combining all the above, we have that

$$\widehat{\sigma}_T^2\widehat{\mathbf{\Omega}}_T\to_p\sigma^2\mathbf{\Omega}^*.$$

By Theorem 3, we have that $\sqrt{T-T_1}(\widehat{\boldsymbol{\alpha}}_T-\boldsymbol{\alpha})\rightsquigarrow\mathcal{N}(0,\sigma^2\mathbf{\Omega}^*)$ as $T\to\infty$. Since $\widehat{\sigma}_T^2\widehat{\mathbf{\Omega}}_T$ is a consistent estimator of $\sigma^2\mathbf{\Omega}^*$, the statements in Corollary 1 hold. $\square$

References

Golub GH, Van Loan CF (2013) *Matrix Computations* (Johns Hopkins University Press), 4th edition.

Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge University Press).

Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press).

Wang BY, Xi BY (1997) Some inequalities for singular values of matrix products. *Linear Algebra and its Applications* 264:109–115.