Causal Bandits: Online Decision-Making in Endogenous Settings

Anonymous Author(s)
Affiliation
Address
email

Abstract

The deployment of Multi-Armed Bandits (MAB) has become commonplace in many economic applications. However, regret guarantees for even state-of-the-art linear bandit algorithms (such as Optimism in the Face of Uncertainty Linear bandit (OFUL)) make strong exogeneity assumptions w.r.t. arm covariates. This assumption is very often violated in many economic contexts and using such algorithms can lead to sub-optimal decisions. In this paper, we consider the problem of online learning in linear stochastic multi-armed bandit problems with endogenous covariates. We propose an algorithm we term *BanditIV*, that uses instrumental variables to correct for this bias, and prove an $\tilde{\mathcal{O}}(k\sqrt{T})$ upper bound for the expected regret of the algorithm. Further, in economic contexts, it is also important to understand how the model parameters behave asymptotically. To this end, we additionally propose ϵ-*BanditIV* algorithm and demonstrate its asymptotic consistency and normality while ensuring the same regret bound. Finally, we carry out extensive Monte Carlo simulations to demonstrate the performance of our algorithms compared to other methods. We show that *BanditIV* and ϵ-*BanditIV* significantly outperform other existing methods.

1 Introduction

The proliferation of user-level data presents at the same time a unique opportunity and challenge in front of decision-makers. Decision-makers want to use individual-level data to tailor their decisions for each user. Further, given the dynamic nature of the platform, decision-makers want to be able to adopt these decisions with incrementally available data. Woodroofe 1979 first proposed a simple model to solve such sequential decision-making problems with covariates. Langford and T. Zhang 2007 later named the model "contextual bandit". Contextual linear bandits have been adopted across a wide variety of applications from advertising (Tang et al. 2015, Aramayo, Schiappacasse, and Goic 2022) to healthcare (Durand et al. 2018), dialogue systems (Liu et al. 2018), and personalized product recommendations (L. Li et al. 2010, Qin, S. Chen, and Zhu 2014). In each round, the decision-maker observes a set of actions with each action characterized by a set of features. Decision-maker selects an action and observes a reward corresponding to that action. The objective of the decision-maker is to achieve cumulative reward close to that of optimal policy in hindsight.

Traditional formulations of contextual bandits make the unconfoundedness assumption i.e., the arm covariates are exogenous. However, in many economic settings, arm features can be correlated with the unobserved noise. For instance, consider the problem of generating product recommendations for consumers. Platform operators would usually run online experiments to uncover the relationship between product features and consumer demand. Generally in such settings, product observable features like price are controlled by product owners (different from the platform operator) which

Submitted to *A causal view on dynamical systems* workshop at NeurIPS 2022. Do not distribute.

could be set in anticipation of consumer demand and hence be correlated with the demand shocks unobserved by the platform operator. In such settings, traditional bandit algorithms might lead to sub-optimal decisions. A common approach to correct the endogeneity bias in the offline setting is to use instrumental variables. Instruments are correlated with the endogenous variable but are otherwise not associated with the outcome variable. The method of instrumental variables uses the variation in the exogenous component of the endogenous variable induced by the variation in the instrumental variable to make inference of causal effects. Thus, to address the issues induced by endogenous features in online settings, we propose an online estimator we term *BanditIV* that uses instrumental variables to learn the relationship between rewards and features.

Next, literature in contextual linear bandit has also primarily focused on minimizing the expected regret of the algorithm (Auer 2002, Dani, Hayes, and Kakade 2008, Chu et al. 2011, and Abbasi-Yadkori, Pál, and Szepesvári 2011). However, in a variety of practical situations, one may also be interested in the inference of model parameters, and algorithms that minimize regret may not guarantee the consistency of the model parameters. Further, knowing the asymptotic distributions, one can easily test the significance of features, and offers a traditional way to select variables. For instance, in the product recommendation example, an online platform might also be interested in understanding the effect of various marketing-mix variables (like price, and promotion) on consumer demand. To this end, we propose a linear bandit algorithm that pursues both regret minimization and consistent estimation of model parameters in endogenous settings. We finally use the martingale central limit theorem to show that our estimator of model parameters is asymptotically normal.

We finally compare our method with existing online algorithms including Optimism in the Face of Uncertainty Linear bandit (OFUL) and Thompson Sampling (TS) through numerical experiments. We find significant advantages of our algorithm in both expected regret and parameter inference.

To summarize, our paper makes the following contributions –

- We study the stochastic linear bandit problem with endogenous features. We propose a new algorithm, we term *BanditIV*, which incorporates instrumental variables to correct for the bias induced by endogenous features and show that the total expected regret of the algorithm is upper bounded by $\tilde{\mathcal{O}}(k\sqrt{T})$.
- Next, as in many economic contexts, researchers might not only be interested in minimizing the regret over the outcomes but also in conducting inference over estimated parameters. To this end, we propose the ϵ-*BanditIV* algorithm with the same upper bound and establish the asymptotic consistency and normality of the estimator.
- Finally, we conduct numerical experiments of *BanditIV* and ϵ-*BanditIV* algorithms on synthetic data. We show *BanditIV* and ϵ-*BanditIV* always outperform other existing linear bandit algorithms including OFUL and TS in terms of both total expected regret and inference.

The rest of the paper is organized as follows. We review related literature in Section 2. We set up the main problem, i.e. endogeneity problem in online stochastic linear bandits in Section 3. We formulate the *BanditIV* and ϵ-*BanditIV* algorithms in Section 4 and show regret bound in Section 4.1. We derive theoretical analysis on consistency and normality of the estimator used in the ϵ-*BanditIV* algorithm in Section 5. We conduct simulations of the proposed algorithms on synthetic data in Section 6.

Notation. Throughout this paper, we use $||\cdot||_p$ to denote the p-norm of a vector or a matrix. We use $||\cdot||_F$ to denote the Frobenius norm for a matrix. For a vector x and positive definite matrix A, we denote $\sqrt{x'Ax}$ as $||x||_A$. We use $\langle\cdot,\cdot\rangle$ as the inner product. We denote the sequence $a_0, a_1, \cdots, a_\infty$ as $\{a_t\}_{t=0}^{\infty}$. For a matrix X, the ith column is denoted as $X^{(i)}$ and the jth element of the ith column is denoted as $X^{(i)(j)}$.

2 Literature review

This paper focuses on Contextual Multi-Armed Bandits (CMAB) approach. Auer 2002 first introduced the contextual bandit setting through the Linear Reinforcement Learning (LinRel) algorithm with linear value functions. Subsequently, the contextual framework was improved by Dani, Hayes, and Kakade 2008 and Chu et al. 2011 through Upper Confidence Bound (UCB) type algorithms.

Abbasi-Yadkori, Pál, and Szepesvári 2011 further modified the analysis of the linear bandit problem and improved the regret bound by a logarithmic factor. More recently, Bastani, Bayati, and Khosravi 2021 studied performance of exploration-free policies in contextual bandit settings and proposed the Greedy-First algorithm which only utilizes observed contexts and rewards to decide whether to follow a greedy algorithm or to explore. This literature often requires independency between contexts and the random error term, which however cannot be always satisfied in the reality. We consider a setting which allows the dependency between contexts and error term. This more general setting is well suited to real-world applications where endogeneity problems happen.

This paper contributes to the stream of literature in inference with MAB. The adaptive nature of data generation in online settings complicates the inference of unknown parameters. Nie et al. 2018 showed that estimates of arm-specific expected rewards in UCB and TS algorithms are biased downwards. This downward bias is due to that arms with random upward fluctuation are sampled more, while those arms with downward fluctuation are sampled less (Hadad et al. 2021). Bibaut et al. 2021 presented that standard estimators no longer follow normal distribution asymptotically so that classic confidence interval fail to provide correct coverage. To address the inference problem in online settings, H. Chen, Lu, and Song 2021 studied the consistency and asymptotic distribution of the online ordinary least square estimator under epsilon-greedy policy. We extend their approach to two stage least square estimation in online settings and contribute to the literature by proving the consistency and asymptotic normality of our estimator.

This paper is also closely related to the stream of literature in instrumental variable methods for endogeneity problems. Griliches 1977, Hausman 1983, Angrist and G. Imbens 1995, and X. Chen, Hong, and Tamer 2005 provided theoretical analysis of instrumental variables in linear models, while Hansen 1982, Ai and X. Chen 2003, Newey and Powell 2003 and Chernozhukov, G. W. Imbens, and Newey 2007 studied instrumental variables in nonlinear models. G. Imbens 2014 reviewed work on instrumental variable methods and discussed applications and restrictions. Different from the standard data analysis framework in offline settings, the endogeneity problem can be exacerbated during the dynamic interaction between data generation and data analysis in online settings (J. Li, Luo, and X. Zhang 2021). This paper builds on existing instrumental variable methods in offline settings and considers the dynamic interaction by utilizing the CMAB framework. For these reasons, the paper contributes to the literature by combining both CMAB models and instrumental variable methods to improve the estimation of causal effect of treatments and hence achieve better performance of decision policies.

3 Problem setting

Let T be the number of rounds and K the number of arms in each round. In each round t, the learner observes K feature vectors $x_{t,a} \in \mathbb{R}^d$, $a \in \{1, \cdots, K\}$, with $||x_{t,a}||_2 \leq L_x$ for samples (one feature vector for each sample), where the subscript a represents the ath arm or sample from the K-sample set. Let $y_{t,a} \in \mathbb{R}$, s.t. $|y_{t,a}| \leq L_y$, be the reward of arm a on round t. After observing the feature vectors, the learner selects an arm a_t and receives reward $y_{t,a}$. Under linear realizability assumption, we have $\mathbb{E}[y_{t,a}|x_{t,a}] = x'_{t,a}\beta_0$ for all t and a, where $\beta_0 \in \mathbb{R}^d$ is an unknown true coefficient vector. We assume $y_{t,a}$ are independent random variables with expectation $x'_{t,a}\beta_0$, as the following.

$$y_{t,a} = \beta'_0 x_{t,a} + e_{t,a} \tag{1}$$

where $e_{t,a} \in \mathbb{R}$ is a 1-subgaussian error term, s.t. $\mathbb{E}[e_{t,a}] = 0$. Standard Contextual Linear Bandit settings have $\mathbb{E}[e_{t,a}x_{t,a}] = 0$, which assumes that the error term is independent with the feature vector. However, this assumption can be violated in cases such as when we have omitted variable in the error term, measurement error in the regressor, simultaneous equation estimation, and etc. Mathematically, when we have the following holds,

$$\mathbb{E}[e_{t,a}x_{t,a}] \neq 0$$

the endogeneity problem occurs. Ordinary Least Square (OLS) estimator is inconsistent in endogenous settings. Econometrics literature uses Instrumental Variable (IV) method to address the endogeneity problem. An instrumental variable $z_{t,a}$ is valid if it satisfies following three conditions (M Wooldridge 2014): (i) it is uncorrelated with the error term $e_{t,a}$, i.e. $\mathbb{E}[z_{t,a}e_{t,a}] = 0$, (ii) it is correlated with the endogenous covariates $x_{t,a}$, (iii) it has no direct effect on the reward $y_{t,a}$.

Consider a valid instrumental variable $z_{t,a} \in \mathbb{R}^k$, with $||z_{t,a}||_2 \leq L_z$, and the minimum eigenvalue of $\mathbb{E}[z_{t,a}z'_{t,a}]$ is positive. We assume $\mathbb{E}[z_{t,a}x'_{t,a}] \in \mathbb{R}^{k\times d}$ has full column rank k.[1]

$$x_{t,a} = \Gamma'_0 z_{t,a} + u_{t,a} \tag{2}$$

where $\Gamma_0 \in \mathbb{R}^{k\times d}$ is an unknown true coefficient vector, $u_{t,a} \in \mathbb{R}^d$ is a 1-subgaussian error term with mean zero and the independency condition $\mathbb{E}[z_{t,a}u'_{t,a}] = 0$ is satisfied by construction. By plugging the equation of $x_{t,a}$ (Equation 2) into Equation 1, we have $y_{t,a} = (\Gamma_0\beta_0)'z_{t,a} + v_{t,a}$ where $v_{t,a} = \beta'_0 u_{t,a} + e_{t,a}$. We denote $\Gamma_0\beta_0$ as δ_0.

Recall that in offline settings, we can use instrumental variables to solve endogeneity problem through methods like TSLS. Consider matrix $Y \in \mathbb{R}^T$ with $y_{t,a}$ as its elements, matrix $X \in \mathbb{R}^{T\times d}$ with $x'_{t,a}$ as its rows, matrix $Z \in \mathbb{R}^{T\times k}$ with $z'_{t,a}$ as its rows, where $t \in \{1, \cdots, T\}$. We briefly illustrate TSLS estimation procedure in offline settings as the following: (i) First, regress the set of feature vectors X on the set of instrumental variable vectors Z using OLS method to obtain an estimated sample features $\hat{X}$; (ii) Then regress Y on the estimated feature vectors $\hat{X}$ to obtain the TSLS estimator $\hat{\beta}_{\hat{X},Y}$ using OLS method again.

Definition 1. (Two-stage Least Square Estimator)

$$\hat{X} = P_Z X, P_Z = Z(Z'Z)^{-1}Z'$$
$$\hat{\beta}_{\hat{X},Y} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

To facilitate further analysis on the TSLS estimator, we also define related OLS estimators as the folowing, where $\hat{\delta} = \hat{\Gamma}\hat{\beta}_{\hat{X},Y}$ (to see this equation, refer Appendix B for the proof).

Definition 2. (First Stage OLS Estimator)

$$\hat{\Gamma} = (Z'Z)^{-1}Z'X$$

Definition 3. (OLS Estimator Based on instrumental variable)

$$\hat{\delta} = (Z'Z)^{-1}Z'Y$$

It's known that TSLS estimator is consistent in offline settings, while it remains unknown for the consistency of the TSLS estimator in online settings. As far as we know, we are the first who combine the contextual linear bandit algorithm with TSLS and prove the consistency of TSLS estimator in online settings. We provide the asymptotic properties in Section 5.

We aim to design an online decision making algorithm that learns the coefficient of main interest β_0 so that we can maximize the total expected reward after pulling arms. We define the total expected regret of an algorithm $\mathcal{A}$ after T rounds as

$$R_T = \sum_{t=1}^{T} \mathbb{E}[y_{t,a_t^*} - y_{t,a_t}] = \sum_{t=1}^{T} \mathbb{E}[\langle \Gamma'_0 z_{t,a_t^*}, \beta_0\rangle - \langle \Gamma'_0 z_{t,a_t}, \beta_0\rangle]$$

where $a_t^* = \arg\max_a \langle \Gamma'_0 z_{t,a}, \beta_0\rangle$ is the best arm at round t according to the true coefficient vectors β_0 and Γ_0, and a_t is the arm selected by the algorithm $\mathcal{A}$ at round t.

We relist key assumptions in this paper as the following,

Assumption 1. $||z_{t,a}||_2 \leq L_z$, $||x_{t,a}||_2 \leq L_x$, $||y_{t,a}||_2 \leq L_y$ for all t and a.

Assumption 2. The minimum eigenvalue of $\mathbb{E}[z_{t,a}z'_{t,a}]$ is larger than a positive constant λ.

Assumption 1 ensures that the $||\cdot||_2$ of instrumental variables, feature vectors, and the reward are bounded by positive constants. Notice that the $||\cdot||_\infty$ of a vector is smaller than the $||\cdot||_2$ of the vector, which implies that the $||\cdot||_\infty$ of the instrumental variables, features , and the reward are also upper bounded by these constants, L_z, L_x, and L_y respectively. Assumption 2 guarantees that, with high probability, the sample second moment $\sum_{s=1}^{t} z_{s,a}z'_{s,a}$ is non-singular so that the OLS estimators $\hat{\Gamma}$ and $\hat{\delta}$ exist. We need these assumptions to bound the total regret and the inference bias.

For the simplicity of notations, we omit the subscript a of $z_{t,a}$, $x_{t,a}$, and $y_{t,a}$ from below. By x_t, we mean one sample or arm chosen from the arm set $\mathcal{X}_t$ at time t; z_t is the instrumental variable related to x_t; y_t is the reward generated from the arm x_t.

[1]The assumption $\mathbb{E}[z_{t,a}x'_{t,a}] \in \mathbb{R}^{k\times d}$ has full column rank k implies that $k \geq d$

4 BanditIV algorithm

We propose the following *BanditIV* Algorithm (Algorithm 1) by extending the TSLS to the online settings. The algorithm is based on existing linear bandit algorithms, especially the OFUL algorithm proposed by Abbasi-Yadkori, Pál, and Szepesvári 2011. The *BanditIV* algorithm takes as input initial regularization parameters $\gamma_z, \gamma_x > 0$, as well as confidence set parameters $\{G_t\}_{t=1}^{\infty}$, $\{B_t\}_{t=1}^{\infty}$. We will specify the confidence set parameters in Section 4.1.

Estimation: The algorithm maintains four matrices U_t, V_t, W_t, Q_t to calculate estimated parameters in each round as described in the pseudo code of Algorithm 1. If the current time t is after the first round, the algorithm utilizes past-choice-related instrumental variables Z_t and past choices X_t to estimate Γ_0 through OLS and obtain an estimated X_t, which is $\hat{X}_t$. Then, based on $\hat{X}_t$ and past observations of rewards Y_t, the algorithm estimates β_0 by OLS again. We denote the estimator for β_0, which is $\hat{\beta}_t$, in the *BanditIV* algorithm as the *BanditIV estimator*.

Confidence Sets: Following the *Optimisim in the Face of Uncertainty principle* (OFU), we need maintain confidence sets for all unknown parameters in the model. As described in Section 3, we have two unknown parameters, Γ_0 and β_0. Thus, we construct two confidence sets $C_{1,s}$, $C_{2,s}$, $s \in \{1, \cdots, T\}$ for the first and the second stage estimation respectively in each round. The idea for the confidence sets are to make the estimation optimistic with the conditions that "with high probability" the true coefficients are in the confidence sets and we can calculate the confidence sets from the past chosen arms X_t, related instrumental variables Z_t, and rewards Y_t.

Execution: The best arm generated by the algorithm is a sample x_t which is related to an instrumental variable z_t that can maximize the estimated reward jointly with a pair of optimistic estimates of two-stage coefficients in the confidence sets. After an arm is chosen, we observe the reward y_t as stated in Equation (1).

Taking the *BanditIV* algorithm as a special case, we further propose the ϵ-*BanditIV* algorithm (Algorithm 2). The ϵ-*BanditIV* algorithm additionally takes a sequence of non-increasing exploration parameters $\{\epsilon_t\}_{t=1}^{\infty}$ as input. In each round, instead of choosing the best arm, we conduct a stochastic decision with probability ϵ_t for choosing a random arm and probability $1 - \epsilon_t$ for choosing the estimated best arm. Notice that when $\epsilon_t = 0$ for $t \in \{1, 2 \cdots, \infty\}$, the ϵ-*BanditIV* algorithm degenerates to the *BanditIV*.

Algorithm 1 BanditIV

Input: γ_z, γ_x, $\{G_t\}_{t=1}^{\infty}$,$\{B_t\}_{t=1}^{\infty}$
Set $U_0 = \gamma_z I \in \mathbb{R}^{k\times k}, V_0 = 0 \in \mathbb{R}^{k\times d}, W_0 = \gamma_x I \in \mathbb{R}^{d\times d}, Q_0 = 0 \in \mathbb{R}^{d\times 1}$
Set $C_{1,s} = \{\Gamma : ||\Gamma^{(i)} - \hat{\Gamma}_s^{(i)}||_{U_s} \leq G_s\}, C_{2,s} = \{\beta : ||\beta - \hat{\beta}_s||_{W_s} \leq B_s\}$, $\forall s \in \{1, 2, \cdots T\}$, $\forall i \in \{1, 2, \cdots d\}$
Nature reveals $\mathcal{Z}_0$. We randomly choose $z_0 \in \mathcal{Z}_0$ and set $Z_0 = z_0$
Nature reveals $\mathcal{X}_0$. From the set $\mathcal{X}_0$, we play x_0 which is related with z_0 and then observe the reward y_0. We set $X_0 = x_0$ and $Y_0 = y_0$
for $t := 1, 2, \cdots, T$, **do**
$U_t = U_{t-1} + z_{t-1}z'_{t-1}, V_t = V_{t-1} + z_{t-1}x'_{t-1}$
$\hat{\Gamma}_t = U_t^{-1}V_t$, $\hat{X}_{t-1} = Z_{t-1}\hat{\Gamma}_t$
$W_t = W_0 + \hat{X}'_{t-1}\hat{X}_{t-1}, Q_t = Q_0 + \hat{X}'_{t-1}Y_{t-1}$
$\hat{\beta}_t = W_t^{-1}Q_t$
Nature reveals $\mathcal{Z}_t$. We choose $z_t = \arg\max_{z\in\mathcal{Z}_t} \max_{\Gamma\in C_{1,t}} \max_{\beta\in C_{2,t}} \langle \Gamma' z, \beta\rangle$ and update $Z_t = [Z'_{t-1} \quad z_t]'$
Nature reveals the set of arms $\mathcal{X}_t$. We play x_t which is related with z_t and observe the reward y_t. Update $X_t = [X'_{t-1} \quad x_t]', Y_t = [Y'_{t-1} \quad y_t]'$
end for

Algorithm 2 ϵ-BanditIV

Input: $\gamma_z, \gamma_x, \{G_t\}_{t=1}^{\infty}, \{B_t\}_{t=1}^{\infty}, \{\epsilon_t\}_{t=1}^{\infty}$
Set $U_0 = \gamma_z I \in \mathbb{R}^{k\times k}, V_0 = 0 \in \mathbb{R}^{k\times d}, W_0 = \gamma_x I \in \mathbb{R}^{d\times d}, Q_0 = 0 \in \mathbb{R}^{d\times 1}$
Set $C_{1,s} = \{\Gamma : ||\Gamma^{(i)} - \hat{\Gamma}_s^{(i)}||_{U_s} \leq G_s\}, C_{2,s} = \{\beta : ||\beta - \hat{\beta}_s||_{W_s} \leq B_s\}, \forall s \in \{1, 2, \cdots T\}$, $\forall i \in \{1, 2, \cdots d\}$
Nature reveals $\mathcal{Z}_0$. We randomly choose $z_0 \in \mathcal{Z}_0$ and set $Z_0 = z_0$
Nature reveals $\mathcal{X}_0$. From the set $\mathcal{X}_0$, we play x_0 which is related with z_0 and then observe the reward y_0. We set $X_0 = x_0$ and $Y_0 = y_0$
for $t := 1, 2, \cdots, T$, **do**
$U_t = U_{t-1} + z_{t-1}z'_{t-1}, V_t = V_{t-1} + z_{t-1}x'_{t-1}$
$\hat{\Gamma}_t = U_t^{-1}V_t, \hat{X}_{t-1} = Z_{t-1}\hat{\Gamma}_t$
$W_t = W_0 + \hat{X}'_{t-1}\hat{X}_{t-1}, Q_t = Q_0 + \hat{X}'_{t-1}Y_{t-1}$
$\hat{\beta}_t = W_t^{-1}Q_t$
Nature reveals $\mathcal{Z}_t$. With probability ϵ_t, we uniformly choose a random $z_t \in \mathcal{Z}_t$. With probability $1 - \epsilon_t$, we choose $z_t = \arg\max_{z\in\mathcal{Z}_t} \max_{\Gamma\in C_{1,t}} \max_{\beta\in C_{2,t}} \langle \Gamma' z, \beta\rangle$ and update $Z_t = [Z'_{t-1} \quad z_t]'$
Nature reveals the set of arms $\mathcal{X}_t$. We play x_t which is related with z_t and observe the reward y_t. Update $X_t = [X'_{t-1} \quad x_t]', Y_t = [Y'_{t-1} \quad y_t]'$
end for

4.1 Regret analysis

In this section, we give upper bounds on the regret of the *BanditIV* algorithm and $\epsilon-$*BanditIV* algorithm. The proofs can be found in Appendix A. We show an $\tilde{\mathcal{O}}(k\sqrt{T})$ upper bound for the total expected regret with parameters of confidence set by Theorem 1 and Corollary 2.

Theorem 1. *The expected cumulative regret of the ϵ-BanditIV at time T, with probability at least $1 - \delta$, is upper-bounded by*

$$R_T \leq B_T\sqrt{2Td\log(\frac{T+d}{d})} + (\frac{2}{\gamma} + \|\beta_0\|_2)G_T\sqrt{2Tk\log(\frac{T+k}{k})} + 2\epsilon_0 T L_y$$

where $B_T = \sqrt{\gamma}||\beta_0||_2 + \sqrt{2\log(\frac{2T}{\delta}) + d\log\left(\frac{5TL_x^2}{d}\right)}$,

$G_T = \sqrt{\gamma_z}||\Gamma_0^{(i)}||_2 + \sqrt{2\log(\frac{2T}{\delta}) + k\log\left(\frac{5TL_y^2}{k}\right)}$.

By setting $\epsilon_t = 0$, we can derive the following corollary directly,

Corollary 2. *The expected cumulative regret of the BanditIV at time T, with probability at least $1 - \delta$, is upper-bounded by*

$$R_T \leq B_T\sqrt{2Td\log(\frac{T+d}{d})} + (\frac{2}{\gamma} + \|\beta_0\|_2)G_T\sqrt{2Tk\log(\frac{T+k}{k})}$$

where $B_t = \sqrt{\gamma}||\beta_0||_2 + \sqrt{2\log(\frac{2T}{\delta}) + d\log\left(\frac{5TL_x^2}{d}\right)}$,

$G_t = \sqrt{\gamma_z}||\Gamma_0^{(i)}||_2 + \sqrt{2\log(\frac{2T}{\delta}) + k\log\left(\frac{5TL_y^2}{k}\right)}$.

Remark 1. Notice that in order to guarantee an $\tilde{\mathcal{O}}(k\sqrt{T})$ upper bound for the total expected regret of the ϵ-BanditIV algorithm, we need a small enough ϵ_t, such as an $\epsilon_t \leq \frac{\sqrt{\log(t)}}{\sqrt{t}}$. The *BanditIV* algorithm which is a special case with $\epsilon_t = 0$ in ϵ-BanditIV algorithm, naturally satisfies this condition.

5 Inference and confidence intervals

Due to the adaptive nature of multi-armed bandits, calculating confidence intervals is not straightforward as the collected data is no longer iid. In this section, we show asymptotic properties of the

BanditIV estimator following H. Chen, Lu, and Song 2021 and Bastani and Bayati 2020. We first present the consistency and then the normality of the *BanditIV* estimator. The proofs are presented in Appendix A.

Proposition 1. *(Tail bound for the BanditIV estimator) In the online decision-making model with ϵ-greedy policy, if the Assumptions 1 and 2 are satisfied, and ϵ_t is non-increasing, then for any $\eta_1, \eta_2 > 0$, $\eta_2 \neq \frac{\lambda_{min}(\Gamma_0)}{(kd)^{\frac{1}{2}}}$,*

$$P(||\hat{\beta} - \beta_0||_1 \leq C_\beta) \geq p_1 p_2^d$$

where $C_\beta = \dfrac{\eta_1 + \eta_2 ||\beta_0||_1}{\lambda_{min}(\Gamma_0) d^{\frac{-1}{2}} - \eta_2 (k)^{\frac{1}{2}}}$,

$$p_1 = 1 - \exp\{-\frac{t\epsilon_t}{8}\} - k\exp\{-\frac{t\epsilon_t \lambda}{32 L_z^2}\} - 2k\exp\{-\frac{t\epsilon_t^2 \lambda^2 \eta_1^2}{128 k^2 \sigma_v^2 L_z^2}\} + 2k^2 \exp\{-\frac{t\epsilon_t^2 \lambda^2 \eta_1^2 + 4t\epsilon_t \lambda k^2 \sigma_v^2}{128 k^2 \sigma_v^2 L_z^2}\},$$

$$p_2 = 1 - \exp\{-\frac{t\epsilon_t}{8}\} - k\exp\{-\frac{t\epsilon_t \lambda}{32 L_z^2}\} - 2k\exp\{-\frac{t\epsilon_t^2 \lambda^2 \eta_1^2}{128 k^2 \sigma_u^2 L_z^2}\} + 2k^2 \exp\{-\frac{t\epsilon_t^2 \lambda^2 \eta_1^2 + 4t\epsilon_t \lambda k^2 \sigma_u^2}{128 k^2 \sigma_u^2 L_z^2}\}$$

Remark 2. If $t\epsilon_t^2 \to \infty$ as $t \to \infty$, then the probability of $||\hat{\beta} - \beta_0|| \leq C_\beta$ goes to 1 for any $\eta_1, \eta_2 > 0$, $\eta_2 \neq \frac{\lambda_{min}(\Gamma_0)}{(kd)^{\frac{1}{2}}}$. We need carefully choose the value for ϵ_t in order to guarantee both the regret bound and the tail bound. For example, an $\epsilon_t = \frac{\sqrt{\log(t)}}{\sqrt{t}}$ or $\epsilon_t = \frac{\sqrt{\log\log(t)}}{\sqrt{t}}$ satisfies both the conditions for the regret bound and the tail bound.

Following from the Propositions 1, we obtain the consistency of the BanditIV estimator easily.

Corollary 3. *(Consistency of the online BanditIV estimator). If Assumptions 1 and 2 are satisfied, ϵ_t is non-increasing and $t\epsilon_t^2 \to \infty$ as $t \to \infty$, then the online BanditIV estimator $\hat{\beta}_t$ is a consistent estimator for β_0.*

Theorem 4. *(Asymptotic normality of the online BanditIV estimator) If Assumptions 1 and 2 are satisfied, ϵ_t is non-increasing and $t\epsilon_t^2 \to \infty$ as $t \to \infty$, the non-increasing function $\epsilon_t \to \epsilon_\infty$ and $t\epsilon_t^2 \to \infty$ as $t \to \infty$. Then*

$$\sqrt{t}(\hat{\beta}_t - \beta_0) \xrightarrow{d} \mathcal{N}_d(0, \mathbb{E}[v_t^2](\Gamma_0'\Gamma_0)^{-1}\Gamma_0'(\int zz' d\mathcal{P}_z)^{-1}\Gamma_0(\Gamma_0'\Gamma_0)^{-1})$$

Theorem 4 provides a theoretical guarantee that the *BanditIV* estimator asymptotically follows a normal distribution with mean zero and variance $\mathbb{E}[v_t^2](\Gamma_0'\Gamma_0)^{-1}\Gamma_0'(\int zz' d\mathcal{P}_z)^{-1}\Gamma_0(\Gamma_0'\Gamma_0)^{-1}$. We can see that this variance depends on the expectation of the error term v_t, which includes the errors in both first stage and the second stage. The variance also depends on the distribution of the instrumental variable, $\mathcal{P}_z$. Notice that, although we need the assumption about ϵ_t to guarantee the consistency of the estimator, the asymptotic variance of the *BanditIV* estimator does not depend on ϵ_t.

To compute confidence intervals of our estimate of β_0 post running our bandits, we use a re-randomization test similar to that of Bojinov, Simchi-Levi, and Zhao 2020 and Farias et al. 2022. For the estimate $\hat{\beta}$, we test the sharp null hypothesis that the $\beta_0 = \tau$ for all t. The sharp null hypothesis implies that the new outcome is $y_t + \tau'(x_t^H - x_t)$ where x_t^H is a newly pulled arm and x_t is a pulled arm by our bandit algorithm.

We conduct exact tests by using the known assignment mechanism to simulate new assignment paths. Algorithm 3 provides the details of how to implement it. In particular, we propose τ by a downward search method based on the estimate from the bandit algorithm. Under the sharp null hypothesis of $\beta_0 = \tau$, a new arm assignment path $1 : T$ leads to a sequence of observed outcomes $y_t + \tau'(x_t^H - x_t)$ for $t \in \{1, \cdots, T\}$. To obtain a confidence interval, we invert a sequence of exact hypothesis tests to identify the region outside where the null hypothesis is violated at the prespecified significance level (G. W. Imbens and Rubin 2015, Bojinov, Simchi-Levi, and Zhao 2020).

Algorithm 3 Sharp Null Hypothesis Test

Require: Fix N_H, total number of samples drawn; Given $\hat{\Gamma}$, X, and Y from the bandit algorithm; Given a prespecified significance level α

for i in $1 : N_H$ **do**

Sample a new assignment path, x_t^H, for $t \in \{1, \cdots, T\}$ according to the assignment mechanism under the null hypothesis $\beta_0 = \tau$

Change the sequence of original outcomes y_t to $y_t + \langle \tau, (x_t^H - x_t) \rangle$ for $t \in \{1, \cdots, T\}$

Compute $\hat{\tau}^{[i]}$ as the original algorithm does, i.e.

$$\hat{X}^H = Z^H \hat{\Gamma}$$

$$\hat{\tau}^{[i]} = \gamma I + ((\hat{X}^H)' \hat{X}^H)^{-1} (\hat{X}^H)' Y^H$$

end for

Compute $\hat{p} = N_H^{-1} \sum_{i=1}^{N_H} \mathbf{1}\{|\hat{\tau}^{[\mathbf{i}]}| > |\hat{\beta}|\}$

Reject the null hypothesis if $\hat{p} < \alpha$

All of these experiments are run on synthetic data constructed as in Section 6. We run 50 rounds in total. In each round, we propose 20 hypotheses for β_0 based on the estimated coefficient of our main interest, $\hat{\beta}$ from a pre-run BanditIV or ϵ-greedy BanditIV and construct confidence intervals for all dimensions of β_0. For each hypothesis of β_0, we run the sharp null hypothesis test where we set N_H as 200. We choose the significance level to be $\alpha = 0.05$. Under BanditIV, the coverage of the test for each dimension of the estimate is listed as follows: 94%, 96%, 96% ; Under ϵ-greedy BanditIV, the coverage of the test for each dimension of the estimate is listed as follows: 96%, 92%, 94%. We can see that, given the significance level as 0.05, the coverage of the test is close to ideal across both BanditIV and ϵ-greedy BanditIV. Thus, the re-randomization tests and corresponding confidence intervals reported here are adequate for inference of the main treatment effect.

6 Numerical experiments

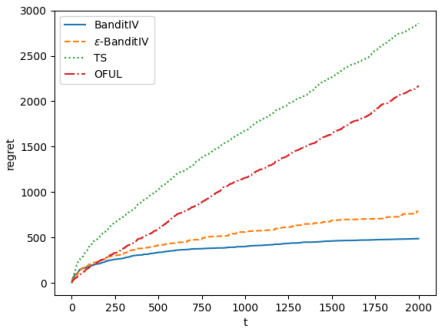
In this section, we construct synthetic data to further validate our algorithm. Referring the simulation set up in Bakhitov and Singh 2021, we consider the model as follows, for $t = 1, \cdots, T$,

$$Y_t = X_t \beta_0 + e_t \rho + \epsilon_t$$
$$X_t = Z_t \Gamma_0 + e_t + u_t$$

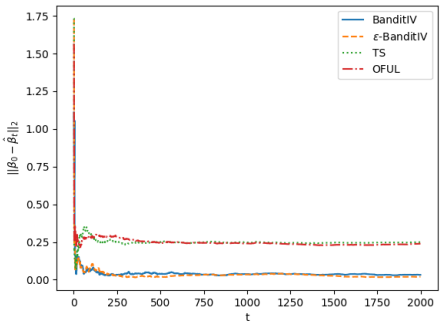
Suppose that all elements of the instrument $Z_t \in \mathbb{R}^{k \times 1}$ are uniformly distributed on the support $[-3, 3]$. $e_t \in \mathbb{R}^{n_t \times d} \sim \mathbb{N}(0, 1)$ is the confounder. $\rho \in \mathbb{R}^{d \times 1}$, is the parameter measuring the degree of endogeneity. When $\rho = 0$, endogeneity disappears, which we can see from Equation 3. $u_t \in \mathbb{R}^{n_t \times d}$, and $\epsilon_t \in \mathbb{R}^{n_t \times 1}$ are i.i.d. normally distributed additional noise components, where n_t is the number of arms at time t.

$$\mathbb{E}[(e_t \rho + \epsilon_t)' X_t] = \mathbb{E}[\rho' e_t' X_t] = \mathbb{E}[\rho' e_t' \mathbb{E}[X_t | e_t]] = \rho' \mathbb{E}[e_t' e_t] \tag{3}$$

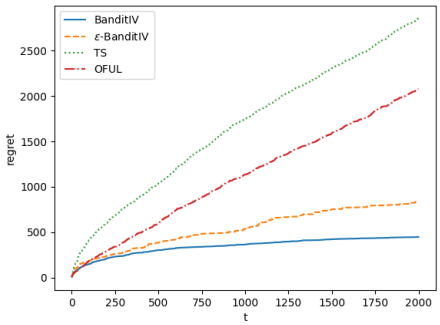
In our simulation, we set $d = 3, k = 5, n_t = 50$ for $t = 1, \cdots, T$. We run $T = 2000$ timesteps for each algorithm. Figures 1 shows the results of regret variation and estimation bias for four algorithms including BanditIV, ϵ-BanditIV, TS and OFUL. We compare algorithm performance under endogeneity case when $\rho = [0.5 \quad 0.5 \quad 0.5]'$ (shown in Figure 1a and 1b) as well as non-endogeneity case when $\rho = [0. \quad 0. \quad 0.]'$ (shown in Figure 1c and 1d). We find that BanditIV and epsilon-greedy BanditIV outperform other two classic algorithms both on regret and inference. This outperformance is more significant under endogeneity cases. We can see that when the exogeneous assumption wrt covariates are satisfied, all the four algorithms achieve approximately zero bias in estimation. However, when endogeneous covariates exist, only BanditIV and ϵ-BanditIV can achieve approximately zero bias in inference. Consistent with our theoretical analysis in Section 4.1, BanditIV reaches lower expected regret than ϵ-BanditIV, but ϵ-BanditIV can obtain a less biased estimation.



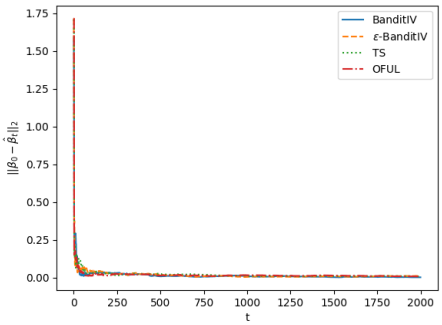
(a) Regret variations



(b) Estimation bias



(c) Regret variations



(d) Estimation bias

Figure 1: Simulation results on synthetic data

7 Conclusion

In this paper, we study the endogeneity problem in online decision making settings where we formulate the decision making process as a Contextual Linear Bandit model. The existence of endogeneity can lead to estimation bias and sub-optimal decision. To correct the bias and optimize the decision, we propose the BanditIV algorithm by utilizing both existing linear bandit algorithms in computer science literature and TSLS method in econometric literature. We present various theoretical properties of the algorithm. We first show that an upper bound for the total regret, and then show the consistency, asymptotic normality of the estimator in the algorithm. On the applied side, we conduct simulations on synthetic data as well as real time bidding data. We find that the BanditIV algorithm outperforms several benchmark linear bandit algorithms, especially when endogeneity problem occurs.

References

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved algorithms for linear stochastic bandits". In: Advances in neural information processing systems 24.

Ai, Chunrong and Xiaohong Chen (2003). "Efficient estimation of models with conditional moment restrictions containing unknown functions". In: Econometrica 71.6, pp. 1795–1843.

Angrist, Joshua and Guido Imbens (1995). Identification and estimation of local average treatment effects.

Aramayo, Nicolás, Mario Schiappacasse, and Marcel Goic (2022). "A Multi-Armed Bandit Approach for House Ads Recommendations". In: Available at SSRN 4107976.

Auer, Peter (2002). "Using confidence bounds for exploitation-exploration trade-offs". In: Journal of Machine Learning Research 3.Nov, pp. 397–422.

Bakhitov, Edvard and Amandeep Singh (2021). "Causal Gradient Boosting: Boosted Instrumental Variable Regression". In: arXiv preprint arXiv:2101.06078.

Bastani, Hamsa and Mohsen Bayati (2020). "Online decision making with high-dimensional covariates". In: Operations Research 68.1, pp. 276–294.

Bastani, Hamsa, Mohsen Bayati, and Khashayar Khosravi (2021). "Mostly exploration-free algorithms for contextual bandits". In: Management Science 67.3, pp. 1329–1349.

Bibaut, Aurélien et al. (2021). "Post-contextual-bandit inference". In: Advances in Neural Information Processing Systems 34, pp. 28548–28559.

Bojinov, Iavor, David Simchi-Levi, and Jinglong Zhao (2020). "Design and analysis of switchback experiments". In: arXiv preprint arXiv:2009.00148.

Carpentier, Alexandra, Claire Vernade, and Yasin Abbasi-Yadkori (2020). "The elliptical potential lemma revisited". In: arXiv preprint arXiv:2010.10182.

Chen, Haoyu, Wenbin Lu, and Rui Song (2021). "Statistical inference for online decision making: In a contextual bandit setting". In: Journal of the American Statistical Association 116.533, pp. 240–255.

Chen, Xiaohong, Han Hong, and Elie Tamer (2005). "Measurement error models with auxiliary data". In: The Review of Economic Studies 72.2, pp. 343–366.

Chernozhukov, Victor, Guido W Imbens, and Whitney K Newey (2007). "Instrumental variable estimation of nonseparable models". In: Journal of Econometrics 139.1, pp. 4–14.

Chu, Wei et al. (2011). "Contextual bandits with linear payoff functions". In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 208–214.

Dani, Varsha, Thomas P Hayes, and Sham M Kakade (2008). "Stochastic linear optimization under bandit feedback". In.

Durand, Audrey et al. (2018). "Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis". In: Machine learning for healthcare conference. PMLR, pp. 67–82.

Farias, Vivek et al. (2022). "Synthetically Controlled Bandits". In: arXiv preprint arXiv:2202.07079.

Griliches, Zvi (1977). "Estimating the returns to schooling: Some econometric problems". In: Econometrica: Journal of the Econometric Society, pp. 1–22.

Hadad, Vitor et al. (2021). "Confidence intervals for policy evaluation in adaptive experiments". In: Proceedings of the National Academy of Sciences 118.15, e2014602118.

Hall, Peter and Christopher C Heyde (2014). Martingale limit theory and its application. Academic press.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". In: Econometrica: Journal of the econometric society, pp. 1029–1054.

Hausman, Jerry A (1983). "Specification and estimation of simultaneous equation models". In: Handbook of econometrics 1, pp. 391–448.

Imbens, Guido (2014). Instrumental variables: an econometrician's perspective. Tech. rep. National Bureau of Economic Research.

Imbens, Guido W and Donald B Rubin (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Langford, John and Tong Zhang (2007). "The epoch-greedy algorithm for contextual multi-armed bandits". In: Advances in neural information processing systems 20.1, pp. 96–1.

Lattimore, Tor and Csaba Szepesvári (2020). Bandit algorithms. Cambridge University Press.

Li, Jin, Ye Luo, and Xiaowei Zhang (2021). "Causal reinforcement learning: An instrumental variable approach". In: arXiv preprint arXiv:2103.04021.

Li, Lihong et al. (2010). "A contextual-bandit approach to personalized news article recommendation". In: Proceedings of the 19th international conference on World wide web, pp. 661–670.

Liu, Bing et al. (2018). "Customized nonlinear bandits for online response selection in neural conversation models". In: Thirty-Second AAAI Conference on Artificial Intelligence.

M Wooldridge, Jeffrey (2014). Introductory econometrics.

Newey, Whitney K and James L Powell (2003). "Instrumental variable estimation of nonparametric models". In: Econometrica 71.5, pp. 1565–1578.

Nie, Xinkun et al. (2018). "Why adaptively collected data have negative bias and how to correct for it". In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1261–1269.

Qin, Lijing, Shouyuan Chen, and Xiaoyan Zhu (2014). "Contextual combinatorial bandit and its application on diversified online recommendation". In: Proceedings of the 2014 SIAM International Conference on Data Mining. SIAM, pp. 461–469.
Tang, Liang et al. (2015). "Personalized recommendation via parameter-free contextual bandits". In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp. 323–332.
Wagenmaker, Andrew et al. (2021). "First-Order Regret in Reinforcement Learning with Linear Function Approximation: A Robust Estimation Approach". In: arXiv preprint arXiv:2112.03432.
Woodroofe, Michael (1979). "A one-armed bandit problem with a concomitant variable". In: Journal of the American Statistical Association 74.368, pp. 799–806.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We don't provide the code but we do provide instructions for the data and the parameters we use in the experiments
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We use normal computer resource.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We don't use existing code, data or models.
   (b) Did you mention the license of the assets? [N/A] We don't use existing code, data or models.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] We don't have any new asset.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We don't use existing code, data or models.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We don't use existing code, data or models.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We don't use crowdsourcing

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We don't use crowdsourcing

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We don't use crowdsourcing

A Appendix

Appendix A. Proof of Main Results

Proof for Theorem 1

Proof. This regret can be decomposed into

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T}(1-\epsilon_t)\mathbb{E}_{u_t}[\langle x_t^*, \beta_0\rangle - \langle x_t, \beta_0\rangle]\mathbf{1}[x_t = \Gamma_0' z_t] + \sum_{t=1}^{T}\epsilon_t \mathbb{E}_{u_t}[\langle x_t^*, \beta_0\rangle - \langle x_t, \beta_0\rangle]\mathbf{1}[x_t \neq \Gamma_0' z_t] \\
&\leq \sum_{t=1}^{T}\langle \Gamma_0' z_t^*, \beta_0\rangle - \langle \Gamma_0' z_t, \beta_0\rangle + 2\epsilon_0 T L_y
\end{aligned}
$$

Next we will upper bound the first term. Let $(\tilde{\Gamma}_t, \tilde{\beta}_t) = \arg\max_{\Gamma\in G_{t-1}, \beta\in C_{t-1}} \langle \Gamma' z_t, \beta\rangle$. By using the *two-stage* optimism we can decompose our regret as follows,

$$
\begin{aligned}
&\sum_{t=1}^{T}\langle \Gamma_0' z_t^*, \beta_0\rangle - \langle \Gamma_0' z_t, \beta_0\rangle \\
&\leq \sum_{t=1}^{T} \max_{\Gamma\in\mathcal{C}_{2,t-1}} \max_{\beta\in\mathcal{C}_{2,t-1}} \langle \Gamma' z_t^*, \beta\rangle - \langle \Gamma_0' z_t, \beta_0\rangle \\
&\leq \sum_{t=1}^{T}\langle \tilde{\Gamma}_t' z_t, \tilde{\beta}_t\rangle - \langle \Gamma_0' z_t, \beta_0\rangle \\
&= \sum_{t=1}^{T}\langle (\tilde{\Gamma}_t - \Gamma_0)' z_t, (\tilde{\beta}_t - \beta_0) + \beta_0\rangle + \langle \hat{x}_t + (\Gamma_0 - \hat{\Gamma}_t)' z_t, \tilde{\beta}_t - \beta_0\rangle \\
&\leq \sum_{t=1}^{T} ||\hat{x}_t||_{W_{t-1}^{-1}} \|\tilde{\beta}_t - \beta_0\|_{W_{t-1}} + \sum_{t=1}^{T} ||(\Gamma_0 - \hat{\Gamma}_t)' z_t||_2 \|\tilde{\beta}_t - \beta_0\|_2 \\
&\quad + \|\beta_0\|_2 \sum_{t=1}^{T} \|(\tilde{\Gamma}_t - \Gamma_0)' z_t\|_2 + \sum_{t=1}^{T} \|(\tilde{\Gamma}_t - \Gamma_0)' z_t\|_2 \|\tilde{\beta}_t - \beta_0\|_2 \\
&= \sum_{t=1}^{T} ||\hat{x}_t||_{W_{t-1}^{-1}} \|\tilde{\beta}_t - \beta_0\|_{W_{t-1}} + \|\beta_0\|_2 \sum_{t=1}^{T} \|(\tilde{\Gamma}_t - \Gamma_0)' z_t\|_2 + 2\sum_{t=1}^{T} ||(\Gamma_0 - \hat{\Gamma}_t)' z_t||_2 \|\tilde{\beta}_t - \beta_0\|_2
\end{aligned}
$$

where the first inequality comes from the Lemma 1, 2 that $\Gamma_0 \in \mathcal{C}_{2,t}, \beta_0 \in \mathcal{C}_{1,t}$ for all t, the second inequality comes from the definition of z_t, the second equality comes from the definition of $\hat{x}_t$ and some careful decomposition, and the last inequality comes from cauchy-swartz inequality.

Now our goal is to upper bounded these three terms, which mainly comes form the estimation error of Γ_0 in the first stage and the estimation error of β_0 in the second stage. For the first term, by applying Cauchy-Schwartz inequality and the standard elliptical potential lemma in Carpentier, Vernade, and Abbasi-Yadkori 2020, we have

$$\begin{aligned}\sum_{t=1}^{T} ||\hat{x}_t||_{W_{t-1}^{-1}} \|\tilde{\beta}_t - \beta_0\|_{W_{t-1}} &\leq \max_t \|\tilde{\beta}_t - \beta_0\|_{W_{t-1}} \sum_{t=1}^{T} ||\hat{x}_t||_{W_{t-1}^{-1}} \\ &\leq \max_t \|\tilde{\beta}_t - \beta_0\|_{W_{t-1}} \sqrt{2Td\log(\frac{T+d}{d})}\end{aligned}$$

For the second term, by using similar arguments, we have that

$$\begin{aligned}\|\beta_0\|_2 \sum_{t=1}^{T} \|(\tilde{\Gamma}_t - \Gamma_0)' z_t\|_2 &\leq \|\beta_0\|_2 \sum_{t=1}^{T}\sum_{i=1}^{d} \mid (\tilde{\Gamma}_t^{(i)} - \Gamma_0^{(i)})' z_t \mid \\ &\leq \|\beta_0\|_2 \sum_{t=1}^{T}\sum_{i=1}^{d} \|\tilde{\Gamma}_t^{(i)} - \Gamma_0^{(i)})\|_{U_{t-1}} \|z_t\|_{U_{t-1}^{-1}} \\ &\leq \|\beta_0\|_2 d \max_{t\in[T], i\in[d]} \|\tilde{\Gamma}_t^{(i)} - \Gamma_0^{(i)})\|_{U_{t-1}} \sum_{t=1}^{T} \|z_t\|_{U_{t-1}^{-1}} \\ &\leq \|\beta_0\|_2 d \max_{t\in[T], i\in[d]} \|\tilde{\Gamma}_t^{(i)} - \Gamma_0^{(i)})\|_{U_{t-1}} \sqrt{2Tk\log(\frac{T+k}{k})}\end{aligned}$$

where the last inequality again comes from standard ellipse potential lemma.

Finally, for the third term, by the definition of W_t, we have $\|\tilde{\beta}_t - \beta_0\|_2 \leq \frac{1}{\gamma}\|\tilde{\beta}_t - \beta_0\|_{W_{t-1}}$. Therefore, by repeating the proof in the second term, we have that,

$$\begin{aligned}&2\sum_{t=1}^{T} ||(\Gamma_0 - \hat{\Gamma}_t)' z_t||_2 \|\tilde{\beta}_t - \beta_0\|_2 \\ &\leq \frac{2}{\gamma} \max_t \|\tilde{\beta}_t - \beta_0\|_{W_{t-1}} \max_{t\in[T], i\in[d]} \|\tilde{\Gamma}_t^{(i)} - \Gamma_0^{(i)})\|_{U_{t-1}} \sqrt{2Tk\log(\frac{T+k}{k})}\end{aligned}$$

Therefore, by combing these three terms and the definition of $\mathcal{C}_{1,t}, \mathcal{C}_{2,t}$, we can upper bound our regret as

$$R_T \leq B_T \sqrt{2Td\log(\frac{T+d}{d})} + (\frac{2}{\gamma} + \|\beta_0\|_2) G_T \sqrt{2Tk\log(\frac{T+k}{k})}$$

□

Lemma 1 (Optimistic estimation of β_0)**.** *With high prob* $1-\delta/2$*, for all* $t \in [T]$*,*

$$||\hat{\beta}_t - \beta_0||_{W_t} \leq B_t$$

where $B_t = \sqrt{\gamma}||\beta_0||_2 + \sqrt{2\log(\frac{2T}{\delta}) + d\log\left(\frac{5TL_x^2}{d}\right)}$*,*

Proof. At any fixed time t, we denote $\hat{x}_s^t = \Gamma_t' z_s$ for all $s \leq t$, so the $\hat{X}_t$ defined in the algorithm is a collection of all $\{\hat{x}_s^t\}_{s\leq t}$. For convenience, we drop the superscript t here. We also denote $\mathbf{e}_t \in \mathbb{R}^t$ as a collection of $\{e_s\}_{s\leq t}$. Now we get the closed-form for $\hat{\beta}_t$ by ridge TSLS estimator as follows,

$$\begin{aligned}\hat{\beta}_t &= W_t^{-1} Q_t \\ &= W_t^{-1} \hat{X}_t' (X_t \beta_0 + \mathbf{e}_t) \\ &= (X_t' P_{Z_t} X_t + \gamma I)^{-1} X_t' P_{Z_t} X_t \beta_0 + W_t^{-1} \hat{X}_t' \mathbf{e}_t \\ &= \beta_0 - \gamma W_t^{-1} \beta_0 + W_t^{-1} \hat{X}_t' \mathbf{e}_t\end{aligned}$$

where we denote $Z_t(Z_t'Z_t)^{-1}Z_t$ as P_{Z_t}. The third equality comes from the definition of $\hat{W}_t$ and the upper bound of $\hat{X}_t$ in Lemma 8 Therefore, we can write the estimation error of $\hat{\beta}$ compared to β_0 as

$$\begin{aligned}
||\hat{\beta}_t - \beta_0||_{W_t} &= ||W_t^{-1}\hat{X}_t'\mathbf{e}_t - \gamma W_t^{-1}\beta_0||_{W_t} \\
&= ||\hat{X}_t'\mathbf{e}_t - \gamma\beta_0||_{W_t^{-1}} \\
&\leq ||\hat{X}_t'\mathbf{e}_t||_{W_t^{-1}} + \gamma||\beta_0||_{W_t^{-1}} \\
&\leq ||\hat{X}_t'\mathbf{e}_t||_{W_t^{-1}} + \gamma||\beta_0||_{(\gamma I)^{-1}} \\
&= ||\hat{X}_t'\mathbf{e}_t||_{W_t^{-1}} + \sqrt{\gamma}||\beta_0||_2
\end{aligned}$$

Finally, by noticing that $\mathbf{e}_t$ is 1-subgaussian and the that $\exp(\langle q, \hat{X}_t'\mathbf{e}_t\rangle - \frac{||q||^2_{\hat{X}_t'\hat{X}_t'}}{2})$ for all $q \in \mathbb{R}^d$ is a supermartingale, according to the Section 20.1 in Lattimore and Szepesvári 2020, we have that, with probability $1 - \delta/2T$,

$$||\hat{X}_t'\mathbf{e}_t||_{W_t^{-1}} \leq \sqrt{2\log(\frac{2T}{\delta}) + \log(\frac{\det(W_t)}{\gamma^d})} \leq B_t.$$

where the last inequality comes from the explicit calculation of W_t detailed in Lemma 8. □

Lemma 2 (Optimistic estimation of Γ_0)**.** *With high prob $1 - \delta/2$, for all $t \in [T]$ and all $i \in [d]$,*

$$||\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}||_{U_t} \leq G_t$$

where $G_t = \sqrt{\gamma_z}||\Gamma_0^{(i)}||_2 + \sqrt{2\log(\frac{2T}{\delta}) + k\log\left(\frac{5TL_y^2}{k}\right)}$.

Proof. The proof steps are similar to the previous lemma. At any fixed time t and dimension $i \in [d]$, we denote $\mathbf{u}_t^i \in \mathbb{R}^t$ as a collection of $\{e_s\}_{s\leq t}$. We again get the closed-form of $\hat{\Gamma}_t$ as

$$\hat{\Gamma}_t = U_t^{-1}V_t = U_t^{-1}(U_t - \gamma_z I)\Gamma_0 + U_t^{-1}Z_t'\mathbf{u}_t^i = \Gamma_0 - \gamma_z U_t^{-1}\Gamma_0 + U_t^{-1}Z_t'\mathbf{u}_t^i$$

And therefore, $||\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}||_{U_t} \leq \|Z_t'\mathbf{u}_t^i\|_{U_t^{-1}} + \sqrt{\gamma_z}||\Gamma_0^{(i)}||_2$.

Finally, again by noticing that $\mathbf{u}_t^i$ is 1-subgaussian and the that $\exp(\langle q, \hat{X}_t'\mathbf{u}_t^i\rangle - \frac{||q||^2_{\hat{X}_t'\hat{X}_t'}}{2})$ for all $q \in \mathbb{R}^d$ is a supermartingale, we have that, with probability $1 - \delta/2T$,

$$||\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}||_{U_t} \leq \sqrt{\gamma_z}||\Gamma_0^{(i)}||_2 + \sqrt{2\log(\frac{2T}{\delta}) + \log(\frac{\det(U_t)}{\gamma_z^k})} \leq G_t$$

where the last inequality comes from the explicit calculation of U_t detailed in Lemma 8 and the union bound over all $t \in [T]$.

□

Lemma 3. *Suppose $\{\mathcal{F}_t : t = 1, \cdots, T\}$ is an increasing filtration of σ-fields. Let $\{W_t : t = 1, \cdots, T\}$ be a sequence of variables such that W_t is $\mathcal{F}_{t-1}$ measurable and $|W_t| \leq L_w$ almost surely for all t. Let $\{v_t : t = 1, \cdots, T\}$ be independent σ_v-subgaussian, and $v_t \perp \mathcal{F}_{t-1}$ for all t. Let $\mathcal{S} = \{s_1, \cdots, s_{|\mathcal{S}|}\} \subseteq \{1, \cdots, T\}$ be an index set where $|\mathcal{S}|$ is the number of elements in $\mathcal{S}$. Then for $\kappa > 0$,*

$$P(\sum_{s\in\mathcal{S}} W_s v_s \geq \kappa) \leq \exp\{-\frac{\kappa^2}{2|\mathcal{S}|\sigma_v^2 L_w^2}\}$$

The proof of this lemma is provided in Lemma 1 of H. Chen, Lu, and Song 2021.

Lemma 4. *(Dependent OLS Tail Inequality). For the online decision making model, if all realizations of z_t satisfy $||z_t||_\infty \leq L_z$ for all t, and $\hat{\Sigma} = \frac{1}{t}\sum_{s=1}^t z_s z_s'$ has minimum eigenvalue $\lambda_{min}(\hat{\Sigma}) > \lambda$ for some $\lambda > 0$ almost surely. Then for any $\eta > 0$,*

$$P(||\hat{\delta}_t - \delta_0||_1 \leq \eta) \geq 1 - 2k\exp\left(-\frac{t\lambda^2\eta^2}{2k^2\sigma_v^2 L_z^2}\right)$$

Proof. Based on the proofs provided by H. Chen, Lu, and Song 2021 and Bastani and Bayati 2020, we make some minor changes. The relation between eigenvalue and l_2 norm of symmetric matrix gives $||\hat{\Sigma}^{-1}||_2 = \lambda_{max}(\hat{\Sigma}^{-1}) = (\lambda_{min}(\hat{\Sigma}))^{-1}$. Therefore,

$$||\hat{\delta}_t - \delta_0||_2 = ||\hat{\Sigma}^{-1}(\frac{1}{t}\sum_{s=1}^{t} z_s v_s)||_2 \leq \frac{1}{t}||\hat{\Sigma}^{-1}||_2||\sum_{s=1}^{t} z_s v_s||_2 \leq \frac{1}{t\lambda}||\sum_{s=1}^{t} z_s v_s||_2$$

Hence, we have

$$\begin{aligned}
P(||\hat{\delta}_t - \delta_0||_2 \leq \eta) &\geq P(||\sum_{s=1}^{t} z_s v_s||_2 \leq t\lambda\eta) \\
&\geq P(|\sum_{s=1}^{t} Z_{j,s} v_s| \leq \frac{t\lambda\eta}{\sqrt{k}}, \cdots, |\sum_{s=1}^{t} Z_{k,s} v_s| \leq \frac{t\lambda\eta}{\sqrt{k}}) \\
&= 1 - P(\bigcup_{j=1}^{k}\{|\sum_{s=1}^{t} Z_{j,s} v_s| > \frac{t\lambda\eta}{\sqrt{k}}\}) \\
&\geq 1 - \sum_{j=1}^{k} P(|\sum_{s=1}^{t} Z_{j,s} v_s| > \frac{t\lambda\eta}{\sqrt{k}})
\end{aligned}$$

Because we know that v_s in the above inequality are i.i.d. subgaussian and $v_s \perp z_s$, we can apply Lemma 3 and have the following

$$P(|\sum_{s=1}^{t} Z_{j,s} v_s| > \frac{t\lambda\eta}{\sqrt{k}}) \leq 2\exp\{-\frac{t\lambda^2\eta^2}{2k\sigma_v^2 L_z^2}\}$$

Thus,

$$P(||\hat{\delta}_t - \delta_0||_1 \leq \eta) \geq P(||\hat{\delta}_t - \delta_0||_2 \leq \frac{\eta}{\sqrt{k}}) \geq 1 - 2k\exp\{-\frac{t\lambda^2\eta^2}{2k^2\sigma_v^2 L_z^2}\}$$

□

Lemma 5. *Let $\{z_t : t = 1, \cdots, T\}$ be a sequence of i.i.d. k-dimension random vectors such that all realizations of z_t satisfy $||z_t||_\infty \leq L_z$ for all t. Denote $\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} z_t z_t'$. If $\Sigma = \mathbb{E}[z_t z_t']$ has minimum eigenvalue $\lambda_{min}(\Sigma) > \lambda$ for some $\lambda > 0$, then*

$$P(\lambda_{min}(\hat{\Sigma}) \leq \frac{\lambda}{2}) \leq k\exp\{-\frac{T\lambda}{8L_z^2}\}$$

Proof. This proof is based on the proof of Lemma 3 in H. Chen, Lu, and Song 2021 with minor changes. First, we have

$$\lambda_{max}(\frac{z_t z_t'}{T}) = \max_{||a||_2=1} a'(\frac{z_t z_t'}{T})a = \frac{1}{T}\max_{||a||_2=1}(a' z_t)^2 \leq \frac{L_z^2}{T}$$

$$\mu_{min} \equiv \lambda_{min}(\mathbb{E}\hat{\Sigma}) = \lambda_{min}(\frac{1}{T}\sum_{t=1}^{t}\mathbb{E}[z_t z_t']) = \lambda_{min}(\Sigma) > \lambda$$

Then, using the Matrix Chernoff bound,

$$P(\lambda_{min}(\hat{\Sigma}) \leq \frac{\lambda}{2}) \leq P(\lambda_{min}(\hat{\Sigma}) \leq \frac{\mu_{min}}{2}) \leq k\exp\{-\frac{T\mu_{min}}{8L_z^2}\} \leq k\exp\{-\frac{T\lambda}{8L_z^2}\}$$

□

Lemma 6.

$$||\hat{\Gamma}(\hat{\beta} - \beta_0)||_1 \geq \frac{\lambda_{min}(\hat{\Gamma})}{\sqrt{d}}||\hat{\beta} - \beta_0||_1$$

where $\lambda_{min}(\hat{\Gamma})$ is the smallest magnitude of a singular value of this matrix.

Proof. By doing Singular Value Decomposition for $\hat{\Gamma}$, we have the following, where the second equality is due to that U is orthonormal.

$$||\hat{\Gamma}(\hat{\beta}-\beta_0)||_2 = ||U\Sigma V'(\hat{\beta}-\beta_0)||_2 = ||\Sigma V'(\hat{\beta}-\beta_0)||_2$$

We assume $rank(\hat{\Gamma}) = min\{k, d\} = d$. Using matrix form, we have

$$\Sigma V'(\hat{\beta}-\beta_0) = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_d \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} V^{(1)(1)} & V^{(2)(1)} & \dots & V^{(d)(1)} \\ V^{(1)(2)} & V^{(2)(2)} & \dots & V^{(d)(2)} \\ \vdots & \vdots & \vdots & \vdots \\ V^{(1)(d)} & V^{(2)(d)} & \dots & V^{(d)(d)} \end{bmatrix} \begin{bmatrix} \hat{\beta}^{(1)} - \beta_0^{(1)} \\ \hat{\beta}^{(2)} - \beta_0^{(2)} \\ \vdots \\ \hat{\beta}^{(d)} - \beta_0^{(d)} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1 \sum_{i=1}^d V^{(i)(1)}(\hat{\beta}^{(i)} - \beta_0^{(i)}) \\ \sigma_2 \sum_{i=1}^d V^{(i)(2)}(\hat{\beta}^{(i)} - \beta_0^{(i)}) \\ \vdots \\ \sigma_d \sum_{i=1}^d V^{(i)(d)}(\hat{\beta}^{(i)} - \beta_0^{(i)}) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Hence,

$$\begin{aligned} ||\Sigma V'(\hat{\beta}-\beta_0)||_1 \geq ||\Sigma V'(\hat{\beta}-\beta_0)||_2 &= \sqrt{\sum_{j=1}^d |\sigma_j \sum_{i=1}^d V^{(i)(j)}(\hat{\beta}^{(i)} - \beta_0^{(i)})|^2} \\ &\geq \lambda_{min}(\hat{\Gamma}) \sqrt{\sum_{j=1}^d |\sum_{i=1}^d V^{(i)(j)}(\hat{\beta}^{(i)} - \beta_0^{(i)})|^2} \\ &= \lambda_{min}(\hat{\Gamma}) ||V'(\hat{\beta}-\beta_0)||_2 \\ &= \lambda_{min}(\hat{\Gamma}) ||(\hat{\beta}-\beta_0)||_2 \end{aligned} \tag{4}$$

where the last equality is due to that V is orthonormal.

Applying Cauchy–Schwarz inequality, we obtain

$$||(\hat{\beta}-\beta_0)||_2 \geq \frac{1}{\sqrt{d}} ||(\hat{\beta}-\beta_0)||_1 \tag{5}$$

Combing inequalities (4) and (5), we complete the proof. □

Lemma 7.

$$\lambda_{min}(\hat{\Gamma}) \geq \lambda_{min}(\Gamma_0) - \eta_2 (kd)^{\frac{1}{2}} \tag{6}$$

Proof. On the event that $||\hat{\Gamma} - \Gamma_0|| \leq \eta_2$, we can rewrite the equation for each element in $\hat{\Gamma}$ using a newly defined matrix $A \in \mathbb{R}^{k \times d} : |A^{(i)(l)}| \leq 1, \forall i \in i \in \{1, 2, \cdots, d\}, \forall l \in \{1, 2, \cdots, k\}$.

$$\hat{\Gamma}^{(i)(l)} = \Gamma_0^{(i)(l)} + \eta_2 A^{(i)(l)} \tag{7}$$

We know that all singular values are non-negative. Here, we assume the singular values of Γ_0 are strictly positive. We can write the singular values of Γ_0 and $\hat{\Gamma}$ as two increasing sequences, $0 < \sigma_\Gamma^{(1)} \leq \sigma_\Gamma^{(2)} \leq \cdots \leq \sigma_\Gamma^{(d)}$ and $\sigma^{(1)} \leq \sigma^{(2)} \leq \cdots \leq \sigma^{(d)}$, respectively.

Using the min-max principle for singular values, we have

$$\sigma_{\Gamma}^{(1)} = \min_{S:dim(S)=1} \max_{\substack{a\in S \\ ||a||_2=1}} ||\Gamma_0 a||_2 = \min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\Gamma_0 a||_2$$

where the first equality is directly derived by applying the min-max theorem. The second equality is from the two facts. First, a should be of an order $d \times 1$ to fit the order of Γ_0. Second, there are only two unique vectors fitting the constraint $||a||_2 = 1$ in the given space S and they generate the same objective values $||\Gamma_0 a||_2$.

Similarly, we can rewrite the smallest singular value of $\hat{\Gamma}$ as the following

$$\sigma^{(1)} = \min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\hat{\Gamma} a||_2 \tag{8}$$

$$\begin{aligned}
\min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\hat{\Gamma} a||_2 &\geq \min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\Gamma_0 a||_2 - \eta_2 ||Aa||_2 \\
&\geq \min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\Gamma_0 a||_2 - \eta_2 ||A||_2 ||a||_2 \\
&= \min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\Gamma_0 a||_2 - \eta_2 ||A||_2 \\
&\geq \min_{\substack{a\in\mathbb{R}^d \\ ||a||_2=1}} ||\Gamma_0 a||_2 - \eta_2 (kd)^{\frac{1}{2}} \\
&= \sigma_{\Gamma}^{(1)} - \eta_2 (kd)^{\frac{1}{2}}
\end{aligned}$$

where the first inequality is derived from the equation (7) and triangle inequality. The third equality is from the constraint that $||a||_2 = 1$. The fourth inequality is from the property of Frobenius norm, which we illustrate as the following

$$||A||_2 \leq ||A||_F = (\sum_{i=1}^{d}\sum_{l=1}^{k} |A^{(i)(l)}|^2)^{\frac{1}{2}} \leq (\sum_{i=1}^{d}\sum_{l=1}^{k} 1)^{\frac{1}{2}} = (kd)^{\frac{1}{2}}$$

□

Proposition 2. *(Tail bounds for the online OLS estimators). In the online decision making model with the ϵ-greedy policy, if Assumptions 1 and 2 are satisfied, and ϵ_t is non-increasing, then for any $\eta_1, \eta_2 > 0$, any $i \in \{1, \cdots, d\}$,*

$$\begin{aligned}
P(||\hat{\delta}_t - \delta_0||_1 \leq \eta_1) \geq 1 - \exp\{-\frac{t\epsilon_t}{8}\} - k\exp\{-\frac{t\epsilon_t\lambda}{32L_z^2}\} - 2k\exp\{-\frac{t\epsilon_t^2\lambda^2\eta_1^2}{128k^2\sigma_v^2L_z^2}\} \\
+ 2k^2\exp\{-\frac{t\epsilon_t^2\lambda^2\eta_1^2 + 4t\epsilon_t\lambda k^2\sigma_v^2}{128k^2\sigma_v^2L_z^2}\}
\end{aligned}$$

$$\begin{aligned}
P(||\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}||_1 \leq \eta_2) \geq 1 - \exp\{-\frac{t\epsilon_t}{8}\} - k\exp\{-\frac{t\epsilon_t\lambda}{32L_z^2}\} - 2k\exp\{-\frac{t\epsilon_t^2\lambda^2\eta_2^2}{128k^2\sigma_u^2L_z^2}\} \\
+ 2k^2\exp\{-\frac{t\epsilon_t^2\lambda^2\eta_2^2 + 4t\epsilon_t\lambda k^2\sigma_u^2}{128k^2\sigma_u^2L_z^2}\}
\end{aligned}$$

Proof. Following the proof of Proposition 3.1 in H. Chen, Lu, and Song 2021, we provide this proof adapting to our setting. Denote $\mathcal{S}_t = \{1, \cdots, t\}$ and define $\hat{\Sigma}(\mathcal{I}) = |\mathcal{I}|^{-1} \sum_{s \in \mathcal{I}} z_s z_s'$ for any $\mathcal{I} \subseteq \mathcal{S}_t$, where $|\mathcal{I}|$ is the number of element in the set $\mathcal{I}$ and $|\mathcal{S}_t| = t$. We have

$$\hat{\delta}_t - \delta_0 = (\frac{1}{t}\sum_{s=1}^{t} z_s z_s')^{-1} \frac{1}{t}\sum_{s=1}^{t} z_s v_s = \{\hat{\Sigma}(\mathcal{S}_t)\}^{-1} \frac{1}{t}\sum_{s=1}^{t} z_s v_s$$

Under ϵ-greedy policy, we pull the estimated optimal arm with probability ϵ_s at each time point s. We denote the collection of time points up to time t when the estimated optimal arm is chosen as $\mathcal{R}_t$. We will first bound the minimum eigenvalue of $\hat{\Sigma}(\mathcal{R}_t)$ and then use it to infer the bound of the minimum eigenvalue of $\hat{\Sigma}(\mathcal{S}_t)$.

We denote the following event,

$$E := \{\lambda_{min}(\hat{\Sigma}(\mathcal{R}_t)) > \frac{\lambda}{4}\}$$

Applying Lemma 5, if $\lambda_{min}(\Sigma) > \lambda$, then we have

$$P(E) \geq 1 - k\exp\{-\frac{|\mathcal{R}_t|\lambda}{16L_z^2}\}$$

Meanwhile, we know that

$$\hat{\Sigma}(\mathcal{S}_t) = \frac{|\mathcal{R}_t|}{|\mathcal{S}_t|}\hat{\Sigma}(\mathcal{R}_t) + \frac{|\mathcal{S}_t| - |\mathcal{R}_t|}{|\mathcal{S}_t|}\hat{\Sigma}(\mathcal{S}_t \backslash \mathcal{R}_t)$$

By Weyl's inequality, on event E,

$$\lambda_{min}(\hat{\Sigma}(\mathcal{S}_t)) \geq \lambda_{min}(\frac{|\mathcal{R}_t|}{|\mathcal{S}_t|}\hat{\Sigma}(\mathcal{R}_t)) + \lambda_{min}(\frac{|\mathcal{S}_t| - |\mathcal{R}_t|}{|\mathcal{S}_t|}\hat{\Sigma}(\mathcal{S}_t \backslash \mathcal{R}_t)) \geq \frac{|\mathcal{R}_t|}{|\mathcal{S}_t|}\lambda_{min}(\hat{\Sigma}(\mathcal{R}_t)) \geq \frac{\lambda|\mathcal{R}_t|}{4t}$$

Applying Lemma 4, we have

$$P(||\hat{\delta}_t - \delta_0||_1 \leq \eta | E) \geq 1 - 2k\exp(-\frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2})$$

Hence,

$$\begin{aligned} P(||\hat{\delta}_t - \delta_0||_1 \leq \eta) &\geq P(||\hat{\delta}_t - \delta_0||_1 \leq \eta | E)P(E) \\ &\geq 1 - 2k\exp(-\frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2}) - k\exp\{-\frac{|\mathcal{R}_t|\lambda}{16L_z^2}\} + 2k^2\exp\{-\frac{|\mathcal{R}_t|\lambda}{16L_z^2} - \frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2}\} \end{aligned}$$

The checking step for that $|\mathcal{R}_t|$ is large enough is exactly the same as that in H. Chen, Lu, and Song 2021. We complete the proof by combining all the results above,

$$\begin{aligned} P(||\hat{\delta}_t - \delta_0||_1 \leq \eta) \geq 1 &- \exp\{-\frac{t\epsilon_t}{8}\} - k\exp\{-\frac{t\epsilon_t\lambda}{32L_z^2}\} - 2k\exp\{-\frac{t\epsilon_t^2\lambda^2\eta^2}{128k^2\sigma_v^2 L_z^2}\} \\ &+ 2k^2\exp\{-\frac{t\epsilon_t^2\lambda^2\eta^2 + 4t\epsilon_t\lambda k^2\sigma_v^2}{128k^2\sigma_v^2 L_z^2}\} \end{aligned}$$

We can prove the inequality for $P(||\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}||_1 \leq \eta_2)$ by similar steps. □

Proposition 1

Proof. By Proposition 2, we have $||\hat{\Gamma} - \Gamma_0||_1 = \max_{1 \leq i \leq d} ||\hat{\Gamma}^{(i)} - \Gamma_0^{(i)}||_1 \leq \eta_2$ with a probability larger than p_2^d and $||\hat{\delta} - \delta_0||_1 \leq \eta_1$ with a probability larger than p_1. On the events that $||\hat{\Gamma} - \Gamma_0||_1 \leq \eta_2$ and $||\hat{\delta} - \delta_0||_1 \leq \eta_1$, we can prove the upper bound of $||\hat{\beta} - \beta_0||_1$ as the following by utilizing triangle inequalities.

$$||\hat{\Gamma}(\hat{\beta} - \beta_0)||_1 - ||(\hat{\Gamma} - \Gamma_0)\beta_0||_1 \leq ||\hat{\Gamma}(\hat{\beta} - \beta_0) + (\hat{\Gamma} - \Gamma_0)\beta_0||_1 = ||\hat{\Gamma}\hat{\beta} - \Gamma_0\beta_0||_1 = ||\hat{\delta} - \delta_0||_1 \leq \eta_1$$

Hence,

$$||\hat{\Gamma}(\hat{\beta} - \beta_0)||_1 \leq \eta_1 + ||(\hat{\Gamma} - \Gamma_0)\beta_0||_1 \leq \eta_1 + ||\hat{\Gamma} - \Gamma_0||_1 ||\beta_0||_1 \leq \eta_1 + \eta_2 ||\beta_0||_1 \tag{9}$$

Combining inequality (9), Lemma 6 and Lemma 7, we complete the proof.

□

Theorem 5. *(Aymptotic normality of the online OLS estimator) If Assumptions 1 and 2 are satisfied, then*

$$\sqrt{t}(\hat{\delta}_t - \delta_0) \xrightarrow{d} \mathcal{N}_k(0, \mathbb{E}[v_t^2](\int zz' d\mathcal{P}_z)^{-1})$$

Proof. In this proof, we mainly use martingale central limit theorem.

By definition of the OLS estimator $\hat{\delta}_t$, we have

$$\sqrt{t}(\hat{\delta}_t - \delta_0) = (\frac{1}{t}\sum_{s=1}^{t} z_s z_s')^{-1}(\frac{1}{\sqrt{t}}\sum_{s=1}^{t} z_s v_s)$$

We define for $1 \leq j \leq t$, $t \geq 1$, and any $q \in \mathbb{R}^k$, the σ-field $\mathcal{F}_{tj}$ as $\mathcal{F}_j$,

$$H_{tj} = \frac{1}{\sqrt{t}}\sum_{s=1}^{j} q' z_s v_s.$$

We have $\mathbb{E}[q' z_s v_s | \mathcal{F}_{s-1}] = 0$. We also notice that $\{H_{tj}, \mathcal{F}_{tj}, 1 \leq j \leq t, t \geq 1\}$ is a martingale array. Hence we are going to prove the convergence of H_{tt}.

We first check the Linderberg condition as follows, for each fixed $\zeta > 0$,

$$\begin{aligned}
&\sum_{s=1}^{t} \mathbb{E}[\frac{1}{t}(q' z_s)^2 v_s^2 I\{|q' z_s v_s| > \zeta\sqrt{t}\}|\mathcal{F}_{t,s-1}] \\
&\leq \frac{||q||_2^2 L_z^2}{t}\sum_{s=1}^{t} \mathbb{E}[v_s^2 I\{v_s^2 > \frac{\zeta^2 t}{||q||_2^2 L_z^2}\}|\mathcal{F}_{s-1}] \\
&\leq ||q||_2^2 L_z^2 \mathbb{E}[v_1^2 I\{v_1^2 > \frac{\zeta^2 t}{||q||_2^2 L_z^2}\}]
\end{aligned}$$

where v_s, $s = 1, \cdots, t$ are i.i.d. from $\mathcal{P}_v$. Notice that $v_1^2 I\{v_1^2 > \frac{\zeta^2 t}{||q||_2^2 L_z^2}\}$ is dominated by v_1^2 with $\mathbb{E}[v_1^2] \leq ||\beta_0||_2^2 + 1$. Also, $v_1^2 I\{v_1^2 > \frac{\zeta^2 t}{||q||_2^2 L_z^2}\}$ converges to 0 almost surely when $t \to \infty$. Thus, applying dominated convergence theorem, we can get, as $t \to \infty$,

$$\sum_{s=1}^{t} \mathbb{E}[\frac{1}{t}(q' z_s)^2 v_s^2 I\{|q' z_s v_s| > \zeta\sqrt{t}\}|\mathcal{F}_{t,s-1}] \to 0$$

Then we are to find the limit of the conditional variance for H_{tt}. Notice that

$$\mathbb{E}[(q' z_s)^2 | \mathcal{F}_{s-1}] = \int q' z_s z_s' q d\mathcal{P}_z$$

and $q'\mathbb{E}[z_s z_s']q \leq ||q||_2^2 L_z^2$. Therefore,

$$\begin{aligned}
&\sum_{s=1}^{t} \mathbb{E}[\frac{1}{t}(q' z_s)^2 v_s^2 | \mathcal{F}_{t,s-1}] \\
&= \frac{\mathbb{E}[v_1^2]}{t}\sum_{s=1}^{t} \mathbb{E}[(q' z_s)^2 | \mathcal{F}_{s-1}] \\
&\xrightarrow{p} \mathbb{E}[v_1^2] q'(\int z_s z_s' d\mathcal{P}_z) q
\end{aligned} \tag{10}$$

Applying Martingale Limit Theorem, we have

$$\frac{1}{\sqrt{t}}\sum_{s=1}^{t} z_s v_s \xrightarrow{d} \mathcal{N}_k(0, \mathbb{E}[v_1^2] \int z_s z_s' d\mathcal{P}_z) \tag{11}$$

Now we are to find the limit of $\frac{1}{t}\sum_{s=1}^{t} q' z_s z_s' q$ for any $q \in \mathbb{R}^k$. We know that $q' z_s z_s' q$ is $\mathcal{F}_s$-measurable for each s and $\mathbb{E}[q' z_s z_s' q] \leq q' \mathbf{1}\mathbf{1}' q L_z^2 < \infty$. Thus, by Theorem 2.19 in Hall and Heyde 2014, we have

$$\frac{1}{t}\sum_{s=1}^{t}(q' z_s z_s' q - \mathbb{E}[q' z_s z_s' q | \mathcal{F}_{s-1}]) \xrightarrow{p} 0 \tag{12}$$

From the result in (10) and (12), we have $\frac{1}{t}\sum_{s=1}^{t} q' z_s z_s' q \xrightarrow{p} q'(\int z_s z_s' d\mathcal{P}_z) q$. Applying Lemma 6 in H. Chen, Lu, and Song 2021 and Continuous Mapping Theorem, we obtain

$$(\frac{1}{t}\sum_{s=1}^{t} z_s z_s')^{-1} \xrightarrow{p} (\int z_s z_s' d\mathcal{P}_z)^{-1} \tag{13}$$

Combining the results in (11) and (13), we complete the proof.

□

Appendix B. Auxiliary Lemmas

Lemma 8. *For each t, we can explicitly upper bound the following terms,*

$$W_t \leq X_t' Z_t' (Z_t' Z_t)^{-1} Z_t X_t, \quad \log(\det(W_t)) \leq d \log\left(\frac{5TL_x^2}{d}\right)$$

$$U_t \leq Z_t' Z_t + \gamma_z I, \quad \log(\det(U_t)) \leq k \log\left(\frac{5TL_y^2}{k}\right)$$

$$\hat{X}_t \leq P_{Z_t} X_t$$

Proof. For the first inequality,

$$\hat{X}_t = Z_t \hat{\Gamma}_t = Z_t (Z_t' Z_t + \gamma_z I)^{-1} Z_t' X_t \leq Z_t (Z_t' Z_t)^{-1} Z_t' X_t.$$

Then for the second inequality,

$$\begin{aligned} W_t &= \hat{X}_t' \hat{X}_t \\ &= (Z_t \hat{\Gamma}_t)'(Z_t \hat{\Gamma}_t) \\ &= (Z_t (Z_t' Z_t + \gamma_z I)^{-1} Z_t' X_t)'(Z_t (Z_t' Z_t + \gamma_z I)^{-1} Z_t' X_t) \\ &= X_t' Z_t (Z_t' Z_t + \gamma_z I)^{-1} Z_t' Z_t (Z_t' Z_t + \gamma_z I)^{-1} Z_t' X_t \\ &\leq X_t' Z_t (Z_t' Z_t)^{-1} Z_t' X_t \end{aligned}$$

Recall that $P_{Z_t} = Z_t'(Z_t' Z_t)^{-1} Z_t$, we finish the upper bound of W_t.Then we have

$$\begin{aligned} \log(\det(W_t)) &\leq d \log\left(\frac{1}{d}\text{trace}(W_t)\right) \\ &\leq d \log\left(\frac{1}{d}\text{trace}((Z_t' Z_t)^{-1} Z_t' X_t X_t' Z_t)\right) \\ &= d \log\left(\frac{1}{d}\text{trace}((Z_t' Z_t)^{-1} Z_t'(Z_t \Gamma_0 + \boldsymbol{u}_t)(Z_t \Gamma_0 + \boldsymbol{u}_t)' Z_t)\right) \\ &= d \log\left(\frac{1}{d}\left(\text{trace}(\Gamma_0 \Gamma_0' Z_t' Z_t) + \text{trace}(\Gamma_0 \Gamma_0' Z_t' \boldsymbol{u}_t \boldsymbol{u}_t' Z_t) + 2\text{trace}(Z_t' \boldsymbol{u} \Gamma_0')\right)\right) \end{aligned}$$

Now because u_t is 1-subgaussian, then we have with high probability, $\boldsymbol{u}_t \boldsymbol{u}_t' < 2I$. Then the above is upper bounded by

$$d \log\left(\frac{1}{d}\left(3\|Z_t \Gamma_0\|_F^2 + 2\|Z_t \Gamma_0\|_F \|\boldsymbol{u}_t\|_F\right)\right) \leq d \log\left(\frac{5TL_x^2}{d}\right)$$

and similarly

$$\log(\det(U_t)) \leq k \log\left(\frac{5TL_y^2}{k}\right)$$

□

Lemma 9. *Consider a sequence of vectors $(x_t)_{t=1}^T, x_t \in \mathbb{R}^d$, and assume that $||x_t||_2 \leq a$ for all t. Let $V_t = \lambda I + \sum_{s=1}^t x_s x_s'$ for some $\lambda > 0$. Then, we will have that $||x_t||_{V_{t-1}^{-1}} > b$ at most*

$$d \log(1 + a^2 T/\lambda)/\log(1+b)$$

times.

Lemma 9 is directly from lemma 6.2 in Wagenmaker et al. 2021.

Lemma 10. *$a_1, a_2, \cdots \in \mathbb{R}^d$ is $\mathcal{F}_t$ adapted and $\mathbb{E}[a_t|\mathcal{F}_{t-1}] = 0$ and let $L_t \in \mathbb{R}^{d\times d}$ be a predictable sequence such that for any $\lambda \in \mathbb{R}^d$,*

$$\mathbb{E}[\exp(\langle \lambda, a_t\rangle)|\mathcal{F}_{t-1}] \leq \exp(||\lambda||_{L_t}^2/2)$$

Let $H_t = \sum_{s=1}^t a_s$, $J_t = \sum_{s=1}^t L_t$.

$$K_t(\lambda) = \exp(\langle \lambda, H_t\rangle - ||\lambda||_{J_t}^2/2)$$

is a super martingale. Moreover, we have that

$$\mathbb{P}(\exists t : ||H_t||_{(J_t+\gamma I)^{-1}} \geq \sqrt{2\log(1/\delta) + \log(\frac{det(J_t+\gamma I)}{\gamma^d})}) \leq \delta$$

Lemma 10 is from Theorem 2 in Abbasi-Yadkori, Pál, and Szepesvári 2011

Proof for "$v^{(i)}$ is $(||\beta_0||_2^2 + 1)$-subgaussian"

Proof.

$$\begin{aligned}
\mathbb{E}[exp(\lambda v^{(i)})] &= \mathbb{E}[exp(\lambda(\sum_{l=1}^d u^{(i)(l)}\beta_0^{(l)} + e^{(i)}))] \\
&= \prod_{l=1}^d \mathbb{E}[exp(\lambda u^{(i)(l)}\beta_0^{(l)})]\mathbb{E}[exp(\lambda e^{(i)})] \\
&\leq \prod_{l=1}^d \mathbb{E}[exp(\frac{\lambda^2 {\beta_0^{(l)}}^2}{2})]\mathbb{E}[exp(\frac{\lambda^2}{2})] \\
&= exp(\frac{\lambda^2}{2}(\sum_{l=1}^d {\beta_0^{(l)}}^2 + 1)) \\
&= exp(\frac{\lambda^2}{2}(||\beta_0||_2^2 + 1))
\end{aligned}$$

where the second equality is due to the independency of distributions of $u^{(i)(l)}$ and $e^{(i)}$. The third inequality is because that $u^{(i)(l)}$ and $e^{(i)}$ are 1-subgaussian with mean zero. □

Proof for "$\hat{\delta} = \hat{\Gamma}\hat{\beta}_{X,Y}$"

Proof. We want to prove the following equation

$$\hat{\Gamma}\hat{\beta}_{X,Y} = \hat{\delta} \tag{14}$$

Multiplying $((Z\hat{\Gamma})'Z\hat{\Gamma})^{-1}(Z\hat{\Gamma})'Z$ to both sides of equation 14, we can obtain the exact form of TSLS estimator which we defined before. Therefore, we can complete the proof.

□