Online Instrumental Variable Regression: Regret Analysis and Bandit Feedback

Riccardo Della Vecchia, Debabrota Basu

Online Instrumental Variable Regression: Regret Analysis and Bandit Feedback

Riccardo Della Vecchia

Debabrota Basu

Équipe Scool, Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189- CRIStAL, F-59000 Lille, France

Abstract

The independence of noise and covariates is a standard assumption in online linear regression and linear bandit literature. This assumption and the following analysis are invalid in the case of endogeneity, i.e., when the noise and covariates are correlated. In this paper, we study the *online setting of instrumental variable (IV) regression*, which is widely used in economics to tackle endogeneity. Specifically, we analyse and upper bound regret of Two-Stage Least Squares (2SLS) approach to IV regression in the online setting. Our analysis shows that Online 2SLS (O2SLS) achieves $\mathcal{O}(d^2 \log^2 T)$ regret after T interactions, where d is the dimension of covariates. Following that, we leverage the O2SLS as an oracle to design OFUL-IV, a linear bandit algorithm. OFUL-IV can tackle endogeneity and achieves $\mathcal{O}(d\sqrt{T}\log T)$ regret. For datasets with endogeneity, we experimentally demonstrate that O2SLS and OFUL-IV incur lower regrets than the state-of-the-art algorithms for both the online linear regression and linear bandit settings.

1. Introduction

Online regression is one of the founding components of online learning (Kivinen et al., 2004), sequential testing (Kazerouni and Wein, 2021), contextual bandits (Foster and Rakhlin, 2020), and reinforcement learning (Ouhamma et al., 2022). Specially, online linear regression is widely used and analysed to design efficient algorithms and to derive corresponding theoretical guarantees, respectively (Greene, 2003; Abbasi-Yadkori et al., 2011b; Hazan and Koren, 2012). In linear regression, the *outcome* (or output variable) $Y \in \mathbb{R}$, and the *input features* (or covariates, or treatments) $\mathbf{X} \in \mathbb{R}^d$ are related by a linear structural equation:

$$Y = \boldsymbol{\beta}^T \mathbf{X} + \eta, \tag{1}$$

where $\boldsymbol{\beta}$ is the *true parameter* and η is the observational noise with variance σ^2. The goal is to estimate $\boldsymbol{\beta}$ from an observational dataset. A common assumption in online linear regression is *exogeneity*, i.e. independence of the noise η and the input features X ($\mathbb{E}[\eta|X] = 0$) (Abbasi-Yadkori et al., 2011b; Ouhamma et al., 2021). In real-life, exogeneity is often violated, and we encounter *endogeneity*, i.e. dependence between noise and covariates ($\mathbb{E}[\eta|X] \neq 0$) (Greene, 2003; Angrist et al., 1996; Zhu et al., 2022).

Endogeneity naturally arises due to plethora of reasons, including causal dependence on multiple variables, omitted explanatory variables, measurement errors etc. (Wald, 1940; Greene, 2003; Mogstad et al., 2021). Another cause of endogeneity in data observed from empirical studies is dependence of the output and the covariates on unobserved confounding

variables (Zhu et al., 2022). *Instrumental Variables* (IVs) are introduced to identify and quantify the causal effects of endogenous covariates (Newey and Powell, 2003). IVs are widely used in economics (Wright, 1928; Mogstad et al., 2021), causal inference (Rubin, 1974; Hernan and Robins, 2020; Harris et al., 2022), bio-statistics and epidemiology (Burgess et al., 2017).

Example 1.1. *(Carneiro et al., 2011; Mogstad et al., 2021) aim to estimate the number of returning students to college using the National Longitudinal Survey of Youth data. The return depends on multiple covariates X, such as whether the individual attended college, her AFQT scores, her family income, her family conditions (mother's years of education, number of siblings, urban residence at age 14 etc.). But often the family conditions have unobserved confounding effects on the college attendance and scores. This endogenous nature of data leads to bias in traditional linear regression estimates, such as Ordinary Least Squares. In order to mitigate this issue, Carneiro et al. (2011); Mogstad et al. (2021) leverage two instrumental variables (Z): average log income in the youth's county of residence at age 17, and the presence of a four-year college in the youth's county of residence at age 14*[1]*. The logic is that a youth might find going to college more attractive when labour market opportunities are weaker and a college is nearby. Using these two instrumental variables, the youth's attendance to college is estimated. Then, in the next stage this estimate of college attendance is used with family conditions to predict the return of the youth to college. This two stage regression approach with instrumental variable produces a more accurate estimate of youths' return to the college than OLS models assuming exogeneity.*

This approach to conduct two stages of linear regression using instrumental variables is called *Two Stage Least Squares Regression* (2SLS) (Angrist and Imbens, 1995; Angrist et al., 1996). 2SLS has become the standard tool in economics, social sciences, and statistics to study the effect of treatments on outcomes involving endogeneity (Mogstad et al., 2021). Recently, in machine learning community, researchers have extended traditional 2SLS techniques to nonlinear structures, non-compliant instruments, and corrupted observations using deep learning (Liu et al., 2020; Xu et al., 2020, 2021; Nareklishvili et al., 2022), graphical models (Stirn and Jebara, 2018), and kernel regression (Zhu et al., 2022), respectively.

In the best of our knowledge, all the existing works assume access to an observational dataset and solves 2SLS for that dataset in an offline setting. Also, the analysis available on the performance of 2SLS is asymptotic, i.e. what can be learned if we have access to infinite number of samples (Singh et al., 2020; Liu et al., 2020; Nareklishvili et al., 2022). In practical applications, this analysis is vacuous as one has access to only finite samples. Additionally, in practice, it is natural to acquire the data sequentially as treatments are chosen on-the-go, and then to learn the structural equation from the sequential data. This setting motivates us to develop and analyse the online extension of 2SLS, referred as O2SLS.

Additionally, in an interactive setting, if a policy maker aims to build more schools at some of the lower income areas as a form of intervention, she observes only the changes corresponding to it. This is referred as bandit feedback in online learning literature and studied under the linear bandit formulation (Abbasi-Yadkori et al., 2011a). This motivates

1. One can argue whether these are sufficient or weak instrumental variables. For simplicity, we assume sufficiency here, i.e. the instruments can decouple the unobserved confounding.

us to further extend O2SLS to *linear bandits, where bandit feedback and endogeneity occur simultaneously.*

In this paper, we investigate these two questions:

1. *What is the upper bound on the loss in performance for deploying parameters estimated by* O2SLS *instead of the true parameters* $\boldsymbol{\beta}$*?*

2. *Can we design efficient algorithms linear bandit with endogeneity by using* O2SLS*?*

Our Contributions. Our investigation has led to

1. *An Analysis of* O2SLS*:* Following online learning literature, we consider *regret*, i.e. the sum of differences between the losses incurred by the estimated parameters $\{\boldsymbol{\beta}_t\}_{t=1}^T$, and the true parameter $\boldsymbol{\beta}$, as the performance metric (Cesa-Bianchi and Lugosi, 2006). In Section 4, we theoretically show that O2SLS achieve $\mathcal{O}(d^2 \log^2 T)$ regret after receiving T samples from the observational data, which is $d \log T$ higher than the regret bound for online linear regression under exogeneity (Gaillard et al., 2019). This is the cost that O2SLS pay for handling endogeneity in two stages. In Section 4.2, we experimentally show that O2SLS incur less error than online ridge in problems with endogeneity. In our knowledge, *we are the first to propose a regret analysis of* O2SLS.

2. OFUL-IV *for Linear Bandits with Endogeneity:* In Section 5, we study the linear bandit problem with endogeneity. *We design an extension of OFUL algorithm used for linear bandit with exogeneity, namely* OFUL-IV, to tackle this problem. OFUL-IV uses O2SLS to estimate the parameters, and corresponding confidence bounds on $\boldsymbol{\beta}$ to balance exploration–exploitation. We show that for bounded outcomes, OFUL-IV achieve $\mathcal{O}(d\sqrt{T} \log T)$ regret after T interactions. We experimentally show that OFUL-IV incur lower regret than OFUL under endogeneity (Section 5.3).

3. *Technical Tools:* To reach our theoretical results, we propose novel technical tools, which can be of parallel interest. (a) We prove that the confidence interval around the parameter $\boldsymbol{\beta}$ at step t is $\mathcal{O}(d \log t)$ (Lemma 4.1), which can be used to build confidence intervals for parameters in cascaded regressions. (b) We also show that the sum of products of the first and second-stage noises, is bounded by $\sqrt{d \log T \log(1/\delta)}$ with probability $1 - \delta$ (Lemma B.4).

2. Related Work

Online Regression without Endogeneity. Our analysis of O2SLS extends the tools and techniques of online linear regression without endogeneity. Analysis of online linear regression began with (Foster, 1991; Littlestone et al., 1991). (Vovk, 1997, 2001) have shown that forward and ridge regressions achieve $\mathcal{O}(dY_{\max}^2 \log T)$ for outcomes with known bound $Y_{\max}$. Bartlett et al. (2015) generalised the analysis further for outcomes with unknown bound but by considering the features known in hindsight. Gaillard et al. (2019) improved the analysis further to propose an optimal algorithm. Ouhamma et al. (2021) further provided a high-probability bound over all possible sequences of bounded input features. But all these works assume independence of the noise and the input features. *In this paper, we analyse online linear regression under endogeneity for the first time.* We do not assume the bound on the outcome variable to be known, and also derive high probability bounds for any bounded sequence of input features.

Linear Bandits without Endogeneity. Linear bandits generalise the setting of online linear regression under bandit feedback (Abbasi-Yadkori et al., 2011a, 2012; Foster and Rakhlin, 2020). To be specific, in bandit feedback, the algorithm observes only the outcomes for the input features that it has chosen to draw during an interaction. Popular algorithm design techniques, such as optimism-in-the-face-of-uncertainty and Thompson sampling, are extended to propose OFUL (Abbasi-Yadkori et al., 2012) and LinTS (Abeille and Lazaric, 2017), respectively. OFUL and LinTS algorithms demonstrate $\mathcal{O}(d\sqrt{T}\log T)$ and $\mathcal{O}(d^{1.5}\sqrt{T}\log T)$ regret guarantees under exogeneity assumption. *Here, we use* O2SLS *as a regression oracle to develop* OFUL-IV *algorithm for linear bandits with endogeneity. We prove that* OFUL-IV *achieves* $\mathcal{O}(d\sqrt{T}\log T)$ *regret.*

Instrument-armed Bandits. Kallus (2018) is the first to study endogeneity, and instrumental variables in stochastic bandit setting. Stirn and Jebara (2018) propose a Thompson sampling-type algorithm for stochastic bandits, where endogeneity arises due to non-compliant actions. But both (Kallus, 2018) and (Stirn and Jebara, 2018) study only the finite-armed bandit setting where arms are independent of each other. In this paper, *we study the linear bandit setting with endogeneity, which requirez different techniques for analysis and algorithm design.*

3. Preliminaries: Instrumental Variables & Offline Two-stage Least Squares (2SLS)

We are given an observational dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ consisting of n pairs of input features and outcomes, such that $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^d$.[2] These inputs and outcomes are stochastically generated using a linear model

$$y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \eta_i, \qquad \text{(Second stage)}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the *unknown true parameter vector* of the linear model, and $\eta_i \sim \mathcal{N}(0, \sigma_2^2)$ is the unobserved error term representing all causes of y_i other than $\boldsymbol{x}_i$. It is assumed that the error terms η_i are independently and identically distributed, and have bounded variance σ^2. The parameter vector $\boldsymbol{\beta}$ quantifies the causal effect on y_i due to a unit change in a component of $\boldsymbol{x}_i$, while retaining other causes of y_i constant. The goal of linear regression is to estimate $\boldsymbol{\beta}$ by *minimising the square loss over the dataset* (Brier, 1950), i.e. $\hat{\boldsymbol{\beta}} \triangleq \operatorname{argmin}_{\boldsymbol{\beta}'} \sum_{i=1}(y_i - \boldsymbol{\beta}'^\top \boldsymbol{x}_i)^2$.

The obtained solution is called the Ordinary Least Square (OLS) estimate of $\boldsymbol{\beta}$ (Wasserman, 2004), and used as a corner stone of online regression (Gaillard et al., 2019) and linear bandit algorithms (Foster and Rakhlin, 2020). Specifically, if the input feature matrix $\mathbf{X}_n \in \mathbb{R}^{n\times d}$ is defined as $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top$, the outcome vector is $\boldsymbol{y}_n \triangleq [y_1, \ldots, y_n]^\top$, and the noise vector is $\boldsymbol{\eta}_n \triangleq [\eta_1, \ldots, \eta_n]^\top$, the OLS estimator is expressed as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \triangleq (\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n)^{-1}\mathbf{X}_n^{\mathrm{T}}\boldsymbol{y}_n = \boldsymbol{\beta} + (\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n)^{-1}\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\eta}_n$$

If $\mathbf{X}_n$ *and* $\boldsymbol{\eta}_n$ *are independent*, the second term has zero expected value conditioned on $\mathbf{X}_n$. Hence, the OLS estimator is asymptotically unbiased, i.e. $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \to \infty$ as $n \to \infty$.

In practice, the input features $\boldsymbol{x}$ and the noise $\boldsymbol{\eta}$ are often correlated (Greene, 2003, Chapter 8). As in Figure 1, this dependence, called endogeneity, is modelled with *a con-*

2. Matrices and vectors are represented with bold capital and bold small letters, e.g. $\mathbf{A}$ and $\mathbf{a}$, respectively.

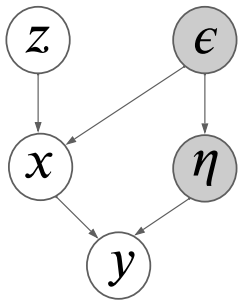


Figure 1: The DAG for 2SLS. The unobserved noises are $\boldsymbol{\epsilon}$ and η (in grey), while $\boldsymbol{z}, \boldsymbol{x}, y$ are observed quantities.

founding unobserved random variable $\boldsymbol{\epsilon}$. To compute an unbiased estimate of $\boldsymbol{\beta}$ under endogeneity, a popular technique is to introduce the Instrumental Variables (IVs) $\boldsymbol{z}$ (Angrist et al., 1996; Newey and Powell, 2003). IVs are chosen such that they are highly correlated with endogenous components of $\boldsymbol{x}$ (relevance condition) but are independent of the noise η (exogeneity condition for $\boldsymbol{z}$). This formulation leads to instrumental variable regression to tackle endogeneity.

In Two-stage Least Squares (2SLS) approach to IV regression (Angrist and Imbens, 1995; Angrist et al., 1996), it is further assumed that IVs, i.e. $\mathbf{Z}_n \triangleq [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^\top$, cause linear effects on the endogenous covariates. Specifically, for the just-identified IVs,

$$\mathbf{X}_n = \mathbf{Z}_n \boldsymbol{\Theta} + \mathbf{E}_n, \tag{First stage}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$ is an unknown first-stage parameter matrix and $\mathbf{E}_n \triangleq [\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n]^\top$ is the unobserved noise matrix leading to confounding in the second stage. This is a "classic" *multiple regression*, where the covariates $\boldsymbol{z}$ are independent of the noise terms $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$ (Wasserman, 2004, Ch. 13). Thus, the first-stage is amenable to OLS regression. This formulation leads us to the 2SLS estimator:

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left(\mathbf{Z}_n^\top \mathbf{X}_n\right)^{-1} \mathbf{Z}_n^\top \boldsymbol{y}_n. \tag{2SLS}$$

As long as $\mathbb{E}[\boldsymbol{z}_i \eta_i] = 0$ in the true model, we observe that

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left(\mathbf{Z}_n^\top \mathbf{X}_n\right)^{-1} \mathbf{Z}_n^\top \mathbf{X}_n \boldsymbol{\beta} + \left(\mathbf{Z}_n^\top \mathbf{X}_n\right)^{-1} \mathbf{Z}_n^\top \boldsymbol{\eta}_n \xrightarrow{p} \boldsymbol{\beta},$$

as $n \to \infty$. This works because IV solves for the unique parameter that satisfies $\frac{1}{n} \mathbf{Z}_n^\top \eta \xrightarrow{p} 0$. Since $\boldsymbol{x}$ and $\boldsymbol{\eta}$ are correlated, 2SLS estimator is not unbiased in finite-time.

Assumption 3.1. *The assumptions for conducting 2*SLS *with just-identified Instrumental Variables are (Greene, 2003):*

1. ***Well behaved data.*** *For every $n \in \mathbb{N}$, the matrices $\mathbf{Z}_n^\top \mathbf{Z}_n$ and $\mathbf{Z}_n^\top \mathbf{X}_n$ are full rank, and thus invertible.*
2. ***Endogeneity of $\boldsymbol{x}$.*** *The second stage input features $\boldsymbol{x}$ and noise η are not independent: $\boldsymbol{x} \not\perp\!\!\!\perp \eta$.*

3. ***Exogeneity of $\boldsymbol{z}$.*** *The IV random variables are independent of the noise in the second stage:* $\boldsymbol{z} \perp\!\!\!\perp \eta$.
4. ***Relevance Condition.*** *The variables* $\boldsymbol{z}$ *and* $\boldsymbol{x}$ *are correlated:* $\boldsymbol{z} \not\perp\!\!\!\perp \boldsymbol{x}$. *This implies that there exists* $\mathfrak{r} > 0$*:*

$$\left\| n \left(\mathbf{Z}_n^\top \mathbf{X}_n \right)^{-1} \right\|_2 \leq \frac{1}{\mathfrak{r}}. \tag{2}$$

4. Online Two-Stage Least Squares Regression

In this section, we describe the problem setting and schematic of *Online Two-Stage Least Squares Regression*, in brief O2SLS. Following that, we provide a theoretical analysis of O2SLS and show that regret of O2SLS is $\mathcal{O}(d^2 \log^2 T)$ (Section 4.1). In Section 4.2, we experimentally show that O2SLS provides an accurate estimate of the true parameter $\boldsymbol{\beta}$, and thus, incur lower regret than Online Ridge Linear Regression, in brief Ridge.

O2SLS. In the online setting of IV regression, the data $(\boldsymbol{x}_1, \boldsymbol{z}_1, y_1), \ldots, (\boldsymbol{x}_T, \boldsymbol{z}_T, y_T), \ldots$ arrives in a stream. Following the 2SLS model (Figure 1), data is generated from

$$\begin{cases} \boldsymbol{x}_t = \boldsymbol{\Theta}^\top \boldsymbol{z}_t + \boldsymbol{\epsilon}_t \\ y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t + \eta_t, \end{cases} \tag{3}$$

such that $\boldsymbol{x}_t \not\perp\!\!\!\perp \eta_t$ and $\boldsymbol{z}_t \perp\!\!\!\perp \eta_t$ for all $t \in \mathbb{N}$. At each step t, the online IV regression algorithm is served with a new input feature $\boldsymbol{x}_t$ and an IV $\boldsymbol{z}_t$, and it aims to predict an outcome $\hat{y}_t \triangleq \boldsymbol{\beta}_t^\top \boldsymbol{x}_t \in \mathbb{R}$. Here, $\boldsymbol{\beta}_t$ is the estimate of the parameter at step t computed using current $(\boldsymbol{x}_t, \boldsymbol{z}_t)$ and the data $\{(\boldsymbol{x}_i, \boldsymbol{z}_i, y_i)\}_{i=1}$ observed so far. Following the prediction, Nature reveals the true outcome y_t. Quality of the prediction is evaluated using a square loss $\ell_t(\boldsymbol{\beta}_t) \triangleq (\widehat{y}_t - y_t)^2$ (Foster, 1991). The online protocol is expressed as:

At each round $t = 1, 2, \ldots, T$
1. $\boldsymbol{z}_t$ is sampled i.i.d. from an unknown distribution
2. $\boldsymbol{x}_t$ is sampled according to Equation (3) given $\boldsymbol{z}_t$
3. we compute an estimate $\boldsymbol{\beta}_\bullet$ and make a prediction $\widehat{y}_t = \boldsymbol{\beta}_\bullet^\top \boldsymbol{x}_t$
4. we observe the true y_t following Equation (3)
5. we incur in a loss $(y_t - \widehat{y}_t)^2 = (y_t - \boldsymbol{\beta}_\bullet^\top \boldsymbol{x}_t)^2$

If the true parameter $\boldsymbol{\beta}$ was known, the square loss incurred at step t by using $\boldsymbol{\beta}$ to predict would be $\ell_t(\boldsymbol{\beta}) \triangleq \left(y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta}\right)^2$. Thus, if the online algorithm is run for T steps, the cost of estimating the parameter from observed data can be quantified as the sum of the instantaneous regrets $\bar{r}_t \triangleq \ell_t(\boldsymbol{\beta}_t) - \ell_t(\boldsymbol{\beta})$ (Cesa-Bianchi and Lugosi, 2006). This quantity is called, *regret* (or *cumulative regret*) of the online IV regression algorithm, and is defined as follows

$$\bar{R}_T \triangleq \sum_{t=1}^{T} \bar{r}_t = \sum_{t=1}^{T} (\ell_t(\boldsymbol{\beta}_t) - \ell_t(\boldsymbol{\beta})). \tag{4}$$

Regret is the cost of not knowing the true $\boldsymbol{\beta}$. Lower is the regret better is the performance of the online algorithm.

Algorithm 1 O2SLS

for $t = 1, 2, \ldots, T$ **do**
  Observe $\boldsymbol{z}_t$, $\boldsymbol{x}_t$
  Compute $\boldsymbol{\beta}_{t-1}$ according to Equation (O2SLS)
  Predict $\widehat{y}_t = \boldsymbol{\beta}_{t-1}^\top \boldsymbol{x}_t$
  Observe y_t and compute loss $\ell_t\left(\boldsymbol{\beta}_{t-1}\right)$
end for

In order to address this problem, we propose an online form of the 2SLS estimator. Thus, modifying Equation (2SLS), we obtain the O2SLS estimator that is computed for the prediction at time t, using information up to time $t-1$:

$$\boldsymbol{\beta}_{t-1} \triangleq \left(\sum_{s=1}^{t-1} \boldsymbol{z}_s \boldsymbol{x}_s^\top\right)^{-1} \sum_{s=1}^{t-1} \boldsymbol{z}_s y_s \tag{O2SLS}$$

We use the O2SLS estimator at step $t-1$ for the prediction $\widehat{y}_t = \boldsymbol{\beta}_{t-1}^\top \boldsymbol{x}_t$. We elaborate O2SLS in Algorithm 1.

Remark 4.1. *We could use $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ that we observe before committing to the estimate $\boldsymbol{\beta}_t$, and use it to predict $\widehat{y}_t$ (Vovk, 2001). Since we cannot use y_t for this estimate, we have to modify 2SLS to incorporate this additional knowledge. We skip this modification and use $\boldsymbol{\beta}_{t-1}$ to predict.*

4.1 Theoretical Analysis

Confidence Interval of $\boldsymbol{\beta}_t$. The central result in our analysis is concentration of the O2SLS estimates $\boldsymbol{\beta}_t$ around $\boldsymbol{\beta}$.

Lemma 4.1 (Confidence Ellipsoid for the Second-stage Parameters)**.** *Let us define the design matrix to be $\mathbf{G}_{\boldsymbol{z},t} = \mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d$ for some $\lambda > 0$. Then, for bounded first stage noise $|\eta_t| \leq L_\eta$, the true parameter $\boldsymbol{\beta}$ belongs to the set*

$$\mathcal{E}_t = \left\{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\mathbf{H}}_t} \leq \sqrt{\mathfrak{b}_t(\delta)}\right\}, \tag{5}$$

with probability at least $1-\delta \in (0,1)$, for all $t \geq 0$. Here, $\mathfrak{b}_t(\delta) \triangleq \frac{dL_\eta^2}{4} \log\left(\frac{1+tL_z^2/\lambda}{\delta}\right)$.

Proof Sketch. **Step 1:** First, we express the estimated $\boldsymbol{\beta}_t$ as a data-dependent perturbation to the true $\boldsymbol{\beta}$. From Equation (2SLS) and (First stage), we obtain that

$$\begin{aligned}
\boldsymbol{\beta}_t &= \boldsymbol{\beta} + \left(\sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{x}_s^\top\right)^{-1} \sum_{s=1}^{t} \boldsymbol{z}_s \eta_s \\
&= \boldsymbol{\beta} + \left(\mathbf{Z}_{t-1}^\top \mathbf{X}_{t-1}\right)^{-1} \mathbf{Z}_{t-1}^\top \boldsymbol{\eta}_{t-1}
\end{aligned} \tag{6}$$

Following this, we show that the difference depends on estimate of the first stage parameter $\widehat{\boldsymbol{\Theta}}_t$ and the data observed till t. Specifically, in the first stage, we estimate the true

parameter $\boldsymbol{\Theta}$ by solving a multiple ridge regression problem.

$$\widehat{\boldsymbol{\Theta}}_t \in \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{s=1}^{t} \left\| \boldsymbol{x}_s^\top - \boldsymbol{z}_s^\top \boldsymbol{\Theta} \right\|_2^2 + \lambda \|\!| \boldsymbol{\Theta} |\!\|_F^2$$

Here, $\|\!|\cdot|\!\|_F^2$ is the Frobenius norm and $\lambda > 0$ is the ridge regression parameter. We can compute $\widehat{\boldsymbol{\Theta}}_t$ by solving d ridge regression problems for each of the columns. Thus, we obtain an estimate of the first stage parameter as

$$\widehat{\boldsymbol{\Theta}}_t = \left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d \right)^{-1} \mathbf{Z}_t^\top \mathbf{X}_t \tag{7}$$

Thus, from Equation (6) and (7), we get

$$\boldsymbol{\beta}_t - \boldsymbol{\beta} = \widehat{\boldsymbol{\Theta}}_t^{-1} \left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d \right)^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t$$

We define $\mathbf{G}_{\boldsymbol{z},t} \triangleq \mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d$ as *the first stage design matrix.* $\mathbf{G}_{\boldsymbol{z},t}$ is always invertible with $\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq \lambda$.

Step 2: Now, for any vector $\boldsymbol{x} \in \mathbb{R}^d$, we can bound $\boldsymbol{x}^\top(\boldsymbol{\beta}_t - \boldsymbol{\beta})$ using Cauchy-Schwarz inequality

$$\boldsymbol{x}^\top(\boldsymbol{\beta}_t - \boldsymbol{\beta}) \leq \left\| \widehat{\boldsymbol{\Theta}}_t^{-\top} \boldsymbol{x} \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} \left\| \mathbf{Z}_t^\top \boldsymbol{\eta}_t \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}}$$

This step decouples the effects of the first stage estimate and the observations. Now, we choose $x = \widehat{\mathbf{H}}_{t-1} \triangleq \widehat{\boldsymbol{\Theta}}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}$, i.e. *the second stage design matrix.* Thus,

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\mathbf{H}}_t} \leq \left\| \mathbf{Z}_t^\top \boldsymbol{\eta}_t \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}}$$

Step 3: Now, using the bound of vector-valued martingales (Abbasi-Yadkori et al., 2011b), we bound the coupling between the second stage noise $\boldsymbol{\eta}_t$ and IVs $\mathbf{Z}_t$ as

$$\left\| \mathbf{Z}_t^\top \boldsymbol{\eta}_t \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} = \left\| \sum_{s=1}^{t} \eta_s \boldsymbol{z}_s \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} \leq \sqrt{\frac{dL_\eta^2}{4} \log\left(\frac{1 + tL_z^2/\lambda}{\delta} \right)}$$

with probably at least $1 - \delta$. Hence, we conclude the proof.

Regret Bound. Now, we state the regret upper bound of O2SLS and a brief proof sketch to achieve it.

Theorem 4.1 (Regret of O2SLS)**.** *If Assumption 3.1 hold true, for bounded first-stage and second-stage noises* $\|\boldsymbol{\epsilon}_t\|_2^2 \leq dL_{\boldsymbol{\epsilon}}^2$, $\eta_t^2 < L_\eta^2$, *the* true *first-stage parameters with bounded* ℓ_2*-norm* $\|\!|\boldsymbol{\Theta}|\!\|_2 \leq L_{\boldsymbol{\Theta}}$, *and bounded IVs* $\|\boldsymbol{z}\|^2 \leq L_z^2$, *regret of* O2SLS *at step* $T > 1$ *satisfies with probability at least* $1 - \delta \in (0, 1)$

$$\bar{R}_T \leq \underbrace{\mathfrak{b}_{T-1}(\delta)}_{\substack{\textit{Estimation} \\ \mathcal{O}(d \log T)}} \underbrace{\left(d(C_1^2 + C_2^2) \left(\frac{C_3 + 1}{\lambda} + \frac{\log(T) + 1}{\lambda_{\min}(\boldsymbol{\Sigma})/2} \right) \right)}_{\substack{\textit{Second Stage Feature norm} \\ \mathcal{O}(d \log T)}}$$

$$+\underbrace{\sqrt{\mathfrak{b}_{T-1}(\delta)}}_{\substack{\textit{Estimation}\\ \mathcal{O}(\sqrt{d\log T})}}\Bigg(\underbrace{L_\eta C_1\sqrt{\left(\frac{C_3+1}{\lambda}+\frac{\log(T)+1}{\lambda_{\min}(\mathbf{\Sigma})/2}\right)\log\left(\frac{\log T}{\delta}\right)}}_{\substack{\textit{First Stage Feature norm}\\ \mathcal{O}(\sqrt{\log T})}}$$
$$+\underbrace{\sqrt{d}L_\eta C_2\left(\frac{C_3+1}{\sqrt{\lambda}}+4\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\lambda_{\min}(\mathbf{\Sigma})}\left(\log\left(T\right)+1\right)}\right)}_{\substack{\textit{Correlated noise}\\ \mathcal{O}(\sqrt{d\log T})}}\Bigg)$$

Here, $\mathfrak{b}_{T-1}(\delta)$ is the confidence bound of O2SLS *estimate around $\boldsymbol{\beta}$ and is defined by Lemma 4.1. C_1, C_2, C_3 are dimension and T-independent positive constants, and $\lambda_{\min}(\mathbf{\Sigma})$ is the minimum eigenvalue of the true covariance matrix of IVs, i.e. $\mathbf{\Sigma} \triangleq \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$.*

Theorem 4.1 entails a regret $\bar{R}_T = \mathcal{O}\left(d^2\log^2(T)\right)$, where d is dimension of IV, and T is the number of interactions. This regret bound is $d\log T$ more than the regret of online ridge regression, which is $\mathcal{O}(d\log T)$ (Gaillard et al., 2019). This is due to the fact that we perform d linear regressions in the first-stage and using the predictions of first stage for the second-stage regression. These two regression steps in cascade induce the proposed regret bound.

Proof Sketch. By Equation (4), regret at step T is

$$\bar{R}_t = \sum_{t=1}^{T}\Big(\left((\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top\boldsymbol{x}_t\right)^2 + 2\eta_t(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top\boldsymbol{x}_t\Big).$$

Thus, using Equation (3), we decompose the regret as

$$\bar{R}_T = \underbrace{\sum_{t=1}^{T}\left((\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top\boldsymbol{x}_t\right)^2}_{(\bullet 1\bullet)} + 2\underbrace{\sum_{t=1}^{T}\eta_t\left(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right)^\top\Theta^\top\boldsymbol{z}_t}_{(\bullet 2\bullet)} + 2\underbrace{\sum_{t=1}^{T}\eta_t\left(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right)^\top\boldsymbol{\epsilon}_t}_{(\bullet 3\bullet)}.$$

The proof proceeds by bounding each of these three terms.

Term 1: Second Stage Regression Error. Term $(\bullet 1\bullet)$ is the error introduced by the second stage regression. First, by applying Cauchy-Schwarz inequality, we decouple the effect of parameter estimation and the feature norms

$$\sum_{t=1}^{T}\left(\left(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right)^\top\boldsymbol{x}_t\right)^2 \le \sum_{t=1}^{T}\left\|\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right\|^2_{\widehat{\mathbf{H}}_{t-1}}\left\|\boldsymbol{x}_t\right\|^2_{\widehat{\mathbf{H}}^{-1}_{t-1}}$$

Following the decomposition, we bound this term by (a) using the confidence bound to control the concentration of $\boldsymbol{\beta}_t$ around $\boldsymbol{\beta}$, and (b) by bounding the sum of feature norms.

Step a: Confidence Intervals of $\boldsymbol{\beta}_t$. As t increases, the O2SLS estimates concentrate around the true parameter $\boldsymbol{\beta}$ (Lemma 4.1). Thus, we derive a confidence interval at step $t-1$ such that with probability at least $1-\delta$,

$$\|\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\|^2_{\widehat{\mathbf{H}}_{t-1}} \le \mathfrak{b}_{t-1}(\delta) = \frac{dL_\eta^2}{4}\log\left(\frac{1+tL_z^2/\lambda}{\delta}\right).$$

This inequality obtained by leveraging the concentration properties of the vector-valued martingales (Theorem E.1).

Step b: Bounding the Second Stage Features with IV norms. Now, we need to bound sum of the feature norms. First, we apply Equation (3) and triangle inequality to get

$$\sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\widehat{\mathbf{H}}^{-1}_{t-1}} \leq \sum_{t=1}^{T} \left\|\boldsymbol{\Theta}^\top \boldsymbol{z}_t\right\|^2_{\widehat{\mathbf{H}}^{-1}_{\boldsymbol{z},t-1}} + \sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^2_{\widehat{\mathbf{H}}^{-1}_{t-1}}$$

From the properties of matrix norm, the sum of IV norms is upper bounded by the product of the minimum eigenvalue of estimated first-stage parameters, maximum eigenvalue of the true first stage parameters, and the sum of minimums eigenvalues of the design matrix of the first stage, i.e. $\sum_{t=1}^{T} \left\|\mathbf{G}^{-1}_{\boldsymbol{z},t-1}\right\|_2$. We bound each of these terms individually to obtain a bound $C_1^2 \left(\frac{C_3+1}{\lambda} + \frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})/2}\right)$. On the other hand, with probability at least $1-\delta$, the norm of first-stage noises is bounded by $dC_2^2 \left(\frac{C_3+1}{\lambda} + \frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})/2}\right)$ (Lemma B.3).

Final step. By combining all together, we conclude that

$$\begin{aligned}
(\bullet 1 \bullet) &\leq \sum_{t=1}^{T} \mathfrak{b}_{t-1}(\delta) \left(\|\boldsymbol{z}_t\|^2_{\widehat{\mathbf{H}}^{-1}_{\boldsymbol{z},t-1}} + \|\boldsymbol{\epsilon}_t\|^2_{\widehat{\mathbf{H}}^{-1}_{t-1}} \right) \\
&\leq \underbrace{\mathfrak{b}_{T-1}(\delta)}_{\substack{\text{Estimation} \\ \mathcal{O}(d \log T)}} \Bigg(\underbrace{d(C_1^2 + C_2^2) \left(\frac{C_3+1}{\lambda} + \frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})/2} \right)}_{\substack{\text{Second Stage Feature norm} \\ \mathcal{O}(d \log T)}} \Bigg)
\end{aligned}$$

The last inequality is true as $\mathfrak{b}_{t-1}$ is non-decreasing in t. Thus, we conclude that Term $(\bullet 1 \bullet)$ is $\mathcal{O}(d^2 \log^2 T)$.

Term 2: Coupling of First-stage Data and Second-stage Parameter Estimation. Now, we bound the second term using concentration inequalities of martingales, similar to the ones used to derive a uniform high probability bound for the confidence intervals.

We observe that $w_t \triangleq \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \Theta^\top \boldsymbol{z}_s$ is a martingale with respect to the filtration $\mathcal{F}_{t-1} = \sigma\left(\boldsymbol{z}_1, \boldsymbol{\epsilon}_1, \eta_1, \ldots, \boldsymbol{z}_{t-1}, \boldsymbol{\epsilon}_{t-1}, \eta_{t-1}, \boldsymbol{z}_t\right)$. We also note that w_t is $\mathcal{F}_{t-1}-$measurable since $\boldsymbol{\beta}_{t-1}$ and $\boldsymbol{z}_t$ are too. Thus, by Theorem E.2 for scalar-valued martingale concentration, we get that with probability at least $1-\delta$

$$\begin{aligned}
(\bullet 2 \bullet) &\leq \left| \sum_{t=1}^{T} \eta_t w_t \right| \\
&\leq \sqrt{2 \left(1 + L_\eta^2 \sum_{t=1}^{T} w_t^2\right) \log\left(\frac{\sqrt{1 + L_\eta^2 \sum_{t=1}^{T} w_t^2}}{\delta} \right)}.
\end{aligned}$$

Now, we focus on bounding the quantity appearing under square root. Thus, by applying Cauchy-Schwarz inequality and a reasoning similar to bounding Term $(\bullet 1 \bullet)$, we get

$$\sum_{t=1}^{T} w_t^2 \leq \mathfrak{b}_{T-1}(\delta) C_1^2 \left(\frac{C_3+1}{\lambda} + \frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})/2} \right).$$

Hence, we get that Term (•2•) is bounded by

$$\underbrace{\sqrt{\mathfrak{b}_{T-1}(\delta)}}_{\substack{\text{Estimation}\\ \mathcal{O}(\sqrt{d\log T})}}\underbrace{\sqrt{C_1^2\left(\frac{C_3+1}{\lambda}+\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})/2}\right)\log\left(\frac{\log T}{\delta}\right)}}_{\substack{\text{First Stage Feature norm}\\ \mathcal{O}(\sqrt{\log T})}},$$

Term (•2•) is $\mathcal{O}(\sqrt{d}\log(T))$ ignoring the log log terms.

Term 3: Coupling of First- and Second-stage Noises. Finally, we bound Term (•3•) containing the correlation between the first- and second-stage noise. This term is referred as the self-fulfilling bias (Li et al., 2021). Similar to the previous terms, we first decouple it into two components

$$\sum_{t=1}^{T}\eta_t\left(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right)^\top\boldsymbol{\epsilon}_t\leq\sum_{t=1}^{T}\left\|\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right\|_{\widehat{\mathbf{H}}_{t-1}}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\mathbf{H}}_{t-1}^{-1}}$$

We know that $\left\|\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right\|_{\widehat{\mathbf{H}}_{t-1}}\leq\sqrt{\mathfrak{b}_{T-1}(\delta)}$ with probability at least $1-\delta$. Now, to bound the norms of the products between η_t and $\boldsymbol{\epsilon}_t$ with respect to $\widehat{\mathbf{H}}_{t-1}^{-1}$, we show that

$$\sum_{t=1}^{T}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\mathbf{H}}_{t-1}^{-1}}\leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}\sum_{t=0}^{T-1}|\eta_{t+1}|\|\boldsymbol{\epsilon}_{t+1}\|_2\sqrt{\lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t}^{-1}\right)}$$

We further show that the minimum eigenvalue of the first stage design matrix grows $\Omega(t)$ (Lemma D.2). Thus, we get that $\lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t}^{-1}\right)$ is $\mathcal{O}(\frac{1}{\sqrt{t}})$, and our aim is to bound $\sum_{t=0}^{T-1}\frac{|\eta_{t+1}|\|\boldsymbol{\epsilon}_{t+1}\|_2}{\sqrt{t}}$. We apply Chernoff bound to obtain

$$\sum_{t=0}^{T-1}\frac{|\eta_{t+1}|\|\boldsymbol{\epsilon}_{t+1}\|_2}{\sqrt{t}}\leq 2L_{\boldsymbol{\epsilon}}L_\eta\sqrt{2d\log\left(\frac{1}{\delta}\right)(\log(T)+1)}.$$

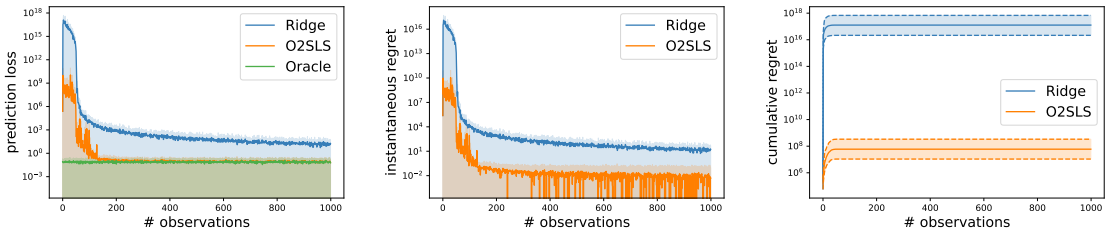
Substituting this result along with the confidence bound, we observe that Term (•3•) is $\mathcal{O}(d\log T)$.

Thus, *we conclude that Term* (•1•) *is the dominant term imposing a regret bound* $\mathcal{O}(d^2\log^2 T)$ *for* O2SLS.

4.2 Experimental Analysis

Now, we aim to compare empirical performance of the online ridge regression (Ridge) with O2SLS estimator. The Ridge estimator is given by $\boldsymbol{\beta}_t^{\mathsf{Ridge}}=\left(\mathbf{X}_t^\top\mathbf{X}_t+\lambda_{\mathsf{Ridge}}\mathbb{I}_d\right)^{-1}\mathbf{X}_t^T\boldsymbol{y}_t$. We choose the regularisation parameters to be $\lambda_{\mathsf{Ridge}}=10^{-3}$ and $\lambda=10^{-3}$. We deploy the experiments in Python3 on a single Intel(R) Core(TM) i7-8665U CPU@1.90GHz. An experiment consisting of 100 runs of an algorithm takes ~ 10 minutes.

We denote the normal distribution as $\mathcal{N}$ and a truncated normal distribution as $\mathcal{N}^{\text{trunc}}$ with a set of of truncation parameters $\vec{L}$. This implies that the normal (multivariate normal) distribution is bounded in the interval $[-L,+L]$ (in an hypercube $[-L,+L]^d$).



(a) Online loss (b) Instantaneous regret (c) Cumulative regret

Figure 2: Performance of O2SLS for online linear regression compared with online ridge regression, and an oracle. O2SLS incurs lower regret than online ridge and also achieves the same prediction loss as the oracle asymptotically. We average our curves over 100 samples and the shaded area indicates one standard deviation. The $y-$axis is in logarithmic scale.

We consider the data is generated using $\boldsymbol{\Theta}_{i,j} \overset{i.i.d}{\sim} \mathcal{N}(0,1), \boldsymbol{z}_t \overset{i.i.d}{\sim} \mathcal{N}_{50}^{\text{trunc}}(0, \mathbb{I}_{50}, 1\vec{0})$, and $\boldsymbol{\epsilon}_t \overset{i.i.d}{\sim} \mathcal{N}_{50}^{\text{trunc}}(0, \mathbb{I}_{50}, 1\vec{0})$. The endogenous noise in the second stage is $\eta_t = (\widetilde{\eta}_t + \sum_{i=1}^{12} \boldsymbol{\epsilon}_{t,i})/13$, where $\widetilde{\eta}_t \overset{i.i.d}{\sim} \mathcal{N}^{\text{trunc}}(0,1,10)$ is a r.v. independent from all the others and $\boldsymbol{\beta} \overset{i.i.d}{\sim} \mathcal{N}_{50}(1\vec{0}, \mathbb{I}_{50})$.

From Figure 2, we observe that in presence of endogeneity the Ridge estimator is not able to recover the true parameter $\boldsymbol{\beta}$, while O2SLS can by 200 observations. Thus, Ridge, which assumes exogeneity, leads to bad online predictions. O2SLS estimator solves this issue. Also, O2SLS achieves almost 10-order less regret by 1000 observations (Fig. 2c).

5. Linear Bandits with Endogeneity

In this section, we formulate Linear Bandits with Endogeneity (LBE) using a two-stage linear model of data generation (Eqn. (3)). Then, we propose an index-based optimistic algorithm, OFUL-IV. Our theoretical analysis in Sec. 5.2 shows that OFUL-IV achieves $\mathcal{O}(d\sqrt{T}\log T)$ regret. In Sec. 5.3, Experimental results show that OFUL-IV achieve lower regret than OFUL that assumes exogeneity.

In bandit setting, we observe $\boldsymbol{x}_t$ and y_t depending on arm (or intervention) $A_t \in \mathcal{A}_t$ drawn at time $t \in \{1, \dots, T\}$.

$$\boldsymbol{x}_t = \boldsymbol{\Theta}^\top \boldsymbol{z}_{t,A_t} + \boldsymbol{\epsilon}_t \qquad \text{(LBE-first)}$$
$$y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t + \eta_t \qquad \text{(LBE-second)}$$

Here, y_t is the reward at round t and each arm a corresponds to *a vector of IVs* $\boldsymbol{z}_{t,a} \in \mathcal{Z}_t \subset \mathbb{R}^d$. Also, a vector of endogenous variables $\boldsymbol{x}_{t,a} \in \mathcal{X}_t \subset \mathbb{R}^d$ is obtained as per (LBE-first). Here, $\mathcal{X}_t$ and $\mathcal{Z}_t$ are sets of IVs and endogenous variables corresponding to $\mathcal{A}_t$. Similar to regression setting, we have two sources of unobserved noises: $\boldsymbol{\epsilon}_t \in \mathbb{R}^d$ are i.i.d. vector error terms at round t which is independent of $\boldsymbol{z}$, and η_t, representing all causes of y_t other than $\boldsymbol{x}_t$. True parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$ are unknown to the agents. This is an extension of the classical stochastic linear bandit (Lattimore and Szepesvári, 2020, Chapter 19). Now, we state the protocol of Linear Bandits with Endogeneity (LBE).

Algorithm 2 OFUL-IV

Input: Initialization parameters $\boldsymbol{\beta}_0, \widehat{\boldsymbol{\Theta}}_0, \mathfrak{b}'_0$
for $t = 1, 2, \ldots, T$ **do**
  Observe $\boldsymbol{z}_{t,a} \in \mathcal{Z}_t$ and $\boldsymbol{x}_{t,a} \in \mathcal{X}_t$ for $a \in \mathcal{A}_t$
  Compute $\boldsymbol{\beta}_{t-1}$ according to Equation (O2SLS)
  Choose action A_t that solves Equation (9)
  Update $\boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}_{t-1}, \widehat{\boldsymbol{\Theta}}_t \leftarrow \widehat{\boldsymbol{\Theta}}_t, \mathfrak{b}'_t \leftarrow \mathfrak{b}'_t$
end for

At each round $t = 1, 2, \ldots, T$, the agent
1. Observes a sample $\boldsymbol{z}_{t,a} \in \mathcal{Z}_t$ and $\boldsymbol{x}_{t,a} \in \mathcal{X}_t$ of contexts for all $a \in \mathcal{A}_t$
2. Chooses an arm $A_t \in \mathcal{A}_t$
3. Obtains the reward y_t computed from (LBE-first)
4. Updates the parameter estimates $\widehat{\boldsymbol{\Theta}}_t$ and $\boldsymbol{\beta}_t$

5.1 OFUL-IV: Algorithm Design

If the agent had full information in hindsight, she could infer the best arm (or intervention) in $\mathcal{A}_t$ as

$$a_t^* = \underset{a \in \mathcal{A}_t}{\operatorname{argmax}} \, \mathbb{E}[\boldsymbol{x}_{t,a}^\top \boldsymbol{\beta}]$$

We denote the corresponding variables as $\boldsymbol{z}_t^*$ and $\boldsymbol{x}_t^*$. Thus, choosing a^* can be shown as choosing $\boldsymbol{z}_t^*$ and $\boldsymbol{x}_t^*$. But the agent does not know them and aims to select $\{a_t\}_{t=1}^T$ leading to minimum regret (Eqn. (4)). Now, we extend the OFUL algorithm minimising regret in linear bandits with exogeneity (Abbasi-Yadkori et al., 2011a). The core idea is that the algorithm maintains a confidence set $\mathcal{B}_{t-1} \subseteq \mathbb{R}^d$ around the parameter $\boldsymbol{\beta}$, which is computed only using the observed data. Then, the algorithm chooses an optimistic estimate of $\widetilde{\boldsymbol{\beta}}_{t-1}$ from that confidence set:

$$\widetilde{\boldsymbol{\beta}}_{t-1} = \underset{\boldsymbol{\beta}' \in \mathcal{B}_{t-1}}{\operatorname{argmax}} \left(\max_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{x}^\top \boldsymbol{\beta}' \right) \tag{8}$$

Then, she chooses the action leading to $\boldsymbol{x}_t = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{x}^\top \widetilde{\boldsymbol{\beta}}_t$, which maximizes the reward according to the estimate $\widetilde{\boldsymbol{\beta}}_t$. In brief, the algorithm chooses the pair $(\boldsymbol{x}_t, \widetilde{\boldsymbol{\beta}}_{t-1}) = \underset{(\boldsymbol{x}, \boldsymbol{\beta}') \in \mathcal{X}_t \times \mathcal{B}_{t-1}}{\operatorname{argmax}} \boldsymbol{x}^\top \boldsymbol{\beta}'$.

In order to tackle endogeneity, we choose to use the O2SLS estimate $\boldsymbol{\beta}_{t-1}$ computed using data observed till $t-1$. Then, we build an ellipsoid $\mathcal{B}_{t-1}$ around it, such that $\mathcal{B}_{t-1} \triangleq \left\{ \boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\mathbf{H}}_t} \leq \sqrt{\mathfrak{b}'_t(\delta)} \right\}$ and $\mathfrak{b}'_t(\delta) \triangleq 2L_\eta^2 \log\left(\frac{\det(\mathbf{G}_{\boldsymbol{z},t})^{1/2} \lambda^{-d/2}}{\delta} \right)$. Given this confidence interval, we optimistically choose the arm

$$A_t = \underset{a \in \mathcal{A}_t}{\operatorname{argmax}} \left\langle \boldsymbol{x}_{t,a}, \boldsymbol{\beta}_{t-1} \right\rangle + \sqrt{\mathfrak{b}'_{t-1}(\delta)} \, \|\boldsymbol{x}_{t,a}\|_{\widehat{\mathbf{H}}_{t-1}^{-1}} . \tag{9}$$

This arm selection index together with the O2SLS estimator yielding $\boldsymbol{\beta}_{t-1}$ construct the OFUL-IV (Algorithm 2).

5.2 Theoretical Analysis

Theorem 5.1. *Under the same assumptions as that of Theorem 4.1, Algorithm 2 incurs a regret*

$$R_T = \mathcal{O}\left(d\sqrt{T\log\left(\frac{T}{\delta}\right)\left(\frac{C_3+1}{\lambda}+\frac{\log(T)+1}{\lambda_{\min}(\mathbf{\Sigma})/2}\right)}\right),$$

with probability $1-\delta$ *and horizon* $T>1$.

Proof Sketch. **Step 1: Optimism.** The regret in this setting $R_T = \sum_{t=1}^{T}\boldsymbol{\beta}^\top \boldsymbol{x}_* - \boldsymbol{\beta}^\top \boldsymbol{x}_t \triangleq \sum_{t=1}^{T} r_t$. Since $(\boldsymbol{x}_t, \widetilde{\boldsymbol{\beta}}_{t-1})$ is optimistic in $\mathcal{X}_t \times \mathcal{B}_t$, and also $\boldsymbol{\beta} \in \mathcal{B}_t$, we obtain

$$r_t \leq (\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta})^\top \boldsymbol{x}_t.$$

Step 2: Decomposition. Now, we decompose regret as

$$(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta})^\top \boldsymbol{x}_t = (\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1})^\top \boldsymbol{x}_t + (\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta})^\top \boldsymbol{x}_t.$$

The first term corresponds to tightness of the confidence interval, while the second term depends on accuracy of the estimate $\boldsymbol{\beta}_{t-1}$.

Step 3: Confidence Bound. Now, we can decouple the impact of parameter and observed data in both the terms using $\left\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1}\right\|_{\widehat{\mathbf{H}}_{t-1}} \|\boldsymbol{x}_t\|_{\widehat{\mathbf{H}}_{t-1}^{-1}}$ and $\left\|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right\|_{\widehat{\mathbf{H}}_{t-1}} \|\boldsymbol{x}_t\|_{\widehat{\mathbf{H}}_{t-1}^{-1}}$, respectively.

By construct of the optimistic confidence interval and concentrations bound of Lemma 4.1, both $\left\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1}\right\|_{\widehat{\mathbf{H}}_{t-1}}$ and $\left\|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right\|_{\widehat{\mathbf{H}}_{t-1}}$ are bounded by $\sqrt{\mathfrak{b}'_{t-1}(\delta)}$. By determinant-trace inequality (Lemma A.1), we get that $\mathfrak{b}'_{T-1}(\delta) \leq \frac{dL_\eta^2}{4}\log\left(\frac{1+TL_z^2/\lambda}{\delta}\right)$.

Final Step. Since the regret $R_T \leq \sqrt{T\sum_{t=1}^{T} r_t^2}$, we obtain

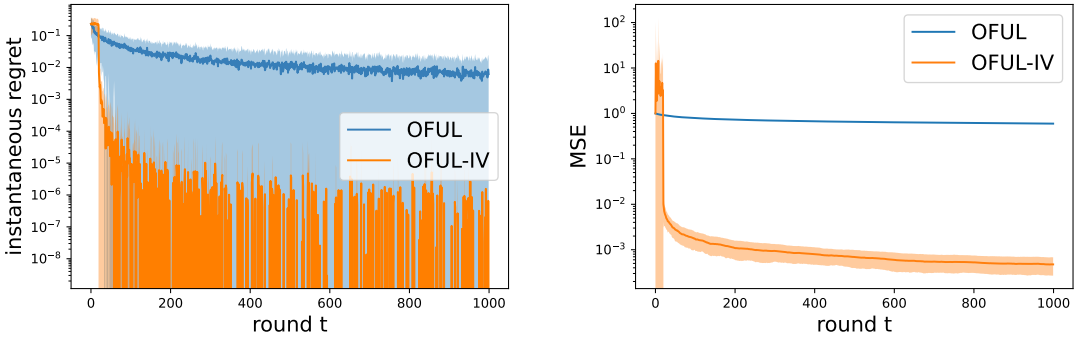
$$R_T \leq L_\eta\sqrt{dT\log\left(\frac{1+TL_z^2/\lambda}{\delta}\right)\left(\sum_{t=1}^{T}\|\boldsymbol{x}_t\|^2_{\widehat{\mathbf{H}}_{t-1}^{-1}}\right)}.$$

Now, we bound the sum of feature norms $\sum_{t=1}^{T}\|\boldsymbol{x}_t\|^2_{\widehat{\mathbf{H}}_{t-1}^{-1}}$ by $d(C_1^2+C_2^2)\left(\frac{C_3+1}{\lambda}+\frac{\log(T)+1}{\lambda_{\min}(\mathbf{\Sigma})/2}\right)$ as we did it in Step 2 while bounding Term ($\bullet 1 \bullet$), which is $\mathcal{O}(d\log T)$.

Thus, we conclude that regret of OFUL-IV is $\mathcal{O}(d\sqrt{T}\log T)$.

5.3 Experimental Analysis

In the LBE setting, the agent chooses actively from a set of arms that correspond to compact and bounded sets of endogenous and instrumental variables. Now, we compare performance of OFUL-IV and OFUL for LBE setting. We implement and run the algorithms in the same setting as before, and with the same regularisation parameters $\lambda_{\mathsf{Ridge}} = 10^{-3}$ and $\lambda = 10^{-3}$. Furthermore, we take $\mathbf{\Theta}_{i,j} \sim \mathcal{N}^{\text{trunc}}(0,1,10), \boldsymbol{z}_{t,a} \overset{i.i.d}{\sim} \mathcal{N}_{50}^{\text{trunc}}(0,\mathbb{I}_{50},\vec{10})$, and $\boldsymbol{\epsilon}_{t,a} \overset{i.i.d}{\sim} \mathcal{N}_{50}^{\text{trunc}}(0,\mathbb{I}_{50},\vec{10})$. The noise in the second stage is $\eta_{t,a} = \frac{1}{13}\left(\widetilde{\eta}_{t,a} + \sum_{i=1}^{12}\boldsymbol{\epsilon}_{t,a,i}\right)$. Here, $\widetilde{\eta}_{t,a} \overset{i.i.d}{\sim} \mathcal{N}^{\text{trunc}}(0,1,10)$ is independent of others, and $\boldsymbol{\beta} \overset{i.i.d}{\sim} \mathcal{N}_{50}(\vec{10},\mathbb{I}_{50})$.



(a) Instantaneous regret

(b) MSE w.r.t. $\boldsymbol{\beta}$

Figure 3: We compare instantaneous regrets (left) of OFUL algorithm and OFUL-IV in a linear bandit setting. We also show the MSE between the parameters estimated by the two algorithms and the true parameter $\boldsymbol{\beta}$ (right). OFUL-IV incurs lower instantaneous regret and MSE loss.

OFUL builds a confidence ellipsoid (Ouhamma et al., 2021) centered at $\boldsymbol{\beta}_{\mathsf{Ridge},t}$ to concentrate around $\boldsymbol{\beta}$, while OFUL-IV uses O2SLS to build an accurate estimate. The estimates obtained by OFUL-IV achieves 3-order less error than those of OFUL (Fig. 3b). Thus, OFUL-IV leads to lower regret than OFUL for linear bandits with endogeneity (Fig. 3a).

6. Conclusions and Future Works

Accuracy of the existing online linear regression estimators depend on independence of the covariates and the noise. In this paper, we study the setting where this assumption is violated. We propose a theoretical analysis of online 2SLS algorithm, which is known to produce unbiased estimates of parameters under endogeneity in the offline setting. We show that online 2SLS (O2SLS) achieves $\mathcal{O}(d^2 \log^2 T)$ regret bound uniformly for all input features. This is $d \log T$ higher than online linear regression with exogeneity assumption (Gaillard et al., 2019; Ouhamma et al., 2021). This is the cost paid by O2SLS due to endogeneity. Following that, we study stochastic linear bandits with endogeneity. We propose OFUL-IV that uses O2SLS to estimate the model parameters. We show that OFUL-IV achieves $\mathcal{O}(d\sqrt{T} \log T)$ regret. We experimentally validate that O2SLS and OFUL-IV achieve better performance under endogeneity, where other online regression and linear bandit algorithms produce inaccurate estimates. Specifically, O2SLS and OFUL-IV incur almost 10 to 7 order lower regret than others.

For simplicity, we consider the just-identified IVs. In future, we will like to extend our algorithms and analysis to weakly or over-identified IVs (Greene, 2003). Additionally, O2SLS and OFUL-IV work if the IVs are already specified. There has been significant work to identify IVs in offline setting (Newey and Powell, 2003; Chen et al., 2020). Still, it is an open question how optimally IVs can be identified online, while O2SLS is performed simultaneously.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011b.

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2012.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *AISTATS*, 2017.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Peter L Bartlett, Wouter M Koolen, Alan Malek, Eiji Takimoto, and Manfred K Warmuth. Minimax fixed-design linear regression. In *Conference on Learning Theory*, pages 226–239. PMLR, 2015.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5): 2333–2355, 2017.

Pedro Carneiro, James J Heckman, and Edward J Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81, 2011.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Jiafeng Chen, Daniel L Chen, and Greg Lewis. Mostly harmless machine learning: learning optimal instruments in linear iv models. *arXiv preprint arXiv:2011.06158*, 2020.

Dean P Foster. Prediction in the worst case. *The Annals of Statistics*, pages 1084–1090, 1991.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Pierre Gaillard, Sébastien Gerchinovitz, Malo Huard, and Gilles Stoltz. Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss. In *Algorithmic Learning Theory*, pages 404–432. PMLR, 2019.

William H Greene. *Econometric analysis*. Pearson Education India, 2003.

Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, and Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR, 2022.

Elad Hazan and Tomer Koren. Linear regression with limited observation. In *29th International Conference on Machine Learning, ICML 2012*, pages 807–814, 2012.

MA Hernan and J Robins. *Causal Inference: What if*. Boca Raton: Chapman & Hill/CRC, 2020.

Nathan Kallus. Instrument-armed bandits. In *Algorithmic Learning Theory*, pages 529–546. PMLR, 2018.

Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

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

Jin Li, Ye Luo, and Xiaowei Zhang. Self-fulfilling bandits: Dynamic selection in algorithmic decision-making. *arXiv preprint arXiv:2108.12547*, 2021.

Nicholas Littlestone, Philip M Long, and Manfred K Warmuth. On-line learning of linear functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 465–475, 1991.

Ruiqi Liu, Zuofeng Shang, and Guang Cheng. On deep instrumental variables estimate. *arXiv preprint arXiv:2004.14954*, 2020.

Magne Mogstad, Alexander Torgovitsky, and Christopher R Walters. The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–98, 2021.

Maria Nareklishvili, Nicholas Polson, and Vadim Sokolov. Deep partial least squares for iv regression. *arXiv preprint arXiv:2207.02612*, 2022.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Reda Ouhamma, Odalric Maillard, and Vianney Perchet. Stochastic online linear regression: the forward algorithm to replace ridge. *arXiv preprint arXiv:2111.01602*, 2021.

Reda Ouhamma, Debabrota Basu, and Odalric-Ambrym Maillard. Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration and planning. *arXiv preprint arXiv:2210.02087*, 2022.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Amandeep Singh, Kartik Hosanagar, and Amit Gandhi. Machine learning instrument variables for causal inference. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 835–836, 2020.

Andrew Stirn and Tony Jebara. Thompson sampling for noncompliant bandits. *arXiv preprint arXiv:1812.00856*, 2018.

Volodya Vovk. Competitive on-line linear regression. *Advances in Neural Information Processing Systems*, 10, 1997.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300, 1940.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *arXiv preprint arXiv:2106.03907*, 2021.

Yuchen Zhu, Limor Gultchin, Arthur Gretton, Matt J Kusner, and Ricardo Silva. Causal inference with treatment measurement error: a nonparametric instrumental variable approach. In *Uncertainty in Artificial Intelligence*, pages 2414–2424. PMLR, 2022.

Appendix

Table of Contents

A. Existing and Useful Results

Notations: We indicate the determinant of matrix $\mathbf{A}$ with $\det(\mathbf{A})$ and its trace with $\mathrm{Tr}(\mathbf{A})$. For a $x \in \mathbb{R}_{\geq 0}$, we indicate the function that takes as input x, and gives as output the least integer greater than or equal to x as $\lceil x \rceil$ (ceiling function). We indicate the identity matrix of dimension d with $\mathbb{I}_d$.

A.1 Norms of Vectors and Matrices

Definition A.1 (ℓ_p-norms)**.** *For a vector $v \in \mathbb{R}^n$, we express its ℓ_p-norm as $\|v\|_p$ for $p \geq 0$. A special case is the Euclidean ℓ_2-norm denoted as $\|\cdot\|_2$, which is induced by classical scalar product on $\mathbb{R}^n$ denoted by $\langle \cdot, \cdot \rangle$.*

Given a rectangular matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ with $n \geq m$, we write its ordered singular values as

$$\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_m(\mathbf{A}) = \sigma_{\min}(\mathbf{A}) \geq 0$$

The minimum and maximum singular values have the variational characterization

$$\sigma_{\max}(\mathbf{A}) = \max_{\boldsymbol{v} \in \mathbb{S}^{m-1}} \|\mathbf{A}\boldsymbol{v}\|_2 \quad \text{and} \quad \sigma_{\min}(\mathbf{A}) = \min_{\boldsymbol{v} \in \mathbb{S}^{m-1}} \|\mathbf{A}\boldsymbol{v}\|_2,$$

where $\mathbb{S}^{d-1} \triangleq \left\{ \boldsymbol{v} \in \mathbb{R}^d \mid \|\boldsymbol{v}\|_2 = 1 \right\}$ is the Euclidean unit sphere in $\mathbb{R}^d$.

Definition A.2 (ℓ_2-operator norm)**.** *The spectral or ℓ_2-operator norm of $\mathbf{A}$ is defined as*

$$|||\mathbf{A}|||_2 \triangleq \sigma_{\max}(\mathbf{A}) \,. \tag{10}$$

Since covariance matrices are symmetric, we also focus on the set of symmetric matrices in $\mathbb{R}^d$, denoted $\mathcal{S}^{d \times d} = \left\{ \mathbf{Q} \in \mathbb{R}^{d \times d} \mid \mathbf{Q} = \mathbf{Q}^{\mathrm{T}} \right\}$, as well as the subset of positive semidefinite matrices given by

$$\mathcal{S}_+^{d \times d} \triangleq \left\{ \mathbf{Q} \in \mathcal{S}^{d \times d} \mid \mathbf{Q} \succeq 0 \right\}.$$

From standard linear algebra, we recall the facts that any matrix $\mathbf{Q} \in \mathcal{S}^{d \times d}$ is diagonalizable via a unitary transformation, and we use $\lambda(\mathbf{Q}) \in \mathbb{R}^d$ to denote its vector of eigenvalues, ordered as

$$\lambda_{\max}(\mathbf{Q}) = \lambda_1(\mathbf{Q}) \geq \lambda_2(\mathbf{Q}) \geq \cdots \geq \lambda_d(\mathbf{Q}) = \lambda_{\min}(\mathbf{Q}).$$

Note that a matrix $\mathbf{Q}$ is positive semidefinite-written $\mathbf{Q} \succeq 0$ for short-if and only if $\lambda_{\min}(\mathbf{Q}) \geq 0$.

Remark A.1 (Rayleigh-Ritz variational characterization of eigenvalues)**.** *We remind also the Rayleigh-Ritz variational characterization of the minimum and maximum eigenvalues-namely*

$$\lambda_{\max}(\mathbf{Q}) = \max_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \boldsymbol{v}^\top \mathbf{Q} \boldsymbol{v} \quad \textit{and} \quad \lambda_{\min}(\mathbf{Q}) = \min_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \boldsymbol{v}^\top \mathbf{Q} \boldsymbol{v}.$$

Remark A.2. *For any symmetric matrix $\mathbf{Q}$, the ℓ_2-operator norm can be written as*

$$|||\mathbf{Q}|||_2 = \max\left\{ \lambda_{\max}(\mathbf{Q}), |\lambda_{\min}(\mathbf{Q})| \right\},$$

by virtue of which it inherits the variational representation $|||\mathbf{Q}|||_2 = \max_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \left| \boldsymbol{v}^\top \mathbf{Q} \boldsymbol{v} \right|$.

Corollary A.1. *Given a rectangular matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $n \geq m$, suppose that we define the m dimensional symmetric matrix $\mathbf{R} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$. We then have the relationship*

$$\lambda_j(\mathbf{R}) = (\sigma_j(\mathbf{A}))^2 \qquad \text{for } j = 1, \ldots, m$$

We now introduce norms that are induced by positive semi-definite matrices in the following way.

Definition A.3. *For any vector $y \in \mathbb{R}^n$ and matrix $\mathbf{A} \in \mathcal{S}_+^{n\times n}$, let us define the norm $\|\boldsymbol{y}\|_{\mathbf{A}} \triangleq \sqrt{\boldsymbol{y}^T\mathbf{A}\boldsymbol{y}} = \sqrt{\langle \boldsymbol{y}, \mathbf{A}\boldsymbol{y}\rangle}$.*

Throughout the paper we will need often to bound matrix induced norms using ℓ_2-norms for operators, the following results shows how this can be done easily for generic matrices. We specialize this result as we need in the text.

Proposition A.1. *Take $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, with $\mathbf{B} \succeq 0$ positive semi-definite and $\boldsymbol{x} \in \mathbb{R}^n$*

$$\|\mathbf{A}\boldsymbol{x}\|_{\mathbf{B}}^2 = \|\boldsymbol{x}\|_{\mathbf{A}^\top\mathbf{B}\mathbf{A}}^2 \leq |||\mathbf{B}|||_2 |||\mathbf{A}|||_2^2 \|\boldsymbol{x}\|_2^2 \tag{11}$$

Proof. The equality holds since we can rewrite

$$\|\mathbf{A}\boldsymbol{x}\|_{\mathbf{B}}^2 = \langle \mathbf{A}\boldsymbol{x}, \mathbf{B}\mathbf{A}\boldsymbol{x}\rangle = \langle \boldsymbol{x}, \mathbf{A}^\top\mathbf{B}\mathbf{A}\boldsymbol{x}\rangle = \|\boldsymbol{x}\|_{\mathbf{A}^\top\mathbf{B}\mathbf{A}}^2. \tag{12}$$

The inequality follows by the definition of ℓ_2-norms, where we further substitute $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, to get

$$\langle \mathbf{A}\boldsymbol{x}, \mathbf{B}\mathbf{A}\boldsymbol{x}\rangle = \frac{\langle \mathbf{A}\boldsymbol{x}, \mathbf{B}\mathbf{A}\boldsymbol{x}\rangle}{\|\mathbf{A}\boldsymbol{x}\|_2^2}\frac{\|\mathbf{A}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\|\boldsymbol{x}\|_2^2 = \frac{\langle \boldsymbol{y}, \mathbf{B}\boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2^2}\frac{\|\mathbf{A}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\|\boldsymbol{x}\|_2^2 \leq |||\mathbf{B}|||_2 |||\mathbf{A}|||_2^2 \|\boldsymbol{x}\|_2^2. \tag{13}$$

We note that the inequality holds trivially for x in the null space of $\mathbf{A}$, therefore, in the previous case, we can safely divide by $\|\mathbf{A}\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_2$. □

A.2 Technical Lemmas

Lemma A.1 (Determinant-Trace Inequality)**.** *Suppose $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|\boldsymbol{z}_s\|_2 \leq L_z$. Let $\mathbf{G}_{\boldsymbol{z},t} = \lambda\mathbb{I} + \sum_{s=1}^t \boldsymbol{z}_s\boldsymbol{z}_s^\top$ for some $\lambda > 0$. Then,*

$$\det\left(\mathbf{G}_{\boldsymbol{z},t}\right) \leq \left(\lambda + tL_z^2/d\right)^d \tag{14}$$

Proof. Let $\alpha_1, \alpha_2, \ldots, \alpha_d$ be the eigenvalues of $\mathbf{G}_{\boldsymbol{z},t}$. Since $\mathbf{G}_{\boldsymbol{z},t}$ is positive definite, its eigenvalues are positive. Also, note that $\det(\mathbf{G}_{\boldsymbol{z},t}) = \prod_{s=1}^d \alpha_s$ and $\mathrm{Tr}\left(\mathbf{G}_{\boldsymbol{z},t}\right) = \sum_{s=1}^d \alpha_s$. By inequality of arithmetic and geometric means,

$$\sqrt[d]{\alpha_1\alpha_2\cdots\alpha_d} \leq \frac{\alpha_1 + \alpha_2 + \cdots + \alpha_d}{d}.$$

Therefore, $\det\left(\mathbf{G}_{\boldsymbol{z},t}\right) \leq \left(\mathrm{Tr}\left(\mathbf{G}_{\boldsymbol{z},t}\right)/d\right)^d$.

Now, it remains to upper bound the trace:

$$\mathrm{Tr}\left(\mathbf{G}_{\boldsymbol{z},t}\right) = \mathrm{Tr}(\lambda\mathbb{I}_d) + \sum_{s=1}^t \mathrm{Tr}\left(\boldsymbol{z}_s\boldsymbol{z}_s^\top\right) = d\lambda + \sum_{s=1}^t \|\boldsymbol{z}_s\|_2^2 \leq d\lambda + tL_z^2$$

and the lemma follows. □

B. Regret Analysis for Online IV Regression: O2SLS

In this section, we elaborate the proofs and techniques to bound the regret of O2SLS.

Theorem B.1 (Regret of O2SLS)**.** *If Assumption 3.1 hold true, then for bounded first and second stage noises* $\|\boldsymbol{\epsilon}_t\|_2^2 \leq dL_{\boldsymbol{\epsilon}}^2$, $|\eta_t|^2 \leq L_\eta^2$, *first-stage parameters with bounded* ℓ_2*-norm* $\|\!|\boldsymbol{\Theta}|\!\|_2 \leq L_{\boldsymbol{\Theta}}$, *and bounded IVs* $\|\boldsymbol{z}\|^2 \leq L_z^2$, *the regret of* O2SLS *at step* T *is*

$$\bar{R}_T \leq \mathfrak{b}_{T-1}(\delta)\left(L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2\left(L_{\boldsymbol{\Theta}}^2 L_{\boldsymbol{z}}^2 + dL_{\boldsymbol{\epsilon}}^2\right)\left(\frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\right) \tag{15}$$

$$+\sqrt{\mathfrak{b}_{T-1}(\delta)}\left(L_\eta L_{\widehat{\boldsymbol{\Theta}}^{-1}} L_{\boldsymbol{\Theta}} L_{\boldsymbol{z}}\sqrt{\left(\frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)}\right. \tag{16}$$

$$\left.+\sqrt{d}(L_{\widehat{\boldsymbol{\Theta}}^{-1}} L_{\boldsymbol{\epsilon}} L_\eta)\left(\frac{C_3+1}{\sqrt{\lambda}} + 4\sqrt{\frac{\log\left(1/\delta\right)}{\lambda_{\min}(\boldsymbol{\Sigma})}(\log\left(T\right)+1)}\right)\right) \tag{17}$$

with probability at least $1-\delta \in [0,1)$.

Here, $\mathfrak{b}_{T-1}(\delta)$ *is the confidence interval defined by Lemma 4.1, the estimates of the first-stage parameters are bounded according to Equation* (55)*, i.e.* $\|\!|\widehat{\boldsymbol{\Theta}}_t^{-1}|\!\|_2 \leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}$, C_3 *is defined in Lemma D.2,* $\lambda > 0$ *is the regularization parameter of the first stage and* $\lambda_{\min}(\boldsymbol{\Sigma})$ *is the minimum eigenvalue of the true covariance matrix of IVs, i.e.* $\boldsymbol{\Sigma} \triangleq \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$.

In order to shorten the result and clarify the dimension and T-dependence, we further define $C_1 \triangleq L_{\widehat{\boldsymbol{\Theta}}^{-1}} L_{\boldsymbol{\Theta}} L_{\boldsymbol{z}}$, $C_2 \triangleq L_{\widehat{\boldsymbol{\Theta}}^{-1}} L_{\boldsymbol{\epsilon}}$. We present the shorten result with these constants in Theorem 4.1. Since these constants are d- and T-independent, and $\mathfrak{b}_{T-1}(\delta)$ is $\mathcal{O}\left(d\log(T)\right)$, we obtain that the regret of O2SLS is $\mathcal{O}\left(d^2\log^2(T)\right)$.

Before proceeding to the proof, we make a remark on the relationship between sub-Gaussian and bounded random variables (r.v.). We leverage this result throughout our analysis, as we assume the IVs, the covariates, and the noises are bounded random variables.

Remark B.1 (Sub-Gaussianity of Bounded Random Variables)**.** *Bounded random variables, for example in* $[a,b]$*, are sub-Gaussian with parameter* $\frac{|b-a|}{2}$*. In our analysis of* O2SLS *and* OFUL-IV*, we consider that the first- and second-stage noises,* η_t *and* $\boldsymbol{\epsilon}_t$*, are bounded random variables. Specifically,* $\eta_t \in [-L_\eta, +L_\eta]$ *and each component of* $\boldsymbol{\epsilon}_t$*, i.e.* $\boldsymbol{\epsilon}_{t,i} \in [-L_{\boldsymbol{\epsilon}}, +L_{\boldsymbol{\epsilon}}]$ *for* $i \in \{1,\dots,d\}$*. This implies that the second-stage noise* η_t *is* L_η*-sub-Gaussian and the components of the first-stage noise* $\boldsymbol{\epsilon}_{t,i}$ *are* $L_{\boldsymbol{\epsilon}}$*-sub-Gaussians.*

Now, we elaborate the detailed proof of Theorem B.1.

Proof. By Equation (4), the instantaneous regret at step t is

$$\bar{r}_t \triangleq \ell_t\left(\boldsymbol{\beta}_{t-1}\right) - \ell_t\left(\boldsymbol{\beta}\right) = \left(y_t - \boldsymbol{\beta}_{t-1}^\top \boldsymbol{x}_t\right)^2 - \left(y_t - \boldsymbol{\beta}^\top \boldsymbol{x}_t\right)^2 = \left(\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t - \eta_t\right)^2 - \eta_t^2 \tag{18}$$

$$= \left(\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t\right)^2 + 2\eta_t\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t \tag{19}$$

Since $\boldsymbol{x}_t = \boldsymbol{\Theta}^\top \boldsymbol{z}_t + \boldsymbol{\epsilon}_t$ by (First stage), the second term can be rewritten as

$$2\eta_t \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t = 2\eta_t \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{\Theta}^\top \boldsymbol{z}_t + 2\eta_t \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{\epsilon}_t \tag{20}$$

Therefore, the cumulative regret or regret by horizon T is

$$\bar{R}_T = \sum_{t=1}^{T} \bar{r}_t = \underbrace{\sum_{t=1}^{T} \left(\boldsymbol{\beta}_{t-1}^\top \boldsymbol{x}_t - \boldsymbol{\beta}^\top \boldsymbol{x}_t\right)^2}_{(\bullet 1 \bullet)} + 2 \underbrace{\sum_{t=1}^{T} \eta_t \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{\Theta}^\top \boldsymbol{z}_t}_{(\bullet 2 \bullet)} + 2 \underbrace{\sum_{t=1}^{T} \eta_t \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{\epsilon}_t}_{(\bullet 3 \bullet)} \tag{21}$$

The proof proceeds by bounding each of the three terms individually.

Term 1: Second-stage Regression Error. The first term $(\bullet 1 \bullet)$ quantifies the error introduced by the second stage regression. We bound it (a) using the confidence bound to control the concentration of $\boldsymbol{\beta}_t$ around $\boldsymbol{\beta}$, and (b) by bounding the sum of feature norms according to the following decomposition.

Step 1: By applying Cauchy-Schwarz inequality, we first decouple the effect of parameter estimation and the feature norms.

$$\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t \leq \left\|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}} \left\|\boldsymbol{x}_t\right\|_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}} \tag{22}$$

and from Lemma B.1, we get

$$\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t \quad \leq \sqrt{\mathfrak{b}_{t-1}(\delta)} \left\|\boldsymbol{x}_t\right\|_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}}, \tag{23}$$

with probability at least $1 - \delta$. This inequality is due to the fact that O2SLS estimates concentrate around the true parameter $\boldsymbol{\beta}$ as t increases.

Since $\mathfrak{b}_t$ is monotonically increasing in t, by Equation (23),

$$(\bullet 1 \bullet) = \sum_{t=1}^{T} \left(\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}_t\right)^2 \leq \mathfrak{b}_{T-1}(\delta) \sum_{t=1}^{T} \left\|\boldsymbol{x}_t\right\|^2_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}}. \tag{24}$$

For any $T > 1$ (Lemma B.1), the confidence interval at step $T - 1$ is

$$\mathfrak{b}_{T-1}(\delta) \triangleq \frac{dL_\eta^2}{4} \log\left(\frac{1 + (T-1)L_z^2/\lambda}{\delta}\right) \tag{25}$$

Step 2: Now, we need to bound the sum of the feature norms

$$\begin{aligned}
\sum_{t=1}^{T} \left\|\boldsymbol{x}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} &= \sum_{t=1}^{T} \left\|\boldsymbol{\Theta}^\top \boldsymbol{z}_t + \boldsymbol{\epsilon}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} && \text{(Equation (First stage))} \\
&\leq \sum_{t=1}^{T} \left\|\boldsymbol{\Theta}^\top \boldsymbol{z}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} + \sum_{t=1}^{T} \left\|\boldsymbol{\epsilon}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \\
& && \text{(Triangle Inequality)}
\end{aligned}$$

$$= \sum_{t=1}^{T} \left\| \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \right\|_2 \left\| \widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \right\|_2^2 \left(\left\| \boldsymbol{\Theta}^\top \boldsymbol{z}_t \right\|_2^2 + \|\boldsymbol{\epsilon}_t\|_2^2 \right) \quad \text{(Proposition A.1)}$$

$$\leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \sum_{t=1}^{T} \left\| \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \right\|_2 \left(\|\boldsymbol{\Theta}\|_2^2 \|\boldsymbol{z}_t\|_2^2 + d L_{\boldsymbol{\epsilon}}^2 \right) \tag{26}$$

$$\leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \left(L_{\boldsymbol{\Theta}}^2 L_{\boldsymbol{z}}^2 + d L_{\boldsymbol{\epsilon}}^2 \right) \sum_{t=1}^{T} \left\| \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \right\|_2 \tag{27}$$

We bound the remaining summation using Lemma D.2:

$$\sum_{t=1}^{T} \left\| \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \right\|_2 = \sum_{t=0}^{C_3} \lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t}^{-1} \right) + \sum_{t=C_3+1}^{T-1} \lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t}^{-1} \right) \tag{28}$$

$$\leq \frac{C_3+1}{\lambda} + \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})} \sum_{t=C_3+1}^{T-1} \frac{1}{t} \tag{29}$$

$$\leq \frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})} \tag{30}$$

and we obtain

$$\sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1} \right)^{-1}} \leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \left(L_{\boldsymbol{\Theta}}^2 L_{\boldsymbol{z}}^2 + d L_{\boldsymbol{\epsilon}}^2 \right) \left(\frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})} \right). \tag{31}$$

Step 3: By combining Equation (24) and (31), we conclude that

$$(\bullet 1 \bullet) \leq \mathfrak{b}_{T-1}(\delta) \sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},T-1} \widehat{\boldsymbol{\Theta}}_{t-1} \right)^{-1}} \tag{32}$$

$$\leq \underbrace{\frac{d L_\eta^2}{4} \log\left(\frac{1 + (T-1)L_z^2/\lambda}{\delta} \right)}_{\mathcal{O}(d \log T)} \underbrace{L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \left(L_{\boldsymbol{\Theta}}^2 L_{\boldsymbol{z}}^2 + d L_{\boldsymbol{\epsilon}}^2 \right) \left(\frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})} \right)}_{\mathcal{O}(d \log T)} \tag{33}$$

$$= \mathcal{O}\left(d^2 \log^2(T) \right) \tag{34}$$

Term 2: Coupling of First-stage Data and Second-stage Parameter Estimation. Now, we bound $(\bullet 2 \bullet)$ using martingale inequalities similar to the ones used for the confidence intervals to derive a uniform high probability bound.

Step 1: Following Theorem E.2, we define

$$w_s \triangleq \left(\boldsymbol{\beta}_{s-1} - \boldsymbol{\beta} \right)^\top \boldsymbol{\Theta}^\top \boldsymbol{z}_s \quad \text{and} \quad \mathcal{F}_{t-1} \triangleq \sigma\left(\boldsymbol{z}_1, \boldsymbol{\epsilon}_1, \eta_1, \ldots, \boldsymbol{z}_{t-1}, \boldsymbol{\epsilon}_{t-1}, \eta_{t-1}, \boldsymbol{z}_t \right).$$

It is immediate to verify that the hypothesis are satisfied, since w_t is $\mathcal{F}_{t-1}$-measurable as $\boldsymbol{\beta}_{t-1}$ and $\boldsymbol{z}_t$ are too. Bearing in mind this substitution we have

$$\left| \sum_{t=1}^{T} \eta_t w_t \right| \leq \sqrt{ 2 \left(1 + L_\eta^2 \sum_{t=1}^{T} w_t^2 \right) \log\left(\frac{\sqrt{1 + L_\eta^2 \sum_{t=1}^{T} w_t^2}}{\delta} \right) }. \tag{35}$$

with probability at least $1-\delta$.

Step 2: Thus, we proceed like for the first term in the *Step 2* for the previous Term ($\bullet 1 \bullet$):

$$\sum_{t=1}^{T}\left\langle \boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}, \boldsymbol{\Theta}^\top \boldsymbol{z}_t\right\rangle^2 \leq \mathfrak{b}_{T-1}(\delta)\sum_{t=1}^{T}\left\|\boldsymbol{\Theta}^\top \boldsymbol{z}_t\right\|^2_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},T-1}\widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}}$$

(Cauchy-Schwarz and Lemma B.1)

$$\leq \mathfrak{b}_{T-1}(\delta)\sum_{t=1}^{T}\left\|\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right\|_2\left\|\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\right\|_2^2\left\|\boldsymbol{\Theta}^\top \boldsymbol{z}_t\right\|_2^2 \quad \text{(Proposition A.1)}$$

$$\leq \mathfrak{b}_{T-1}(\delta)L^2_{\widehat{\boldsymbol{\Theta}}^{-1}}\sum_{t=1}^{T}\left\|\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right\|_2\|\boldsymbol{\Theta}\|_2^2\|\boldsymbol{z}_t\|_2^2 \tag{36}$$

$$\leq \mathfrak{b}_{T-1}(\delta)L^2_{\widehat{\boldsymbol{\Theta}}^{-1}}L^2_{\boldsymbol{\Theta}}L^2_{\boldsymbol{z}}\sum_{t=1}^{T}\left\|\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right\|_2 \tag{37}$$

$$= \underbrace{\mathfrak{b}_{T-1}(\delta)}_{\mathcal{O}(d\log T)}\underbrace{L^2_{\widehat{\boldsymbol{\Theta}}^{-1}}L^2_{\boldsymbol{\Theta}}L^2_{\boldsymbol{z}}\left(\frac{C_3+1}{\lambda}+2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)}_{\mathcal{O}(\log T)} \tag{38}$$

$$= \mathcal{O}\left(d\log^2 T\right) \tag{39}$$

where in the first inequality we also used the fact that $\mathfrak{b}_{t-1}(\delta)$ is monotonically increasing in t, to take the radii outside the summation.

Step 3: Thus, substituting inside Equation (35), the order of ($\bullet 2 \bullet$) is $\mathcal{O}\left(\sqrt{d}\log(T)\log\left(\frac{\sqrt{d}\log(T)}{\delta}\right)\right)$.

Term 3: Coupling of First- and Second-stage Noises. We bound the term ($\bullet 3 \bullet$) containing the self-fulfilling bias, i.e. the correlation between the first- and second-stage noise.

$$\sum_{t=1}^{T}\eta_t\left(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right)^\top\boldsymbol{\epsilon}_t \leq \sum_{t=1}^{T}\left\|\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}\right\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1}\widehat{\boldsymbol{\Theta}}_{t-1}}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}}$$

(Cauchy-Schwarz)

$$\leq \sqrt{\mathfrak{b}_{T-1}(\delta)}\sum_{t=1}^{T}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \quad \text{(Lemma B.1)}$$

$$\leq \underbrace{\sqrt{\mathfrak{b}_{T-1}(\delta)}}_{\mathcal{O}(\sqrt{d\log T})}\underbrace{\sqrt{d}(L_{\widehat{\boldsymbol{\Theta}}^{-1}}L_{\boldsymbol{\epsilon}}L_{\eta})\left(\frac{C_3+1}{\sqrt{\lambda}}+4\sqrt{\frac{\log\left(1/\delta\right)}{\lambda_{\min}(\boldsymbol{\Sigma})}}(\log\left(T\right)+1)\right)}_{\mathcal{O}(\sqrt{d\log T})}$$

(Lemma B.4)

$$= \mathcal{O}\left(d\log T\right) \tag{40}$$

where in the second inequality we also used the fact that $\mathfrak{b}_{t-1}(\delta)$ is monotonically increasing in t, to take the radii outside the summation.

□

Remark B.2. *In Appendix F, we describe that the first stage regression in* O2SLS *can be expressed as running d independent ridge regressions for each column of* $\boldsymbol{\Theta}$ *(Equation* (94)*). Since the standard analysis of each of the ridge regressions assume independent and sub-Gaussian noise added in the linear model (cf. Theorem F.1; (Ouhamma et al., 2021)), we assume that each component of the first stage noise, i.e.* $\boldsymbol{\epsilon}_{t,i}$*, corresponding to the i-th ridge regression is bounded in* $[-L_{\boldsymbol{\epsilon}}, L_{\boldsymbol{\epsilon}}]$*. Thus, we obtain that* $\|\boldsymbol{\epsilon}_t\|_2^2 = \sum_{i=1}^d \boldsymbol{\epsilon}_{t,i}^2 \le dL_{\boldsymbol{\epsilon}}^2$*. For the rest of the paper, we use this fact that* $\|\boldsymbol{\epsilon}_t\|_2^2$ *is bounded by* $dL_{\boldsymbol{\epsilon}}^2$*.*

Lemma B.1 (Confidence Ellipsoid for the Second-stage Parameters)**.** *Let us define the design matrix to be* $\mathbf{G}_{\boldsymbol{z},t} = \mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d$ *for some* $\lambda > 0$*. Then, for bounded first stage noise* $|\eta_t| \le L_\eta$*, the true parameter* $\boldsymbol{\beta}$ *belongs to the set*

$$\mathcal{E}_t = \left\{ \boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\mathbf{H}}_t} \le \sqrt{\mathfrak{b}_t(\delta)} \right\}, \tag{41}$$

with probability at least $1 - \delta \in (0,1)$*, for all* $t \ge 0$*. Here,* $\mathfrak{b}_t(\delta) \triangleq \frac{dL_\eta^2}{4} \log\left(\frac{1 + tL_z^2/\lambda}{\delta}\right)$.

Proof. We can rewrite

$$\boldsymbol{\beta}_t - \boldsymbol{\beta} = \left(\mathbf{Z}_t^\top \mathbf{X}_t\right)^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t = \widehat{\boldsymbol{\Theta}}_t^{-1} \left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d\right)^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t = \widehat{\boldsymbol{\Theta}}_t^{-1} \mathbf{G}_{\boldsymbol{z},t}^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t. \tag{42}$$

Take $\boldsymbol{x} \in \mathbb{R}^d$ and by the Cauchy–Schwarz inequality we have

$$\overbrace{\boldsymbol{x}^\top \boldsymbol{\beta}_t - \boldsymbol{x}^\top \boldsymbol{\beta}}^{\text{(I)}} = \boldsymbol{x}^\top \widehat{\boldsymbol{\Theta}}_t^{-1} \mathbf{G}_{\boldsymbol{z},t}^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t = \left\langle \widehat{\boldsymbol{\Theta}}_t^{-\top} \boldsymbol{x}, \mathbf{G}_{\boldsymbol{z},t}^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t \right\rangle \tag{43}$$

$$\le \left\| \widehat{\boldsymbol{\Theta}}_t^{-\top} \boldsymbol{x} \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} \left\| \mathbf{G}_{\boldsymbol{z},t}^{-1} \mathbf{Z}_t^\top \boldsymbol{\eta}_t \right\|_{\mathbf{G}_{\boldsymbol{z},t}} = \underbrace{\left\| \widehat{\boldsymbol{\Theta}}_t^{-\top} \boldsymbol{x} \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}}}_{\text{(II)}} \underbrace{\left\| \mathbf{Z}_t^\top \boldsymbol{\eta}_t \right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}}}_{\text{(III)}}. \tag{44}$$

The choice we will make for $\boldsymbol{x}$ is the following

$$\boldsymbol{x} \triangleq \widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t} \widehat{\boldsymbol{\Theta}}_t (\boldsymbol{\beta}_t - \boldsymbol{\beta}) = \widehat{\mathbf{H}}_t (\boldsymbol{\beta}_t - \boldsymbol{\beta}), \tag{45}$$

which leads to the following rewriting for the single terms in which we decomposed our problem.

Term (I): From a simple substitution and the definition of a norm induced by a matrix we have

$$\boldsymbol{x}^\top \boldsymbol{\beta}_t - \boldsymbol{x}^\top \boldsymbol{\beta} = \langle \boldsymbol{x}, \boldsymbol{\beta}_t - \boldsymbol{\beta} \rangle = \left\langle \widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t} \widehat{\boldsymbol{\Theta}}_t (\boldsymbol{\beta}_t - \boldsymbol{\beta}), \boldsymbol{\beta}_t - \boldsymbol{\beta} \right\rangle \tag{46}$$

$$= \|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|^2_{\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t} \widehat{\boldsymbol{\Theta}}_t}. \tag{47}$$

Term (II): First, we rewrite the following term

$$\left\|\widehat{\boldsymbol{\Theta}}_t^{-\top}\boldsymbol{x}\right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} = \left\langle \widehat{\boldsymbol{\Theta}}_t^{-\top}\boldsymbol{x}, \widehat{\boldsymbol{\Theta}}_t^{-\top}\boldsymbol{x}\right\rangle_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} = \boldsymbol{x}^\top \widehat{\boldsymbol{\Theta}}_t^{-1}\mathbf{G}_{\boldsymbol{z},t}^{-1}\widehat{\boldsymbol{\Theta}}_t^{-\top}\boldsymbol{x} = \|\boldsymbol{x}\|_{\widehat{\boldsymbol{\Theta}}_t^{-1}\mathbf{G}_{\boldsymbol{z},t}^{-1}\widehat{\boldsymbol{\Theta}}_t^{-\top}} \quad (48)$$

$$= \|\boldsymbol{x}\|_{\left(\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t\right)^{-1}}, \quad (49)$$

and, once again, we substitute the definition of $\boldsymbol{x}$ in Equation (45):

$$\|\boldsymbol{x}\|_{\left(\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t\right)^{-1}} = \|\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{\beta}_t - \boldsymbol{\beta})\|_{\left(\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t\right)^{-1}} \quad (50)$$

$$= \|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t}. \quad (51)$$

Terms (III): We bound the last term using Theorem E.1 for the first inequality, and Lemma A.1 in the second inequality:

$$\left\|\mathbf{Z}_t^\top \boldsymbol{\eta}_t\right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} = \left\|\sum_{s=1}^{t}\eta_s \boldsymbol{z}_s\right\|_{\mathbf{G}_{\boldsymbol{z},t}^{-1}} \leq \sqrt{2(L_\eta/2)^2 \log\left(\frac{\det\left(\mathbf{G}_{\boldsymbol{z},t}\right)^{1/2}\lambda^{-d/2}}{\delta}\right)} \leq \sqrt{\frac{dL_\eta^2}{4}\log\left(\frac{1+tL_z^2/\lambda}{\delta}\right)}. \quad (52)$$

Finally, from our initial decomposition, dividing on both sides by $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t}$, we get

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t}\widehat{\boldsymbol{\Theta}}_t} \leq \sqrt{\frac{dL_\eta^2}{4}\log\left(\frac{1+tL_z^2/\lambda}{\delta}\right)}. \quad (53)$$

□

Remark B.3. *We note that the ellipsoid bound has the following order in d and t while neglecting the constants:*

$$\mathfrak{b}_t(\delta) = \mathcal{O}\left(d\log(t)\right) \quad (54)$$

Lemma B.2 (Bounding the First-stage Parameters)**.** *Given the relevance condition in Assumption 3.1 and a regularization parameter $\lambda > 0$, we have the following upper bound for the inverse of the estimated parameter (Equation (94)) in the first-stage regression:*

$$\left\|\widehat{\boldsymbol{\Theta}}_t^{-1}\right\|_2 \leq \frac{\lambda + L_{\boldsymbol{z}}^2}{\mathfrak{r}} \triangleq L_{\widehat{\boldsymbol{\Theta}}^{-1}} \quad (55)$$

Proof. By the sub-multiplicativity of the matrix norms we have

$$\left\|\widehat{\boldsymbol{\Theta}}_t^{-1}\right\|_2 = \left\|\left(\mathbf{Z}_t^\top \mathbf{X}_t\right)^{-1}\left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda\mathbb{I}\right)\right\|_2 \quad (56)$$

$$\leq \left\|\left(\mathbf{Z}_t^\top \mathbf{X}_t\right)^{-1}\right\|_2 \left\|\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda\mathbb{I}\right\|_2 \quad (57)$$

Then, by the sub-additivity of the norms

$$\left\|\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}\right\|_2 \leq \left\|\mathbf{Z}_t^\top \mathbf{Z}_t\right\|_2 + \left\|\lambda \mathbb{I}\right\|_2 \leq \max_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \left\langle \boldsymbol{v}, \sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{z}_s^\top \boldsymbol{v} \right\rangle + \lambda \tag{58}$$

$$= \max_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \sum_{s=1}^{t} \langle \boldsymbol{v}, \boldsymbol{z}_s \rangle^2 + \lambda \leq \sum_{s=1}^{t} \|\boldsymbol{z}_s\|_2^2 + \lambda \tag{59}$$

$$\leq t L_{\boldsymbol{z}}^2 + \lambda \tag{60}$$

Then, we note that the quantity $\left\|\left(\mathbf{Z}_t^\top \mathbf{X}_t\right)^{-1}\right\|_2 \leq \frac{1}{t\mathfrak{r}}$ by the definition of relevance, which implies

$$\left\|\widehat{\boldsymbol{\Theta}}_t^{-1}\right\|_2 \leq \frac{t L_{\boldsymbol{z}}^2 + \lambda}{t\mathfrak{r}} = \frac{L_{\boldsymbol{z}}^2 + \lambda/t}{\mathfrak{r}} \leq \frac{L_{\boldsymbol{z}}^2 + \lambda}{\mathfrak{r}} \tag{61}$$

□

Lemma B.3 (Bounding the Impact of First-stage Noise)**.** *For bounded first and second stage noises* $\|\boldsymbol{\epsilon}_t\|_2^2 \leq d L_{\boldsymbol{\epsilon}}^2$, $|\eta_t|^2 < L_\eta^2$, *and first-stage parameter estimates satisfying* $\left\|\widehat{\boldsymbol{\Theta}}_t^{-1}\right\|_2 \leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}$ *(Lemma B.2), we have that*

$$\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \leq d L_{\boldsymbol{\epsilon}}^2 L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \left(\frac{C_3 + 1}{\lambda} + 2 \frac{\log(T) + 1}{\lambda_{\min}(\boldsymbol{\Sigma})} \right). \tag{62}$$

with probability at least $1 - \delta$.

Proof. The proof follows using chain of inequalities from Proposition A.1 and Lemma B.2,

$$\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \leq \sum_{t=1}^{T} \left\|\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right\|_2 \left\|\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\right\|_2^2 \|\boldsymbol{\epsilon}_t\|_2^2 \leq d L_{\boldsymbol{\epsilon}}^2 L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \sum_{t=1}^{T} \lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right) \tag{63}$$

Then, we use the high probability bound that we introduce in Lemma D.2, plus the estimate $\sum_{k=1}^{n} \frac{1}{k} < \log(n) + 1$:

$$d L_{\boldsymbol{\epsilon}}^2 L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \sum_{t=1}^{T} \lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right) \leq d L_{\boldsymbol{\epsilon}}^2 L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \left(\sum_{t=0}^{C_3} \frac{1}{\lambda} + \sum_{t=C_3+1}^{T-1} \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma}) t} \right) \tag{64}$$

$$\leq d L_{\boldsymbol{\epsilon}}^2 L_{\widehat{\boldsymbol{\Theta}}^{-1}}^2 \left(\frac{C_3 + 1}{\lambda} + 2 \frac{\log(T) + 1}{\lambda_{\min}(\boldsymbol{\Sigma})} \right). \tag{65}$$

□

Lemma B.4 (Concentration of Correlated First and Second-stage Noise)**.** *For bounded first- and second-stage noise, and first-stage parameter estimates satisfying* $\left\|\widehat{\boldsymbol{\Theta}}_t^{-1}\right\|_2 \leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}$ *(Lemma B.2), we show that*

$$\sum_{t=1}^{T} \|\eta_t \boldsymbol{\epsilon}_t\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \leq \sqrt{d} L_{\widehat{\boldsymbol{\Theta}}^{-1}} L_{\boldsymbol{\epsilon}} L_\eta \left(\frac{C_3 + 1}{\sqrt{\lambda}} + 4 \sqrt{\frac{\log(1/\delta)}{\lambda_{\min}(\boldsymbol{\Sigma})} (\log(T) + 1)} \right), \tag{66}$$

with probability at least $1 - \delta \in [0, 1)$.

Proof. We exploit again the results of Proposition A.1

$$\sum_{t=1}^{T}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \leq \sum_{t=1}^{T}\sqrt{\left\|\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right\|_2}\left\|\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\right\|_2\|\eta_t\boldsymbol{\epsilon}_t\|_2 = L_{\widehat{\boldsymbol{\Theta}}^{-1}}\sum_{t=1}^{T}|\eta_t|\|\boldsymbol{\epsilon}_t\|_2\sqrt{\lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\right)} \tag{67}$$

Now we concentrate the eigenvalues of the inverse of the design matrix according to Lemma D.2, and with probability at least $1-\delta$

$$\sum_{t=1}^{T}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} = L_{\widehat{\boldsymbol{\Theta}}^{-1}}\left(\sum_{t=0}^{C_3}|\eta_{t+1}|\|\boldsymbol{\epsilon}_{t+1}\|_2\sqrt{\lambda_{\max}\big(\mathbf{G}_{\boldsymbol{z},t}^{-1}\big)} + \sum_{t=C_3+1}^{T-1}|\eta_{t+1}|\|\boldsymbol{\epsilon}_{t+1}\|_2\sqrt{\lambda_{\max}\big(\mathbf{G}_{\boldsymbol{z},t}^{-1}\big)}\right) \tag{68}$$

$$\leq L_{\widehat{\boldsymbol{\Theta}}^{-1}}\left(\sqrt{\frac{d}{\lambda}}(C_3+1)L_{\boldsymbol{\epsilon}}L_\eta + \sqrt{\frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}}\sum_{t=C_3+1}^{T-1}\frac{|\eta_t|\|\boldsymbol{\epsilon}_t\|_2}{\sqrt{t}}\right). \tag{69}$$

We proceed using Chernoff bound on $\sum_{t=1}^{T-1}\frac{|\eta_t|\|\boldsymbol{\epsilon}_t\|_2}{\sqrt{t}}$ which upper bounds the corresponding sum in the second term above. We start noting that by sub-Gaussianity and boundedness of $\eta_t, \boldsymbol{\epsilon}_t$

$$\mathbb{E}\left[e^{\mu\sum_{s=1}^{t}|\eta_s|\|\boldsymbol{\epsilon}_s\|_2/\sqrt{s}}\right] = \mathbb{E}\left[e^{\mu\sum_{s=1}^{t-1}|\eta_s|\|\boldsymbol{\epsilon}_s\|_2/\sqrt{s}}\mathbb{E}\left[e^{\mu|\eta_t|\|\boldsymbol{\epsilon}_t\|_2/\sqrt{t}}\;\middle|\;\mathcal{F}_{t-1}\right]\right] \tag{70}$$

$$\leq e^{2\mu^2L_\eta^2dL_{\boldsymbol{\epsilon}}^2/t}\mathbb{E}\left[e^{\mu\sum_{s=1}^{t-1}|\eta_s|\|\boldsymbol{\epsilon}_s\|_2/\sqrt{s}}\right] \quad \text{(boundedness of } |\eta_s|,\ \|\boldsymbol{\epsilon}_s\|_2\text{)}$$

$$\leq e^{2\mu^2L_\eta^2dL_{\boldsymbol{\epsilon}}^2\sum_{s=1}^{t}1/s} \quad \text{(iterating the same procedure)}$$

$$\leq e^{2\mu^2L_\eta^2dL_{\boldsymbol{\epsilon}}^2(\log(t)+1)}. \tag{71}$$

At this point the probability of devitation of our quantity of interest using the Chernoff's method reads

$$\mathbb{P}\left[\sum_{t=1}^{T-1}\frac{|\eta_t|\|\boldsymbol{\epsilon}_t\|_2}{\sqrt{t}} \geq \delta\right] \leq \inf_{\mu}\left\{e^{2\mu^2L_\eta^2dL_{\boldsymbol{\epsilon}}^2(\log(T)+1)-\mu\delta}\right\} \tag{72}$$

$$= e^{-\frac{\delta^2}{8L_\eta^2dL_{\boldsymbol{\epsilon}}^2(\log(T)+1)}} \quad \text{(the infimum } \mu^* = \delta/4L_\eta^2dL_{\boldsymbol{\epsilon}}^2(\log(T)+1)\text{)}$$

Therefore, with probability bigger than $1-\delta$:

$$\sum_{t=1}^{T}\frac{|\eta_t|\|\boldsymbol{\epsilon}_t\|_2}{\sqrt{t}} \leq 2L_{\boldsymbol{\epsilon}}L_\eta\sqrt{2d\log\left(1/\delta\right)\left(\log\left(T\right)+1\right)}\,. \tag{73}$$

Finally, putting all together we have

$$\sum_{t=1}^{T}\|\eta_t\boldsymbol{\epsilon}_t\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1}\mathbf{G}_{\boldsymbol{z},t-1}^{-1}\widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}} \leq \sqrt{d}L_{\widehat{\boldsymbol{\Theta}}^{-1}}L_{\boldsymbol{\epsilon}}L_\eta\left(\frac{C_3+1}{\sqrt{\lambda}} + 4\sqrt{\frac{\log\left(1/\delta\right)}{\lambda_{\min}(\boldsymbol{\Sigma})}(\log\left(T\right)+1)}\right). \tag{74}$$

□

C. Regret Analysis for IV Linear Bandits: OFUL-IV

Theorem C.1. *Under the same assumptions as that of Theorem 4.1, Algorithm 2 incurs a regret*

$$R_T = \sum_{t=1}^{T} r_t \leq 2\sqrt{T}\sqrt{\mathfrak{b}_{T-1}(\delta)}\sqrt{L^2_{\widehat{\boldsymbol{\Theta}}^{-1}}\left(L^2_{\boldsymbol{\Theta}}L^2_{\boldsymbol{z}} + dL^2_{\boldsymbol{\epsilon}}\right)\left(\frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)} \tag{75}$$

with probability $1-\delta$ *and horizon* $T > 1$.

Proof. The instantaneous regret reads

$$\begin{aligned}
r_t &= \langle \boldsymbol{\beta}, \boldsymbol{x}_* \rangle - \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle && (76)\\
&\leq \left\langle \widetilde{\boldsymbol{\beta}}_{t-1}, \boldsymbol{x}_t \right\rangle - \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle && \text{(since } (\boldsymbol{x}_t, \widetilde{\boldsymbol{\beta}}_{t-1}) \text{ is optimistic inside } \mathcal{X}_t \times \mathcal{B}_t)\\
&= \left\langle \boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}, \boldsymbol{x}_t \right\rangle + \left\langle \widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1}, \boldsymbol{x}_t \right\rangle && \text{(summing and subtracting } \boldsymbol{\beta}_{t-1})\\
&\leq \left\|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}} \left\|\boldsymbol{x}_t\right\|_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}}\\
&\qquad + \left\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1}\right\|_{\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}} \left\|\boldsymbol{x}_t\right\|_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}}\\
&\qquad \text{(Cauchy-Schwarz inequality)}\\
&\leq 2\sqrt{\mathfrak{b}_{t-1}(\delta)} \left\|\boldsymbol{x}_t\right\|_{\left(\widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}\right)^{-1}} && \text{(Lemma B.1)}
\end{aligned}$$

The last inequality uses the concentration of $\boldsymbol{\beta}_t$ around the true value $\boldsymbol{\beta}$, and the fact that we choose $\widetilde{\boldsymbol{\beta}}_{t-1}$ inside $\mathcal{B}_{t-1}$. In both cases, the two norms are bounded by the radius of the ellipsoid, i.e. $\sqrt{\mathfrak{b}_{t-1}(\delta)}$.

Since we already know in this case how to concentrate the sum of the features norms $\sum_{t=1}^{T} \left\|\boldsymbol{x}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}}$ from Equation (31), we bound the regret using Cauchy-Schwarz inequality in the first step to obtain

$$\begin{aligned}
R_T &\leq \sqrt{T\sum_{t=1}^{T} r_t^2} && \text{(Cauchy-Schwarz inequality)}\\
&\leq 2\sqrt{T\sum_{t=1}^{T} \mathfrak{b}_{t-1}(\delta) \left\|\boldsymbol{x}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}}} && (77)\\
&\leq 2\sqrt{T}\sqrt{\mathfrak{b}_{T-1}(\delta)}\sqrt{\sum_{t=1}^{T} \left\|\boldsymbol{x}_t\right\|^2_{\widehat{\boldsymbol{\Theta}}_{t-1}^{-1} \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \widehat{\boldsymbol{\Theta}}_{t-1}^{-\top}}} && (78)\\
&\leq 2\sqrt{T}\underbrace{\sqrt{\mathfrak{b}_{T-1}(\delta)}}_{\mathcal{O}(\sqrt{d\log T})}\underbrace{\sqrt{L^2_{\widehat{\boldsymbol{\Theta}}^{-1}}\left(L^2_{\boldsymbol{\Theta}}L^2_{\boldsymbol{z}} + dL^2_{\boldsymbol{\epsilon}}\right)\left(\frac{C_3+1}{\lambda} + 2\frac{\log(T)+1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)}}_{\mathcal{O}(\sqrt{d\log T})} && \text{(Equation (31))}\\
&= \mathcal{O}\left(d\sqrt{T}\log T\right) && (79)
\end{aligned}$$

In the Equation (78), we use the fact that the radius $\mathfrak{b}_{t-1}(\delta)$ is monotonically increasing in t. $\square$

D. Concentration of The Minimum Eigenvalue of The Design Matrix

The aim of the section is to find a concentration result for the minimum eigenvalue of the design matrix, which, in turn, gives us a concentration of the ℓ_2-norm of the inverse of the design matrix $\left\|\left\|\mathbf{G}_{\boldsymbol{z},t}^{-1}\right\|\right\|_2$.

We start by staging two know results that we use in order to derive Lemma D.2. Lemma D.1 is a direct corollary of Weyl's theorem for eigenvalues (see for example Exercise 6.1 in (Wainwright, 2019)).

Lemma D.1. *For two symmetric matrices* $\mathbf{A}$ *and* $\mathbf{B}$

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \left\|\left\|\mathbf{A} - \mathbf{B}\right\|\right\|_2. \tag{80}$$

The following, is a classical concentration result for the covariance matrix using the ℓ_2-norm for matrices, for a proof of this result we refer the reader to Corollary 6.20 in (Wainwright, 2019).

Theorem D.1 (Estimation of covariance matrices)**.** *Let* $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t$ *be i.i.d. zero-mean random vectors with covariance* $\boldsymbol{\Sigma}$ *such that* $\|\boldsymbol{z}_s\|_2 \leq L_{\boldsymbol{z}}$ *almost surely. Then for all* $\delta > 0$*, the sample covariance matrix* $\widehat{\boldsymbol{\Sigma}}_t = \frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_s \boldsymbol{z}_s^\top$ *satisfies*

$$\mathbb{P}\left[\left\|\left\|\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right\|\right\|_2 \geq \delta\right] \leq 2d \exp\left(-\frac{t\delta^2}{(2L_{\boldsymbol{z}}^2 \left\|\left\|\boldsymbol{\Sigma}\right\|\right\|_2 + \delta)}\right), \tag{81}$$

this means that with probability at least $1 - \delta$

$$\left\|\left\|\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right\|\right\|_2 \leq \frac{4L_{\boldsymbol{z}}^2}{t}\log\left(\frac{2d}{\delta}\right) + 2\sqrt{\frac{2L_{\boldsymbol{z}}^2}{t}\log\left(\frac{2d}{\delta}\right)\left\|\left\|\boldsymbol{\Sigma}\right\|\right\|_2}. \tag{82}$$

Now, we use this concentration bound together with the bound on the difference of the minimum eigenvalues of two symmetric matrices in order to bound the maximum eigenvalue of the inverse of the design matrix.

Lemma D.2 (Well-behavedness of First-stage Design Matrix)**.** *Let* $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t$ *be i.i.d. zero-mean random vectors with covariance* $\boldsymbol{\Sigma}$ *such that* $\|\boldsymbol{z}_s\|_2 \leq L_{\boldsymbol{z}}$ *almost surely. We denote the regularized design matrix as* $\mathbf{G}_{\boldsymbol{z},t} = \lambda \mathbb{I}_d + \sum_{s=1}^t \boldsymbol{z}_s \boldsymbol{z}_s^\top$*. For all* $\delta > 0$ *and regularization parameter* $\lambda > 0$*, we observe that*

$$\left\|\left\|\mathbf{G}_{\boldsymbol{z},t}^{-1}\right\|\right\|_2 = \lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t}^{-1}\right) \leq \begin{cases} \frac{1}{\lambda} \text{ if } t \leq C_3 \\ \frac{2}{t\,\lambda_{\min}(\boldsymbol{\Sigma})} \text{ if } t > C_3 \end{cases}. \tag{83}$$

Here, $C_3 > 0$ *is a constant defined by Equation* (91) *and* $\lambda_{\min}(\boldsymbol{\Sigma})$ *is the minimum eigenvalue of the true covariance matrix of* $\boldsymbol{z}$*, i.e.* $\boldsymbol{\Sigma} \triangleq \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$*.*

Proof. First, we aim to find a lower bound for the smallest eigenvalue of the design, matrix where we set the regularization parameter λ to zero. We denote the 'non-regularized' design matrix as $\mathbf{G}_{\boldsymbol{z},t}^{\lambda=0}$. For $t \geq 1$, we observe that $\mathbf{G}_{\boldsymbol{z},t}^{\lambda=0}/t \triangleq \widehat{\boldsymbol{\Sigma}}_t$.

Thus, by applying Equation (80), we obtain

$$\left|\lambda_{\min}\left(\mathbf{G}_{\boldsymbol{z},t}^{\lambda=0}/t\right) - \lambda_{\min}(\boldsymbol{\Sigma})\right| \leq \frac{4L_{\boldsymbol{z}}^2}{t}\log\left(\frac{2d}{\delta}\right) + 2\sqrt{\frac{2L_{\boldsymbol{z}}^2}{t}\log\left(\frac{2d}{\delta}\right)\left\|\boldsymbol{\Sigma}\right\|_2}\,. \tag{84}$$

Further substituting $A \triangleq 2L_{\boldsymbol{z}}^2 \log\left(\frac{2d}{\delta}\right)$ leads to the following lower bound for the minimum eigenvalue

$$\lambda_{\min}\left(\mathbf{G}_{\boldsymbol{z},t}^{\lambda=0}\right) \geq \max\left\{0, t\left(\lambda_{\min}(\boldsymbol{\Sigma}) - 2A/t - 2\sqrt{A\,\lambda_{\max}(\boldsymbol{\Sigma})/t}\right)\right\}. \tag{85}$$

Here, $\lambda_{\max}(\boldsymbol{\Sigma})$ and $\lambda_{\min}(\boldsymbol{\Sigma})$ is the maximum and minimum eigenvalues of the true covariance matrix of $\boldsymbol{z}$, i.e. $\boldsymbol{\Sigma} \triangleq \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$. By well-behavedness assumption of the IV, both of them are positive and bounded reals.

Now, from the variational definition of the minimum eigenvalues, we have

$$\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq \lambda_{\min}\left(\mathbf{G}_{\boldsymbol{z},t}^{\lambda=0}\right) + \lambda,$$

which implies that $\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq \lambda$ for all $t \geq 0$, with equality for $t = 0$. Thus, we have

$$\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq \max\left\{\lambda, \lambda + t\left(\lambda_{\min}(\boldsymbol{\Sigma}) - 2A/t - 2\sqrt{A\,\lambda_{\max}(\boldsymbol{\Sigma})/t}\right)\right\}. \tag{86}$$

Let us consider the second term inside maximum of Equation (86), and we split it in the following way

$$\lambda + t\,\lambda_{\min}(\boldsymbol{\Sigma}) - 2A - 2\sqrt{tA\,\lambda_{\max}(\boldsymbol{\Sigma})} = t\underbrace{\frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2}}_{\text{Term (A)}} + \underbrace{\left(\frac{t}{2}\lambda_{\min}(\boldsymbol{\Sigma}) - 2\sqrt{tA\,\lambda_{\max}(\boldsymbol{\Sigma})} + \lambda - 2A\right)}_{\text{Term (B)}} \tag{87}$$

Now we study for which values Term (B) is non-negative. The corresponding second order polynomial equation is obtained substituting $u = \sqrt{t}$, and reads

$$u^2\,\lambda_{\min}(\boldsymbol{\Sigma}) - 4u\sqrt{A\,\lambda_{\max}(\boldsymbol{\Sigma})} + 2(\lambda - 2A) = 0\,, \tag{88}$$

which has two solutions given by

$$u_\pm = \frac{2\sqrt{A\,\lambda_{\max}(\boldsymbol{\Sigma})} \pm \sqrt{4A\,\lambda_{\max}(\boldsymbol{\Sigma}) + 2(2A - \lambda)\,\lambda_{\min}(\boldsymbol{\Sigma})}}{\lambda_{\min}(\boldsymbol{\Sigma})}\,. \tag{89}$$

In particular for $t > \lceil u_+ \rceil$, Term (A) ≥ 0, and Equation (86) reads

$$\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq \max\left\{\lambda, t\,\lambda_{\min}(\boldsymbol{\Sigma})/2\right\}. \tag{90}$$

Therefore, for $t > \lceil 2\lambda/\lambda_{\min}(\boldsymbol{\Sigma})\rceil$ and $t > \lceil u_+ \rceil$ we have that $\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq t\,\lambda_{\min}(\boldsymbol{\Sigma})/2$.

Putting the results together, we conclude that

$$\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq t\,\lambda_{\min}(\boldsymbol{\Sigma})/2 \quad \text{for} \quad t > C_3 \triangleq \max\left\{\lceil 2\lambda/\lambda_{\min}(\boldsymbol{\Sigma})\rceil, \lceil u_+ \rceil\right\}, \tag{91}$$

while for $t \leq C_3$, we retain the trivial lower bound of the minimum eigenvalue, i.e. λ.

In summary, we have

$$\lambda_{\min}(\mathbf{G}_{\boldsymbol{z},t}) \geq \begin{cases} \lambda \text{ if } t \leq C_3 \\ t\,\lambda_{\min}(\boldsymbol{\Sigma})/2 \text{ if } t > C_3 \end{cases} \iff \lambda_{\max}\left(\mathbf{G}_{\boldsymbol{z},t}^{-1}\right) \leq \begin{cases} \frac{1}{\lambda} \text{ if } t \leq C_3 \\ \frac{2}{t\,\lambda_{\min}(\boldsymbol{\Sigma})} \text{ if } t > C_3 \end{cases}. \tag{92}$$

□

E. Concentration of Scalar and Vector-valued Martingales

We look for deviations of the vector martingales $\sum_{s=1}^{t} \eta_s \boldsymbol{z}_s$ and the scalar valued martingale $\sum_{t=1}^{T} \eta_t \left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\right)^\top \boldsymbol{\Theta}^\top \boldsymbol{z}_t$ from their expected values. These results are required in the proof of Theorem B.1. The first martingale is vector-valued while the second is scalar-valued. For the vector-valued martingale, we want to bound its deviations when its values are weighted by the inverse of its design matrix $\mathbf{G}_{\boldsymbol{z},t}^{-1}$ like it appears in $\|\boldsymbol{s}_t\|^2_{\mathbf{G}_{\boldsymbol{z},t}^{-1}}$. The design matrix $\mathbf{G}_{\boldsymbol{z},t}^{-1}$ is itself derived from the martingale. Hence, it is called the 'self-normalized bound'.

The following theorems were introduced in (Abbasi-Yadkori et al., 2011a, 2012) for the two cases. We state and prove them here for completeness. We leverage the fact that the IVs, the covariates, and the noises are bounded random variables, and thus sub-Gaussian.

E.1 Vector Martingales

Lemma E.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be arbitrary and consider for any $t \geq 0$*

$$m_t^{\boldsymbol{\mu}} \triangleq \prod_{s=1}^{t} \exp\left(\frac{\eta_s \langle \boldsymbol{\mu}, \boldsymbol{z}_s \rangle}{\sigma_2} - \frac{1}{2} \langle \boldsymbol{\mu}, \boldsymbol{z}_s \rangle^2\right).$$

Let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Then, $m_\tau^{\boldsymbol{\mu}}$ is a.s. well-defined and $\mathbb{E}\left[m_\tau^{\boldsymbol{\mu}}\right] \leq 1$.

Proof. We claim that $\left\{m_t^{\boldsymbol{\mu}}\right\}_{t=0}^{\infty}$ is a supermartingale. Let

$$d_t^{\boldsymbol{\mu}} \triangleq \exp\left(\frac{\eta_t \langle \boldsymbol{\mu}, \boldsymbol{z}_t \rangle}{\sigma_2} - \frac{1}{2} \langle \boldsymbol{\mu}, \boldsymbol{z}_t \rangle^2\right).$$

Observe that by conditional R-sub-Gaussianity of η_t we have $\mathbb{E}\left[d_t^{\boldsymbol{\mu}} \mid \mathcal{F}_{t-1}\right] \leq 1$. Clearly, $d_t^{\boldsymbol{\mu}}$ is $\mathcal{F}_t$-measurable, as is $m_t^{\boldsymbol{\mu}}$. Further,

$$\mathbb{E}\left[m_t^{\boldsymbol{\mu}} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[m_1^{\boldsymbol{\mu}} \cdots d_{t-1}^{\boldsymbol{\mu}} d_t^{\boldsymbol{\mu}} \mid \mathcal{F}_{t-1}\right] = d_1^{\boldsymbol{\mu}} \cdots d_{t-1}^{\boldsymbol{\mu}} \mathbb{E}\left[d_t^{\boldsymbol{\mu}} \mid \mathcal{F}_{t-1}\right] \leq m_{t-1}^{\boldsymbol{\mu}},$$

showing that $\left\{m_t^{\boldsymbol{\mu}}\right\}_{t=0}^{\infty}$ is indeed a supermartingale and in fact $\mathbb{E}\left[m_t^{\boldsymbol{\mu}}\right] \leq 1$.

Now, we argue that $m_\tau^{\boldsymbol{\mu}}$ is well-defined. By the convergence theorem for nonnegative supermartingales, $M_\infty^{\boldsymbol{\mu}} = \lim_{t\to\infty} m_t^{\boldsymbol{\mu}}$ is almost surely well-defined. Hence, $m_\tau^{\boldsymbol{\mu}}$ is indeed well-defined independently of whether $\tau < \infty$ holds or not. Next, we show that $\mathbb{E}\left[m_\tau^{\boldsymbol{\mu}}\right] \leq 1$. For this let $Q_t^{\boldsymbol{\mu}} = M_{\min\{\tau,t\}}^{\boldsymbol{\mu}}$ be a stopped version of $\left(m_t^{\boldsymbol{\mu}}\right)_t$. By Fatou's Lemma, $\mathbb{E}\left[m_\tau^{\boldsymbol{\mu}}\right] = \mathbb{E}\left[\liminf_{t\to\infty} Q_t^{\boldsymbol{\mu}}\right] \leq \liminf_{t\to\infty} \mathbb{E}\left[Q_t^{\boldsymbol{\mu}}\right] \leq 1$, showing that $\mathbb{E}\left[m_\tau^{\boldsymbol{\mu}}\right] \leq 1$ indeed holds. □

Next lemma uses the "method of mixtures" technique, (Lattimore and Szepesvári, 2020) Chapter 20.

Lemma E.2. *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Then, for any $\delta > 0$, with probability $1 - \delta$*

$$\|\boldsymbol{s}_\tau\|^2_{\mathbf{G}_{\boldsymbol{z},\tau}^{-1}} \leq 2\sigma_2^2 \log\left(\frac{\det\left(\mathbf{G}_{\boldsymbol{z},\tau}\right)^{1/2} \lambda^{-d/2}}{\delta}\right).$$

Proof. We decompose $\mathbf{G}_{\boldsymbol{z},t}$ according to the following notation in order to ease the notation

$$\mathbf{G}_{\boldsymbol{z},t} \triangleq \lambda \mathbb{I}_d + \sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{z}_s^\top = \mathbf{V} + \mathbf{V}_t$$

where $\mathbf{V}$ and $\mathbf{V}_t$ are defined by $\mathbf{V} \triangleq \lambda \mathbb{I}_d$ and $\mathbf{V}_t \triangleq \sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{z}_s^\top$. We can rewrite for example $m_t^{\boldsymbol{\mu}}$ as follows $m_t^{\boldsymbol{\mu}} = \exp\left(\frac{\langle \boldsymbol{\mu}, \boldsymbol{s}_t \rangle}{\sigma_2} - \frac{1}{2}\|\boldsymbol{\mu}\|_{\mathbf{V}_t}^2\right)$.

Let $\boldsymbol{\mu}$ be a Gaussian random variable which is independent of all the other random variables and whose covariance is $\mathbf{V}^{-1}$. Define

$$m_t \triangleq \mathbb{E}\left[m_t^{\boldsymbol{\mu}} \mid \mathcal{F}_\infty\right],$$

where $\mathcal{F}_\infty$ is the tail σ-algebra of the filtration i.e. the σ-algebra generated by the union of the all events in the filtration. Clearly, we still have $\mathbb{E}[m_\tau] = \mathbb{E}\left[\mathbb{E}\left[m_\tau^{\boldsymbol{\mu}} \mid \boldsymbol{\mu}\right]\right] \leq 1$. Let us calculate m_t. Let f denote the density of $\boldsymbol{\mu}$ and for a positive definite matrix $\mathbf{P}$ let $c(\mathbf{P}) = \sqrt{(2\pi)^d / \det(\mathbf{P})} = \int \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x}\right) d\boldsymbol{x}$. Then,

$$\begin{aligned}
m_t &= \int_{\mathbb{R}^d} \exp\left(\langle \boldsymbol{\mu}, \boldsymbol{s}_t \rangle - \frac{1}{2}\|\boldsymbol{\mu}\|_{\mathbf{V}_t}^2\right) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&= \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|\boldsymbol{\mu} - \mathbf{V}_t^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}_t}^2 + \frac{1}{2}\left\|\boldsymbol{s}_t\right\|_{\mathbf{V}_t^{-1}}^2\right) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&= \frac{1}{c(\mathbf{V})} \exp\left(\frac{1}{2}\left\|\boldsymbol{s}_t\right\|_{\mathbf{V}_t^{-1}}^2\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\{\left\|\boldsymbol{\mu} - \mathbf{V}_t^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}_t}^2 + \|\boldsymbol{\mu}\|_{\mathbf{V}}^2\right\}\right) d\boldsymbol{\mu}
\end{aligned}$$

Elementary calculation shows that if $\mathbf{P}$ is positive semi-definite and $\mathbf{Q}$ is positive definite

$$\|\boldsymbol{x} - a\|_{\mathbf{P}}^2 + \|x\|_{\mathbf{Q}}^2 = \left\|\boldsymbol{x} - (\mathbf{P}+\mathbf{Q})^{-1}\mathbf{P}a\right\|_{\mathbf{P}+\mathbf{Q}}^2 + \|a\|_{\mathbf{P}}^2 - \|\mathbf{P}a\|_{(\mathbf{P}+\mathbf{Q})^{-1}}^2.$$

Therefore,

$$\begin{aligned}
\left\|\boldsymbol{\mu} - \mathbf{V}_t^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}_t}^2 + \|\boldsymbol{\mu}\|_{\mathbf{V}}^2 &= \left\|\boldsymbol{\mu} - (\mathbf{V}+\mathbf{V}_t)^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}+\mathbf{V}_t}^2 + \left\|\mathbf{V}_t^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}_t}^2 - \left\|\boldsymbol{s}_t\right\|_{(\mathbf{V}+\mathbf{V}_t)^{-1}}^2 \\
&= \left\|\boldsymbol{\mu} - (\mathbf{V}+\mathbf{V}_t)^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}+\mathbf{V}_t}^2 + \left\|\boldsymbol{s}_t\right\|_{\mathbf{V}_t^{-1}}^2 - \left\|\boldsymbol{s}_t\right\|_{(\mathbf{V}+\mathbf{V}_t)^{-1}}^2,
\end{aligned}$$

which gives

$$\begin{aligned}
m_t &= \frac{1}{c(\mathbf{V})} \exp\left(\frac{1}{2}\left\|\boldsymbol{s}_t\right\|_{(\mathbf{V}+\mathbf{V}_t)^{-1}}^2\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|\boldsymbol{\mu} - (\mathbf{V}+\mathbf{V}_t)^{-1}\boldsymbol{s}_t\right\|_{\mathbf{V}+\mathbf{V}_t}^2\right) d\boldsymbol{\mu} \\
&= \frac{c(\mathbf{V}+\mathbf{V}_t)}{c(\mathbf{V})} \exp\left(\frac{1}{2}\left\|\boldsymbol{s}_t\right\|_{(\mathbf{V}+\mathbf{V}_t)^{-1}}^2\right) = \left(\frac{\det(\mathbf{V})}{\det(\mathbf{V}+\mathbf{V}_t)}\right)^{1/2} \exp\left(\frac{1}{2}\left\|\boldsymbol{s}_t\right\|_{(\mathbf{V}+\mathbf{V}_t)^{-1}}^2\right) \\
&= \left(\frac{\det(\mathbf{V})}{\det(\mathbf{V}+\mathbf{V}_t)}\right)^{1/2} \exp\left(\frac{1}{2}\left\|\boldsymbol{s}_t\right\|_{(\mathbf{V}+\mathbf{V}_t)^{-1}}^2\right)
\end{aligned}$$

Now, from $\mathbb{E}[m_\tau] \leq 1$, we obtain

$$\begin{aligned}\mathbb{P}\left[\|\boldsymbol{s}_\tau\|^2_{(\mathbf{V}+\mathbf{V}_\tau)^{-1}} > 2\log\left(\frac{\det(\mathbf{V}+\mathbf{V}_\tau)^{1/2}}{\delta\det(\mathbf{V})^{1/2}}\right)\right] &= \mathbb{P}\left[\frac{\exp\left(\frac{1}{2}\|\boldsymbol{s}_\tau\|^2_{(\mathbf{V}+\mathbf{V}_\tau)^{-1}}\right)}{\delta^{-1}\left(\det(\mathbf{V}+\mathbf{V}_\tau)/\det(\mathbf{V})\right)^{\frac{1}{2}}} > 1\right] \\ &\leq \mathbb{E}\left[\frac{\exp\left(\frac{1}{2}\|\boldsymbol{s}_\tau\|^2_{(\mathbf{V}+\mathbf{V}_\tau)^{-1}}\right)}{\delta^{-1}\left(\det(\mathbf{V}+\mathbf{V}_\tau)/\det(\mathbf{V})\right)^{\frac{1}{2}}}\right] \\ &= \mathbb{E}[m_\tau]\,\delta \leq \delta\end{aligned}$$

and substituting back the definition of $\mathbf{G}_{\boldsymbol{z},t}$ gives

$$\mathbb{P}\left[\|\boldsymbol{s}_\tau\|^2_{\mathbf{G}^{-1}_{\boldsymbol{z},\tau}} > 2\log\left(\frac{\det(\mathbf{G}_{\boldsymbol{z},\tau})^{1/2}}{\delta\lambda^{d/2}}\right)\right] \leq \delta.$$

□

Theorem E.1 (Self-Normalized Bound for Vector-Valued Martingales)**.** *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally σ_2-sub-Gaussian for some $\sigma_2 \geq 0$ i.e. $\forall\lambda \in \mathbb{R}$ holds*

$$\mathbb{E}\left[e^{\lambda\eta_t} \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2\sigma_2^2}{2}\right).$$

Let $\{\boldsymbol{z}_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{z}_t$ is $\mathcal{F}_{t-1}$-measurable. For any $t \geq 0$, define $\boldsymbol{s}_t = \sum_{s=1}^t \eta_s\boldsymbol{z}_s$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

$$\|\boldsymbol{s}_t\|^2_{\mathbf{G}^{-1}_{\boldsymbol{z},t}} \leq 2\sigma_2^2\log\left(\frac{\det(\mathbf{G}_{\boldsymbol{z},t})^{1/2}\lambda^{-d/2}}{\delta}\right)$$

Proof. We will use a stopping time construction, which goes back at least to (Freedman, 1975). Define the bad event

$$B_t(\delta) = \left\{\omega \in \Omega : \|\boldsymbol{s}_t\|^2_{\mathbf{G}^{-1}_{\boldsymbol{z},t}} > 2\sigma_2^2\log\left(\frac{\det(\mathbf{G}_{\boldsymbol{z},t})^{1/2}\det(\mathbf{V})^{-1/2}}{\delta}\right)\right\}$$

We are interested in bounding the probability that $\bigcup_{t\geq 0} B_t(\delta)$ happens. Define $\tau(\omega) = \min\{t \geq 0 : \omega \in B_t(\delta)\}$, with the convention that $\min\emptyset = \infty$. Then, τ is a stopping time. Further, $\bigcup_{t\geq 0} B_t(\delta) = \{\omega : \tau(\omega) < \infty\}$ Thus,

$$\begin{aligned}\mathbb{P}\left[\bigcup_{t\geq 0} B_t(\delta)\right] = \mathbb{P}[\tau < \infty] &= \mathbb{P}\left[\|\boldsymbol{s}_\tau\|^2_{\mathbf{G}^{-1}_{\boldsymbol{z},\tau}} > 2\sigma_2^2\log\left(\frac{\det(\mathbf{G}_{\boldsymbol{z},\tau})^{1/2}\det(\mathbf{V})^{-1/2}}{\delta}\right), \tau < \infty\right] \\ &\leq \mathbb{P}\left[\|\boldsymbol{s}_\tau\|^2_{\mathbf{G}^{-1}_{\boldsymbol{z},\tau}} > 2\sigma_2^2\log\left(\frac{\det(\mathbf{G}_{\boldsymbol{z},\tau})^{1/2}\det(\mathbf{V})^{-1/2}}{\delta}\right)\right] \leq \delta.\end{aligned}$$

□

E.2 Scalar Martingales

Lemma E.3. *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration such that w_t is $\mathcal{F}_{t-1}$ measurable and η_t is $\mathcal{F}_t$ measurable and is conditionally σ_2-sub-Gaussian. Let τ be a stopping time w.r.t. to this filtration i.e. the event $\{\tau \leq t\}$ belongs to $\mathcal{F}_t$. The following sequence of random variables is a martingale with respect to $\mathcal{F}_t$: $s_t = \sum_{s=1}^{t} \eta_s w_s$. Furthermore, for any $\delta > 0, \sigma_2 > 0$, with probability at least $1-\delta$:*

$$|s_\tau| \leq \sigma_2 \sqrt{2\left(1/\sigma_2^2 + \sum_{t=1}^{\tau} w_t^2\right) \log\left(\frac{\sqrt{1+\sigma_2^2 \sum_{t=1}^{\tau} w_t^2}}{\delta}\right)}$$

Proof. The fact that it is a martingale follows from the conditional sub-Gaussianity. Then for $\lambda \in \mathbb{R}^d, t > 0$ we define

$$m_t^\lambda = \exp\left(\frac{\lambda\, s_t}{\sigma_2} - \frac{\lambda^2}{2}\sum_{s=1}^{t} w_s^2\right)$$
$$d_t^\lambda = \frac{\lambda \eta_t w_t}{\sigma_2} - \frac{\lambda^2}{2} w_t^2$$

Since τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ we can show that m_τ^λ is well-defined almost surely and $\mathbb{E}\left[m_\tau^\lambda\right] \leq 1$. We start by proving that $\left\{m_t^\lambda\right\}_{t=0}^{\infty}$ is a supermartingale. Clearly, d_t^λ is $\mathcal{F}_t$-measurable, as is m_t^λ. Further,

$$\mathbb{E}\left[m_t^\lambda \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[m_1^\lambda \cdots d_{t-1}^\lambda d_t^\lambda \mid \mathcal{F}_{t-1}\right] = d_1^\lambda \cdots d_{t-1}^\lambda \mathbb{E}\left[d_t^\lambda \mid \mathcal{F}_{t-1}\right] \leq m_{t-1}^\lambda,$$

showing that $\left\{m_t^\lambda\right\}_{t=0}^{\infty}$ is indeed a supermartingale. Next we show that m_τ^λ is always well-defined and $\mathbb{E}\left[m_\tau^\lambda\right] \leq 1$. First define $\widetilde{M} = m_\tau^\lambda$ and note that $\widetilde{M}(\omega) = M_{\tau(\omega)}^\lambda(\omega)$. Thus, when $\tau(\omega) = \infty$, we need to argue about $M_\infty^\lambda(\omega)$. By the convergence theorem for non-negative supermartingales, $\lim_{t\to\infty} m_t^\lambda(\omega)$ is well-defined, which means m_τ^λ is well-defined, independently of whether $\tau < \infty$ holds or not. Now let $Q_t^\lambda = M_{\min\{\tau,t\}}^\lambda$ be a stopped version of m_t^λ. We proceed by using Fatou's Lemma to show that $\mathbb{E}\left[m_\tau^\lambda\right] = \mathbb{E}\left[\liminf_{t\to\infty} Q_t^\lambda\right] \leq \liminf_{t\to\infty} \mathbb{E}\left[Q_t^\lambda\right] \leq 1$.

Let $\Lambda \sim \mathbb{N}\left(0, \sigma_2^2\right)$ be a Gaussian random variable and define $m_t = \mathbb{E}\left[m_t^\Lambda \mid F^\infty\right]$. Clearly, we still have $\mathbb{E}\left[m_t\right] = \mathbb{E}\left[\mathbb{E}\left[m_t^\Lambda \mid \Lambda\right]\right] \leq 1$. Let us calculate m_t. We will need the density λ which is $f(\lambda) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\lambda^2/2\sigma_2^2}$. Now, it is easy to write m_t explicitly

$$m_t = \mathbb{E}\left[m_t^\Lambda \mid \mathcal{F}_\infty\right] = \int_{-\infty}^{\infty} m_t^\lambda f(\lambda) d\lambda = \sqrt{\frac{1}{2\pi\sigma_2^2}} \int_{\infty}^{\infty} \exp\left(\frac{\lambda s_t}{\sigma_2} - \frac{\lambda^2}{2}\sum_{s=1}^{t} w_s^2\right) e^{-\lambda^2/2\sigma_2^2} d\lambda$$
$$= \exp\left(\frac{s_t^2}{2\sigma_2^2\left(1/\sigma_2^2 + \sum_{s=1}^{t} w_s^2\right)}\right) \sqrt{\frac{1}{1+\sigma_2^2\sum_{s=1}^{t} w_s^2}}$$

where we have used that $\int_{-\infty}^{\infty} \exp\left(a\lambda - b\lambda^2\right) = \exp\left(a^2/(4b)\right)\sqrt{\pi/b}$.

To finish the proof, we use Markov's inequality and the fact that $\mathbb{E}[m_\tau] \leq 1$:

$$\begin{aligned}
&\mathbb{P}\left[|s_\tau| \geq R\sqrt{2\left(1/\sigma_2^2 + \sum_{t=1}^{\tau} w_t^2\right)\log\left(\frac{\sqrt{1+\sigma_2^2\sum_{t=1}^{\tau} w_t^2}}{\delta}\right)}\right] \\
&= \mathbb{P}\left[\frac{\left(\sum_{t=1}^{\tau}\eta_t w_t\right)^2}{2\sigma_2^2\left(1/\sigma_2^2+\sum_{t=1}^{\tau} w_t^2\right)} \geq \log\left(\frac{\sqrt{1+\sigma_2^2\sum_{t=1}^{\tau} w_t^2}}{\delta}\right)\right] \\
&= \mathbb{P}\left[\exp\left(\frac{s_\tau^2}{2\sigma_2^2\left(1/\sigma_2^2+\sum_{t=1}^{\tau} w_t^2\right)}\right) \geq \frac{\sqrt{1+\sigma_2^2\sum_{t=1}^{\tau} w_t^2}}{\delta}\right] \\
&= \mathbb{P}\left[m_\tau \geq \frac{1}{\delta}\right] \leq \frac{\mathbb{E}[m_\tau]}{1/\delta} \leq \delta
\end{aligned}$$

□

Theorem E.2 (Self-normalized Bound for Scalar Valued Martingales)**.** *Under the same assumptions as the previous theorem, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,*

$$\left|\sum_{s=1}^{t}\eta_t w_t\right| \leq \sigma_2\sqrt{2\left(1/\sigma_2^2+\sum_{s=1}^{t} w_s\right)\log\left(\frac{\sqrt{1+\sigma_2^2\sum_{s=1}^{t} w_s^2}}{\delta}\right)} \tag{93}$$

Proof. Define the "bad" event

$$B_t(\delta) = \left\{\omega \in \Omega : \frac{\left(\sum_{s=1}^{t}\eta_s w_s\right)^2}{1/\sigma_2^2+\sum_{t=1}^{\tau} w_t^2} > 2\sigma_2^2\ln\left(\frac{\sqrt{1+\sigma_2^2\sum_{s=1}^{t} w_s^2}}{\delta}\right)\right\}$$

We are interested in bounding the probability that $\bigcup_{t\geq 0} B_t(\delta)$ happens. Define $\tau(\omega) = \min\{t \geq 0 : \omega \in B_t(\delta)\}$, with the convention that $\min\emptyset = \infty$. Then, τ is a stopping time. Further, $\bigcup_{t\geq 0} B_t(\delta) = \{\omega : \tau(\omega) < \infty\}$ Thus, by the previous theorem it holds that

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{t\geq 0} B_t(\delta)\right] &= \mathbb{P}[\tau < \infty] = \mathbb{P}\left[\frac{\left(\sum_{s=1}^{t}\eta_s w_s\right)^2}{1/\sigma_2^2+\sum_{t=1}^{\tau} w_t^2} > 2\sigma_2^2\ln\left(\frac{\sqrt{1+\sigma_2^2\sum_{s=1}^{t} w_s^2}}{\delta}\right) \text{ and } \tau < \infty\right] \\
&= \mathbb{P}\left[\frac{\left(\sum_{s=1}^{t}\eta_s w_s\right)^2}{1/\sigma_2^2+\sum_{t=1}^{\tau} w_t^2} > 2\sigma_2^2\ln\left(\frac{\sqrt{1+\sigma_2^2\sum_{s=1}^{t} w_s^2}}{\delta}\right)\right] \leq \delta
\end{aligned}$$

□

F. Parameter Estimation and Concentration in First Stage

There are many ways of addressing the regression problem in the first stage and they fundamentally reduce to a choice for the regularizer in the regression. If we do not introduce such regularizer, we are left with a system of multiple regressions that can be solved with standard OLS (estimator). Another choice is to introduce a Frobenius norm regularizer. We introduce the parameter $\lambda > 0$ and a regularization term $\lambda \|\|\boldsymbol{\Theta}\|\|_F^2$, which is used to penalize the model complexity. By choosing the Frobenius norm, the system of equations decouples again but each with a regularizer term. Thus, we end up with d independent linear equations that we try to fit separately. More interesting settings could try to solve the optimization problem jointly by a regularizer that couples the equations (e.g. (Wainwright, 2019) provides concentration results for such settings). This will be interesting to investigate in future works.

We indicate with $\underline{\boldsymbol{\theta}}_j$ the j-th column of the matrix $\boldsymbol{\Theta}$, then

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}}_t &\in \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{s=1}^{t} \left\| \boldsymbol{x}_s^\top - \boldsymbol{z}_s^\top \boldsymbol{\Theta} \right\|_2^2 + \lambda \|\|\boldsymbol{\Theta}\|\|_F^2 \\
&= \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{s=1}^{t} \left(\sum_{j=1}^{d} \left(\boldsymbol{x}_{s,j} - \boldsymbol{z}_s^\top \underline{\boldsymbol{\theta}}_j \right)^2 + \lambda \sum_{j=1}^{d} \left\| \underline{\boldsymbol{\theta}}_j \right\|_2^2 \right) \\
&= \operatorname*{argmin}_{\{\underline{\boldsymbol{\theta}}_j\}_{j=1}^d} \sum_{j=1}^{d} \left(\sum_{s=1}^{t} \left(\boldsymbol{x}_{s,j} - \boldsymbol{z}_s^\top \underline{\boldsymbol{\theta}}_j \right)^2 + \lambda \left\| \underline{\boldsymbol{\theta}}_j \right\|_2^2 \right)
\end{aligned}$$

Clearly, we can compute separately the columns $\widehat{\boldsymbol{\theta}}_{t,j}$ of $\widehat{\boldsymbol{\Theta}}_t$ as

$$\widehat{\boldsymbol{\theta}}_{t,j} \in \operatorname*{argmin}_{\underline{\boldsymbol{\theta}}_j} \sum_{s=1}^{t} \left(\boldsymbol{x}_{s,j} - \boldsymbol{z}_s^\top \underline{\boldsymbol{\theta}}_j \right)^2 + \lambda \left\| \underline{\boldsymbol{\theta}}_j \right\|_2^2 \tag{94}$$

with solution

$$\widehat{\boldsymbol{\theta}}_{t,j} = \left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d \right)^{-1} \mathbf{Z}_t^\top \boldsymbol{x}_{t,j}$$

where $\boldsymbol{x}_{t,j}$ is a vector with components $x_{1,j}, \dots, x_{t,j}$. The solution to the independent quadratic optimization problems is in matrix notation equal to

$$\widehat{\boldsymbol{\Theta}}_t = \left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbb{I}_d \right)^{-1} \mathbf{Z}_t^\top \mathbf{X}_t$$

From the decomposition of the problem into multiple independent regressions, we understand that it is enough to concentrate the individual columns of $\widehat{\boldsymbol{\Theta}}_t$ around the ones of $\boldsymbol{\Theta}$ and then to use a union bound to put things together.

Theorem F.1 (Confidence Ellipsoid for Columns in First Stage)**.** *Define $\boldsymbol{x}_t = \boldsymbol{z}_t^\top \underline{\boldsymbol{\theta}}_j + \epsilon_{t,j}$ with $\epsilon_{t,j}$ is $L_{\boldsymbol{\epsilon}}$-sub-Gaussian and assume that $\left\| \underline{\boldsymbol{\theta}}_j \right\|_2 \leq S$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$, $\underline{\boldsymbol{\theta}}_j$ lies in the set*

$$\mathcal{E}_t = \left\{ \underline{\boldsymbol{\theta}}_j \in \mathbb{R}^d : \left\| \widehat{\underline{\boldsymbol{\theta}}}_{t,j} - \underline{\boldsymbol{\theta}}_j \right\|_{\mathbf{G}_{\boldsymbol{z},t}} \leq L_{\boldsymbol{\epsilon}} \sqrt{2 \log \left(\frac{\det \left(\mathbf{G}_{\boldsymbol{z},t} \right)^{1/2} \det(\lambda \mathbb{I}_d)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right\}$$

Furthermore, if for all $t \geq 1, \|\boldsymbol{z}_t\|_2 \leq L_z$ *then with probability at least* $1-\delta$*, for all* $t \geq 0$*,*

$$\left\|\widehat{\underline{\boldsymbol{\theta}}}_{t,j} - \underline{\boldsymbol{\theta}}_j\right\|_{\mathbf{G}_{\boldsymbol{z},t}} \leq L_{\boldsymbol{\epsilon}}\sqrt{d\log\left(\frac{1+tL_z^2/\lambda d}{\delta}\right)} + \lambda^{1/2}S$$

Corollary F.1 (Confidence Ellipsoid for First Stage)**.** *Under the conditions of the previous theorem, for any* $\delta > 0$*, with probability at least* $1-\delta$*, for all* $t \geq 0$

$$\left\|\widehat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta}\right\|_F^2 = \sum_{j=1}^{d}\left\|\widehat{\underline{\boldsymbol{\theta}}}_{t,j} - \underline{\boldsymbol{\theta}}_j\right\|_2^2 \leq d\left(L_{\boldsymbol{\epsilon}}\sqrt{d\log\left(\frac{1+tL_z^2/\lambda d}{\delta}\right)} + \lambda^{1/2}S\right)$$