Stochastic Online Instrumental Variable Regression: Regrets for Endogeneity and Bandit Feedback

Riccardo Della Vecchia, Debabrota Basu

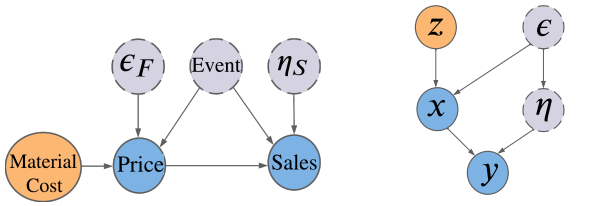
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRIStAL, F-59000 Lille, France
ric.della-vecchia@gmail.com, debabrota.basu@inria.fr

Abstract

Endogeneity, i.e. the dependence of noise and covariates, is a common phenomenon in real data due to omitted variables, strategic behaviours, measurement errors etc. In contrast, the existing analyses of stochastic online linear regression with unbounded noise and linear bandits depend heavily on exogeneity, i.e. the independence of noise and covariates. Motivated by this gap, we study the *over- and just-identified Instrumental Variable (IV) regression, specifically Two-Stage Least Squares, for stochastic online learning*, and propose to use an online variant of Two-Stage Least Squares, namely O2SLS. We show that O2SLS achieves $\mathcal{O}(d_{\boldsymbol{x}} d_{\boldsymbol{z}} \log^2 T)$ identification and $\widetilde{\mathcal{O}}(\gamma\sqrt{d_{\boldsymbol{z}} T})$ oracle regret after T interactions, where $d_{\boldsymbol{x}}$ and $d_{\boldsymbol{z}}$ are the dimensions of covariates and IVs, and γ is the bias due to endogeneity. For $\gamma = 0$, i.e. under exogeneity, O2SLS exhibits $\mathcal{O}(d_{\boldsymbol{x}}^2 \log^2 T)$ oracle regret, which is of the same order as that of the stochastic online ridge. Then, we leverage O2SLS as an oracle to design OFUL-IV, a stochastic linear bandit algorithm to tackle endogeneity. OFUL-IV yields $\widetilde{\mathcal{O}}(\sqrt{d_{\boldsymbol{x}} d_{\boldsymbol{z}} T})$ regret that matches the regret lower bound under exogeneity. For different datasets with endogeneity, we experimentally show efficiencies of O2SLS and OFUL-IV.

1 Introduction

Online regression is a founding component of online learning (Kivinen, Smola, and Williamson 2004), sequential testing (Kazerouni and Wein 2021), contextual bandits (Foster and Rakhlin 2020), and reinforcement learning (Ouhamma, Basu, and Maillard 2022). Especially, online linear regression is widely used and analysed to design efficient algorithms with theoretical guarantees (Greene 2003; Abbasi-Yadkori, Pál, and Szepesvári 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 structural equation: $Y = \boldsymbol{\beta}^T \mathbf{X} + \eta$, 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.* Two common assumptions in the analysis of linear regression are (i) bounded observations and covariates (Vovk 1997; Bartlett et al. 2015; Gaillard et al. 2019), and (ii) *exogeneity*, i.e. conditional mean independence of the noise η and the input features $\mathbf{X}$ ($\mathbb{E}[\eta|\mathbf{X}] = \mathbf{0}$) (Abbasi-Yadkori, Pál, and Szepesvári 2011b; Ouhamma, Maillard, and Perchet 2021). Under exogeneity, researchers have studied scenarios where the observational noise is unbounded and has only bounded variance σ^2. But this setting asks for a different technical analysis than the bounded adversarial setting popular in online regression literature. For example, (Ouhamma, Maillard, and Perchet 2021) analyse online forward and online ridge regressions in the unbounded stochastic setting.



(a) Graphical Model of Example 1.1 (b) DAG of 2SLS at t

Figure 1: Relations of IVs (orange), Covariates (blue), and Outcome (blue) in Online Two-stage Regression. Unobserved and Observed variables are in dotted and solid circles.

Example 1.1 (Learning Price-Sales Dynamics)**.** *A market analyst wants to learn how price of a food item at a given day t affects the sales of the item, given a daily stream of data of food prices and sales. The analyst decides to run* online linear regression *to learn the relation* $\mathrm{Price} \to \mathrm{Sales}$ *(Fig. 1a). The analyst finds the β to be positive. But now if a big festival happens in a city, a restaurant owner can increase price of the food knowing still the sales will increase. This event contradicts the analyst's learnt parameter. Thus, as she incorporates the price-sales data of this city, the estimated β changes to negative. Because her algorithm assumes exogeneity of noise and covariates, while in reality the omitted variable* Event_t *affects both price and sales (Fig. 1a). Hence, the covariate* Price_t*, gets correlated with the noise,* $\eta_t = \rho_s \mathrm{Event}_t + \eta_{S,t}$. Price_t *is an endogeneous variable.*

Similar to this example, in real-life, *endogeneity*, i.e. dependence between noise and covariates ($\mathbb{E}[\eta\mathbf{X}] \neq \mathbf{0}$) (Greene 2003; Angrist, Imbens, and Rubin 1996) is often observed due to omitted explanatory variables, strategic behaviours during data generation, measurement errors, the dependence of the output and the covariates on unobserved confounding variables etc. (Wald 1940; Mogstad, Torgovitsky, and Wal-

ters 2021; Zhu et al. 2022). Motivated by such scenarios, *we analyse online linear regression aiming to estimate $\boldsymbol{\beta}$ accurately from endogenous observational data, where noise is stochastic and unbounded.*

Instrumental Variable (IV) Regression. Historically, IVs were introduced to identify and quantify the causal effects of endogenous covariates (Newey and Powell 2003). IVs are widely used in economics (Wright 1928; Mogstad, Torgovitsky, and Walters 2021), causal inference (Rubin 1974; Hernan and Robins 2020; Harris et al. 2022), bio-statistics and epidemiology (Burgess, Small, and Thompson 2017). One popular framework is to choose IVs which heavily influence the covariates but are independent of the exogenous noise. For example, $\mathrm{MaterialPrice}$ is an instrumental variable in Example 1.1 as per the DAG in Figure 1b. Given such an IV(s), one aims to learn the relations between IVs and covariates in one stage, and then uses the predicted values of covariates to learn the relation between covariates and the outcome variables in the next stage. This approach of conducting two stages of linear regression using IVs is called *Two Stage Least Squares Regression* (2SLS) (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996). 2SLS has become the standard tool in economics, social sciences, and statistics to study the effect of treatments on outcomes involving endogeneity. Recently, in machine learning, researchers extended classical 2SLS to nonlinear structures, non-compliant instruments, corrupted observations with deep learning (Liu, Shang, and Cheng 2020; Xu et al. 2020; Xu, Kanagawa, and Gretton 2021), graphical models (Stirn and Jebara 2018), and kernel regression (Zhu et al. 2022).

Thus, we are interested in studying the 2SLS approach for online learning. But the existing analyses of 2SLS are asymptotic, i.e. what can be learned if we have access to an infinite number of samples in an offline setting (Singh, Hosanagar, and Gandhi 2020; Liu, Shang, and Cheng 2020; Nareklishvili, Polson, and Sokolov 2022). In 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 (Venkatraman et al. 2016). This motivates us to analyse the online extension of 2SLS, named O2SLS.

Instrumental Variable Bandits. As stated in Example 1.1, the goal of the analyst might be to aid a restaurant owner or food supplier to strategically decide on a price of the food item from a price range and also corresponding raw material cost among multiple suppliers to improve the sales as much as possible. At this point, the estimation of the first- and second-stage parameters aid in sequential decision making, leading to the maximisation of accumulated outcomes over time. This is exactly a linear bandit problem. In this case, the analyst only observes the outcome corresponding to the choice of price and material cost. This is referred to as bandit feedback in online learning literature and studied under the linear bandit formulation (Abbasi-Yadkori, Pál, and Szepesvári 2011a). This motivates us to extend O2SLS to *linear bandits, where both bandit feedback and endogeneity occur*. In this paper, we investigate two questions:

1. *What is the upper bound on the loss in performance for deploying parameters estimated by O2SLS instead of the true parameters $\boldsymbol{\beta}$? How does estimating the true parameters $\boldsymbol{\beta}$ influence different performance metrics under endogeneity?*

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

Our Contributions. Our investigation has led to

1. *A Non-asymptotic Analysis of O2SLS:* First, we identify three notions of regret: *identification regret*, *oracle regret*, and *population regret*. Though all of them are of same order under exogeneity, we show that the relations are more nuanced under endogeneity and unbounded noise (see Appendix B.9). We focus specifically on the identification regret, i.e. the sum of differences between the estimated parameters $\{\boldsymbol{\beta}_t\}_{t=1}^T$, and the true parameter $\boldsymbol{\beta}$, and *oracle 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 $\boldsymbol{\beta}$. In Section 3, we theoretically show that O2SLS achieve $\mathcal{O}(d_{\boldsymbol{x}} d_{\boldsymbol{z}} \log^2 T)$ identification regret and $\mathcal{O}(d_{\boldsymbol{x}} d_{\boldsymbol{z}} \log^2 T + \gamma\sqrt{d_{\boldsymbol{z}} T \log T})$ oracle regret after receiving T samples from the observational data. Identification regret of O2SLS is $d_{\boldsymbol{z}}$ multiplicative factor higher than regret of online linear regression under exogeneity, and oracle regret is $\mathcal{O}(\gamma\sqrt{d_{\boldsymbol{z}} T \log T})$ additive factor higher. These are the costs that O2SLS pay for tackling endogeneity in two stages. In our knowledge, *we are the first to propose a non-asymptotic regret analysis of O2SLS with stochastic and unbounded noise*. Due to two-stages and endogeneity, we cannot rely on standard techniques and need to prove concentration bounds for dependent r.v. (Appendix B.6) leading to a novel analysis for Thm. 3.5. We also experimentally demonstrate efficiency of O2SLS on synthetic and real-data with endoegeneity (Section 5, Appendix D).

2. OFUL-IV *for Linear Bandits with Endogeneity:* In Section 4, we study the linear bandit problem with endogeneity. *We design an extension of OFUL (Optimism in Face of Uncertainty for Linear) 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. Lemma 3.3 derives a new confidence ellipsoid around O2SLS estimator $\boldsymbol{\beta}_t$ with a new design matrix $\hat{\boldsymbol{H}}_t$. Following existing derivations for ridge would ask algorithm to know tight upper bounds on the hidden paramenter unlike us. We show that OFUL-IV achieve $\mathcal{O}(\sqrt{d_{\boldsymbol{x}} d_{\boldsymbol{z}} T} \log T)$ regret after T interactions. We further experimentally validate that OFUL-IV incurs lower regret under endeogeneity than OFUL (Abbasi-Yadkori, Pál, and Szepesvári 2011a) (Section 5, AppendixD).

Related Works: 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, Long, and Warmuth 1991). (Vovk 1997, 2001) show that forward and ridge regressions achieve $\mathcal{O}(d_{\boldsymbol{x}} Y_{\max}^2 \log T)$ for outcomes with bound $Y_{\max}$. (Bartlett et al. 2015) generalise the analysis further by considering the features known in hindsight. (Gaillard et al. 2019) improve the analysis further to propose an optimal algorithm and a lower bound. *These works perform an adversarial analysis with bounded outcomes, covariates, and observational noise, while we focus*

on the stochastic setting. (Ouhamma, Maillard, and Perchet 2021) study the stochastic setting with bounded input features and unbounded noise. But they need to assume independence of noise and input features. *In this paper, we analyse online* 2SLS *under endogeneity and unbounded (stochastic) noise.* We do not know the bound on the outcome and derive high probability bounds for any bounded sequence of features. Previously, (Venkatraman et al. 2016) studied `O2SLS` for system identification but provided only asymptotic analysis.

Related Works: Linear Bandits without Endogeneity. Linear bandits generalise the setting of online linear regression under bandit feedback (Abbasi-Yadkori, Pál, and Szepesvári 2011a; Abbasi-Yadkori, Pal, and Szepesvari 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 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, Pal, and Szepesvari 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 under exogeneity. *Here, we use* `O2SLS` *as a regression oracle to develop* OFUL-IV *for linear bandits with endogeneity. We prove that* OFUL-IV *achieves* $\mathcal{O}(d\sqrt{T}\log T)$ *regret.*

Related Works: Instrument-Armed Bandits. (Kallus 2018) is the first to study endogeneity and instrumental variables in a 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 stochastic linear bandits with endogeneity requiring different techniques of analysis and algorithm design.* (Krishnamurthy, Wu, and Syrgkanis 2018) studies a linear contextual bandit setup close to ours, but they assume arm-independent and bounded noise, and thus, yielding significantly different analysis.

Notations. Matrices and vectors are denoted by bold capital and bold small letters (e.g. $\mathbf{A}$ and $\mathbf{a}$). $\|\cdot\|_p$ is l_p norm of a vector. $\sigma_{\min}(\cdot)$ is minimum singular value and $|||\cdot|||_2$ is operator norm of a matrix. For any vector $y \in \mathbb{R}^n$ and a positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, let us define the norm $\|\boldsymbol{y}\|_{\mathbf{A}} \triangleq \sqrt{\boldsymbol{y}^T\mathbf{A}\boldsymbol{y}} = \sqrt{\langle \boldsymbol{y}, \mathbf{A}\boldsymbol{y}\rangle}$.

2 Background

We are given an observational dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ consisting of n pairs of input features and outcomes. 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 $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^{d_{\boldsymbol{x}}}$, $\boldsymbol{\beta} \in \mathbb{R}^{d_{\boldsymbol{x}}}$ is the *unknown true parameter vector* of the linear model, and $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$ is the *unobserved error* representing all causes of y_i other than $\boldsymbol{x}_i$. It is assumed that the error terms η_i are independent and identically distributed with bounded variance σ_η^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 regression is to estimate $\boldsymbol{\beta}$ by *minimising the square loss over dataset* (Brier 1950): $\widehat{\boldsymbol{\beta}} \triangleq \operatorname{argmin}_{\boldsymbol{\beta}'} \sum_{i=1}^n (y_i - \boldsymbol{\beta}'^\top \boldsymbol{x}_i)^2$.

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

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

If $\mathbb{E}[\eta_i | \boldsymbol{x}_j] = 0$*, i.e. under exogeneity*, the OLS estimator is unbiased: $\mathbb{E}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] = \boldsymbol{\beta}$. Furthermore, it is also asymptotically consistent since $\mathbf{X}_n^\top \boldsymbol{\eta}_n \xrightarrow{p} 0$ implies $\widehat{\boldsymbol{\beta}}_{\text{OLS}} \xrightarrow{p} \boldsymbol{\beta}$ (Greene 2003). In practice, the input features $\boldsymbol{x}$ and the noise η are often correlated (Greene 2003, Chapter 8). This dependence, called *endogeneity*, is modelled with *a confounding unobserved random variable* $\boldsymbol{\epsilon}$. In Figure 1b, we draw the DAG between the variables under endogeneity, where each arrow implies conditional dependence of the child on the parent. To allow consistent estimation of $\boldsymbol{\beta}$ under endogeneity, the Instrumental Variables (IVs) $\boldsymbol{z}$ are introduced (Angrist, Imbens, and Rubin 1996; Newey and Powell 2003). IVs are chosen such that they are highly correlated with endogenous components of $\boldsymbol{x}$ (relevance condition) but are uncorrelated with the noise η. This leads to Two-stage Least Squares (2SLS) approach to IV regression (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996). We further assume that IVs, $\mathbf{Z}_n \triangleq [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^\top \in \mathbb{R}^{n\times d_{\boldsymbol{z}}}$, cause linear effects on the endogenous covariates

$$\mathbf{X}_n = \mathbf{Z}_n\boldsymbol{\Theta} + \mathbf{E}_n. \qquad \text{(First Stage)}$$

$\boldsymbol{\Theta} \in \mathbb{R}^{d_{\boldsymbol{z}}\times d_{\boldsymbol{x}}}$ is the unknown first-stage parameter matrix and $\mathbf{E}_n \triangleq [\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n]^\top \in \mathbb{R}^{n\times d_{\boldsymbol{x}}}$ is the unobserved noise matrix leading to confounding in the second stage. First-stage is a "classic" *multiple regression* with $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{z}] = 0$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{I}_{d_{\boldsymbol{x}}})$ (Wasserman 2004, Ch. 13). Thus, OLS in first-stage yields an estimate $\widehat{\boldsymbol{\Theta}}_n \triangleq (\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top \mathbf{X}_n$, leading to the 2SLS estimator:

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

Here, $\widehat{\mathbf{X}}_n = \widehat{\boldsymbol{\Theta}}_n^\top \mathbf{Z}_n$ is the predicted covariate from first-stage. Given $\mathbb{E}[\boldsymbol{z}_i \eta_i] = 0$ in the true model, we observe that $\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left(\widehat{\mathbf{X}}_n^\top \widehat{\mathbf{X}}_n\right)^{-1} \widehat{\mathbf{X}}_n^\top \mathbf{X}_n\boldsymbol{\beta} + \left(\widehat{\mathbf{X}}_n^\top \widehat{\mathbf{X}}_n\right)^{-1} \widehat{\mathbf{X}}_n^\top \boldsymbol{\eta}_n \xrightarrow{p} \boldsymbol{\beta}$, as $n \to \infty$. Since $\boldsymbol{x}$ and $\boldsymbol{\eta}$ are correlated, 2SLS estimator is not unbiased in finite-time. When $d_{\boldsymbol{x}} = d_{\boldsymbol{z}}$, we call $\boldsymbol{z}$ *just-identified IVs*. When $d_{\boldsymbol{x}} < d_{\boldsymbol{z}}$, we call $\boldsymbol{z}$ *over-identified IVs*. In this paper, we analyse both of the conditions.

Assumption 2.1. The assumptions for conducting 2SLS are (Greene 2003; Hernan and Robins 2020):

1. **Endogeneity of X.** The second stage input features $\mathbf{X}$ and noise η are correlated: $\mathbb{E}[\eta\mathbf{X}] \neq \mathbf{0}$.
2. **Exogeneity of Z.** IV random variables and the noise in the second stage are uncorrelated: $\mathbb{E}[\eta\mathbf{Z}] = \mathbf{0}$.

3. **Relevance Condition**. The variables $\mathbf{Z}$ and $\mathbf{X}$ are correlated: $\mathbb{E}[\mathbf{Z}\mathbf{X}^\top] = \mathbf{\Sigma_{ZX}}$. $\mathbf{\Sigma_{ZX}} \in \mathbb{R}^{d_{\boldsymbol{z}} \times d_{\boldsymbol{x}}}$ with rank $d_{\boldsymbol{x}}$. For non-asymptotic analysis, we further use an empirical version of relevance condition. Specifically, we assume , the empirical cross-covariance to have maximum rank at any $t > 0$, i.e. $\sigma_{\min}\left(\mathrm{Cov}(\mathbf{Z}_t, \mathbf{X}_t)\right) = \sigma_{\min}\left(\frac{1}{t}\sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{x}_s^\top\right) \geq \mathfrak{r} > 0$.

3 O2SLS: Online Two-Stage Least Squares

In this section, we describe the problem setting and schematic of *Online Two-Stage Least Squares Regression*, in brief O2SLS. Following that, we first define two notions of regret: *identification* and *oracle*, measuring accuracy of estimating the true parameter and the predicted outcomes, respectively. We provide a theoretical analysis of O2SLS and upper bound the two types of regret (Section 3.2).

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 2SLS model (Fig. 1b), data is generated endogeneously

$$\boldsymbol{x}_t = \mathbf{\Theta}^\top \boldsymbol{z}_t + \boldsymbol{\epsilon}_t \qquad y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t + \eta_t, \tag{1}$$

such that $\boldsymbol{x}_t$ and η_t are correlated while $\boldsymbol{z}_t$ and η_t are not, 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 $\boldsymbol{z}_t$. At time $t+1$, the algorithm aims to yield an estimate of the parameter $\boldsymbol{\beta}_t$ and predict an outcome $\widehat{y}_{t+1} \triangleq \boldsymbol{\beta}_t^\top \boldsymbol{x}_{t+1} \in \mathbb{R}$ using the data $\{(\boldsymbol{x}_s, \boldsymbol{z}_s, y_s)\}_{s=1}^t$ observed so far. Following the prediction, Nature reveals the true outcome y_{t+1}.

To address endogeneity in this problem, we propose an online form of the 2SLS estimator. Modifying Eq. (2SLS), we obtain the O2SLS estimator that is computed for the prediction at time $t+1$, using information up to time t:

$$\begin{aligned}\boldsymbol{\beta}_t &\triangleq \left(\widehat{\mathbf{X}}_t^\top \widehat{\mathbf{X}}_t\right)^{-1} \widehat{\mathbf{X}}_t^\top \boldsymbol{y} = \left(\sum_{s=1}^{t} \widehat{\boldsymbol{x}}_s \widehat{\boldsymbol{x}}_s^\top\right)^{-1} \sum_{s=1}^{t} \widehat{\boldsymbol{x}}_s y_s \\ &= (\mathbf{I} - \mathbf{G}_{\widehat{\boldsymbol{x}},t}^{-1} \widehat{\boldsymbol{x}}_t \widehat{\boldsymbol{x}}_t^\top)\boldsymbol{\beta}_{t-1} + \mathbf{G}_{\widehat{\boldsymbol{x}},t}^{-1} \widehat{\boldsymbol{x}}_t y_t. \qquad \text{(O2SLS)}\end{aligned}$$

Here, $\mathbf{G}_{\widehat{\boldsymbol{x}},t} \triangleq \sum_{s=1}^{t} \widehat{\boldsymbol{x}}_s \widehat{\boldsymbol{x}}_s^\top = \mathbf{G}_{\widehat{\boldsymbol{x}},t-1} + \widehat{\boldsymbol{x}}_t \widehat{\boldsymbol{x}}_t^\top$. We need the predicted value of covariate $\widehat{\boldsymbol{x}}_t \triangleq \widehat{\mathbf{\Theta}}_{t-1} \boldsymbol{z}_t$ and output y_t for the iterative update of $\boldsymbol{\beta}_t$ given the estimates at previous time $\boldsymbol{\beta}_t$ and $\mathbf{G}_{\widehat{\boldsymbol{x}},t}$. To predict the covariate $\widehat{\boldsymbol{x}}_t$, we leverage an iterative update rule of $\widehat{\mathbf{\Theta}}_t$ from $\widehat{\mathbf{\Theta}}_{t-1}$ by using $\boldsymbol{z}_t$, $\boldsymbol{x}_t$ (Appendix B). In brief, for $\lambda > 0$,

$$\begin{aligned}\widehat{\mathbf{\Theta}}_t &= \left(\sum_{s=1}^{t-1} \boldsymbol{z}_s \boldsymbol{z}_s^\top + \lambda \mathbf{I}_{d_{\boldsymbol{z}}}\right)^{-1} \sum_{s=1}^{t-1} \boldsymbol{z}_s \boldsymbol{x}_s^\top \\ &= (\mathbf{I} - \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \boldsymbol{z}_t \boldsymbol{z}_t^\top)\widehat{\mathbf{\Theta}}_{t-1} + \mathbf{G}_{\boldsymbol{z},t-1}^{-1} \boldsymbol{z}_t \boldsymbol{x}_t^\top. \end{aligned} \tag{2}$$

Finally, we use the O2SLS estimator at step $t+1$ for the prediction $\widehat{y}_{t+1} = \boldsymbol{\beta}_t^\top \boldsymbol{x}_{t+1}$, as in Algorithm 1.

Computational Complexity. For second-stage, we compute $\mathbf{G}_{\widehat{\boldsymbol{x}},t}^{-1}$ from $\mathbf{G}_{\widehat{\boldsymbol{x}},t-1}^{-1}$ in $O(d_{\boldsymbol{x}}^2)$ time using the Sherman–Morrison formula. We store $\mathbf{G}_{\widehat{\boldsymbol{x}},t}^{-1}$ after each step, which requires $O(d_{\boldsymbol{x}}^2)$ memory. However, the first-stage update has $O(d_{\boldsymbol{z}}^2 + d_{\boldsymbol{z}} d_{\boldsymbol{x}})$ time complexity and requires $O(d_{\boldsymbol{z}}^2)$ memory. Thus, O2SLS exhibits quadratic space and time complexity.

Algorithm 1: O2SLS

Input: Initialisation parameters $\boldsymbol{\beta}_0, \widehat{\mathbf{\Theta}}_0, \lambda$
for $t = 1, \ldots, T$ **do**
  Observe $\boldsymbol{z}_t$ generated i.i.d. by Nature, and $\boldsymbol{x}_t$ sampled from Eq. (1) given $\boldsymbol{z}_t$
  Compute first-stage and second-stage estimates $\boldsymbol{\beta}_{t-1}$ and $\widehat{\mathbf{\Theta}}_{t-1}$ as per Equation (2) and Equation (O2SLS)
  Predict $\widehat{y}_t = \boldsymbol{\beta}_{t-1}^\top \boldsymbol{x}_t$
  Observe y_t generated by Nature

Remark 3.1 (*Proper Online Learning*). As in improper online learning algorithms, 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 avoid this modification, and follow a proper online learning approach to use $\boldsymbol{\beta}_{t-1}$ to predict.

3.1 Two Regrets for Estimation & Prediction

To analyse the online regression algorithms, it is essential to define proper performance metrics, specifically *regrets*. Regret quantifies what an online (or sequential) algorithm cannot achieve as it does not have access to the whole dataset rather observes it step by step. Here, we discuss and define different regrets that we leverage in our analysis of O2SLS.

In econometrics and bio-statistics, where 2SLS is popularly used the focus is the accurate identification of the underlying structural model $\boldsymbol{\beta}$. Identifying $\boldsymbol{\beta}$ leads to understanding of the underlying economic or biological causal relations and their dynamics. In ML, (Venkatraman et al. 2016) applied O2SLS for online linear system identification. Thus, given sequences of estimators $\{\boldsymbol{\beta}_t\}_{t=1}^T$ and covariates $\{\boldsymbol{x}_t\}_{t=1}^T$, the cost of identifying the true parameter $\boldsymbol{\beta}$ can be quantified by

$$\widetilde{R}_T(\boldsymbol{\beta}) \triangleq \sum\nolimits_{t=1}^{T} (\boldsymbol{x}_t^\top \boldsymbol{\beta}_{t-1} - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2. \tag{3}$$

We refer to $\widetilde{R}_T(\boldsymbol{\beta})$ as *identification regret* over horizon T. In the just identified setting that we are considering, the identification regret is equivalent to the regret of counterfactual prediction (Eqn. 5, (Hartford et al. 2016)). Counterfactual predictions are important to study the causal questions: what would have changed in the outcome if Treatment a had been used instead of treatment b? A modern application of IVs is to facilitate such counterfactual predictions (Hartford et al. 2016; Bennett, Kallus, and Schnabel 2019; Zhu et al. 2022).

Alternatively, one might be interested in evaluating and improving the quality of prediction obtained using an estimator $\{\boldsymbol{\beta}_t\}_{t=1}^T$ with respect to an underlying oracle (or expert), which is typically the case in statistical learning theory and forecasting (Foster 1991; Cesa-Bianchi and Lugosi 2006). If the oracle has access to the true parameters $\boldsymbol{\beta}$, the cost in terms of prediction that the estimators pay with respect to the oracle is $\bar{r}_t \triangleq (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta}_{t-1})^2 - (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2$. Thus, the

regret in terms of the quality of prediction is

$$\overline{R}_T(\boldsymbol{\beta}) \triangleq \sum\nolimits_{t=1}^{T} (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta}_{t-1})^2 - (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2. \quad (4)$$

We refer to $\overline{R}_T(\boldsymbol{\beta})$ as the *oracle regret*. It is studied for stochastic analysis of online regression (Ouhamma, Maillard, and Perchet 2021) and bandits (Foster and Rakhlin 2020).

As O2SLS is interesting for learning causal structures (Hartford et al. 2016; Bennett, Kallus, and Schnabel 2019), we focus on the identification regret. To compare with the existing results in online linear regression, we also analyse the oracle regret of O2SLS. Though they are of similar order (in T) under exogeneity, we show them to significantly differ for O2SLS under endogeneity (Section 3.2).

Remark 3.2 (**Hardness of Exogeneity vs. Endogeneity for Prediction**). In online learning theory focused on Empirical Risk Minimisation (ERM), another type of regret is considered where the oracle has access to the best offline estimator $\boldsymbol{\beta}_T \triangleq \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{t=1}^{T} (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2$ given the observations over T steps (Cesa-Bianchi and Lugosi 2006). Thus, the new formulation of regret becomes $R_T = \sum_{t=1}^{T} (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta}_{t-1})^2 - \min_{\boldsymbol{\beta}} \sum_{t=1}^{T} (y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2$. We refer to it as the *population regret*. *Under exogeneity, (Ouhamma, Maillard, and Perchet 2021) show that oracle regret and population regret differs by $o(\log^2 T)$.* We show that under endogeneity, their expected values differ by $\Omega(T)$. Thus, we avoid studying population regret, and detail it in App. B.9.

3.2 Theoretical Analysis

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

Lemma 3.3 (Confidence Ellipsoid for the Second-stage Parameters)**.** *Let us define the* design matrix of IVs *to be $\mathbf{G}_{\boldsymbol{z},t} \triangleq \mathbf{Z}_t^\top \mathbf{Z}_t + \mathbf{G}_{\boldsymbol{z},0} = \sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{z}_s^\top + \mathbf{G}_{\boldsymbol{z},0}$ with $\mathbf{G}_{\boldsymbol{z},0} = \lambda \mathbf{I}_{d_{\boldsymbol{z}}}$ for some $\lambda > 0$. Then, for some conditionally σ_η-sub-Gaussian second stage noise η_t (formally $\forall \lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \eta_t} | \mathbf{Z}_{t-1}, \boldsymbol{\eta}_{t-1}, \mathbf{E}_{t-1}] \leq e^{\frac{\lambda^2 \sigma_\eta^2}{2}}$), for bounded IVs $\|\boldsymbol{z}\|_2^2 \leq L_z^2$, and for all $t > 0$, the true parameter $\boldsymbol{\beta}$ belongs to the confidence set*

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

with probability at least $1 - \delta \in (0, 1)$. Here, $\mathfrak{b}_t(\delta) \triangleq d_{\boldsymbol{z}} \sigma_\eta^2 \log\left(\frac{1 + t L_z^2 / \lambda d_{\boldsymbol{z}}}{\delta}\right)$, $\widehat{\mathbf{H}}_t \triangleq \widehat{\boldsymbol{\Theta}}_t^\top \mathbf{G}_{\boldsymbol{z},t} \widehat{\boldsymbol{\Theta}}_t$, and $\widehat{\boldsymbol{\Theta}}_t$ is the estimate of the first-stage parameter at time t.

Lemma 3.3 extends the well-known elliptical lemma for OLS and Ridge estimators under exogeneity to the O2SLS estimator under endogeneity. It shows that the size of the confidence intervals induced by O2SLS estimate at time T is $\mathcal{O}(\sqrt{d_{\boldsymbol{z}} \log T})$, which is of the same order as that of the exogenous elliptical lemma (Abbasi-Yadkori, Pál, and Szepesvári 2011a) but while applied on IVs.

Identification Regret Bound. Now, we state the identification regret upper bound of O2SLS.

Theorem 3.4 (Identification Regret of O2SLS)**.** *If Assumption 2.1 holds true, for bounded IVs $\|\boldsymbol{z}\|_2^2 \leq L_z^2$ and bounded first-stage parameters $\|\!|\boldsymbol{\Theta}|\!\|_2 \leq L_{\boldsymbol{\Theta}}^2$, the conditionally σ_η-sub-Gaussian second stage noise η_t and the component-wise conditionally $\sigma_{\boldsymbol{\epsilon}}$-sub-Gaussian first stage noise $\boldsymbol{\epsilon}_t$ (formally $\forall \lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \boldsymbol{\epsilon}_{t,i}} | \mathbf{Z}_{t-1}, \boldsymbol{\eta}_{t-1}, \mathbf{E}_{t-1}] \leq e^{\frac{\lambda^2 \sigma_{\boldsymbol{\epsilon}}^2}{2}}$), the regret of O2SLS satisfies with probability at least $1 - \delta$,*

$$\widetilde{R}_T \leq \sum_{t=1}^{T} \underbrace{\|\boldsymbol{\beta}_t - \boldsymbol{\beta}\|_{\widehat{\mathbf{H}}_t}^2}_{\text{Estimation}} \underbrace{\|\boldsymbol{x}_t\|_{\widehat{\mathbf{H}}_t^{-1}}^2}_{\text{Second-stage norm}} = \mathcal{O}\left(d_{\boldsymbol{x}} d_{\boldsymbol{z}} \log^2 T\right).$$

Regret bound of Theorem 3.4 is $d_{\boldsymbol{z}} \log T$ more than the regret of online ridge regression, i.e. $\mathcal{O}(d_{\boldsymbol{x}} \log T)$ (Gaillard et al. 2019). This is because we perform $d_{\boldsymbol{x}}$ linear regressions in the first-stage and use the predictions of the first-stage for the second-stage regression. These two regression steps in cascade induce the proposed regret bound.

Oracle Regret Bound. Now, we bound the oracle regret, i.e. the goodness of predictions yielded by O2SLS. Detailed proof is deferred to Appendix B.

Theorem 3.5 (Oracle Regret of O2SLS)**.** *Under the same hypothesis of Theorem 3.4, Oracle Regret of O2SLS at step $T > 1$, i.e. $\overline{R}_T$ is upper bounded by (ignoring $\log \log$ terms)*

$$\underbrace{\widetilde{R}_T}_{\substack{\text{Identif.}\\ \text{Regret}\\ \mathcal{O}(d_{\boldsymbol{x}} d_{\boldsymbol{z}} \log^2 T)}} + \underbrace{\sqrt{\mathfrak{b}_{T-1}(\delta)}}_{\substack{\text{Estimation}\\ \mathcal{O}(\sqrt{d_{\boldsymbol{z}} \log T})}} \Big(\underbrace{C_1 \sqrt{f(T)}}_{\substack{\text{First-stage}\\ \text{feature norm}\\ \mathcal{O}(\sqrt{\log T})}} + \underbrace{C_2 \sqrt{2 d_{\boldsymbol{x}} f(T)} + \sqrt{d_{\boldsymbol{x}}} C_3}_{\substack{\text{Correlated noise}\\ \text{Concentration term}\\ \mathcal{O}(\sqrt{d_{\boldsymbol{x}} \log T})}} + \underbrace{\gamma C_4 \sqrt{T}}_{\substack{\text{Correlated noise}\\ \text{Bias term}\\ \mathcal{O}(\gamma \sqrt{T})}} \Big)$$

with probability at least $1 - \delta \in (0, 1)$. Here, $\gamma \triangleq \|\boldsymbol{\gamma}\|_2 = \|\mathbb{E}[\eta_s \boldsymbol{\epsilon}_s]\|_2$. C_1, C_2, C_3, C_4 are $d_{\boldsymbol{z}}$, $d_{\boldsymbol{x}}$ and T-independent positive constants. $f(T) = \mathcal{O}(\log T)$ is in Corollary B.9.

Discussion. 1. *Hardness of Endogeneity vs. Exogeneity.* Under exogeneity and unbounded stochastic noise, the oracle regret of online linear regression is $\mathcal{O}(d^2 \log^2 T)$ (Ouhamma, Maillard, and Perchet 2021). If we take the just-identified IVs, i.e. $d_{\boldsymbol{x}} = d_{\boldsymbol{z}} = d$, due to endogeneity, O2SLS incurs an additive factor of $\mathcal{O}(\gamma \sqrt{d_{\boldsymbol{z}} T \log T})$ in the oracle regret. This term appears due to the correlation between the second and the first-stage noises, and it is proportional to the degree of correlation between the noises in these two stages. Thus, the bias due to the correlation of noises dominates. In 2SLS literature, this phenomenon is called the *self-fulfilling bias* (Li, Luo, and Zhang 2021). But we did not find any explicit bound on it in a stochastic and non-asymptotic analysis.

2. *Tightness of Analysis.* In the existing literature, we do not have any lower bound for online regression with endogeneity. But we can indicate tightness of the proposed analysis by observing two things. (i) Our identification regret under endoegeneity is of the same order as that of exogeneous case (Abbasi-Yadkori, Pal, and Szepesvari 2012). (ii) If we assume $\gamma = 0$, i.e. the noises are independent, and we retrieve also the oracle regret of the same order as that of the exogenous case (Ouhamma, Maillard, and Perchet 2021).

3. *Bounded vs. Unbounded Noise.* For bounded noise and covariates, the adversarial setting would subsume the endogeneous setting where the responses are sampled according to

dependent bounded noise since covariates and responses are adversarially chosen sequences. If we consider the adversarial linear regression for bounded responses $y_t \leq Y$ and covariates, the the Vovk-Azoury-Warmuth Regressor (VAWR) achieves almost optimal $\mathcal{O}(Y^2 \log T)$ regret (Orabona 2019). Thus, the $\gamma\sqrt{T}$ term disappears as we obtain in our oracle regret bound with unbounded and stochastic noise. This contrast indicates non-triviality of our analysis for the unbounded noise and unbounded outcomes.

4 Linear Bandits with Endogeneity: OFUL-IV

Example 1.1 (Pricing with Price-Sales Dynamics)**.** *Let us revisit the price-sales dynamics. Now, on a day t, a restaurant owner wants to decide on a price x_t among a feasible set of prices $\mathcal{X}_t$ to increase the sales y_t. The owner also has access to a set of suppliers such that each price correspond to compatible a material cost z_t. If the restaurant owner wants the analyst to develop an algorithm to decide on the price dynamically such that the total sales over a year is maximised, her problem is exactly a linear bandit with endoegeneity, where a price corresponds to an arm (action).*

We formulate *stochastic Linear Bandits with Endogeneity* (**LBE**) with a two-stage linear model of data generation (Eq. (1)). Here, we illustrate the interactive protocol of LBE.

At each round $t = 1, 2, \ldots, T$, the agent

1. Observes a sample $\boldsymbol{x}_{t,a} \in \mathcal{X}_t$ of contexts $\forall a \in \mathcal{A}_t$
2. Chooses an arm $A_t \in \mathcal{A}_t$
3. Observes the sample $\boldsymbol{z}_{t,A_t} \in \mathcal{Z}_t$
4. Obtains a reward y_t
5. Updates the parameter estimates $\widehat{\boldsymbol{\Theta}}_t$ and $\boldsymbol{\beta}_t$

Here, $\mathcal{X}_t \subset \mathbb{R}^{d_{\boldsymbol{x}}}$ and $\mathcal{Z}_t \subset \mathbb{R}^{d_{\boldsymbol{z}}}$ are the sets of covariates and IVs corresponding to $\mathcal{A}_t$, i.e. the set of feasible actions at time t. Similar to regression under endoegenity, $\boldsymbol{\epsilon}_t$ and η_t are correlated while $\boldsymbol{\epsilon}_t$ and $\boldsymbol{z}_t$ are not (Eq. (1)). True parameters $\boldsymbol{\beta} \in \mathbb{R}^{d_{\boldsymbol{x}}}$ and $\boldsymbol{\Theta} \in \mathbb{R}^{d_{\boldsymbol{z}} \times d_{\boldsymbol{x}}}$ are unknown to the agents. This is an extension of stochastic linear bandit (Lattimore and Szepesvári 2020, Ch. 19) to the endeogenous setting.

Remark 4.1 (Alternative Protocol of LBE). We can alternatively represent the protocol, where at every step t, the agent observes covariates $\boldsymbol{x}_{t,a}$ and IVs $\boldsymbol{z}_{t,a}$ for all the arms $a \in \mathcal{A}_t$. Then, the agent can use all these information to choose an arm A_t and observes the corresponding outcome y_t. But we do not require $\boldsymbol{z}_{t,a}$ except $\boldsymbol{z}_{t,A_t}$ to update the parameters and select arms. Thus, the alternative is reducible to the protocol above, while asking for less information.

OFUL-IV: Algorithm Design. If the agent had full information in hindsight, she could infer the best arm (a.k.a. action or intervention) in $\mathcal{A}_t$ as $a_t^* = \operatorname{argmax}_{a \in \mathcal{A}_t} \mathbb{E}[\boldsymbol{x}_{t,a}^\top \boldsymbol{\beta}]$. Choosing a_t^* is equivalent to choosing $\boldsymbol{z}_{t,a_t^*}$ and $\boldsymbol{x}_{t,a_t^*}$. But the agent does not know them and aims to select $\{a_t\}_{t=1}^T$ to minimise regret: $R_T \triangleq \mathbb{E}[\sum_{t=1}^T \boldsymbol{\beta}^\top (\boldsymbol{x}_{t,a_t^*} - \boldsymbol{x}_{t,a_t})]$. Now, we extend the OFUL algorithm minimising regret in linear bandits with exogeneity (Abbasi-Yadkori, Pál, and Szepesvári 2011a). The core idea is that the algorithm maintains a confidence set $\mathcal{B}_{t-1} \subset \mathbb{R}^{d_{\boldsymbol{x}}}$ around the estimated parameter $\boldsymbol{\beta}_{t-1}$, which is computed only using the observed data. In order to tackle endogeneity, we choose to use the O2SLS estimate $\boldsymbol{\beta}_{t-1}$ computed using data observed till $t-1$ (Equation (O2SLS)). Then, we build an ellipsoid $\mathcal{B}_{t-1}$ around it, such that $\mathcal{B}_{t-1} \triangleq \left\{\boldsymbol{\beta} \in \mathbb{R}^{d_{\boldsymbol{x}}} : \|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}\|_{\widehat{\mathbf{H}}_{t-1}} \leq \sqrt{\mathfrak{b}'_{t-1}(\delta)}\right\}$, where $\mathfrak{b}'_{t-1}(\delta) \triangleq 2\sigma_\eta^2 \log\left(\det(\mathbf{G}_{\boldsymbol{z},t-1})^{1/2} \lambda^{-d_{\boldsymbol{z}}/2} / \delta\right)$ and $\widehat{\mathbf{H}}_{t-1} = \widehat{\boldsymbol{\Theta}}_{t-1}^\top \mathbf{G}_{\boldsymbol{z},t-1} \widehat{\boldsymbol{\Theta}}_{t-1}$. Then, the algorithm chooses an optimistic estimate $\widetilde{\boldsymbol{\beta}}_{t-1}$ from that confidence set: $\widetilde{\boldsymbol{\beta}}_{t-1} \triangleq \operatorname{argmax}_{\boldsymbol{\beta}' \in \mathcal{B}_{t-1}} \left(\max_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{x}^\top \boldsymbol{\beta}'\right)$. Then, she chooses the action A_t corresponding to $\boldsymbol{x}_{t,A_t} = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{x}^\top \widetilde{\boldsymbol{\beta}}_{t-1}$, which maximises the reward as per the estimate $\widetilde{\boldsymbol{\beta}}_{t-1}$. In brief, the algorithm chooses the pair $(\boldsymbol{x}_{t,A_t}, \widetilde{\boldsymbol{\beta}}_{t-1}) = \operatorname{argmax}_{(\boldsymbol{x},\boldsymbol{\beta}') \in \mathcal{X}_t \times \mathcal{B}_{t-1}} \boldsymbol{x}^\top \boldsymbol{\beta}'$. Given confidence interval, we optimistically choose A_t by solving

$$\operatorname*{argmax}_{a \in \mathcal{A}_t} \left\{ \left\langle \boldsymbol{x}_{t,a}, \boldsymbol{\beta}_{t-1} \right\rangle + \sqrt{\mathfrak{b}'_{t-1}(\delta)}\, \|\boldsymbol{x}_{t,a}\|_{\widehat{\mathbf{H}}_{t-1}^{-1}} \right\} \tag{6}$$

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

Algorithm 2: OFUL-IV

Input: Initialisation parameters $\boldsymbol{\beta}_0, \widehat{\boldsymbol{\Theta}}_0, \mathfrak{b}'_0, \lambda$
for $t = 1, 2, \ldots, T$ **do**
  Observe $\boldsymbol{x}_{t,a} \in \mathcal{X}_t$ for $a \in \mathcal{A}_t$
  Compute $\boldsymbol{\beta}_{t-1}$ according to Equation (O2SLS)
  Choose action A_t that solves Equation (6)
  Observe $\boldsymbol{z}_{t,A_t}$ and y_t
  Update $\boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}_{t-1}, \widehat{\boldsymbol{\Theta}}_t \leftarrow \widehat{\boldsymbol{\Theta}}_{t-1}, \mathfrak{b}'_t \leftarrow \mathfrak{b}'_{t-1}$

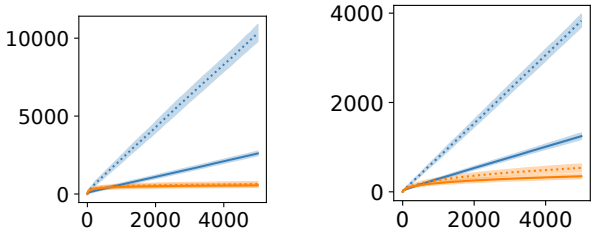
Theorem 4.2 (Regret Upper Bound of OFUL-IV)**.** *Under the assumptions and notations of Thm. 3.4 and 3.5, for horizon $T > 1$ and with probability $1 - \delta$, Algorithm 2 incurs regret*

$$R_T \leq \sqrt{4T} \underbrace{\sqrt{\mathfrak{b}_{T-1}(\delta)}}_{\textit{Estimation}} \underbrace{\sqrt{\sum_{t=1}^T \|\boldsymbol{x}_{t,A_t}\|^2_{\widehat{H}_t^{-1}}}}_{\textit{Second-stage feature norm}} = \tilde{\mathcal{O}}(\sqrt{d_{\boldsymbol{x}} d_{\boldsymbol{z}} T})$$

For just-identified IVs, i.e. $d_{\boldsymbol{x}} = d_{\boldsymbol{z}} = d$, OFUL-IV achieves regret of similar order under endogeneity as OFUL achieves under exogeneity, i.e. $\widetilde{\mathcal{O}}(d\sqrt{T})$, and matches the lower bounds for linear bandits w.r.t. d and T (Lattimore and Szepesvári 2020). The proof details are in Appendix C.

1. *Handling Unbounded Outcomes.* We do not need bounded outcomes ($\boldsymbol{\beta}^\top \boldsymbol{x} \in [-1, 1]$) as OFUL (Abbasi-Yadkori, Pál, and Szepesvári 2011a). Similar to (Ouhamma, Maillard, and Perchet 2021), which removes this assumption for exoegeneity, our analysis shows that the bounded outcome assumption is not needed even under endogeneity.

2. *Two-Stage OFUL-IV vs. One-Stage OFUL under Endogeneity.* An alternative proposal than using a two-stage approach for LBE is to reduce it to a one-level linear bandit, i.e. $y_t = (\boldsymbol{\Theta}\boldsymbol{\beta})^\top \boldsymbol{z}_t + \boldsymbol{\beta}^\top \boldsymbol{\epsilon}_t + \eta_t$. This leads to a composite unknown parameter $\boldsymbol{\Theta}\boldsymbol{\beta}$, and a new sub-Gaussian noise with component-wise variance $\sigma_{new}^2 = 2(\|\boldsymbol{\beta}\|_2^2 \sigma_{\boldsymbol{\epsilon}}^2 + \sigma_\eta^2)$.



(a) **Regression:** id. regret. (b) **LBE:** cumulative regret.

Figure 2: (Left) **Regression:** Identification regrets of Online Ridge (*blue*) and O2SLS (*orange*) over $T = 5000$ steps, and for $\rho = 1$ (solid) and 2 (dotted). *With the increase in ρ, O2SLS performs better.* (Right) **LBE:** Cumulative regrets of OFUL (*blue*) and OFUL-IV (*orange*) over $T = 5000$ steps with $\rho = 1, 2$. OFUL-IV *incurs lower regrets, and improvements w.r.t.* OFUL *increases with ρ. Here, $d_{\boldsymbol{x}} = 8, d_{\boldsymbol{z}} = 16$.*

	Online Regression	Linear Bandit
1st Stage	$\boldsymbol{x}_t = \boldsymbol{\Theta}\, \boldsymbol{z}_t + \boldsymbol{\epsilon}_t$	$\boldsymbol{x}_{t,a} = \boldsymbol{\Theta} \boldsymbol{z}_{t,a} + \boldsymbol{\epsilon}_t$
2nd Stage	$y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t + \eta_t$	$y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_{t,A_t} + \eta_t$

Table 1: Experimental settings. ρ controls endogeneity since $\eta_t = \rho \epsilon_{t,1} + \widetilde{\eta}_t$ and $\widetilde{\eta}_t$ is exogenous noise.

Now, if we apply OFUL to this setup, we observe that the confidence bound around the new hidden parameter $\boldsymbol{\Theta}\boldsymbol{\beta}$ is proportional to σ_{new}. Thus, the confidence interval blows up by $\|\boldsymbol{\beta}\|_2$. Also, using this interval in Equation (6) requires us to know either $\|\boldsymbol{\beta}\|_2^2$ or a tight upper bound on it. In real-life, we cannot assume access to such knowledge. Additionally, $\|\boldsymbol{\beta}\|_2^2$ grows linearly with $d_{\boldsymbol{x}}$. Thus, this approach can incur significantly higher regret for high-dimensional covariates. In contrast, OFUL-IV does not require to know either $\|\boldsymbol{\beta}\|_2$ or any bound on it. Additionally, the confidence interval of OFUL-IV depends on σ_η^2, which is significantly smaller than σ_{new}^2. These observations demonstrate the benefit of using OFUL-IV under endogeneity than OFUL with a single-stage reduction. Single-stage reduction might be helpful if the confounding noise is covariate dependent but is independent of the unknown parameter (Krishnamurthy, Wu, and Syrgkanis 2018). Numerical results in Fig. 3 further validates this benefit of OFUL-IV over an one-stage reduction.

5 Experimental Analysis

We present experimental results for both online regression and linear bandit in Figure 2a, 2b and 3. We compare the performance of O2SLS and Online Ridge Regression (Ridge). For LBEs, we compare the performance of OFUL-IV and OFUL (Abbasi-Yadkori, Pál, and Szepesvári 2011a).

Setup. We induce endogeneity in the problem in the following arbitrary way: by settings $\eta_t = \rho \epsilon_{t,1} + \widetilde{\eta}_t$ where $\epsilon_{t,1}$ indicates the first component of the vector $\boldsymbol{\epsilon}_t$. Then, we control the level of endogeneity of the two stages through ρ.

We choose $d_{\boldsymbol{x}} = \{2, 5, 8\}$ and $d_{\boldsymbol{z}} = \{4, 10, 16\}$ respectively. In our experiments, we choose arbitrarily $\boldsymbol{\beta}$ as a normalised vector with equal negative entries; therefore, the values in the components are uniquely determined by the dimension $d_{\boldsymbol{x}}$. We choose $\boldsymbol{\Theta} = \mathbf{I}_{d_{\boldsymbol{z}}, d_{\boldsymbol{x}}}$ which has ones on the entries $i = j$ and zeros for $i \neq j$. Then, we sample at each time t (and also for every arm a for the LBE setting) the vectors $\boldsymbol{z}_t \sim \mathcal{N}_{d_{\boldsymbol{z}}}(\vec{0}, \mathbf{I}_{d_{\boldsymbol{z}}})$ ($\boldsymbol{z}_{t,a} \sim \mathcal{N}_{d_{\boldsymbol{z}}}(\vec{0}, \mathbf{I}_{d_{\boldsymbol{z}}})$), the vector noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}_{d_{\boldsymbol{x}}}(\vec{0}, \mathbf{I}_{d_{\boldsymbol{x}}})$, and the scalar noise $\eta_t = \widetilde{\eta}_t + \rho \cdot \epsilon_{t,1}$ where $\widetilde{\eta}_t \sim \mathcal{N}_1(0, 1)$. We run the algorithms with the same regularisation parameters, i.e. $\lambda = 0.1$. We repeat our experiments 20 times. We average the results, and for each algorithm, we report the mean and standard deviation of the cumulative regret (shaded areas correspond to one standard deviation). For further experiments and results with both synthetic and real data, we refer to Appendix D.

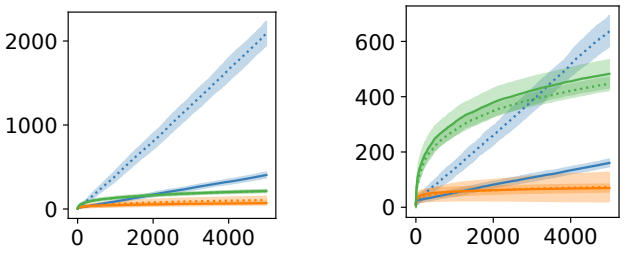


Figure 3: Regrets due to one-stage OFUL (green), OFUL (blue), and OFUL-IV (orange) for different norms of the hidden parameters ($S = 2$ (left), 5 (right)). Only OFUL-IV (orange) is independent of S and incurs the lowest regret across $\rho = 1$ (solid) and 2 (dotted). Here, $d_{\boldsymbol{x}} = 2, d_{\boldsymbol{z}} = 4$.

Summary of Results. *1. Regression.* O2SLS outperforms Ridge in all the settings, and the performance-gain increases with increasing values of ρ, i.e. the level of endogeneity. *2. Bandits.* OFUL builds a confidence ellipsoid centered at $\boldsymbol{\beta}_t^{\text{Ridge}}$, while OFUL-IV uses O2SLS to build an accurate estimate and an ellipsoid containing $\boldsymbol{\beta}$. Figure 2 indicates that OFUL-IV incurs lower regret than OFUL. *3. Independence of* OFUL-IV *from S.* Figure 3 shows that only OFUL-IV's regret is independent of the maximum norm of $\boldsymbol{\beta}$, i.e. S. But OFUL (oblivious to endogeneity) and OFUL with one-stage reduction incur higher regret with increasing S as per theory.

6 Conclusion and Future Works

We study online IV regression, specifically online 2SLS, for unbounded noise and endogenous data. Our analysis shows that O2SLS incurs $\mathcal{O}(d_{\boldsymbol{x}} d_{\boldsymbol{z}} \log^2 T)$ identification regret, and $\mathcal{O}(\gamma \sqrt{d_{\boldsymbol{z}} T \log T})$ oracle regret as the correlation between the noises in the two-stages dominate the identification regret. For just-identified IVs, identification regrets are of the same order for exogeneity and endogeneity. We propose OFUL-IV for linear bandits with endogeneity that uses O2SLS for estimation. OFUL-IV achieves $\mathcal{O}(\sqrt{d_{\boldsymbol{x}} d_{\boldsymbol{z}} T \log T})$ regret. We experimentally show that O2SLS and OFUL-IV are more accurate than Ridge and OFUL, respectively. In future, it would be interesting to extend the analysis of online IV regression and bandits with endogeneity to non-parametric (Newey and Powell 2003) and non-linear (Xu et al. 2020) settings.

Acknowledgements

This work is supported by the CHIST-ERA project CausalXRL (ANR-21-CHR4-0007). D. Basu also acknowledges the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01) and the PEPR project FOUNDRY (ANR23-PEIA-0003).

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011a. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

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

Abbasi-Yadkori, Y.; Pal, D.; and Szepesvari, C. 2012. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, 1–9. PMLR.

Abeille, M.; and Lazaric, A. 2017. Linear Thompson Sampling Revisited. In *AISTATS*.

Angrist, J. D.; and Imbens, G. W. 1995. 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.

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

Bartlett, P. L.; Koolen, W. M.; Malek, A.; Takimoto, E.; and Warmuth, M. K. 2015. Minimax fixed-design linear regression. In *Conference on Learning Theory*, 226–239. PMLR.

Bennett, A.; Kallus, N.; and Schnabel, T. 2019. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32.

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78: 1–3.

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

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

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

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

Gaillard, P.; Gerchinovitz, S.; Huard, M.; and Stoltz, G. 2019. Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss. In *Algorithmic Learning Theory*, 404–432. PMLR.

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

Harris, K.; Ngo, D. D. T.; Stapleton, L.; Heidari, H.; and Wu, S. 2022. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, 8502–8522. PMLR.

Hartford, J.; Lewis, G.; Leyton-Brown, K.; and Taddy, M. 2016. Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*.

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

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

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

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

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

Krishnamurthy, A.; Wu, Z. S.; and Syrgkanis, V. 2018. Semiparametric contextual bandits. In *International Conference on Machine Learning*, 2776–2785. PMLR.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Li, J.; Luo, Y.; and Zhang, X. 2021. Self-fulfilling Bandits: Dynamic Selection in Algorithmic Decision-making. *arXiv preprint arXiv:2108.12547*.

Littlestone, N.; Long, P. M.; and Warmuth, M. K. 1991. Online learning of linear functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, 465–475.

Liu, R.; Shang, Z.; and Cheng, G. 2020. On deep instrumental variables estimate. *arXiv preprint arXiv:2004.14954*.

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

Nareklishvili, M.; Polson, N.; and Sokolov, V. 2022. Deep Partial Least Squares for IV Regression. *arXiv preprint arXiv:2207.02612*.

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

Orabona, F. 2019. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

Ouhamma, R.; Basu, D.; and Maillard, O.-A. 2022. Bilinear Exponential Family of MDPs: Frequentist Regret Bound with Tractable Exploration and Planning. *arXiv preprint arXiv:2210.02087*.

Ouhamma, R.; Maillard, O.; and Perchet, V. 2021. Stochastic Online Linear Regression: the Forward Algorithm to Replace Ridge. *arXiv preprint arXiv:2111.01602*.

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

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

Stirn, A.; and Jebara, T. 2018. Thompson Sampling for Noncompliant Bandits. *arXiv preprint arXiv:1812.00856*.

Venkatraman, A.; Sun, W.; Hebert, M.; Bagnell, J.; and Boots, B. 2016. Online instrumental variable regression with applications to online linear system identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

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

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

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

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

Xu, L.; Chen, Y.; Srinivasan, S.; de Freitas, N.; Doucet, A.; and Gretton, A. 2020. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*.

Xu, L.; Kanagawa, H.; and Gretton, A. 2021. Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation. *arXiv preprint arXiv:2106.03907*.

Zhu, Y.; Gultchin, L.; Gretton, A.; Kusner, M. J.; and Silva, R. 2022. Causal inference with treatment measurement error: a nonparametric instrumental variable approach. In *Uncertainty in Artificial Intelligence*, 2414–2424. PMLR.