Causal Inference and Decision Making on Digital Platforms

Jingwen Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Yong Tan, Chair

Stephanie Lee

Amandeep Singh

Mingwen Yang

Program Authorized to Offer Degree:

Business Administration

University of Washington

Abstract

Causal Inference and Decision Making on Digital Platforms

Jingwen Zhang

Chair of the Supervisory Committee:
Yong Tan
Department of Information Systems and Operations Management

This thesis examines advanced methodological approaches for addressing complex inference and decision-making challenges in digital and economic environments characterized by uncertainty, measurement error, and endogeneity. The development of robust methodological solutions in this domain is increasingly critical as organizations rely on algorithmic decision-making for high-stakes business decisions, platforms design experimentation systems for continuous optimization, and researchers leverage AI-generated variables for econometric analyses. Traditional methods often fail in these contexts, potentially leading to biased inferences, suboptimal decisions, and misguided strategies that can significantly impact business outcomes and research validity. Through three interconnected essays, this thesis demonstrates novel frameworks that enhance both theoretical understanding and practical applications in causal inference and decision making for online and offline settings, providing essential methodological tools for researchers and practitioners navigating the complexities of modern digital economics.

The first essay introduces the ϵ-BanditIV algorithm to address endogeneity issues in online dynamic decision-making contexts. By incorporating instrumental variables into the Multi-Armed Bandit (MAB) framework, this approach achieves both optimal regret minimization and asymptotically consistent parameter estimation, establishing a methodological foundation for unbiased causal inference in adaptive experimentation settings with endogenous covariates.

The second essay addresses the critical challenge of measurement error in machine learning

(ML) or artificial intelligence (AI) generated regressors within partially linear models. By developing estimators that utilize Two-Stage Least Square (TSLS) and Generalized Method of Moments (GMM) under the Double Machine Learning (DML) framework, this work provides a robust solution for debiasing predictions, advancing causal inference where ML/AI tools are used to generate feature variables for econometric analysis.

The third essay investigates decision-making processes in livestream e-commerce, focusing on hosts' product selection and presentation timing behaviors. Using structural models that combine online learning frameworks with survival analysis, this research reveals how livestream hosts balance exploration-exploitation trade-offs and optimize product transitions to maximize sales performance, providing both theoretical insights and practical recommendations for livestream commerce optimization.

Together, these essays contribute to advancing methodological approaches across econometrics, machine learning, and information systems, offering frameworks that address fundamental challenges in causal inference and decision-making. The first two essays develop theoretical solutions to estimation challenges that threaten causal validity in both sequential learning environments and static models with ML/AI-generated variables. The third essay then demonstrates how structural modeling of online learning processes can reveal opportunities for improved decision-making in practice. This complementary relationship between theoretical advances in causal inference and empirical applications in decision optimization creates a cohesive contribution with significant implications for marketing, e-commerce, and digital platform design in environments characterized by dynamic interactions, endogeneity concerns, and imperfect information.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my advisor, Professor Yong Tan. Prof. Tan has been an exceptional mentor and role model for me, both in research and in life. He has consistently supported me throughout my Ph.D. journey, offering valuable guidance whenever I faced challenges in research or teaching. His unwavering encouragement and insightful suggestions have been instrumental in my academic development. I feel truly fortunate to have had Prof. Tan as my advisor.

My sincere thanks also go to my committee members, Professor Stephanie Lee, Professor Mingwen Yang, Professor Amandeep Singh, and Professor Shan Liu. Professor Lee and Professor Singh, as my coauthors, have shared their research expertise and collaborated with me in ways that have greatly enriched my academic experience. Professor Yang, with whom I've had the privilege to work as a Teaching Assistant, has provided invaluable mentorship in teaching and serves as another role model in research for me. I am particularly grateful to Professor Liu for her support as my Graduate School Representative during my dissertation exams and throughout the committee process. Their collective guidance and support have been crucial throughout my Ph.D. journey.

I would also like to extend my appreciation to all the other faculties and staff at the Foster School of Business for their support and assistance throughout my doctoral studies. My gratitude also goes to my fellow doctoral students who have shared this challenging yet rewarding journey with me, offering both academic insights and friendship along the way.

I am deeply grateful to my parents and grandparents for their unconditional love, unwavering belief in my abilities, and constant encouragement. Their support has been my foundation throughout this endeavor. I would also like to thank my friends who have stood by me, providing both moral support and welcome distractions when needed.

DEDICATION

to those who believe in the pursuit of knowledge.

Chapter 1

INTRODUCTION

The digital transformation of economic activities has created unprecedented opportunities for data-driven decision-making while simultaneously introducing complex methodological challenges in causal inference and optimization. As businesses, platforms, and policymakers increasingly rely on algorithms and automated systems to guide their strategies, addressing issues such as endogeneity, measurement error, and exploration-exploitation trade-offs has become crucial for ensuring accurate analysis and effective decision-making. Traditional econometric methods often struggle to account for the dynamic nature of these environments, where decisions are sequential, covariates may be endogenous, and many key variables are generated through imperfect machine learning or artificial intelligence systems. These complexities call for innovative methodological approaches that can preserve causal validity while adapting to the unique characteristics of digital and online settings. This thesis addresses these challenges through three interconnected essays that develop novel frameworks for robust causal inference and online decision-making. Each essay tackles a specific methodological problem at the intersection of econometrics, machine learning, and online learning, together forming a comprehensive approach to inference and decision-making in complex digital environments.

Chapter 3 examines a critical limitation in the application of Multi-Armed Bandit (MAB) algorithms: the prevalent assumption of exogeneity in arm covariates. In many practical applications, arm characteristics are often correlated with unobserved error terms, leading to biased estimation and suboptimal decisions when traditional bandit algorithms are employed. To address this challenge, this essay develops ϵ-BanditIV, an algorithm that incorporates instrumental variables into the online learning framework to correct for endogeneity bias. The approach achieves two simultaneous objectives: (1) it minimizes expected regret with an upper bound of $\tilde{O}(k\sqrt{T})$, where k is

the dimension of the instrumental variable and T is the number of rounds, and (2) it ensures the asymptotic consistency and normality of parameter estimates, enabling valid statistical inference. Through extensive Monte Carlo simulations and empirical applications to iOS app pricing and Real-Time Bidding data, this essay demonstrates that ϵ-BanditIV significantly outperforms existing methods in endogenous settings. The developed framework provides both theoretical contributions to the bandit literature and practical tools for marketers to simultaneously optimize decisions and conduct valid causal inference in adaptive experimentation settings.

Chapter 4 addresses an increasingly common challenge in information systems research: bias introduced by measurement errors in machine learning (ML) or artificial intelligence (AI) generated regressors. As researchers increasingly leverage ML/AI technologies to predict feature variables from unstructured data for use in econometric models, the imperfections in these predictions can lead to biased estimation and inaccurate conclusions. This essay examines the problem of debiasing ML/AI-generated regressors specifically in partially linear regression models, which have gained popularity for their ability to effectively combine traditional econometric analysis with powerful machine learning approaches. The developed methodology extends existing debiasing frameworks to semi-parametric settings and proposes estimators utilizing Two-Stage Least Square (TSLS) and Generalized Method of Moments (GMM) under the Double Machine Learning (DML) framework. Through theoretical analysis, the essay establishes the asymptotic consistency and normality of the proposed estimators. Monte Carlo simulations and empirical applications demonstrate that these methods outperform alternative approaches in correcting measurement error bias. The work advances causal inference in information systems research by providing robust tools for addressing the measurement error problems that commonly arise when using ML/AI-generated variables in economic analysis.

Chapter 5 investigates the strategic decision-making processes of hosts in livestream e-commerce, an increasingly important sales channel in digital markets. Focusing on two fundamental decisions that hosts must navigate—which products to present and when to transition between different items—this essay develops a novel structural modeling approach that combines online learning frameworks with survival analysis. Employing customized Multi-Armed Bandit models, the re-

search examines how hosts balance exploration and exploitation in product selection within the dynamic livestream environment. The analysis reveals that hosts explore with consistently high probability, with exploration behavior significantly influenced by platform activity levels, social network characteristics, previous livestream experience, and demographic factors. Using a stratified Cox proportional hazards model, the essay further demonstrates how product set characteristics and livestream environment features impact hosts' presentation timing decisions. Based on these empirical findings, simulation experiments evaluate alternative product selection strategies, showing that modified approaches such as Thompson Sampling and adaptive ϵ-Greedy strategies can significantly improve sales performance. The effectiveness of these strategies varies substantially across host characteristics, suggesting the value of targeted, customized support systems for different host segments.

This thesis makes several important contributions to the fields of econometrics, machine learning, and information systems. First, it advances the theoretical understanding of causal inference in complex digital environments by developing frameworks that address endogeneity, measurement error, and strategic learning. Second, it provides practical methodological tools that can be implemented by researchers, platforms, and practitioners to improve decision-making and estimation in real-world settings. Third, it offers specific insights into important domains including adaptive experimentation, ML/AI integration in economic analysis, and livestream commerce optimization.

The remainder of this thesis is organized as follows: Chapter 2 shows related literature. Chapters 3, 4, and 5 present the three essays in their entirety. Chapter 6 synthesizes the findings, discusses their broader implications, and outlines directions for future research. Finally, the appendices provide supplementary technical details, proofs, and additional empirical results to support the main analyses presented in the essays.

Chapter 2

LITERATURE REVIEW

This thesis contributes to several interconnected streams of literature spanning econometrics, machine learning, information systems, and marketing. By developing methodological frameworks that address endogeneity, measurement error, and strategic decision-making in digital environments, this work advances both theoretical understanding and practical applications across these domains. The first essay relates primarily to instrumental variables and online learning literature, developing a novel approach to address endogeneity in contextual multi-armed bandits. The second essay builds on measurement error correction in ML/AI-generated variables and extends this to partially linear models. The third essay connects to livestreaming commerce and online learning models, specifically exploring how hosts make dynamic product selection decisions. The following literature review synthesizes the relevant research that informs and contextualizes these three essays.

2.1 Causal Inference with Endogeneity

2.1.1 Instrumental Variables and Endogeneity Correction

This thesis relates closely to the literature on instrumental variable methods for addressing endogeneity. Seminal works by Griliches (1977), Hausman (1983), Angrist and Imbens (1995), and Chen et al. (2005) provided theoretical analysis of instrumental variables in linear models. These foundational approaches have proven essential for establishing causal relationships when explanatory variables are correlated with error terms.

However, unlike standard offline analysis, endogeneity issues can be exacerbated by the dynamic interactions between data generation and analysis that arise in online settings (Li et al., 2021). The first essay in this thesis builds on offline instrumental variable techniques while ac-

counting for the complications of online learning. It demonstrates how instrumental variables can be used in online settings to improve decision-making, for instance, in the estimation of price effects and ad exposure effects, thereby enhancing decision policy performance.

Prior work by Ngo et al. (2021) and Kallus (2018) utilized arm pulls as an instrumental variable for a second stage regression after running a standard Multi-Armed Bandit (MAB) to address the noncompliance problem, where the treatment assigned may differ from the treatment applied. The approach in this thesis deviates significantly from theirs as it focuses on non-independent endogenous arm covariates with fully compliant treatment.

2.1.2 Measurement Error in ML/AI-Generated Variables

As an innovative endeavor in the information systems literature, researchers have begun addressing bias stemming from measurement error in ML/AI-generated variables. Yang et al. (2018) examined a simulation-based method, MC-SIMEX, to correct estimation bias arising from regressing outcomes on ML/AI-generated variables in the Ordinary Least Square (OLS) model. Qiao and Huang (2021) theoretically studied consistent estimators for cases where ML/AI-generated variables are discrete valued and outcome models are generalized linear models.

Fong and Tyler (2021) constructed a Generalized Method of Moments (GMM) estimator utilizing a sample splitting method to rectify bias in linear regression models and provided a theoretical guarantee of consistency and asymptotic normality. Wei and Malik (2022) proposed a two-step learning strategy for linear regression models, discrete choice models, instrumental variables, and general structural models, demonstrating its performance through simulations.

More recently, Yang et al. (2022) introduced an innovative method to address measurement error bias by exploiting random forest algorithms, extending their utility beyond prediction to create instrumental variables that facilitate bias correction. Burtch et al. (2023) proposed a three-part "EnsembleIV" method that mitigates measurement error by generating candidate instruments from ensemble learning techniques, transforming these instruments to satisfy exclusion conditions, and selecting strong instruments for unbiased instrumental variable regression estimates. Allon et al. (2023) considered a general framework of measurement errors in parametric models, showing how

bias would be introduced by measurement error in ML algorithms and providing two consistent correction methods.

The second essay in this thesis expands upon this existing theoretical framework by proposing two-step estimators and GMM estimators for partially linear models, while keeping linear models as a special case. This work is particularly innovative in addressing measurement error in ML/AI-generated regressors within the increasingly popular class of partially linear models.

2.2 Online Learning and Multi-Armed Bandits

2.2.1 Interactive Learning and Inference with MABs

Multi-armed bandit (MAB) algorithms have garnered surging attention for adaptive experimentation in digital marketing. Examples include applications in advertising (Schwartz et al., 2017; Tang et al., 2015), healthcare (Durand et al., 2018), pricing (Misra et al., 2019), and personalized product recommendations (Li et al., 2010; Qin et al., 2014; Aramayo et al., 2022). The first and third essays in this thesis build upon the contextual multi-armed bandit framework (Auer, 2002), which utilizes contextual information of arms in the learning and decision-making process.

Since the formulation of the contextual bandit problem, a wide range of variants have been investigated with corresponding solutions (Yang and Zhu, 2002; Dani et al., 2008; Rigollet and Zeevi, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agarwal et al., 2012; Perchet and Rigollet, 2013; Qian and Yang, 2016; Bastani and Bayati, 2020). However, the extant literature has relied on the independence between the contexts and unobserved error term, ignoring the potential correlation between them. The first essay extends this literature by accounting for potential endogeneity using instrumental variables.

The theoretical performance of bandit algorithms is evaluated by examining their regret performance, defined as the difference between the total reward obtained by consistently selecting the optimal arm and the actual total reward obtained by the algorithm's decisions. Early work by Vovk (1997) used linear regressions under the bandit framework and showed an $\mathcal{O}(d \log T)$ regret assuming bounded outcomes. Both SupLinRel by Auer (2002) and SupLinUCB by Chu et al. (2011)

achieved a bounded regret of $\mathcal{O}(\log^{3/2}(KT\log T)\sqrt{dT})$, where K is the number of arms.

To address issues with large K, more practical linear bandit algorithms independent of arm numbers were developed. Dani et al. (2008) used confidence ellipsoids to achieve regrets upper bounded by $\tilde{\mathcal{O}}(d\sqrt{T})$ and $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. State-of-the-art performance is generally attributed to linear bandit algorithms by Rusmevichientong and Tsitsiklis (2010) and Abbasi-Yadkori et al. (2011), which achieve regrets of $\tilde{\mathcal{O}}(d\sqrt{T})$ with bounded action sets. By comparison, the ε-BanditIV algorithm developed in the first essay achieves an upper bound of $\tilde{\mathcal{O}}(k\sqrt{T})$, where k denotes the number of instruments, matching state-of-the-art performance in the just-identified case.

2.2.2 Learning and Exploration Models in Practice

The third essay of this thesis applies online learning frameworks, especially Multi-Armed Bandits (MAB), to understand hosts' product assortment decisions under uncertainty in livestream e-commerce. Two specific approaches within this framework are applied: ε-Greedy and Thompson Sampling (TS).

The ε-Greedy bandits provide a straightforward yet effective method for addressing the exploration-exploitation dilemma. In this approach, the algorithm exploits the current best-known action with probability $1-\varepsilon$, while exploring random actions with probability ε (Sutton, 2018). This strategy ensures continuous exploration of the action space while increasingly focusing on high-reward actions.

Thompson sampling offers a Bayesian approach to the MAB problem, maintaining a posterior distribution over unknown parameters and sampling from this distribution to make decisions (Agrawal and Goyal, 2013). Thompson sampling has shown excellent empirical performance and has theoretical guarantees in many settings. In online decision-making, it has been successfully applied to dynamic pricing (Bastani et al., 2022; Ferreira et al., 2018; Jain et al., 2024), transportation (Lagos et al., 2024), personalized recommendations (Zhou et al., 2023), and vaccination testing (Wang et al., 2024).

The contextual MAB framework is utilized in the third essay to model hosts' product selection decisions. Contextual MABs extend the traditional MAB framework by incorporating relevant

contextual information to guide decision-making, allowing for more nuanced and adaptive decisions.

2.3 Partially Linear Models and Double Machine Learning

Partially linear models have gained increasing popularity in econometrics theory and real-world applications. These models facilitate the effective utilization of various ML algorithms for causal inference, as detailed by Chernozhukov et al. (2018). More generally, partially linear specifications significantly enhance the compatibility of causal models with recent substantial advancements in ML.

Partially linear models can alleviate endogeneity concerns resulting from misspecifications of linear models (e.g., omitting high-order or interaction terms) in empirical research (Yu et al., 2020; Wang et al., 2021). They are especially useful in high-dimensional causal inference where the number of control variables can be substantial. For instance, when studying user behaviors on digital platforms, there may be thousands of variables of user profiling that need to be controlled. Partially linear models can adeptly integrate dimension-reduction ML algorithms (e.g., random forest, LASSO, kernel ridge regression), effectively addressing challenges posed by the curse of dimensionality (Chernozhukov et al., 2018).

Under the assumption of a partially linear model, Double Machine Learning (DML) harnesses intricate black-box ML models to generate consistent estimates of parameters of interest from observational data with high-dimensional covariates (Fingerhut et al., 2022). The DML method has been extensively employed in empirical contexts for causal inference. Yu et al. (2020) exploited a DML framework to discern the influence of negative emotional expression on the diffusion of online content. Dube et al. (2020) used a DML estimator to identify how task vacancy durations are influenced by task rewards and provided empirical evidence of monopsony in a large online labor market. Knaus (2021) conducted DML estimation to investigate the impact of engaging in musical practice on child development, while Knaus (2022) employed a DML-based approach for assessing labor market policies. Ye et al. (2023) integrated deep learning with DML estimation.

The second essay of this thesis extends the DML framework to address measurement error in

ML/AI-generated regressors, thereby advancing causal inference in information systems research with partially linear models.

2.4 Adaptive Experimentation and Market Applications

2.4.1 Adaptive Experimentation in Business

There has been significant recent research on how digital advertising impacts outcomes of interest like demand (Gordon et al., 2021; Zantedeschi et al., 2017; Rafieian and Yoganarasimhan, 2021). Product recommendations have gained traction as a form of advertising in digital markets to directly stimulate consumer demand. Extensive literature has studied product recommendations in digital markets, tracing back to early work by Ansari et al. (2000), Koren et al. (2009), Agrawal et al. (2013), and Dzyabura and Tuzhilin (2013).

To address the challenge of scaling advertisements and personalized recommendations cost-effectively, adaptive experimentation has gained popularity. Adaptive experiments can reduce costs compared to static designs by requiring smaller sample sizes, allowing for early stopping when results are conclusive, dropping ineffective treatment arms, optimizing resource allocation to more promising treatments, and avoiding failed experiments.

Dzyabura and Hauser (2011) proposed an active machine learning method to select questions and learn consumers' heuristic decision rules adaptively. Nuara et al. (2018) modeled online joint advertising bid optimization into a combinatorial-bandit problem and used Gaussian Process regressions to estimate the ad effect. Recent work (Zhao et al., 2013; Wang et al., 2016, 2017) has explored product recommendation through adaptive experimentation.

For treatment effect estimation, Dimakopoulou et al. (2021) proposed doubly adaptive Thompson sampling to conduct causal inference of true means of arm rewards sequentially. Simchi-Levi and Wang (2023) mathematically derived conditions for a Pareto optimal solution where both the statistical power of treatment inference and efficiency are optimal in Multi-armed bandit experimental design.

The first essay of this thesis contributes to this literature by addressing endogeneity concerns in

adaptive experimentation settings, which have been largely ignored in previous work despite their prevalence in marketing applications.

2.4.2 Livestreaming Commerce

The third essay examines decision-making in livestreaming e-commerce, a rapidly evolving field that has attracted significant scholarly attention. Recent studies have explored various aspects of this novel sales channel, each contributing to our understanding of this dynamic marketplace.

Coupon targeting strategies have been examined by Liu (2023), who developed dynamic approaches to optimize promotional efforts in livestream sales. Lin et al. (2022) examined complex negotiation dynamics between brands and key opinion leaders in livestream commerce. The critical role of influencer selection in livestream marketing was studied by Gu et al. (2024a), offering insights into factors determining influencer effectiveness.

Lin et al. (2021) and Bharadwaj et al. (2022) investigated the impact of livestream sellers' emotional displays on consumer engagements and purchasing decisions, highlighting the importance of affective factors in this interactive sales environment. Lo et al. (2022) focused on determinants influencing consumers' impulsive buying behaviors within the context of livestreaming. Xie et al. (2022) examined the effect of presentation duration on sales, comparing outcomes across single-brand and multi-brand livestreams. Gu et al. (2024b) examined how live streamers should balance entertainment content versus product information to maximize commission.

Despite this growing body of research, the third essay addresses a significant gap in understanding how livestream hosts navigate the complex decision-making process regarding product assortment and presentation timing. It introduces a novel approach to modeling the dynamic decision-making process of livestream hosts using a contextual Multi-Armed Bandit framework, which captures the sequential nature of decisions in an environment characterized by uncertainty.

2.5 Research Gaps and Contributions

The literature review reveals several key research gaps that this thesis addresses:

1. **Endogeneity in Online Learning**: Despite the widespread application of MAB algorithms in marketing and recommendation systems, existing approaches fail to account for endogeneity in arm covariates, potentially leading to biased estimates and suboptimal decisions. The first essay directly addresses this gap by developing an instrumental variable approach within the MAB framework.

2. **Measurement Error in ML/AI-Generated Variables**: As researchers increasingly use ML/AI tools to generate variables for econometric analysis, the measurement error in these predictions can bias subsequent analyses. While some recent work has begun addressing this in linear models, the second essay extends these approaches to the increasingly popular partially linear models.

3. **Empirical Understanding of Exploration-Exploitation Decisions**: While theoretical models of exploration-exploitation trade-offs are well-developed, there is limited empirical research on how decision-makers navigate these trade-offs in practice. The third essay fills this gap by applying structural models to understand product selection and timing decisions in livestream commerce.

By addressing these gaps, this thesis contributes to advancing methodological approaches in causal inference and decision-making under uncertainty, with significant implications for research and practice in digital marketing, e-commerce, and platform economics.

Chapter 3

CAUSAL BANDITS: ONLINE DECISION-MAKING IN ENDOGENOUS SETTINGS

3.1 Problem Setting

3.1.1 Problem Introduction

Let T be the number of rounds and K the number of arms in each round. In each round t, the learner observes a feature vector $x_{t,a} \in \mathbb{R}^d$, $a \in \{1, \cdots, K\}$, with $\|x_{t,a}\|_2 \leq L_x$, for each arm. We use the subscript a to represent the a^{th} arm from the K-arm set and L_x to represent a constant upper bound for $\|x_{t,a}\|_2$. After observing the feature vectors, the learner selects an arm a and receives a reward $y_{t,a} \in \mathbb{R}$, with $|y_{t,a}| \leq L_y$ where L_y is a constant upper bound for $|y_{t,a}|$. Under the linear realizability assumption, we specifically assume $y_{t,a}$ as the following.

$$y_{t,a} = \beta_0' x_{t,a} + e_{t,a} \tag{3.1}$$

where $e_{t,a} \in \mathbb{R}$ is a 1-subgaussian error term, s.t. $\mathbb{E}[e_{t,a}] = 0$, and $\beta_0 \in \mathbb{R}^d$ is an unknown true coefficient vector. We denote the cumulative distribution of $e_{t,a}$ as $\mathcal{P}_e$. Unlike traditional bandit algorithms, we have a twofold goal in this problem: (i) to estimate the main coefficient of interest β_0 (ii) to achieve the largest reward by selecting the optimal arm.

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

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

the endogeneity problem occurs. The ordinary least square (OLS) estimator for the coefficient of interest β_0 is inconsistent in endogenous settings. Econometrics literature uses Instrumental Variable (IV) method to address the endogeneity problem. A variable $z_{t,a}$ is a valid instrumental variable if it satisfies the following two conditions (M Wooldridge, 2014): (i) it is uncorrelated with the error term $e_{t,a}$, i.e. $\mathbb{E}[z_{t,a}e_{t,a}] = 0$, (ii) it is correlated with the endogenous covariates $x_{t,a}$, and it has no direct effect on the reward $y_{t,a}$. Consider a valid instrumental variable $z_{t,a} \in \mathbb{R}^k$, with $\|z_{t,a}\|_2 \leq L_z$ where L_z is a constant upper bound for $\|z_{t,a}\|_2$. We assume $\mathbb{E}[z_{t,a}x'_{t,a}] \in \mathbb{R}^{k\times d}$ has full column rank d.[1] and the minimum eigenvalue of $\mathbb{E}[z_{t,a}z'_{t,a}]$ is positive. Now, consider the projection of $x_{t,a}$ on $z_{t,a}$:

$$x_{t,a} = \Gamma'_0 z_{t,a} + u_{t,a} \tag{3.2}$$

where $\Gamma_0 \in \mathbb{R}^{k\times d}$ is an unknown true coefficient vector, each element in $u_{t,a} \in \mathbb{R}^d$ is a 1-subgaussian error term with mean zero. Note that Γ_0 is a linear projection matrix and the independency condition $\mathbb{E}[z_{t,a}u'_{t,a}] = 0$ is satisfied by construction and Γ_0 does not capture the true causal relationship between $x_{t,a}$ and $z_{t,a}$. By plugging the equation of $x_{t,a}$ (Equation (3.2)) into Equation (3.1), we have

$$y_{t,a} = (\Gamma_0\beta_0)'z_{t,a} + v_{t,a} \tag{3.3}$$

where $v_{t,a} = \beta'_0 u_{t,a} + e_{t,a}$. We can prove that each element of $v_{t,a}$ is $(\|\beta_0\|_1 + 1)$-subgaussian (we provide the proof in Appendix A.0.7). We denote $\Gamma_0\beta_0$ as ϑ_0, and $(\|\beta_0\|_1 + 1)$ as σ_v, for the simplicity of analysis.

It is known that Two-Stage Least Squares (2SLS) estimator is consistent in offline settings. In this essay, we adopt the standard two-stage least square procedure to the online setting and demonstrate how one can use instrumental variables to address the issue of endogeneity in linear contextual bandits. For readers who are not familiar with Two-Stage Least Squares estimator, we provide a brief review in the following section.

---

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

We aim to design an online decision-making algorithm that learns the coefficient of main interest β_0 so that we can maximize the total expected reward after pulling arms under the endogeneity problem. We define the total expected regret of an algorithm after T rounds as

$$R_T = \sum_{t=1}^{T} \mathbb{E}[y_{t,a_t^*} - y_{t,a_t}] = \sum_{t=1}^{T} \mathbb{E}[\langle \Gamma_0' z_{t,a_t^*}, \beta_0 \rangle - \langle \Gamma_0' z_{t,a_t}, \beta_0 \rangle]$$

where $a_t^* = \arg\max_a \langle \Gamma_0' z_{t,a}, \beta_0 \rangle$ is the best arm at round t according to the true coefficient vectors β_0 and Γ_0, and a_t is the arm selected by the algorithm at round t.

3.1.2 Preliminaries and Assumptions

Two-Stage Least Squares estimator is an instrumental variable method commonly used in offline settings to correct estimation bias in the endogeneity problem. Consider vector $Y \in \mathbb{R}^T$ with $y_{t,a}$ as its elements, matrix $X \in \mathbb{R}^{T \times d}$ with $x_{t,a}'$ as its rows, matrix $Z \in \mathbb{R}^{T \times k}$ with $z_{t,a}'$ as its rows, where $t \in \{1, \cdots, T\}$. We briefly illustrate Two-Stage Least Squares estimation procedure in offline settings as the following: (i) First, regress the set of feature vectors X on the set of instrumental variable vectors Z using OLS method to obtain a set of estimated sample features $\hat{X}$; (ii) Then regress Y on the estimated feature vectors $\hat{X}$ to obtain the estimator $\hat{\beta}_{\hat{X},Y}$ using OLS method again.

Definition 3.1.1 (Two-Stage Least Squares Estimator)**.**

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

To facilitate further analysis on the Two-Stage Least Squares estimator, we also define related OLS estimators as the following, where $\hat{\vartheta}_{Z,Y} = \hat{\Gamma}_{Z,X}\hat{\beta}_{\hat{X},Y}$ (to see this equation, refer to Appendix A.0.7 for the proof).

Definition 3.1.2 (First Stage OLS Estimator)**.**

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

Definition 3.1.3 (OLS Estimator Based on Instrumental Variable)**.**

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

We relist key assumptions in this essay as the following,

Assumption 3.1.1. $\|z_{t,a}\|_2 \leq L_z$, $\|x_{t,a}\|_2 \leq L_x$, $\|y_{t,a}\|_2 \leq L_y$ for all t and a.

Assumption 3.1.2. The minimum eigenvalue of $\mathbb{E}[z_{t,a}z'_{t,a}]$ is larger than a positive constant λ.

Assumption 3.1.1 ensures that the $\|\cdot\|_2$ of instrumental variables, feature vectors, and the reward are bounded by positive constants. Notice that the $\|\cdot\|_\infty$ of a vector is smaller than the $\|\cdot\|_2$ of the vector, which implies that the $\|\cdot\|_\infty$ of the instrumental variables, features, and the reward are also upper bounded by these constants, L_z, L_x, and L_y respectively. Assumption 3.1.2 guarantees that, with high probability, the sample second moment $\sum_{s=1}^{t} z_{s,a}z'_{s,a}$ is non-singular so that the OLS estimators $\hat{\Gamma}_{Z,X}$ and $\hat{\vartheta}_{Z,Y}$ exist. We need these assumptions to bound the total regret and the inference bias.

3.2 BanditIV Algorithm

In this section, we propose the ϵ-*BanditIV* Algorithm (Algorithm 1) by incorporating instrumental variables in online settings. The algorithm is based on existing linear bandit algorithms, especially the OFUL algorithm proposed by Abbasi-Yadkori et al. (2011).

The ϵ-*BanditIV* algorithm takes as input initial regularization parameters $\gamma_z, \gamma_x > 0$, confidence set parameters $\{G_t\}_{t=1}^{\infty}$, $\{B_t\}_{t=1}^{\infty}$, as well as a sequence of non-increasing exploration parameters $\{\epsilon_t\}_{t=1}^{\infty}$. We will specify the confidence set parameters in Section 3.2.1.

Estimation: The algorithm maintains four matrices U_t, V_t, W_t, Q_t to calculate estimated parameters in each round as described in the pseudocode of Algorithm 1. If the current time t is after

the first round, the algorithm utilizes past-choice-related instrumental variables Z_{t-1} and past-choice-related feature variables X_{t-1} to estimate Γ_0 through ridge OLS and obtain an estimated X_{t-1}, which is $\hat{X}_{t-1}$. Then, based on $\hat{X}_{t-1}$ and past observations of rewards Y_{t-1}, the algorithm estimates β_0 by ridge OLS again. We denote the estimator for β_0 when $\gamma_z = \gamma_x = 0$, in the ϵ-*BanditIV* algorithm as the ϵ-*BanditIV estimator*.

Confidence Sets: Following the *Optimism in the Face of Uncertainty principle* (OFU), we need to maintain confidence sets for all unknown parameters in the model. As described in Section 3.1, we have two unknown parameters, Γ_0 and β_0. Thus, we construct two confidence sets $C_{1,s}$, $C_{2,s}$, $s \in \{1, \cdots, T\}$ for the first and the second stage estimation respectively in each round. The idea for the confidence sets is to make the estimation optimistic with the conditions that "with high probability" the true coefficients are in the confidence sets, and we can calculate the confidence sets from the past-choice-related feature variables X_{t-1}, related instrumental variables Z_{t-1}, and rewards Y_{t-1}.

Execution: In each round, we conduct a stochastic decision with probability ϵ_t for choosing a random arm and probability $1 - \epsilon_t$ for choosing the estimated best arm. The estimated best arm generated by the algorithm is an arm a_t which is related to an instrumental variable z_t that can maximize the estimated reward jointly with a pair of optimistic estimates of two-stage coefficients in the confidence sets. To maximize the estimated reward, the algorithm chooses a pair of optimistic estimates $(\tilde{\Gamma}_t, \tilde{\beta}_t) = \arg\max_{\Gamma \in C_{1,t} \beta \in C_{2,t}} (\max_{a \in \mathcal{A}_t} \langle \Gamma_t' z_a, \beta \rangle)$ and then chooses an arm a_t such that $a_t = \arg\max_{a \in \mathcal{A}_t} \langle \tilde{\Gamma}_t' z_a, \tilde{\beta}_t \rangle$, where we denote the arm set at time t as $\mathcal{A}_t$. Equivalently, the estimated best arm a_t, is stated in Equation (3.4). After an arm is chosen, we observe the reward y_{t,a_t} as stated in Equation (3.1).

$$a_t = \arg \max_{a \in \mathcal{A}_t} \max_{\Gamma \in C_{1,t}} \max_{\beta \in C_{2,t}} \langle \Gamma' z_a, \beta \rangle \tag{3.4}$$

Note that when $\epsilon_t = 0$ for $t \in \{1, 2 \cdots, \infty\}$, the algorithm chooses the estimated best arm in each round. To simplify the analysis, we denote the algorithm when $\epsilon_t = 0$ for $t \in \{1, 2 \cdots, \infty\}$ as *BanditIV*.

Algorithm 1 ϵ-BanditIV

Input: $\gamma_z, \gamma_x, \{G_t\}_{t=1}^{\infty}$,$\{B_t\}_{t=1}^{\infty}$, $\{\epsilon_t\}_{t=1}^{\infty}$
Set $U_0 = \gamma_z I \in \mathbb{R}^{k\times k}, V_0 = 0 \in \mathbb{R}^{k\times d}, W_0 = \gamma_x I \in \mathbb{R}^{d\times d}, Q_0 = 0 \in \mathbb{R}^{d\times 1}$
Set $C_{1,s} = \{\Gamma : \|\Gamma^{(i)} - \hat{\Gamma}_s^{(i)}\|_{U_s} \leq G_s\}, C_{2,s} = \{\beta : \|\beta - \hat{\beta}_s\|_{W_s} \leq B_s\}$, $\forall s \in \{1, 2, \cdots T\}$, $\forall i \in \{1, 2, \cdots d\}$
Nature reveals $\mathcal{Z}_0$ and $\mathcal{X}_0$. We randomly choose and play $a_0 \in \mathcal{A}_0$ and then observe the reward y_{a_0}. We set $Z_0 = z_{a_0}$, $X_0 = x_{a_0}$ and $Y_0 = y_{a_0}$.
for $t := 1, 2, \cdots, T$, **do**
$U_t = U_{t-1} + z_{t-1}z'_{t-1}, V_t = V_{t-1} + z_{t-1}x'_{t-1}$
$\hat{\Gamma}_t = U_t^{-1}V_t, \hat{X}_{t-1} = Z_{t-1}\hat{\Gamma}_t$
$W_t = W_0 + \hat{X}'_{t-1}\hat{X}_{t-1}, Q_t = Q_0 + \hat{X}'_{t-1}Y_{t-1}$
$\hat{\beta}_t = W_t^{-1}Q_t$
Nature reveals $\mathcal{Z}_t$ and $\mathcal{X}_t$. With probability ϵ_t, we uniformly choose a random $a_t \in \mathcal{A}_t$. With probability $1 - \epsilon_t$, we choose $a_t = \arg\max_{a\in\mathcal{A}_t} \max_{\Gamma\in C_{1,t}} \max_{\beta\in C_{2,t}} \langle \Gamma' z_a, \beta\rangle$
We observe the reward y_{a_t} and update $Z_t = [Z'_{t-1} \quad z_{a_t}]', X_t = [X'_{t-1} \quad x_{a_t}]', Y_t = [Y'_{t-1} \quad y_{a_t}]'$.
end for

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

3.2.1 Regret Analysis

In this section, we give upper bounds on the regret of the *ϵ-BanditIV* algorithm as well as the *BanditIV* algorithm. The proofs can be found in Appendix A.0.6. We first show an $\tilde{\mathcal{O}}(k\sqrt{T})$ upper bound for the total expected regret with parameters of confidence set by Theorem 3.2.1 and Corollary 3.2.1, where $\tilde{\mathcal{O}}(\cdot)$ omits constants and the polylogarithmic factors in k and T. Then, by Lemma 3.2.1 and Lemma 3.2.2, we present that the true coefficients are in the confidence sets with high probability, and we also give accurate definitions for confidence set parameters.

Theorem 3.2.1. *The expected cumulative regret of the ϵ-BanditIV at time T, with probability at*

least $1-\delta$*, is upper-bounded by*

$$R_T \leq 3B_T\sqrt{2Td\log(\frac{T+d\gamma_x}{d\gamma_x})} + \tilde{G}_T\sqrt{2Tk\log(\frac{T+k\gamma_z}{k\gamma_z})} + \\ + 2B_T\frac{4\|\Gamma_0\|_2 L_z}{\sqrt{\lambda}(\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z}G_T)}(\sqrt{T} - \sqrt{T^\dagger - 1}) + 2L_y((1-\epsilon_1)T^\dagger + \epsilon_1 T)$$

where $B_T = \sqrt{\gamma_x}\|\beta_0\|_2 + \sqrt{2\log(\frac{6T}{\delta}) + d\log(\frac{2TkL_z^2}{d\gamma_x}\|\Gamma_0\|_F^2 + \frac{4kT}{\gamma_x}\log(\frac{6Td}{\delta})\log(\frac{T+k\gamma_z}{k\gamma_z}) + 1)}$,
$\tilde{G}_T = \sqrt{\gamma_z}\|\Gamma_0\beta_0\|_2 + \|\beta_0\|_1\sqrt{2\log(\frac{3T}{\delta}) + 2\log(\|\beta_0\|_1) + k\log(\frac{TL_z^2}{\gamma_z} + 1)}$,
$G_T = \sqrt{\gamma_z}\max_{i\in\{1,\cdots,d\}}\|\Gamma_0^{(i)}\|_2 + \sqrt{2\log(\frac{6Td}{\delta}) + k\log\left(\frac{TL_z^2}{\gamma_z} + 1\right)}$,
and $T^\dagger = \frac{6L_z^2}{\lambda}\log(\frac{6k}{\delta})$.

From Theorem 3.2.1, we can see that the upper bound is $\tilde{\mathcal{O}}(k\sqrt{T})$, which means that when the number of rounds T increases, the cumulative expected regret increases sublinearly, i.e., the difference between the true optimal reward and the reward generated from our algorithm is not increasing proportionally with time. Sublinear regret is a desirable characteristic for bandit algorithms. Having sublinear regret implies that the algorithm is learning from its exploration and improving its decision-making capabilities over time.

The classical OFUL algorithm proposed in Abbasi-Yadkori et al. (2011) achieves $\tilde{\mathcal{O}}(d\sqrt{T})$ without assuming endogenous covariates, where d is the dimension of the covariates. This result is at the state of the art. By Theorem 3.2.1, the regret of our algorithm matches the regret of OFUL in respect of T implying that our algorithm obtains a regret at the state of the art in stochastic contextual linear bandits while accommodating more general settings, especially endogeneity cases.

Remark 3.2.1. The upper bound for the expected regret increases with the dimensions of the feature vector and the instrumental variable. Thus, if we have feature vectors of a large dimension, we may need to utilize feature selection techniques to reduce the dimension and hence improve the regret bound in practice.

Remark 3.2.2. Notice that when ϵ_t increases, the regret bound also increases, which implies that the regret bound of an ϵ-*BanditIV* algorithm with a positive ϵ_t will always be larger than the regret

bound of the *BanditIV* algorithm.

Remark 3.2.3. In order to guarantee an $\tilde{\mathcal{O}}(k\sqrt{T})$ upper bound for the total expected regret of the ϵ-BanditIV algorithm, we need a small enough ϵ_t, such as $\epsilon_t \leq \frac{\sqrt{\log(t)}}{\sqrt{t}}$ or $\epsilon_t \leq \frac{\sqrt{\log\log(t)}}{\sqrt{t}}$. To see the validity of these examples, note that the total expected regret can be written as upper bounded by $\tilde{\mathcal{O}}(k\sqrt{T}) + 2L_y \sum_{t=1}^{T} \epsilon_t$, where $2L_y \sum_{t=1}^{T} \epsilon_t$ can be upper bounded by $\tilde{\mathcal{O}}(\sqrt{T})$ if we have $\epsilon_t \leq \tilde{\mathcal{O}}(\frac{1}{\sqrt{t}})$. The *BanditIV* algorithm which is a special case with $\epsilon_t = 0$ in ϵ-BanditIV algorithm, naturally satisfies this condition.

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

Corollary 3.2.1. The expected cumulative regret of the BanditIV at time T, with probability at least $1 - \delta$, is upper-bounded by

$$R_T \leq 3B_T\sqrt{2Td\log(\frac{T+d\gamma_x}{d\gamma_x})} + \tilde{G}_T\sqrt{2Tk\log(\frac{T+k\gamma_z}{k\gamma_z})} + \\ + 2B_T\frac{4\|\Gamma_0\|_2 L_z}{\sqrt{\lambda}(\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z}G_T)}(\sqrt{T} - \sqrt{T^\dagger - 1}) + 2L_y T^\dagger$$

Lemma 3.2.1 (Optimistic estimation of β_0)**.** With high prob $1 - \frac{\delta}{3}$, for all $t \in \{1, 2, \cdots, T\}$,

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

where $B_t = \sqrt{\gamma_x}\|\beta_0\|_2 + \sqrt{2\log(\frac{6t}{\delta}) + d\log(\frac{2tkL_z^2}{d\gamma_x}\|\Gamma_0\|_F^2 + \frac{4kt}{\gamma_x}\log(\frac{6td}{\delta})\log(\frac{t+k\gamma_z}{k\gamma_z}) + 1)}$.

Lemma 3.2.2 (Optimistic estimation of Γ_0)**.** With high prob $1 - \frac{\delta}{6}$, for all $t \in \{1, 2, \cdots, T\}$ and all $i \in \{1, 2, \cdots, d\}$,

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

where $G_t = \sqrt{\gamma_z}\max_{i\in\{1,\cdots,d\}}\|\Gamma_0^{(i)}\|_2 + \sqrt{2\log(\frac{6td}{\delta}) + k\log\left(\frac{tL_z^2}{\gamma_z} + 1\right)}$

We construct double confidence intervals for the two-stage estimators to guarantee the regret bound. In every round, the algorithm chooses estimates for Γ_0 and β_0 from the confidence sets as we describe in Section 3.2. Lemmas 3.2.1 and 3.2.2 show that β_0 and Γ_0 with high probability fall in ellipsoids with center at $\hat{\beta}_t$ and $\hat{\Gamma}_t$ respectively, i.e., in the double confidence sets.

3.3 Inference and Confidence Intervals

Due to the adaptive nature of multi-armed bandits, calculating confidence intervals is not straightforward as the collected data is no longer i.i.d. anymore. In this section, we show asymptotic properties of the ϵ-*BanditIV* estimator following Chen et al. (2021) and Bastani and Bayati (2020). We first present the consistency of the online OLS estimator and the ϵ-*BanditIV* estimator, then we demonstrate the normality of the ϵ-*BanditIV* estimator. The proofs are presented in Appendix A.0.6.

Proposition 3.3.1 (Tail bounds for the online OLS estimators)**.** In the online decision-making model with the ϵ-greedy policy, if Assumptions 3.1.1 and 3.1.2 are satisfied, and ϵ_t is non-increasing, then for any $\eta_1, \eta_2 > 0$, any $i \in \{1, \dots, d\}$,

$$P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta_1) \geq 1 - \exp(-\frac{t\epsilon_t}{8}) - k\exp(-\frac{t\epsilon_t\lambda}{12L_z^2}) - 2k\exp(-\frac{t\epsilon_t^2\lambda^2\eta_1^2}{128k^2\sigma_v^2L_z^2})$$
$$P(\|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_1 \leq \eta_2) \geq 1 - \exp(-\frac{t\epsilon_t}{8}) - k\exp(-\frac{t\epsilon_t\lambda}{12L_z^2}) - 2k\exp(-\frac{t\epsilon_t^2\lambda^2\eta_2^2}{128k^2\sigma_u^2L_z^2})$$

where $\sigma_u = 1$ and $\sigma_v = \|\beta_0\|_1 + 1$.

To simplify the notations, we use $\hat{\vartheta}_t$ and $\hat{\Gamma}_t$ in this section to denote the OLS estimators $\hat{\vartheta}_{Z_t,Y_t}$ and $\hat{\Gamma}_{Z_t,X_t}$ defined in Section 3.1 respectively. Proposition 3.3.1 states that both the OLS estimator based on the instrumental variable and the first stage OLS estimator are consistent if $t\epsilon_t^2 \to \infty$ as $t \to \infty$. Based on this result, we derive the tail bound for the ϵ-*BanditIV* estimator as the following.

Proposition 3.3.2 (Tail bound for the ϵ-BanditIV estimator)**.** In the online decision-making model with ϵ-greedy policy, if the Assumptions 3.1.1 and 3.1.2 are satisfied, and ϵ_t is non-increasing, then

for any $\eta_1, \eta_2 > 0$, $\eta_2 \neq \frac{\lambda_{min}(\Gamma_0)}{\sqrt{d}}$,

$$P(\|\hat{\beta}_t - \beta_0\|_1 \leq C_\beta) \geq 1 - (P_1 + dP_2)$$

where $C_\beta = \dfrac{\sqrt{2}(\eta_1 + \eta_2\|\beta_0\|_1)}{\lambda_{min}(\Gamma_0) - \eta_2\sqrt{d}}$, $\lambda_{min}(\Gamma_0)$ denotes the smallest nonzero singular value of the matrix Γ_0, and

$$P_1 = \exp(-\frac{t\epsilon_t}{8}) + k\exp(-\frac{t\epsilon_t\lambda}{12L_z^2}) + 2k\exp(-\frac{t\epsilon_t^2\lambda^2\eta_1^2}{128k^2\sigma_v^2L_z^2}),$$
$$P_2 = \exp(-\frac{t\epsilon_t}{8}) + k\exp(-\frac{t\epsilon_t\lambda}{12L_z^2}) + 2k\exp(-\frac{t\epsilon_t^2\lambda^2\eta_2^2}{128k^2\sigma_u^2L_z^2})$$

Remark 3.3.1. Following the Proposition 3.3.1, if $t\epsilon_t^2 \to \infty$ as $t \to \infty$, then the probability of $\|\hat{\beta}_t - \beta_0\| \leq C_\beta$ goes to 1 for any $\eta_1, \eta_2 > 0$, $\eta_2 \neq \frac{\lambda_{min}(\Gamma_0)}{\sqrt{d}}$. Combining the Theorem 3.2.1 and Proposition 3.3.2, in order to guarantee both the regret bound and the tail bound, we need to carefully choose the value for ϵ_t. Some reasonable examples can be $\epsilon_t = \frac{\sqrt{\log(t)}}{\sqrt{t}}$ or $\epsilon_t = \frac{\sqrt{\log\log(t)}}{\sqrt{t}}$.

It's important to note that although BanditIV makes use of instrumental variables, the exploitative nature of bandit algorithms produces biased estimates. The ϵ-BanditIV overcomes this limitation. One might prefer using BanditIV when the primary concern is maximizing cumulative reward, whereas ϵ-BanditIV is more suitable when interpretability or establishing causal relationships also matters to the decision-maker. A situation where one could prefer ϵ-BanditIV is in clinical trials of a new procedure, where maximizing the cumulative reward (e.g., mortality outcomes) is important, but understanding how patient features affect those outcomes is also crucial.

Remark 3.3.2. The probability of $\|\hat{\beta}_t - \beta_0\| \leq C_\beta$ decreases with the dimension of instrumental variables, the variance of error terms in both stages of the regression model, and L_z. Moreover, the probability of $\|\hat{\beta}_t - \beta_0\| \leq C_\beta$ increases with the lower bound of the minimum eigenvalue of $\mathbb{E}[z_t z_t']$.

Following from the Propositions 3.3.1 and 3.3.2, we obtain the consistency of the OLS estimators $\hat{\vartheta}_t$ and $\hat{\Gamma}_t$, and the *ϵ-BanditIV* estimator $\hat{\beta}_t$ easily.

Corollary 3.3.1 (Consistency of the online OLS estimators)**.** If Assumptions 3.1.1 and 3.1.2 are satisfied, ϵ_t is non-increasing and $t\epsilon_t^2 \to \infty$ as $t \to \infty$, then the online OLS estimators $\hat{\Gamma}_t$ is a consistent estimator for Γ_0 and $\hat{\vartheta}_t$ is a consistent estimator for ϑ_0.

Corollary 3.3.2 (Consistency of the ϵ-BanditIV estimator)**.** If Assumptions 3.1.1 and 3.1.2 are satisfied, ϵ_t are non-increasing and $t\epsilon_t^2 \to \infty$ as $t \to \infty$, then the online ϵ-BanditIV estimator $\hat{\beta}_t$ is a consistent estimator for β_0.

Theorem 3.3.1 (Asymptotic normality of the online OLS estimator)**.** *If Assumptions 3.1.1 and 3.1.2 are satisfied, then*

$$\sqrt{t}(\hat{\vartheta}_t - \vartheta_0) \xrightarrow{d} \mathcal{N}_k(0, \mathbb{E}[v_t^2](\int zz' d\mathcal{P}_z)^{-1})$$

Based on the above results, we can derive the normality of the ϵ-*BanditIV* estimator by applying Slutsky Theorem. We also provide a consistent estimator for the variance of the normal distribution. The proof for the consistency of the variance estimator can be found in Appendix A.0.6.

Theorem 3.3.2 (Asymptotic normality of the ϵ-*BanditIV* estimator)**.** *If Assumptions 3.1.1 and 3.1.2 are satisfied, ϵ_t is non-increasing and $t\epsilon_t^2 \to \infty$ as $t \to \infty$. Then*

$$\sqrt{t}(\hat{\beta}_t - \beta_0) \xrightarrow{d} \mathcal{N}_d(0, S)$$

where $S = \mathbb{E}[v_t^2](\Gamma_0' \int zz' d\mathcal{P}_z \Gamma_0)^{-1}$.

A consistent estimator for S is given by

$$\sum_{s=1}^{t} \hat{v}_s^2 (\hat{\Gamma}_t' \sum_{s=1}^{t} z_s z_s' \hat{\Gamma}_t)^{-1}$$

where $\hat{v}_s = y_s - (\hat{\Gamma}_s \hat{\beta}_s)' z_s$.

Theorem 3.3.2 provides a theoretical guarantee that the ϵ-*BanditIV* estimator asymptotically follows a normal distribution with mean zero and variance S. We can see that this variance depends

on the expectation of the error term v_t, which includes the errors in both the first stage and the second stage. The variance also depends on the distribution of the instrumental variable, $\mathcal{P}_z$. Notice that, although we need the assumption about ϵ_t to guarantee the consistency of the estimator, the asymptotic variance of the ϵ-*BanditIV* estimator does not depend on ϵ_t.

To compute confidence intervals of our estimate of β_0 post running our bandits, we provide two ways as the following. As the first way, we use directly Theorem 3.3.2. The data set is constructed as illustrated in Section 3.4 with $\rho = 5$, $k = 1$,and $d_{exo} = 1$. We run 1,000 trials in total. In each trial, we take the final estimation $\hat{\beta}$ from a pre-run ϵ-BanditIV with 2,000 rounds as the mean of the confidence interval and calculate the estimated standard deviation as the theorem provides to construct one confidence interval. Among the 1,000 trials, we get a 95.5% coverage for the 95% confidence interval, which is close to the ideal.

As the second way, we use a re-randomization test similar to that of Bojinov et al. (2020) and Farias et al. (2022). For the estimate $\hat{\beta}$, we test the sharp null hypothesis that the $\beta_0 = \tau$ for all t. The sharp null hypothesis implies that the new outcome is $y_t + \tau(p_t^H - p_t)$ where p_t^H is a newly pulled arm in the post test and p_t is a pulled arm by the pre-run ϵ-*BanditIV*.

We conduct exact tests by using the known assignment mechanism to simulate new assignment paths. Algorithm 2 provides the details of how to implement it, where we stack $[x_t', p_t]', t \in \{1, ..., T\}$, together denoted as X. In particular, we propose τ by a downward search method based on the estimate from the bandit algorithm. Under the sharp null hypothesis of $\beta_0 = \tau$, a new arm assignment path leads to a sequence of observed outcomes $y_t + \tau(p_t^H - p_t)$ for $t \in \{1, \cdots, T\}$. To obtain a confidence interval, we invert a sequence of exact hypothesis tests to identify the region outside where the null hypothesis is violated at the prespecified significance level (Imbens and Rubin, 2015; Bojinov et al., 2020).

Consistent with the setting illustrated in the first way, all of these experiments are run on synthetic data constructed as in Section 3.4, and we consider the case $k = 1$, $d_{exo} = 1$ and $\rho = 5$. We run 50 trials and in each trial, we propose 20 hypotheses for β_0 based on $\hat{\beta}$ from a pre-run ϵ-*BanditIV* with 2,000 rounds and construct confidence intervals for the estimate of β_0. For each hypothesis of β_0, we run the sharp null hypothesis test where we set N_H as 200. We choose

Algorithm 2 Sharp Null Hypothesis Test

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

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

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

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

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

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

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

end for

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

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

the significance level to be $\alpha = 0.05$. Under ϵ-*BanditIV*, the coverage of the test is 96%. We can see that, given the significance level as 0.05, the coverage of the test is close to ideal. Thus, the re-randomization tests and corresponding confidence intervals reported here are adequate for inference of the main treatment effect.

3.4 Numerical Experiments

In this section, we will construct synthetic data to evaluate the empirical performance of our algorithms. We will focus on the problem of a platform operator who aims to recommend products to its customers in an online setting, with recommendations being made at each time period. To construct the data generation process, we extend the previous simulation setup of Bakhitov and Singh (2021), adapting it to the online setting.

Consider a platform operator that wants to recommend products to customers. For every period $t = 1, \cdots, T$, the platform operator has to choose among $J_t = 50$ products (or arms) where each product (or arm) is associated with features $x_{t,a} \in \mathbb{R}^{d_{exo} \times 1}$ and price $p_{t,a} \in \mathbb{R}$. For our simulation, we exogenously draw each element of $x_{t,a}$ from a normal distribution with mean 0 and variance 1.

We simulate the reward ($y_{t,a} \in \mathbb{R}$) the platform operator gets if it selects arm $a \in \mathcal{A}_t$ as follows –

$$y_{t,a} = \omega_0' x_{t,a} + \beta_0 p_{t,a} + \rho \xi_{1,t,a} + \xi_{2,t,a},$$

where $\xi_{1,t,a}, \xi_{2,t,a} \in \mathbb{R}$ are drawn independently from a normal distribution with mean 0 and standard deviation 0.1; $\rho \in \mathbb{R}$ is the degree of endogeneity; we set all elements in $\omega_0 = 1$ and set the parameter $\beta_0 = -1$ and assume it is the main parameter of interest. Now to model the endogeneity in the data-generating process, we simulate $p_{t,a}$ in the following way

$$p_{t,a} = \Gamma_0' z_{t,a} + \xi_{1,t,a} + \xi_{3,t,a} \tag{3.5}$$

where $\xi_{3,t,a} \in \mathbb{R}$ is drawn from a normal distribution with mean 0 and standard deviation 0.1; we exogenously draw each instrumental variable in $z_{t,a} \in \mathbb{R}^{k \times 1}$ from a normal distribution with mean 0 and variance 1, and we set all elements in $\Gamma_0 \in \mathbb{R}^{k \times 1}$ as 1.

In our simulations, we will vary the endogeneity degree ρ to see how it affects the performance of algorithms. We also vary the dimension of exogenous variable x_t and the dimension of instrumental variable z_t to see whether these dimensions affect the algorithm performance. We show the simulation results in Figures 3.1, 3.2, and 3.3.

We have two objectives in the simulation: (i) to achieve the maximum reward through selecting optimal arms, and (ii) to obtain an accurate estimation of the causal relation between the endogenous variable p_t and the reward y_t, which is -1 in this setting. We compare the performance of our algorithm on these two metrics with three other bandit algorithms – ϵ-Greedy, Linear TS, and OFUL. We adopt these algorithms as benchmarks because they have good theoretical and empirical performance, making them the most popular families of algorithms in both practical application and research. Many recent studies in bandit algorithms have also employed these benchmarks for comparison purposes (Dimakopoulou et al., 2019; Hao et al., 2020).

We run $T = 2,000$ time steps for each algorithm and collect the regret and estimation bias over time. We use the absolute difference between the estimated coefficient of p_t and the true

value of the coefficient, -1, to measure the estimation bias. Across all simulations in this section, the dimension of the endogenous variable is fixed at 1. In Figure 3.1, we set both the dimensions of the exogenous variable and the instrumental variable as 1. Then, in Figure 3.2, we change the dimension of the exogenous variable to 2, and in Figure 3.3, we set the dimension of the instrumental variable as 2. With the instrument dimension equal to or larger than the endogeneity dimension, the endogenous effect becomes just-identified or overidentified. The first, second, and third columns of each figure present results for $\rho = 0, 5, 10$ respectively. We use blue lines to represent *BanditIV* results, orange lines to represent ϵ-*BanditIV* results, green lines to represent Linear TS results, red lines to represent OFUL results, and purple lines to represent ϵ-Greedy results. Broadly, two persistent results emerge: first, when there is no endogeneity, i.e., $\rho = 0$, all algorithms perform similarly in respect of both regret and estimation bias (we do notice that *BanditIV* and ϵ-*BanditIV* can have a higher regret especially compared to Linear TS and OFUL when $\rho = 0$, and provide a discussion at the end of this section); but as the endogeneity issue becomes more severe, i.e., ρ increases, our proposed algorithms outperform more significantly than ϵ-Greedy, Linear TS and OFUL both on regret and inference. Second, *BanditIV* reaches lower actual regret than ϵ-*BanditIV*, but ϵ-*BanditIV* yields slightly less biased estimation. On average, ϵ-*BanditIV* is approximately 6.9% less biased than $BanditIV$ in our simulation. To further see how a noisier instrumental variable can affect the algorithm performances, we show simulation results with various values of the variance of the errors $\xi_{1,t}$ and $\xi_{3,t}$ in Appendix A.0.2.

Note that the estimation bias in β_0 could be either upward or downward, depending on the sign of ρ.[2] For instance, if ρ is positive, the bias will be upward ($\hat{\beta} \geq -1$), while if ρ is negative, the bias will be downward ($\hat{\beta} \leq -1$). Note that, in the specific case where $x_{t,a} = 0$ and if ρ is selected to be negative, then regardless of the degree of endogeneity (and thus the magnitude of bias), ϵ-Greedy, TS and OFUL will all choose the arm with the smallest value of $p_{t,a}$ with the highest probabilty, i.e., the arm that maximizes $\hat{y}_{t,a} = 0 + \hat{\beta} \cdot p_{t,a}$. This might result in the same arm getting selected across varying endogeneity levels, thus resulting in the same regret.

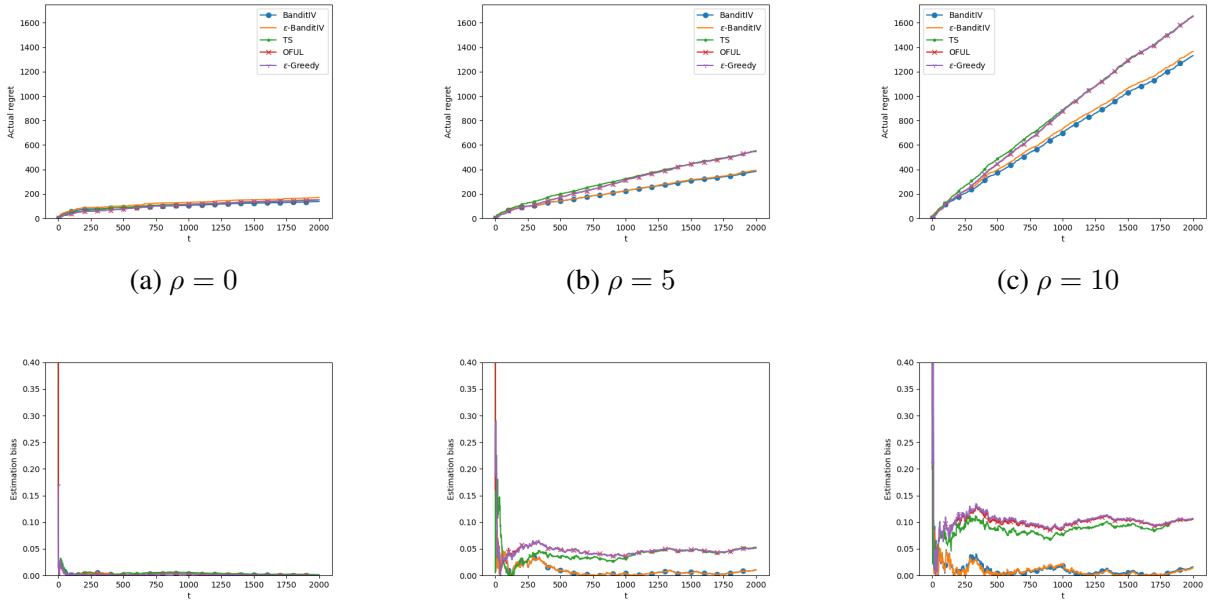
[2]Bandit algorithms like TS and OFUL implicitly estimate a ridge or ordinary least squares regression. Thus, one can use the directional bias in static settings as intuition to predict the direction of bias in online settings.

In addition, when there is no endogeneity ($\rho = 0$), the performance of *BanditIV* and ϵ-*BanditIV* can be slightly worse compared to state of the art traditional approaches like OFUL and Linear TS. This is because *BanditIV* and ϵ-*BanditIV* introduce an additional estimation step to account for potential endogeneity, which can introduce extra variance in the estimators when endogeneity is absent, particularly in smaller samples.

However, this tradeoff is a well-known phenomenon in instrumental variable (IV) methods more broadly: when endogeneity is present, IV-based approaches provide crucial bias correction, but when endogeneity is absent, the additional estimation step can slightly increase variance. This underscores an important practical consideration—if a practitioner is confident that no endogeneity exists in their application, state of the art traditional bandit methods such as OFUL and Linear TS remain well-suited. On the other hand, in settings where endogeneity is uncertain or likely, *BanditIV* and ϵ-*BanditIV* provide a robust alternative.

3.5 Real-world Applications

In this section, we will apply our proposed algorithms to real-world data to test their practical performance. It is important to note that evaluating the real-world applicability of bandit algorithms is challenging, as it requires deploying them in practice, which is often not feasible. As a result, most of the extant literature on online learning relies on pseudo-real-world demonstrations of the algorithms' abilities. Following the extant literature, we will conduct our evaluations in a two-step process. First, we will use real-world data to estimate the parameters of a model that captures the underlying contextual primitives. This model will serve as a representation of the real-world environment in which the algorithms will be tested. In the second step, the estimated model from the first step will be used to simulate the real world. The algorithms will then be run in this "simulated" real world to assess their performance and applicability.

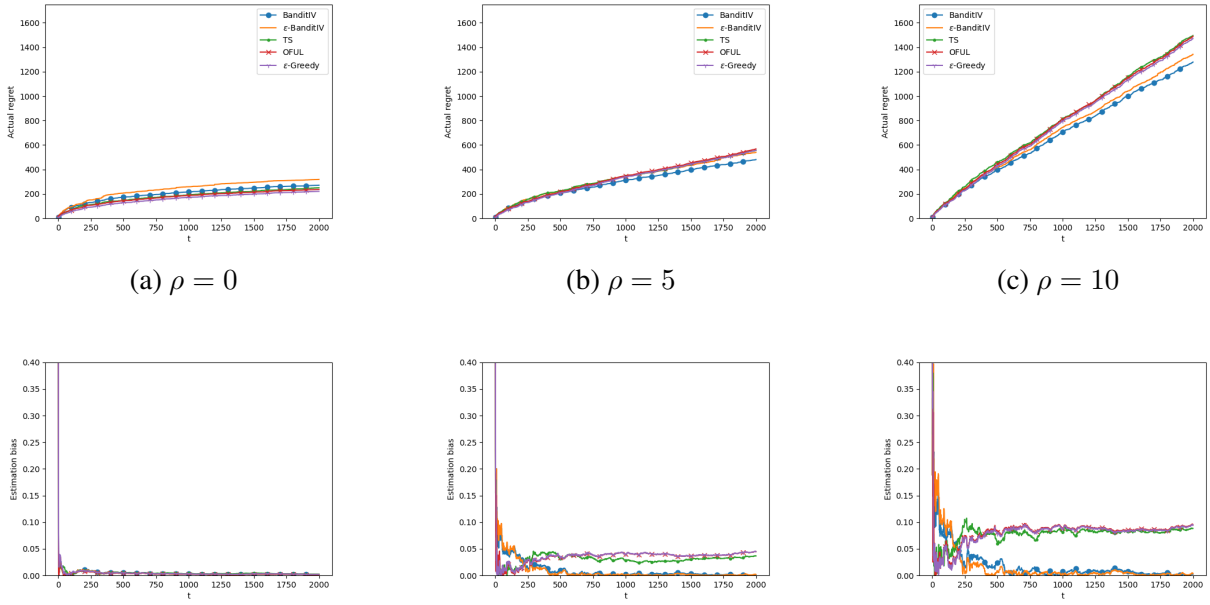


(a) $\rho = 0$ (b) $\rho = 5$ (c) $\rho = 10$

(d) $\rho = 0$ (e) $\rho = 5$ (f) $\rho = 10$

Figure 3.1: Simulation results on synthetic data with endogeneity when $k = 1, d_{exo} = 1$

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the actual regret. We use the absolute difference between the estimated coefficient of p_t and the true value of the coefficient -1 to measure the estimation bias and the actual cumulative regret to measure the regret. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm. Our findings indicate that the ϵ-$BanditIV$ algorithm is, on average, approximately 6.9% less biased than $BanditIV$.

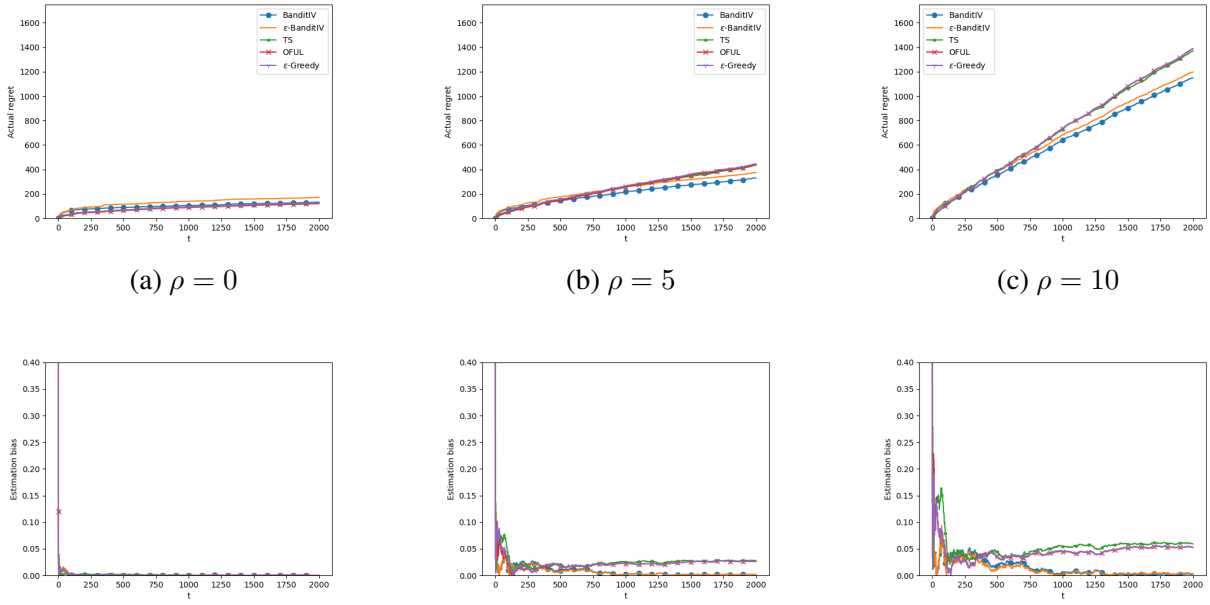


(a) $\rho = 0$ (b) $\rho = 5$ (c) $\rho = 10$

(d) $\rho = 0$ (e) $\rho = 5$ (f) $\rho = 10$

Figure 3.2: Simulation results on synthetic data with endogeneity when $k = 1, d_{exo} = 2$

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the actual regret. We use the absolute difference between the estimated coefficient of p_t and the true value of the coefficient -1 to measure the estimation bias and the actual cumulative regret to measure the regret. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm. Our findings indicate that the ϵ-$BanditIV$ algorithm is, on average, approximately 6.9% less biased than $BanditIV$.



(a) $\rho = 0$ (b) $\rho = 5$ (c) $\rho = 10$

(d) $\rho = 0$ (e) $\rho = 5$ (f) $\rho = 10$

Figure 3.3: Simulation results on synthetic data with endogeneity when $k = 2, d_{exo} = 1$

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the actual regret. We use the absolute difference between the estimated coefficient of p_t and the true value of the coefficient -1 to measure the estimation bias and the actual cumulative regret to measure the regret. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm. Our findings indicate that the ϵ-$BanditIV$ algorithm is, on average, approximately 6.9% less biased than $BanditIV$.

3.5.1 Mobile App Recommendation

We will first examine the problem of product recommendations, focusing on the challenge faced by Apple in recommending apps to its users. For this exercise, we will slightly simplify the problem faced by Apple. Consider Apple, which wants to recommend one app to its users every day. Apple wants to choose among a set of apps $\mathcal{A}_t$, where each app is associated with specific features $x_{t,a}$ and price $p_{t,a}$, and recommend them to a fixed userbase of M customers. Given the vast number of apps on the Apple app store, it is extremely difficult for customers to discover non-top apps. Next, we assume that if an app is not recommended to the user base under consideration, it remains undiscoverable and receives 0 downloads.

If Apple recommends app $a \in \mathcal{A}_t$ to the users, we model the downloads it receives as follows:

$$y_{t,a} = \omega_0' x_{t,a} + \beta_0 p_{t,a} + \rho \xi_{1,t,a} + \xi_{2,t,a},$$

where $\xi_{1,t,a}$ and $\xi_{2,t,a}$ capture demand shocks unobservable to Apple, and ρ captures the degree of correlation between price ($p_{t,a}$) and unobservable consumer demand shocks $\xi_{1,t,a}$. Note that generally, for growing platforms, customer engagement numbers are more important than immediate revenue. For instance, 100 customers downloading a free app might be more valuable to platform growth than one customer buying an expensive app.[3] In addition, in practical recommendation settings where contextual bandits are commonly used (Li et al. (2010); Nguyen et al. (2014)), the objective is to learn the relationship between contextual features (e.g., user and product attributes) and user engagement metrics (e.g., click-through rate, conversion rate, or ratings). Following this literature, we use demand (number of downloads) as a proxy for engagement and use the number of downloads as a measure of reward ($y_{t,a}$).

[3]For paid apps, Apple generates revenue through two sources: (i) upfront payments and (ii) in-app purchases. Some apps may charge a higher upfront price and lower in-app purchase prices, while others might do the opposite. To use revenues as a measure of reward, additional assumptions on handling in-app purchases would be needed. If one ignores the revenue through in-app purchases, the upfront revenue, i.e., $p*Downloads$, can also be used as the reward in our framework.

To model the correlational structure, we assume $p_{t,a}$ is set in the following way:

$$p_{t,a} = \Gamma_0' z_{t,a} + \xi_{1,t,a} + \xi_{3,t,a},$$

where $\xi_{3,t,a}$ captures exogenous shocks.

Estimates of data-generating process from real-world data We use a dataset from Apple's iOS app store on paid mobile apps. The dataset records daily downloads for 926 unique iOS paid apps during the period from Jan 2, 2013, to Sep 30, 2014. For each app, we observe the values of the feature vector (*Age*, *Version*, *VersionAge*, *Rating*, and *Price*[4]) for each day during the study period. The pools of potential apps to recommend vary over time due to the entries and exits of apps in the market. To calculate the values of parameters ω_0 and β_0, we estimate an instrumental variable regression with the number of downloads as the dependent variable. We also collect the name of the country where the headquarters of each app is located. For each country, we then collect the daily exchange rate (*ExchangeRate*) with USD as the base currency and use that as an instrument for app prices. Similar instruments have been used in prior literature, such as in Clements and Ohashi (2005). We report the results of the ordinary least squares (Model I) and two-stage least squares (Model II) using our instrumental variable in Table 3.1. We report the first stage regression result of Model II in Table 3.2. We find that the coefficient of price in Model I is positive, which indicates the presence of endogeneity. Our instrumental variable is able to correct for this in Model II. Therefore, we use values from Model II for parameters ω_0, β_0 and Γ_0 (see Equations (3.6) and (3.7)). Then we draw $\xi_{1,t}$ exogenously from a normal distribution with a mean of 0 and a standard deviation of 0.2. We draw $\xi_{2,t}$ and $\xi_{3,t}$ independently from a normal distribution with a mean of 0 and a standard deviation of 0.01.

[4] *Age* is the number of days an app has been available in the market; *Version* is a version index of an app and increments with each new release; *VersionAge* is the number of days a particular version of app has been available in the market; *Rating* is the average review rating an app has received up to the current date; *Price* is the amount users must pay to download the app.

Table 3.1: Regression Coefficients from Observational Data

	Dependent Variable: *Downloads*	
	Model Ⅰ	Model Ⅱ
Price	0.545***	-20.785***
	(2.708)	(-28.118)
Rating	5.061***	1.888***
	(12.175)	(6.099)
Age	-0.018***	-0.015***
	(-15.253)	(-20.886)
Version	0.397***	0.596***
	(6.433)	(14.111)
VersionAge	-0.022***	0.018***
	(-3.643)	(4.049)
Const.	25.062***	88.708***
	(12.692)	(31.312)
Observations	100,086	100,086
R^2	0.005	-0.107

Notes. t-statistics are in brackets. *p<0.1, **p<0.05, ***p<0.01

Table 3.2: First-Stage Regression Result

Variables	*Price*
ExchangeRate	0.012***
	(31.069)
Rating	-0.120***
	(-17.328)
Age	0.000***
	(8.388)
Version	0.017***
	(21.638)
VersionAge	0.002***
	(16.073)
Const.	2.654***
	(86.413)
Observations	100,086
R^2	0.048
F-statistic	648.0

Notes. t-statistics are in brackets. *p<0.1, **p<0.05, ***p<0.01

$$\begin{aligned} \textit{Downloads}_{t,a} =& 88.708 - 20.785\textit{Price}_{t,a} + 0.018\textit{VersionAge}_{t,a} + 0.596\textit{Version}_{t,a} - \\ & 0.015\textit{Age}_{t,a} + 1.888\textit{Rating}_{t,a} + \rho\xi_{1,t,a} + \xi_{2,t,a} \end{aligned} \tag{3.6}$$

$$\begin{aligned} \textit{Price}_{t,a} =& 2.654 + 0.012\textit{ExchangeRate}_{t,a} + 0.002\textit{VersionAge}_{t,a} + 0.017\textit{Version}_{t,a} - \\ & 0.000\textit{Age}_{t,a} - 0.120\textit{Rating}_{t,a} + \xi_{1,t,a} + \xi_{3,t,a} \end{aligned} \tag{3.7}$$

Additionally, to demonstrate substantial intertemporal variation in app prices within our synthetic dataset, for each app, we calculate the relative price range during our study period using the formula $\frac{Price_{i,max} - Price_{i,min}}{Price_{i,min}}$, where we use subscript i to denote an app. Figure 3.4 shows the distribution of this price variation across apps. We can see that apps exhibit considerable price variation across time.

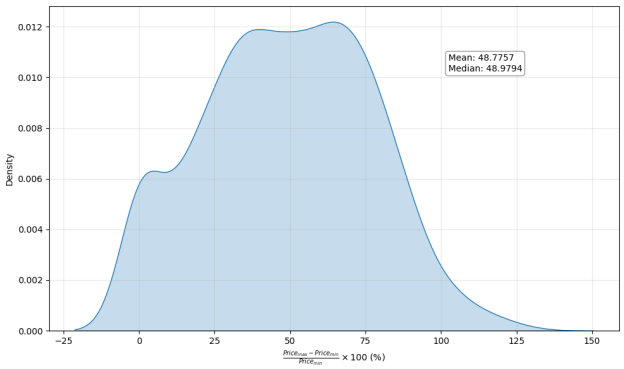


Figure 3.4: Price Variation of Mobile Apps

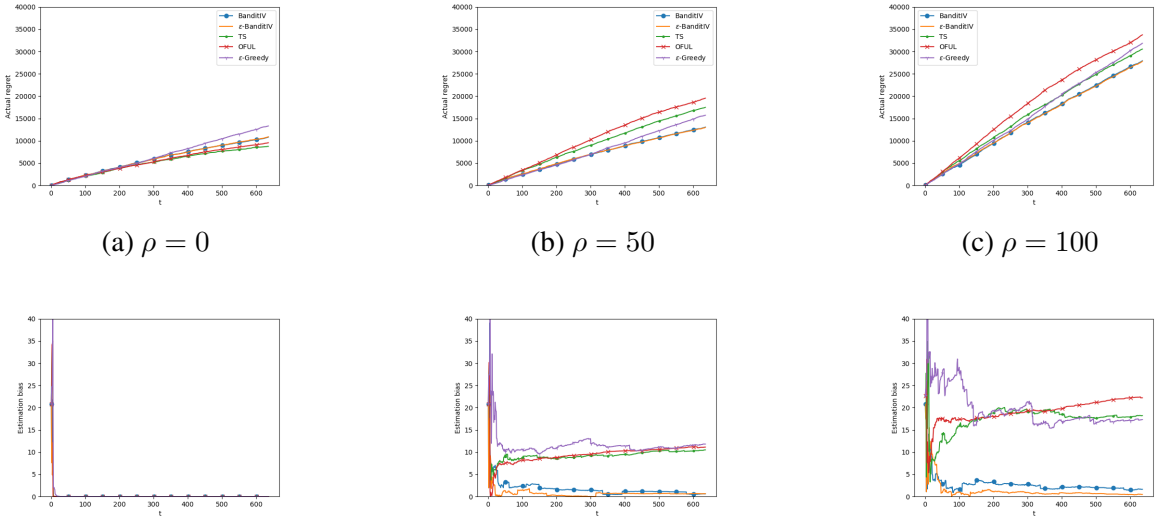
We run synthetic simulations for 637 days mimicking the time window the dataset covers. In Figure 3.5, we show the regret and estimation bias for four algorithms – *BanditIV*, ϵ-*BanditIV*, ϵ-Greedy, Linear TS and OFUL on the Apple iOS data. We find that *BanditIV* and ϵ-*BanditIV* significantly outperform the other three classic algorithms both on regret and estimation bias, especially under higher endogeneity cases. When there are no endogeneity issues, i.e., $\rho = 0$, all of the algorithms have close-to-zero estimation bias. Meanwhile, we notice that the Linear TS and OFUL algorithms have slightly lower regrets. As discussed in Section 3.4, the *BanditIV* approaches incorporate an additional estimation procedure to account for potential endogeneity. When endo-

geneity is absent, especially in smaller sample sizes, this extra step can introduce unnecessary variance into the estimators which leads to higher regret. When endogeneity increases, including the cases where $\rho = 50, 100$, the estimation bias from classic algorithms increases dramatically while ϵ-*BanditIV* algorithms still keep extremely low estimation bias and achieve lower regret. To conclude, our proposed algorithms consistently outperform state-of-art algorithms like ϵ-Greedy, Linear TS, and OFUL, in both aspects of regret and inference especially when endogeneity issues exist.

Note that that our proposed framework is not intended as a standalone replacement for existing recommender systems but rather as a component in modern recommendation pipelines. Large-scale recommender systems typically employ an ensemble of methods—including collaborative filtering, matrix factorization, and contextual bandits—to address challenges such as dynamic pricing, and new-user or new-item scenarios (Koren et al., 2009; Li et al., 2010). Companies such as Instacart (link), Netflix (link), and Expedia (link) integrate contextual bandits as part of their recommendation system pipelines.

Contextual bandits are specifically adopted in recommendation settings to learn new user preferences and new item (or product) feature effects in real time. This is useful in scenarios where limited interactions or newly introduced items require rapid adaptation (Li et al., 2010; Nguyen et al., 2014). Our method augments existing contextual bandit-based recommendation systems by explicitly addressing endogeneity within the bandit framework—a crucial factor in real-world settings where user engagement and item exposure can be influenced by confounded variables.

In endogenous settings, if traditional bandit algorithms are used without accounting for endogeneity, the learned policy may be systematically biased, leading to suboptimal recommendations. This, in turn, results in higher regret, as demonstrated in our experiments (see Figure 3.5). The consequence of higher regret is lower user engagement—undermining the fundamental goal of a recommendation system, which is to maximize user interaction with relevant products. Our method addresses this issue by incorporating endogeneity correction within the bandit framework, making it particularly valuable in real-world recommendation applications where various factors (such as price) can distort the learning process.



(a) $\rho = 0$ (b) $\rho = 50$ (c) $\rho = 100$

(d) $\rho = 0$ (e) $\rho = 50$ (f) $\rho = 100$

Figure 3.5: Simulation results on mobile app data with different endogeneity degrees

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the regret. We use actual cumulative regret to measure the difference between the numbers of downloads from the true optimal at each time and from the focal algorithm. We use the absolute difference between the true value of the price effect and the estimated price effect by the focal algorithm to measure the estimation bias. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm.

3.5.2 Real-Time Bidding Auctions

RTB is a dominant auction mechanism for selling ad exposures in display markets. A consumer entering a publisher's website generates an ad impression opportunity for advertisers. An auction takes place for every single impression opportunity. In RTB, an advertiser makes a buying decision in the auction immediately after the consumer arrives at the publisher's website. Buying decisions of impressions involve programmatic automation. The advertiser's ad is served to the consumer only if the advertiser wins the auction of the impression.

An advertiser makes buying decisions based on impression-specific information, like the width, height, and visibility of an ad slot, and the user's profile. Generally, the advertiser's objective is twofold: first, to measure the marginal value of the ad exposure; second, to achieve the largest profit through optimizing the bidding price. Notice that the advertiser has to bid in an auction to win the impression opportunity. Either overbidding or underbidding may bring significant costs to the advertiser. The advertiser can obtain a higher winning probability with a very high bid, but may not cover the paying price with the profit of ad exposure. On the contrary, the advertiser will be less likely to win with a low bid.

We will model the advertiser's ad bidding problem as a linear contextual bandit. Consider an advertiser that in every period t bids on customer impressions. Each impression is characterized by a set of features denoted by q_t. If the advertiser chooses to bid an amount $b_t \in \mathcal{A}_t$, where $\mathcal{A}_t$ represents the set of all potential bids the advertiser considers, it realizes some profit y_{t,b_t}. We model the profit (y_{t,b_t}) the advertiser gets if it wins the auction as the sum of a base revenue ($\beta^{(0)}$), the revenue generated from the ad exposure ($\beta^{(1)}$) subtracted by the paying price ($B_{CP,t}$), and a random error (e_{t,b_t}).

$$y_{t,b_t} = \beta^{(0)} + (\beta^{(1)} - B_{CP,t})\mathbb{I}\{B_{CP,t} < b_t\} + e_{t,b_t}, \tag{3.8}$$

We assume a second-price auction and use $B_{CP,t}$ to denote the highest bidding price among the advertiser's competitors at time t. The base revenue is a part of the revenue the advertiser will always receive, regardless of whether it wins the auction or not. We use $\beta^{(0)}$ to represent the

expected base revenue. The marginal revenue of the ad exposure, $\beta^{(1)}$, is our main coefficient of interest and is also referred to as the treatment effect, i.e., the effect of ad exposures on revenue. To determine whether the advertiser wins the auction or not, we use an indicator function, $\mathbb{I}\{B_{CP,t} < b_t\}$, where b_t is the advertiser's bidding price at time t. The unobserved demand shock at time t is denoted by the error term e_{t,b_t}.

Note that, even if the advertiser were to randomize the bids (b_t) it places, it still cannot randomize the ad exposure ($\mathbb{I}\{B_{CP,t} < b_t\}$). This is due to the fact that the ad exposure ($\mathbb{I}\{B_{CP,t} < b_t\}$) depends on the "best" competing bid and cannot be controlled by the focal advertiser. In the above data-generating process, the arm covariate $\mathbb{I}\{B_{CP,t} < b_t\}$ can be correlated with the error term e_{t,b_t}. Certain competing advertisers might be able to anticipate consumer demand shocks (captured by e_{t,b_t}) better than the focal advertiser and adjust their bids accordingly. This could create a potential correlation between $B_{CP,t}$ and e_{t,b_t}, and thus a correlation between $\mathbb{I}\{B_{CP,t} < b_t\}$ and e_{t,b_t}. Our assumption is that only the "best" competing bid needs to be correlated with demand shocks, which could occur even if one of the competitors has access to more market trends or customer information compared to the focal bidder. We do not assume that every bidder can observe the market demand shocks equally. Our model accounts for the possibility of information asymmetry among bidders, which is a realistic representation of real-world bidding scenarios. To model the competing bid ($B_{CP,t}$), we use the following data-generating process:

$$B_{CP,t} = \psi^{(0)} q_t + \psi^{(1)} \nu_t + \xi_{1,t} + \xi_{3,t}, \tag{3.9}$$

where q_t captures the characteristics of the impression (including ad slot width, height, visibility, and floor price), ν_t captures certain exogenous market shocks uncorrelated with consumer demand shocks (e_{t,b_t}), $\xi_{1,t}$ captures demand shocks unobservable to the focal advertiser which are correlated with consumer demand shocks (e_{t,b_t}), and $\xi_{3,t}$ captures exogenous shocks unobservable to the focal advertiser.

To model the correlational structure, we rewrite the error term (e_{t,b_t}) as the sum of two error terms: $\xi_{1,t}$ and ξ_{2,t,b_t}. We assume that $\mathbb{I}\{B_{CP,t} < b_t\}$ is correlated with the first error term, $\xi_{1,t}$,

while uncorrelated with the second error term, ξ_{2,t,b_t}. The parameter ρ indicates the degree of endogeneity, with larger $|\rho|$ implying higher endogeneity.

$$e_{t,b_t} = \rho\xi_{1,t} + \xi_{2,t,b_t} \tag{3.10}$$

Estimates of data-generating process from real-world data We now use real data from iPinYou (Liao et al., 2014) to calibrate the parameters of our data-generating process. The iPinYou RTB dataset is publicly available and includes data on ad bidding, impressions, clicks, and final conversions. For this study, we focus on the impression data for the *Telecom* product category. The auction mechanism used in this setting was a second-price auction. As a result, for every auction on an impression, we observe the highest competing bidding price ($B_{CP,t}$), which is also the bidder's paying price, in the dataset. To estimate the parameters of the data-generating process for simulating the best competing bid (Equation (3.9)), we first regress the observed values of the best competing bids on impression features, such as ad slot width, height, visibility, and floor price. The parameter results are reported in Table 3.3, and these values are used to construct $\psi^{(0)}$. Next, we synthetically simulate an instrumental variable, ν_t, by drawing samples from a uniform distribution within the range [6, 6.2] and assume the value of $\psi^{(1)}$ to be 1. We draw $\xi_{1,t}$ exogenously from a normal distribution with a mean of 0 and a variance of 1, and $\xi_{3,t}$ from a normal distribution with a mean of 0 and a standard deviation of 0.5. To calibrate the parameters for simulating the profit (y_{t,b_t}), we need to estimate the values of $\beta^{(0)}$ and $\beta^{(1)}$, where $\beta^{(1)}$ represents the effect of ad exposures on revenue. As the iPinYou dataset does not directly provide revenue data for advertisers, we use the average paying price for advertisers as an approximation, which reflects their valuation for the ad exposure, and set $\beta^{(1)} = 288$. For $\beta^{(0)}$, we consider multiple values: $\beta^{(0)} \in [100, 1000, 10000]$. We provide results with $\beta^{(0)} \in [1000, 10000]$ in Appendix A.0.4. Finally, the error term ξ_{2,t,b_t} is drawn from a normal distribution with a mean of 0 and a variance of 1.

The variable ν_t will serve as the instrumental variable to address the endogeneity bias. An instrumental variable must satisfy two requirements: (i) $\mathbb{E}[\nu_t e_t] = 0$, meaning that the instrument

Table 3.3: Regression Coefficients from Observational Data

Variable	$B_{CP,t}$
Slot Width	0.029***
	(43.638)
Slot Height	0.077***
	(65.136)
Slot Visibility	-8.305***
	(-165.663)
Floor Price	0.607***
	(274.459)
Const.	73.097***
	(138.719)
Observations	354,320
R^2	0.311

Notes. t-statistics are in brackets.
*p<0.1, **p<0.05, ***p<0.01

is uncorrelated with the error term, (ii) ν_t affects $\mathbb{I}\{B_{CP,t} < b_t\}$, indicating that the instrument has a direct effect on the endogenous variable, and ν_t does not affect y_{t,b_t} directly.

One potential example of exogenous data that can be used as an instrumental variable in RTB auctions is the intensity of political ads (Sinkinson and Starc (2019)) at a given time. As demonstrated by Sinkinson and Starc (2019), political ads have a displacement effect on commercial ads, yet they do not directly affect the revenue generated from commercial ads. Therefore, the variation in political ad intensity can represent exogenous variation in an advertiser's likelihood of winning an auction and serve as a potential instrument in the RTB problem.

It is important to note that the relationship between $\mathbb{I}\{B_{CP,t} < b_t\}$ and ν_t is not linear, while both our algorithms require a linear relationship between the endogenous variable and the instrumental variable. This discrepancy could lead to misspecification bias and adversely affect the performance of our algorithms. Despite this limitation, in our simulations, we will execute our algorithms while ignoring the non-linear relationship. Our simulations do reveal a small bias due

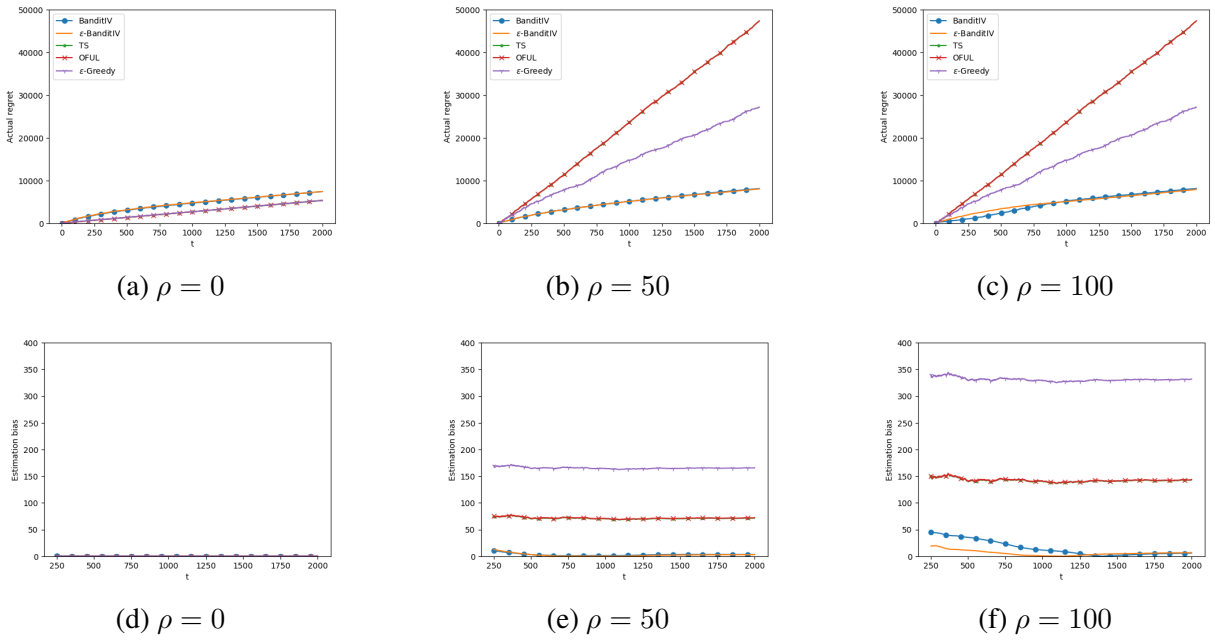
to this assumption.

We run our simulation for 2,000 time periods. In each time period, the advertiser is shown an impression opportunity. The impression characteristics (q_t) are randomly sampled from our dataset. At every time period t, the advertiser must choose a bid b_t from a set of available prices $\mathcal{A}_t$. Each price $b_{t,a} \in \mathcal{A}_t$ is regarded as an arm, and we define the minimum and maximum values in set $\mathcal{A}$ to correspond to the lowest and highest paying prices found in the dataset. Unlike traditional bandit settings, the arm features are not revealed at period t. In this case, the arm feature is the value $\mathbb{I}\{B_{CP,t} < b_t\}$. At time t, the advertiser does not know the value of $B_{CP,t}$. However, we assume that the advertiser can compute the value of $\mathbb{E}[B_{CP,t}]$ using the true values of $\psi^{(0)}$ and $\psi^{(1)}$. Thus, in this setup, regardless of the bandit algorithm employed, when deciding which arm (or bid) to choose, the advertiser makes that decision by computing profit using $\mathbb{E}[B_{CP,t}]$ instead of $B_{CP,t}$. The actual values of $B_{CP,t}$ are realized by the advertisers only after the auction has taken place. Once the advertiser bids with the chosen arm, b_t, they observe whether they have won the auction and the realized revenue.

In our simulation, to predict the value of $\mathbb{I}\{B_{CP,t} < b_t\}$, the classical algorithms including ϵ-Greedy, Linear TS and OFUL first conduct a linear projection of historical observations $\{\mathbb{I}\{B_{CP,s} < b_s\}\}_{s=0}^{t-1}$ on $\{\hat{B}_{CP,s}\}_{s=0}^{t-1}$ and $\{b_s\}_{s=0}^{t-1}$ and then predict $\mathbb{I}\{B_{CP,t} < b_t\}$ based on the projection, $\mathbb{E}[B_{CP,t}]$, and b_t. In contrast, the *BanditIV* and ϵ-*BanditIV* algorithms incorporate a projection step onto the instruments, that is, a linear projection of $\{\mathbb{I}\{B_{CP,s} < b_s\}\}_{s=0}^{t-1}$ on $\{\nu_s\}_{s=0}^{t-1}$, $\{q_s\}_{s=0}^{t-1}$ and $\{b_s\}_{s=0}^{t-1}$, and then predict $\mathbb{I}\{B_{CP,t} < b_t\}$ based on the projection, ν_t, q_t, and b_t.

In Figure 3.6, we show the regret and estimation bias for four algorithms – *BanditIV*, ϵ-*BanditIV*, ϵ-Greedy, Linear TS and OFUL on the RTB data. We compare algorithm performance under different endogeneity cases. We find that *BanditIV* and ϵ-*BanditIV* significantly outperform the other three classic algorithms both on regret and estimation bias, especially under higher endogeneity cases. When there are no endogeneity issues, i.e., $\rho = 0$, all of the algorithms have close-to-zero estimation bias while the *BanditIV* and the ϵ-*BanditIV* algorithms have slightly higher regrets. When endogeneity increases, including the cases where $\rho = 50, 100$, the estimation bias from classic algorithms increases dramatically while *BanditIV* algorithms keep extremely low es-

timation bias and regret.



(a) $\rho = 0$ (b) $\rho = 50$ (c) $\rho = 100$

(d) $\rho = 0$ (e) $\rho = 50$ (f) $\rho = 100$

Figure 3.6: Simulation results on RTB data with different endogeneity degrees and $\beta^{(0)} = 100$

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or regret. We use actual cumulative regret to measure the difference between the revenue from the true optimal at each time and from the focal algorithm. We use the absolute difference between the true value of the ad effect and the estimated ad effect by the focal algorithm to measure the estimation bias. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm.

3.6 Discussion

In this essay, we study the endogeneity problem in online decision-making settings where we formulate the decision-making process as a contextual linear bandit model. We discuss how many economic settings can be characterized by endogeneity and extant methods can lead to estimation bias and suboptimal decisions. To correct the bias and optimize online decisions, we propose the ϵ-*BanditIV* algorithm by utilizing both existing linear bandit algorithms and instrumental variables. We present the theoretical properties of our algorithm. We first show an upper bound for the

total expected regret, and then show the consistency, asymptotic normality of the estimator in the algorithm. On the applied side, we conduct extensive simulations on synthetic data. We find that the ϵ-*BanditIV* algorithm outperforms several benchmark linear bandit algorithms when the endogeneity problem occurs. Finally, we use datasets on both the mobile app recommendation and the RTB system to demonstrate how ϵ-*BanditIV* can be effective in estimating the causal impact of price and advertising respectively while maintaining close to oracle regret.

Chapter 4

DEBIASING ML- OR AI-GENERATED REGRESSORS IN PARTIALLY LINEAR MODELS

4.1 Problem Setting

In this section, we introduce the problem settings of ML/AI-generated variables in general models. Let Y_i denote an $\mathbb{R}$-valued data observation of the dependent variable, and $\{D_i^*, Z_i\}$ denote a set of data observations of independent variables. Assume that the variable Z is $\mathbb{R}^{d_Z}$-valued and we already have all data observation of this variable. Nevertheless, the variable D^* is $\mathbb{R}^{d_D}$-valued and we need to extract D_i^* from an unstructured dataset through methods like human annotation. We are interested in the causal effect of D^* on Y. Ideally, we need a dataset $\{Y_i, D_i^*, Z_i\}_{i=1}^n$, where n is the size of the whole dataset. Yet, due to reasons like high human annotation cost, we can only observe $\{Y_i, D_i^*, Z_i\}_{i=1}^{n_l}$ for a small subset of the data with a size n_l, where we use subscript l to represent "labeled". We use ι_i to indicate whether the data observation is in the labeled set (if in the labeled set, $\iota_i = 1$). To simplify our notation, we denote the subset of the labeled data as $\{Y_i, D_i^*, Z_i\}^{n_l}$. Note that all the data observations are i.i.d. drawn from unknown population distributions.

Consider that an ML/AI algorithm is used to predict the independent variable D_i^* for the whole dataset. We denote the predicted value for D_i^* as D_i, which is also $\mathbb{R}^{d_D}$-valued. As a result, we can observe the data $\{Y_i, D_i, Z_i\}_{i=1}^n$ through this ML/AI algorithm. Let $X_i \coloneqq (D_i', Z_i')'$ and $X_i^* \coloneqq ((D_i^*)', Z_i')'$. The true model is described in (4.1), where $\phi_o(\cdot)$ is a general function, and we will specify the function form later in Sections 4.2. The term ε_i is a random error independent of X_i^* and X_i. That is,

$$Y_i = \phi_o\left(X_i^*\right) + \varepsilon_i, \qquad E\left[\varepsilon \mid X^*, X\right] = 0. \tag{4.1}$$

The second part in Equation (4.1) is the identification assumption. Loosely speaking, it assumes

the gap between ML/AI-generated variable D and the true variable D^* will not affect the causal relationship between Y and X^*, provided one can observe the observations for X^*. This condition is relatively mild because the causal relationship in the subsequent econometric model is usually irrelevant to the training objective. It typically involves using the true label and input, such as unstructured data, to train the ML/AI algorithm and is not related to the estimation of the causal parameter.

Suppose we are interested in the causal effect of D^* on Y; the parameter of interest can be written as follows:

$$\theta_o = E\left[\frac{\partial Y}{\partial D^*}\right]. \tag{4.2}$$

The model specification and estimation strategy are detailed in Section 4.2. With a labeled dataset, one can estimate θ_o without bias. However, as we assume that $n_l \ll n$, this estimation strategy can cause a large variance due to the limited finite sample, and hence low statistical power. Combining with the unlabeled dataset, we could have a much larger dataset. But naively replacing D_i^* by D_i in estimating θ_o may introduce bias, thus not ideal for empirical researchers. Therefore, in the following section, we propose debiased estimation strategies for a general partially linear model specification, with the linear model as a special case.

4.2 Models and Estimators

4.2.1 Illustration Using Linear Model

First, we consider a linear model, i.e., $\phi_o(\cdot)$ is a linear function. The model can be written as:

$$Y_i = \beta_o' X_i^* + \varepsilon_i, \quad E\left[\varepsilon \mid X^*, X\right] = 0. \tag{4.3}$$

where $X_i^* := ((D_i^*)', Z_i')'$ is an $\mathbb{R}^{d_D+d_Z}$-valued random variable. It is known that running Ordinary Least Squares (OLS) of Y_i on X_i using the entire dataset directly can result in estimation bias. As illustrated in Greene (2003), Carroll et al. (1995), and Yang et al. (2018), if we consider classical measurement error, i.e., $X_i = X_i^* + \omega_i$, where ω_i is independent of ε_i and X_i^*, we can see an

attenuation bias as well as asymptotic inconsistency generated by the OLS estimation on the whole dataset. If we consider nonclassical measurement error, like $X_i = \varpi X_i^* + \omega_i$, where $\varpi < 1$ represents a multiplicative systematic error and ω_i is independent of ε_i and X_i^*, we can see an amplification bias as well as asymptotic inconsistency by the OLS estimation on the whole dataset (Yang et al., 2018). One way to obtain a potentially unbiased estimation of β is to run an OLS of Y_i on the true labels X_i^* using the labeled dataset only as in Equation (4.4).

Estimator 1 (OLS on the labeled dataset only)

$$\hat{\beta}_{ols} = \left(\sum_{i=1}^{n} \iota_i X_i^* X_i^{*\prime} \right)^{-1} \left(\sum_{i=1}^{n} \iota_i X_i^* Y_i \right). \tag{4.4}$$

We can easily obtain unbiasedness and asymptotic consistency for the estimation of β_o through Estimator 1. However, the drawback of Estimator 1 is that, in our setting, the size of the labeled dataset is small, which may still lead to unfavorable finite sample properties (finite-sample bias and unreliable standard errors). Another way to correct the estimation bias for β_o is to run a Two-Stage Least Square (TSLS) on the full dataset as described in Equation (4.6).

Estimator 2 (TSLS on the full dataset)

1. Estimate the projection coefficient matrix $\hat{\Gamma}$ by performing a linear projection of X_i^* on X_i using the labeled dataset $\{X_i^*, X_i\}^{n_l}$. Then, predict $\{\hat{X}_i^*\}_{i=1}^n$ using $\{X_i\}_{i=1}^n$ and $\hat{\Gamma}$. That is,

$$\hat{\Gamma} = (\sum_{i=1}^{n} \iota_i X_i X_i')^{-1} (\sum_{i=1}^{n} \iota_i X_i X^{*\prime}), \quad \hat{X}_i^* = \hat{\Gamma}' X_i \quad \text{for } i = 1, \ldots, n. \tag{4.5}$$

   Notice that $\hat{\Gamma}$ is an estimator for $\Gamma_o = E[XX']^{-1} E[XX^{*\prime}]$. Let $\xi_i = X_i^* - \Gamma_o' X_i$, we have $E[\xi X'] = 0$ is satisfied by the definition of Γ_o, provided that $E[XX']$ is positive definite.

2. Run an OLS of Y_i on $\hat{X}_i^*$ using $\left\{ Y_i, \hat{X}_i^* \right\}_{i=1}^n$. The TSLS estimator of β_o is given by

$$\hat{\beta}_{tsls} = \left(\sum_{i=1}^{n} \hat{X}_i^* \hat{X}_i^{*\prime} \right)^{-1} \left(\sum_{i=1}^{n} \hat{X}_i^* Y_i \right). \tag{4.6}$$

Estimator 2 in Equation (4.6) is equivalent to the estimator in the first example in Qiao and Huang (2021). We provide the proof for the equivalence in Appendix (B.6). The Estimator 2 is consistent but not the most efficient given the current moment condition in (4.3). Intuitively, TSLS estimator only uses the moment condition $E[\varepsilon X] = 0$, while does not jointly account for the variance for both moments: $E[\varepsilon X] = 0$ and $E[\xi X'] = 0$. Therefore, we construct a GMM estimator that uses both moment conditions to gain efficiency.

Let $g_1(b) := E[X^*(Y - (X^*)'b)]$, $g_2(b) := E[X(Y - (X^*)'b)]$, $\hat{g}_1(b) := \frac{1}{n_l}\sum_{i=1}^{n} \iota_i X_i^*(Y_i - (X_i^*)'b)$ and $\hat{g}_2(b) := \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \hat{X_i^*}'b)$. We have that $g_1(\beta_o) = g_2(\beta_o) = 0$ by (4.3). To simplify the analysis, we denote $g(b) := [(g_1(b))' \quad (g_2(b))']'$ and $\hat{g}(b) := [(\hat{g}_1(b))' \quad (\hat{g}_2(b))']'$. The GMM estimator is given as follows:

Estimator 3 (GMM on the full dataset)

Denote $\tilde{\beta}$ as a consistent estimator for β_o, e.g., $\hat{\beta}_{ols}$ as described in Estimator 1. We first follow (4.5) to get $\hat{X_i^*}$ for $i = 1, \ldots, n$. Let $\lambda_l := \lim_{n, n_l \to \infty} \frac{n_l}{n} \in (0, \infty)$. Then the GMM estimator can be written as

$$\hat{\beta}_{gmm} = \underset{b}{\operatorname{argmin}}\, \hat{g}(b)^T W \hat{g}(b), \tag{4.7}$$

where $\hat{g}(b) = \begin{pmatrix} \hat{g}_1(b) \\ \hat{g}_2(b) \end{pmatrix}$ and W is a pre-fixed positive definite weighting matrix, which is given by

$$W^{-1} = \frac{1}{n_l}\sum_{i=1}^{n} h_i h_i', \text{ where } h_i = \begin{pmatrix} \iota_i X_i^*(Y_i - (X_i^*)'\tilde{\beta}) \\ \lambda_l X_i(Y_i - (\hat{\Gamma}'X_i)'\tilde{\beta}) - \iota_i X_i(X_i^* - \hat{\Gamma}'X_i)'\tilde{\beta} \end{pmatrix}. \tag{4.8}$$

Let

$$B := \begin{pmatrix} \frac{1}{n_l}\sum_{i=1}^{n} \iota_i X_i^* Y_i \\ \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \end{pmatrix}, \quad A := \begin{pmatrix} \frac{1}{n_l}\sum_{i=1}^{n} \iota_i X_i^*(X_i^*)' \\ \frac{1}{n}\sum_{i=1}^{n} X_i \hat{X_i^*}' \end{pmatrix}. \tag{4.9}$$

We can also write $\hat{g}(b) = B - Ab$ and

$$\hat{\beta}_{gmm} = (A'WA)^{-1}A'WB. \tag{4.10}$$

Assumption 4.2.1 (Hansen (2022); Newey and McFadden (1994))**.** Recall that $\|\cdot\|$ denotes the Euclidean norm. Assume the followings hold[1]: (a) The full dataset $\{Y_i, X_i\}_{i=1}^n$ are i.i.d. samples drawn from $f_{Y,X}$ and the labeled dataset $\{X_i^*, X_i\}_{i=1}^{n_l}$ are i.i.d. samples drawn from $f_{X^*,X}$; (b) $\lambda_l = \lim_{n_l \to \infty} n_l/n \in [0, 1]$;

All the three linear estimators exhibit consistency and asymptotic normality, which are key properties for valid causal inference. See B.4 for the theoretical results.

Remark 4.2.1. Notice that here D_i can be continuous and is not limited to one dimension, which is different from the setting in Qiao and Huang (2021). Both TSLS estimator and GMM estimator can achieve lower variances as $\lambda_l \to 0$, which implies that, asymptotically adding extra unlabeled data can improve estimation efficiency for both TSLS estimator and GMM estimator.

Remark 4.2.2 (Efficiency Gain). Our proposed GMM estimator, being a weighted average of the OLS and TSLS estimators where the weight inversely correlates with estimator variance, ensures that it has the smallest variance given current model assumptions.

4.2.2 Partially Linear Model

To accommodate more flexible forms, especially high dimensional features, while maintaining the linear form for the main effect, we consider the following partially linear model in this section,

$$Y_i = \theta_o D_i^* + s_o(Z_i) + \varepsilon_i, \quad E(\varepsilon | D, D^*, Z) = 0, \tag{4.11}$$

$$D_i^* = r_o(Z_i) + u_i, \quad E(u|Z) = 0, \tag{4.12}$$

$$D_i = m_o(Z_i) + e_i, \quad E(e|Z) = 0, \tag{4.13}$$

[1]For conciseness and to maintain focus, we present the standard regularization assumptions in Appendix B.3. These assumptions are usually readily satisfied in empirical applications.

where we assume the dimension of D as $d_D = 1$ in this section and $s_o(\cdot), r_o(\cdot), m_o(\cdot)$ are arbitrary nonparametric functions of Z with high dimensional parameters. Our parameter of main interest is θ_o. To simplify the analysis, we denote $l_o(Z) := E(Y|Z)$, $\nu_i := Y_i - l_o(Z_i) = \theta_o u_i + \varepsilon_i$, and $\eta_o := (l_o(Z), r_o(Z), m_o(Z))$.

Chernozhukov et al. (2018) provided a consistent, approximately unbiased, and asymptotically normally distributed estimator, DML, for θ_o in the partially linear model. Thus, we first revisit the DML method as Estimator 4 and highlight its limitation in our context. Let $g_{1,dml}(\theta_o) := E[(Y - l_o(Z) - \theta_o(D^* - r_o(Z)))(D^* - r_o(Z))]$. We have $g_{1,dml}(\theta_o) = 0$ by (4.11).

Estimator 4 (DML on the labeled dataset only)

1. Take a K-fold random partition $(I_{k,l})_{k=1}^K$ of observations with true labels such that the size of each fold $I_{k,l}$ is $N_l = n_l/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define I_l as the index set of the dataset with true labels, and $I_{k,l}^c := I_l \backslash I_{k,l}$.

2. For each $k \in [K]$, construct an ML estimator $\hat{\eta}_{k,l} = \hat{\eta}((W_i)_{i \in I_{k,l}^c})$ of η_o, where the randomness of $\hat{\eta}_{k,l}$ depends only on the subset of data indexed by $I_{k,l}^c$.

3. Construct the estimator $\hat{\theta}_{dml}$ as the solution to (if not exactly, as close as)

$$\hat{g}_{1,dml}(\hat{\theta}_{dml}) := \frac{1}{K} \sum_{k=1}^{K} \hat{E}_{k,l}[\psi(W; \hat{\theta}_{dml}, \hat{\eta}_{k,l})] = 0,$$
$$\psi(W_i; \theta, \eta) = (Y_i - l(Z_i) - \theta(D_i^* - r(Z_i)))(D_i^* - r(Z_i)), \tag{4.14}$$

   where $\hat{E}_{k,l}$ is the empirical expectation over the kth fold of the data with true labels, i.e., $\hat{E}_{k,l}[\psi(W)] = N_l^{-1} \sum_{i \in I_{k,l}} \psi(W_i)$.

Notice that D_i^* is only available for the labeled dataset, and Estimator 4 is calculated on the labeled dataset only. Similar to the drawbacks of Estimator 1, a small size of the labeled dataset can cause large finite-sample bias and standard errors through Estimator 4. To improve the finite sample property and lower the variance, we propose a two-step DML estimator by utilizing both

labeled and unlabeled datasets. Let $g_{2,dml}(\theta_o) := E[(Y - l_o(Z) - \theta_o(D^* - r_o(Z))(D - m_o(Z))] = 0$, and $\Gamma_o^{plr} := \left(E\left[(D - m_o(Z))^2\right]\right)^{-1} \left(E\left[(D^* - r_o(Z)(D - m_o(Z))\right]\right)$.

Estimator 5 (Two-step DML on the full dataset)

1. Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[n] = \{1, \ldots, n\}$ such that the size of each fold I_k is $N = n/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_k^c := \{1, \ldots, N\} \backslash I_k$. For each $k \in [K]$, construct an ML estimator $(\hat{l}_k, \hat{m}_k) = (\hat{l}((W_i)_{i \in I_k^c}), \hat{m}((W_i)_{i \in I_k^c}))$ of (l_o, m_o), where the randomness of $(\hat{l}_k, \hat{m}_k)$ depends only on the subset of data indexed by I_k^c.

2. Take a K-fold random partition $(I_{k,l})_{k=1}^K$ of observations with true labels such that the size of each fold $I_{k,l}$ is $N_l = n_l/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define I_l as the index set of the dataset with true labels, $I_{k,l}^c := I_l \backslash I_{k,l}$. For each $k \in [K]$, construct an ML estimator $\hat{r}_k = \hat{r}((W_i)_{i \in I_{k,l}^c})$ of r_o, where the randomness of $\hat{r}_k$ depends only on the subset of data indexed by $I_{k,l}^c$.

3. Construct the two-step estimator $\hat{\theta}_{tsdml}$ as the solution to (if not exactly, as close as) [2]

$$\frac{1}{K} \sum_{k=1}^{K} \hat{E}_{k,l}[\psi_1(W; \hat{\Gamma}_{tsdml}, \hat{\eta}_k)] = 0, \text{ and} \tag{4.15}$$

$$\hat{g}_{2,dml}(\hat{\theta}_{tsdml}) := \frac{1}{K} \sum_{k=1}^{K} \hat{E}_k[\psi_2(W; \hat{\theta}_{tsdml}, \hat{\Gamma}_{tsdml}, \hat{\eta}_k)] = 0, \text{ where} \tag{4.16}$$

$$\psi_1(W_i; \Gamma, \eta) = (D_i^* - r(Z_i) - \Gamma(D_i - m(Z_i)))(D_i - m(Z_i)), \text{ and} \tag{4.17}$$

$$\psi_2(W_i; \theta, \Gamma, \eta) := (Y_i - l(Z_i) - \theta\Gamma(D_i - m(Z_i)))(D_i - m(Z_i)). \tag{4.18}$$

Here, $\hat{E}_{k,l}$ denotes the empirical expectation over the kth fold of the data with true labels,

[2] This is called the two-step DML estimator because both ψ_1 and ψ_2 possess the Neyman Orthogonality property that their pathwise derivatives with respect to η vanish when evaluated at the true parameter η_0, i.e.,

$$\partial_\eta E\left[\psi_1\left(W; \Gamma_o^{plr}, \eta_o\right)\right][\eta - \eta_0] = 0, \quad \partial_\eta E\left[\psi_2\left(W; \theta_o, \Gamma_o^{plr}, \eta_o\right)\right][\eta - \eta_0] = 0.$$

We provide the proof of the Neyman Orthogonality property in B.9.

i.e., $\hat{E}_{k,l}[\psi(W)] = N_l^{-1}\sum_{i\in I_{k,l}}\psi(W_i)$, and $\hat{E}_k$ is the empirical expectation over the kth fold of the data, i.e., $\hat{E}_k[\psi(W)] = N^{-1}\sum_{i\in I_k}\psi(W_i)$.

To improve estimation efficiency, we propose the following GMM-DML estimator, $\hat{\theta}_{gdml}$.

Estimator 6 (GMM DML on the full dataset)

$$\hat{\theta}_{gdml} = \underset{\theta}{\operatorname{argmin}}\, \hat{g}_{dml}(\theta)^T W_{dml}\hat{g}_{dml}(\theta), \tag{4.19}$$

where $\hat{g}_{dml}(\theta) = \begin{pmatrix}\hat{g}_{1,dml}(\theta)\\ \hat{g}_{2,dml}(\theta)\end{pmatrix}$ and W_{dml} is a pre-fixed positive definite weighting matrix:

$$W_{dml} = E\left[\begin{pmatrix}\psi(W;\theta_o,\eta_o)\\ \lambda_l\psi_2(W;\theta_o,\Gamma_o^{plr},\eta_o) - \psi_1\left(W;\Gamma_o^{plr},\eta_o\right)\theta_o\end{pmatrix}\begin{pmatrix}\psi(W;\theta_o,\eta_o)\\ \lambda_l\psi_2(W;\theta_o,\Gamma_o^{plr},\eta_o) - \psi_1\left(W;\Gamma_o^{plr},\eta_o\right)\theta_o\end{pmatrix}'\right].$$

We will give an estimator for W_{dml} later. By rewriting (4.19) as follows, we can derive the closed-form expression for the GMM estimator. Let $\hat{\Gamma}_{tsdml}$ be the solution to (4.15) and let

$$B_{plr} = \begin{pmatrix}\frac{1}{K}\sum_{k=1}^K \hat{E}_{k,l}[(Y_i - \hat{l}(Z_i))(D_i^* - \hat{r}(Z_i))]\\ \frac{1}{K}\sum_{k=1}^K \hat{E}_k[(Y_i - \hat{l}(Z_i))(D_i - \hat{m}(Z_i))]\end{pmatrix}, A_{plr} = \begin{pmatrix}\frac{1}{K}\sum_{k=1}^K \hat{E}_{k,l}[(D_i^* - \hat{r}(Z_i))^2]\\ \frac{1}{K}\sum_{k=1}^K \hat{E}_k[\hat{\Gamma}_{tsdml}(D_i - \hat{m}(Z_i))^2]\end{pmatrix}. \tag{4.20}$$

We can have $\hat{g}_{dml}(\theta) = B_{plr} - A_{plr}\theta$ and $\frac{\partial}{\partial\theta}\hat{g}_{dml}(\theta) = -A_{plr}$. Through the first-order condition, we have

$$\hat{\theta}_{gdml} = (A_{plr}' W_{dml} A_{plr})^{-1} A_{plr}' W_{dml} B_{plr}. \tag{4.21}$$

Assumption 4.2.2. Assume the followings hold[3]: (a) The full dataset $\{Y_i, X_i\}_{i=1}^n$ are i.i.d. samples drawn from $f_{Y,X}$ and the labeled dataset $\{X_i^*, X_i\}_{i=1}^{n_l}$ are i.i.d. samples drawn from $f_{X^*,X}$; (b)

[3]For conciseness and to maintain focus, we present the standard regularization assumptions in Appendix B.3. These assumptions are usually readily satisfied in empirical applications.

Equations (4.11), (4.12), and (4.13) hold; (c) $\lambda_l = \lim_{n_l \to \infty} n_l/n \in [0,1]$

Assumption 4.2.2.(b) outlines the moment conditions required for the partially linear model.

Theorem 4.2.1 (Consistency). *Under Assumption 4.2.2, any estimator* $\hat{\theta} \in \{\hat{\theta}_{dml}, \hat{\theta}_{tsdml}, \hat{\theta}_{gdml}\}$ *is a consistent estimator for* θ_o*, i.e.,*

$$\hat{\theta} \xrightarrow{p} \theta_o, \quad \text{as } n \to \infty. \tag{4.22}$$

Theorem 4.2.2 (Asymptotic Normality). *Under Assumption 4.2.2, any estimator* $\hat{\theta} \in \{\hat{\theta}_{dml}, \hat{\theta}_{tsdml}, \hat{\theta}_{gdml}\}$ *follows asymptotically normality, i.e.,*

$$\sqrt{n_l}(\hat{\theta} - \theta_o) \xrightarrow{d} \mathcal{N}(0, V_{plr}) \tag{4.23}$$

where we denote the variance V_{plr} *for* $\hat{\theta}_{dml}, \hat{\theta}_{tsdml}, \hat{\theta}_{gdml}$ *as* $V_{dml}, V_{tsdml}, V_{gdml}$ *as follows:*

$V_{dml} = (E[u^2])^{-1}\Xi_{dml}(E[u^2])^{-1}$, *where* $\Xi_{dml} = E[\varepsilon^2 u^2]$.

$V_{tsdml} = J_o^{-1}\Xi_{tsdml}J_o^{-1}$, *where* $J_o = E\left[\Gamma_o^{plr} e^2\right]$ *and*

$$\Xi_{tsdml} = \lambda_l E[(\nu - \theta_o\Gamma_o^{plr}e)^2 e^2] + E[(u - \Gamma_o^{plr}e)^2 e^2 \theta_o^2] - 2\lambda_l E[(\nu - \theta_o\Gamma_o^{plr}e)(u - \Gamma_o^{plr}e)e^2\theta_o].$$

$V_{gdml} = (\triangle_{dml}'\Xi_{gdml}^{-1}\triangle_{dml})^{-1}$, *where* $\triangle_{dml} = (-E[u^2] \quad -E[ue])'$ *and* $\Xi_{gdml} = \begin{pmatrix} \Xi_{dml} & \Xi_{dml,tsdml} \\ \Xi_{dml,tsdml} & \Xi_{tsdml} \end{pmatrix}$

with $\Xi_{dml,tsdml} = \lambda_l E[(\nu - \theta_o u)(\nu - \theta_o\Gamma_o^{plr}e)eu] - E[(\nu - \theta_o u)(u - \Gamma_o^{plr}e)ue\theta_o]$.

We provide the proofs of Theorems 4.2.1 and 4.2.2 in Appendix (B.10).

Based on Theorem 4.2.1, we are able to construct the confidence intervals for the scalar parameter θ_o. We provide the confidence intervals in Appendix (B.5)

An estimator for the optimal weighting matrix is given by

$$W_{dml}^{-1} = \frac{1}{K}\sum_{k=1}^{K} h_{k,dml}h_{k,dml}', \text{ where } h_{k,dml} = \begin{pmatrix} \hat{E}_{k,l}[\psi(W;\hat{\theta},\hat{\eta})] \\ \hat{\lambda}_l\hat{E}_k[\psi_2(W;\hat{\theta},\hat{\Gamma},\hat{\eta})] - \hat{E}_{k,l}[\psi_1\left(W;\hat{\Gamma},\hat{\eta}\right)\hat{\theta}] \end{pmatrix},$$

and $\hat{\lambda}_l = \frac{n_l}{n}$. The estimators $\hat{\theta}$ and $\hat{\Gamma}$ are some consistent estimator of θ_o and Γ_o^{plr} (such as the DML estimators on the labeled set, $\hat{\theta}_{dml}$ and $\hat{\Gamma}_{tsdml}$), the estimator $\hat{\eta}$ is an estimator for η_o under Assumption 4.2.2.

Remark 4.2.3. While the partially linear model involves more complex estimation steps because it does not assume a fully parametric model, the underlying intuition remains consistent with the linear model discussed in section 4.2.1. Asymptotically, we can further reduce the estimation variance and improve efficiency through the Two-step DML estimator or the GMM DML estimator as we add more unlabeled data. Moreover, the GMM DML estimator combines the DML estimator on the labeled dataset only and the Two-step DML estimator with a weight inversely correlated with the estimator variances. As a result, the GMM DML estimator can achieve the lowest estimation variance compared with the DML estimator on the labeled dataset only and the Two-step DML estimator.

Remark 4.2.4. We can further write the asymptotic variance of the Two-step DML estimator V_{tsdml} into the following form:

$$\begin{aligned} V_{tsdml} &= J_o^{-2}\left(\lambda_l E\left[(\theta_o\gamma+\varepsilon)^2e^2\right] + E\left[\gamma^2e^2\right]\theta_o^2 - 2\lambda_l E\left[(\theta_o\gamma+\varepsilon)\gamma e^2\theta_o\right]\right) \\ &= J_o^{-2}\left(\underbrace{\lambda_l E\left[\varepsilon^2e^2\right]}_{\text{variance for estimating } \theta_o \text{ using labeled dataset}} + \underbrace{(1-\lambda_l)\theta_o^2 E\left[\gamma^2e^2\right]}_{\text{variance due to missing true labels}}\right), \end{aligned} \tag{4.24}$$

where γ is the residual after projecting u on the space of e, that is: $u = \Gamma_o^{plr}e + \gamma$. Recall that $e := D - m_o(Z)$, $u := D^* - r_o(Z)$, and $\varepsilon := Y_i - \theta_o D_i^* - s_o(Z_i)$.

When the asymptotic proportion of the labeled dataset λ_l becomes closer to 1, the asymptotic variance tends to be driven more by the first component as outlined in (4.24), making it more akin to the variance observed in a classical Two-Stage Least Squares (TSLS) method with balanced sample sizes (i.e., $n_l = n$). When λ_l becomes closer to zero, the asymptotic variance tends to be driven more by the second component, which is the variance due to missing true labels. Moreover, $\theta_o^2 E\left[\gamma^2e^2\right]$ contains the asymptotic variance for estimating the linear projection coefficient, that is, estimating Γ_o^{plr}.

Remark 4.2.5. The effect of ML/AI-generated measurement error on asymptotic variance is captured by the second-order moment of γ, as detailed in (4.24). Notably, this ML/AI-generated measurement error is distinct from γ; γ represents the residual from projecting this error onto the space of D and Z. In other words, it embodies the component that cannot be linearly explained by D and Z. The intuition is that if we write the ML/AI-generated variable as $D = D^* + \omega$, with ω denotes the ML/AI-generated measurement error. Directly plugging in the ML/AI-generated variable could lead to a biased estimator for θ_o due to endogeneity—D being correlated with the measurement error ω. Our Two-step DML approach amends this by establishing orthogonality through projection: projecting ω onto the space of D and Z. This allows us to construct a moment condition for estimation without endogeneity. It is interesting to note that the component of the ML/AI-generated measurement error that can be explained by D and Z does not influence the asymptotic variance, and is captured by the projection coefficient Γ_o. It is the residual, γ, that contributes to the asymptotic variance. Consequently, a higher variance in γ leads to an increased asymptotic variance for the Two-step DML estimator, $\hat{\theta}_{tsdml}$.

4.3 Numerical Experiments

We utilize synthetic data to validate the effectiveness of our proposed estimator in this section. We first simulate measurement error in the variable of interest and then set up a linear regression model as well as a partially linear model for the outcome and feature vectors in Section 4.3.1 and Section 4.3.2. We examine the performance of our proposed estimators compared with other existing estimators.

Based on the measurement error simulation setup in Yang et al. (2018), we focus on the systematic independent nonclassical measurement error in this section, i.e.,

$$D_i = 0.9D_i^* + \varsigma_{1,i}. \tag{4.25}$$

For each observation i, D_i^* is the true label. D_i is the ML/AI-generated label for D_i^* and maintains a linear relation with D_i^*. The measurement error of the ML/AI algorithm here includes both a

systematic error 0.1 and a random term $\varsigma_{1,i}$.

4.3.1 Linear Case

Following the setup of case 1 in Sections 3.1 and 4.1 in Qiao and Huang (2021), we assume the true Data Generating Process (DGP) is as follows,

$$Y_i = 1 + 2D_i^* + 3Z_i + \varsigma_{2,i}, \tag{4.26}$$

where all variables stated in Equations (4.25) and (4.26) are scalars. The parameter of our interest is the coefficient before D_i^* in Equation (4.26), which equals 2 here. Z_i denotes the observations of the control variable, and Y_i denotes the observations of the outcome variable. We generate D_i^* and Z_i through two independent normal distributions. We set both the mean and variance of the distribution of D_i^* as 1. We set the mean and variance of the distribution of Z_i as 2 and 0.5, respectively. The two random noise terms $\varsigma_{1,i}$ and $\varsigma_{2,i}$ are i.i.d. drawn from two independent normal distributions. The noise term $\varsigma_{2,i}$ follows a standard normal distribution. Meanwhile, we vary the mean μ_1 and variance σ_1 of the distribution of $\varsigma_{1,i}$ to see the effect of measurement error on the parameter estimation.

The sample size of the whole dataset is 1,000. We observe (Y_i, D_i, Z_i) for the entire dataset, and we can also observe (Y_i, D_i^*, D_i, Z_i) for a smaller labeled dataset. We vary the proportion of the labeled data $\frac{n_l}{n}$ to see the effect of these parameters on the main effect estimation. We apply our proposed method to gain estimators from this simulated dataset and compare the performance of our estimators with other existing estimators, including OLS on the labeled data only (Estimator 1) and OLS on the full dataset.

We repeat the estimation for 1,000 runs and plot the empirical distributions of the estimators from all runs in Figure B.1. In the plots, we denote the direct OLS estimation of Y_i on (D_i, Z_i) on the full dataset as "ols_full" with a blue line, Estimator 1 as "ols_labelonly" with an orange line, our proposed Estimator 2 as "2step_full" with a green line and Estimator 3 as "gmm_full" with a red line. We use a red dotted vertical line to represent the true value of the main effect.

We can see from the plots the direct OLS estimation of Y_i on (D_i, Z_i) on the full dataset always has an upward bias, while Estimators 1, 2, and 3 generally have a much lower bias. Comparing the variance among estimators with low bias, our proposed Estimators 2 and 3 always have lower variance than Estimator 1 does in our experiments. The outperformance of Estimators 2 and 3 remains significant even when the proportion of the labeled dataset in the entire dataset increases.

The effect of the measurement error in ML/AI algorithms. The mean of the measurement error does not influence performance (neither bias nor efficiency). The intuition is the effect of the mean of the measurement error is captured by the constant term in the second step econometric model and thus does not affect our estimation of the parameter of interest.

However, the variance of the measurement error does affect performance by introducing bias and reducing efficiency. For instance, when comparing Figures B.1 (a) and 1 (b), it appears that an increase in variance reduces the bias in the direct/naive OLS estimation using ML-predicted labels (ols_full). Yet, this observation does not imply that higher variance always results in less bias or a superior estimator. This is because, in our context, the ML-predicted label tends to underpredict the true label ($D_i = 0.1D_i^* + \varsigma_{1,i}$), leading to an amplification bias. Meanwhile, the variance in the measurement error introduces an attenuation bias. As the variance increases, the attenuation bias also increases, further offsets the amplification bias. Consequently, we observe a reduction in bias with increased variance in $\varsigma_{1,i}$, but this does not necessarily indicate a less biased estimator. Instead, it reveals the intricate impact of the ML/AI algorithm measurement error on the estimation of the second-step parameter of interest, potentially leading to misleading results.

Additionally, we note that an increase in the variance of the measurement error reduces the efficiency of both the GMM and TSLS estimators, as depicted in Figures B.1 (a) and 1 (b). Loosely speaking, a higher variance in the measurement error dilutes the information about the true label contained in the predicted label, thereby diminishing the efficiency of both the GMM and TSLS estimators. Nevertheless, it is noted that these estimators remain unbiased despite the increased variance in measurement error.

The effect of labeled data proportion. As we increase the proportion of the labeled dataset, the OLS estimator, the TSLS estimator, and the GMM estimator all achieve lower variance while

keeping the estimation unbiased (see Figures B.1 (a) and 1 (e)). Increasing the labeled data proportion while keeping the whole dataset size the same leads to an increased labeled dataset size and a reduced unlabeled dataset size. An increased sample size for the labeled dataset naturally reduces the variation of OLS estimator and the first step in the TSLS estimator, which further reduces the variance of the GMM estimator.

Meanwhile, we observe that the reduction in the variance of the GMM estimation is slightly larger than the reduction in the variance of the TSLS estimator (see Figure B.1 (g)). This is because the benefit attained in the estimation efficiency for the OLS estimator is larger than TSLS estimator. As the GMM estimator is a weighted average of the OLS and the TSLS estimators with a weight inversely correlated with the estimator variance, GMM estimator can achieve a larger efficiency improvement due to the OLS estimation efficiency improvement.

4.3.2 Partially Linear Case

In the partially linear case, the true DGP is specified as follows:

$$Y_i = 1 + 2D_i^* + 3Z_i + 4Z_i^2 + 3Z_i^3 + \varsigma_{2,i}. \tag{4.27}$$

Here, the variable of interest, D_i^*, maintains a linear relationship with the outcome Y_i. However, the control variables exhibit increased complexity compared to those in the linear DGP. The parameters for D_i^*, Z_i, $\varsigma_{1,i}$, and $\varsigma_{2,i}$ are identical to those outlined in Section 4.3.1.

Again, we set the sample size of the whole dataset as 1,000. We can only observe (Y_i, D_i, Z_i) for the whole dataset, while we can observe (Y_i, D_i^*, D_i, Z_i) for the labeled dataset. We vary the proportion of the labeled data $\frac{n_l}{n}$, the mean μ_1 and variance σ_1 of $\varsigma_{1,i}$ to see the effect of these three parameters on the main effect estimation. We estimate the nuisance functions, denoted as η, utilizing two machine learning (ML) algorithms: Random Forest (RF) and Gradient Boosting (GB). Subsequently, we calculate our proposed estimators. We compare the performance of our estimators (Estimator 5 and Estimator 6) with other existing estimators. This includes the DML approach applied solely to the labeled data (Estimator 4), and the DML estimator which regresses

Y on (D, Z) using the full dataset.

We repeat the estimation for 1,000 runs and plot the empirical distributions for all estimators examined from all runs in Figure B.2 with GB and Figure B.3 with RF. In the plots, we denote the direct DML estimation of Y_i on (D_i, Z_i) on the full dataset with GB as "dml_full_gb" with a blue line, Estimator 4 with GB or RF as "dml_labelonly_gb" or "dml_labelonly_rf" respectively with an orange line, Estimator 5 with GB or RF as "2step dml_full_gb" or "2step dml_full_rf" respectively with a green line, Estimator 6 with GB or RF as "gmm dml_full_gb" or "gmm dml_full_rf" respectively with a red line. Again, we use a red dotted vertical line to represent the true value of the main effect. As observed in Figure B.1, the plots in Figures B.2 and B.3 indicate that the direct DML estimation of Y_i on (D_i, Z_i) across the entire dataset also consistently exhibits an upward bias. In contrast, Estimators 4, 5, and 6 demonstrate significantly lower bias. Comparing the variance among estimators with low bias, we found that our proposed Estimators 5 and 6 consistently exhibit lower variance than Estimator 4 in our experiments. The outperformance of Estimator 5 and Estimator 6 is significant when we increase the proportion of the labeled dataset in the whole dataset. The results with RF and the results with GB are quite consistent.

The partially linear regression model employs a more complex functional form than standard linear specifications. Simply plugging in predicted values from an ML/AI model into the double machine learning estimator is insufficient for accurately recovering treatment effects in partially linear contexts. Doing so can introduce bias and sometimes can also inflate the variance around estimates. Moreover, we can conclude from the figures that when we increase the proportion of the labeled dataset while keeping the whole dataset size the same, we can generally reduce the variances of the DML estimator on the labeled dataset only, the Two-step DML estimator, and the GMM DML estimator (see Figures B.2 (a) and B.2 (e)).

4.4 Real-world Application on App Review

To further evaluate our estimators, we follow the existing literature (Yang et al., 2018) to test an empirical question in this section: How does review sentiment affect review helpfulness? Understanding the effect of review sentiment on review helpfulness can help enhance the overall

usefulness and quality of reviews on the platform. Therefore, it has been a central question in the digital marketing and information systems disciplines (Yin et al., 2017; Qiao and Huang, 2021; Yu et al., 2023). Review helpfulness is usually indicated by helpfulness votes; the sentiment of a textual review can be identified using natural language processing methods (Yang et al., 2018). In Sections 4.4.1 and 4.4.2, we adopt a commonly used ML algorithm, XGBoost, and a recently emerging generative AI, ChatGPT, to predict sentiment scores for review content. We demonstrate the performance of our proposed Two-step and GMM estimators by comparing them with a labeled-only estimator and an estimator that uses predicted labels (i.e., by XGBoost or ChatGPT).

To conduct this empirical exercise and examine the performance of our estimators, we adopt an open dataset of Tinder App reviews from Kaggle[4]. This dataset collects 590,780 user reviews for the online dating app[5], *Tinder*, posted on the Google Play Store. This dataset contains variables mainly including user name, user profile image, review content, user rating for the app, number of helpfulness votes to the review, app version the review regards to, and review time. We preprocess the data as follows: first, we drop all records that miss comment content or user name, which reduces the dataset size to 589,438 observations; second, we construct several new covariates based on user reviews and review time, including word count of a review, whether it is a weekend for a review time, whether it is a holiday for a review time, year of a review time, month of a review time, time bin in a day for a review time[6]. We use the number of helpfulness votes as the dependent variable. Following Yang et al. (2018), we treat the user rating as the ground truth of the review sentiment. Due to the subjective nature of sentiment, one might argue that the star rating might not be the perfect ground truth for sentiment, while the rating provides an objective label to evaluate the estimators (Yang et al., 2018). Alternatively, one can view the relationship of interest as review rating and review helpfulness. Notably, in many cases, the rating is not available (Godes and Mayzlin, 2004). Then, researchers may use the review text and ML algorithms to achieve a noisy

[4]https://www.kaggle.com/datasets/shivkumarganesh/tinder-google-play-store-review?resource=download

[5]Note that this dataset keeps adding new records. We downloaded the dataset on July 3, 2023.

[6]We divide each day into three time periods. The value of the time-period variable is 0 if the review time belongs to 0:00-8:00, 1 if it belongs to 8:00 - 16:00, or 2 if it belongs to 16:00 - 24:00.

prediction of the review rating and then to understand its effect on review helpfulness.

We use the true labels on the full dataset to conduct a benchmark estimation for the sentiment effect, as illustrated in Equation (4.28), where s_o is a random forest predicting model and θ_o is the parameter of our interest. Mimicking the practical constraints where we only have the true labels on a small proportion of the full dataset, we estimate the sentiment effect through our proposed estimators (Two-step and GMM estimators) by following the steps outlined in Section 4.2.2 for Estimators 5 and 6. Additionally, we provide the estimator that utilized the labeled dataset, following the steps described in Section 4.2.2 for Estimator 4. We also provide an estimator that utilizes the predicted label from the full dataset, following the steps similar to Estimator 4, but replacing the true label with predicted labels. In Appendix B.3, we provide a detailed explanation of how Assumption 2 is satisfied in this context.

We use the bootstrap method to calculate standard errors for all estimators. We compare estimated effects and standard errors from all the estimations with the true effect which is estimated by random forest using all the true labels.

$$\textit{ReviewHelpfulness}_i = \theta_o \textit{ReviewSentiment}_i + s_o(\textit{ControlVariables}_i) + \varepsilon_i \tag{4.28}$$

4.4.1 Correcting ML-predicted Variable Bias

In this section, we use review content to predict sentiment labels through machine learning algorithms and then use the predicted sentiment variable to conduct estimations of the effect θ_o through our proposed estimators and other estimators. Specifically, we extract TF-IDF features from review content and then adopt XGBoost algorithm to fit regression models of the user rating on TF-IDF features. We use out-of-sample predictions for the user ratings based on the fitted XGBoost model as ML-predicted sentiment. There might be a concern that ML-predicted sentiment of a review is correlated with user ratings of other reviews if the fitting model uses all reviews' user ratings. To avoid such concern, we utilize sample splitting before we fit the models. We use a part of the data to fit the XGBoost model and use another part of the data to predict the review sentiment and conduct further process.

It is noticeable that XGBoost is an ensemble supervised learning algorithm which combines multiple decision tree models' predictions. The predicting performance of XGBoost is dependent on parameters. We will show how the estimation of θ_o is affected by the parameter and number of estimators in XGBoost in this section. Number of estimators is the number of trees in the ensemble. Increasing the number of estimators would lead to improvement of the model fitting but can also potentially cause the overfitting issue.

We use the bootstrap method to calculate the means and standard deviations of estimations. Bootstrapping allows us to empirically validate our estimations by resampling the dataset. To further illustrate our estimation method, we list the whole process with bootstrap in the following.

1. **Construct Dataset:** Randomly draw 200,000 samples from the original dataset with replacement as a superior dataset. Shuffle the superior dataset and split it with a proportion ratio of 5%:95%. We use the 5% dataset with user ratings and use the 95% dataset without user ratings to construct a *Pre-Full Dataset*. The way we construct the Pre-Full Dataset is to mimic the situation where we only have true sentiment labels (user ratings) for a small proportion (5%) of the data. We call the 5% dataset with user ratings as the *Labeled Dataset* and the 95% dataset without user ratings as the *Unlabeled Dataset*. We extract TF-IDF features from review content for the whole superior dataset. For a visual representation of our data splitting methodology, please refer to Figure B.4 in Appendix;

2. **Train Sentiment via ML:** Train an XGBoost regression model of user ratings on the TF-IDF features. To avoid violation with the iidness of a review's predicted user rating and another review's TF-IDF features, we use 50% of the Labeled Dataset to do the training and only keep those data not used in the training to conduct further prediction.

3. **Predict Sentiment:** Predict user ratings for the remaining 50% (not used in the training) of the Labeled dataset and the Unlabeled Dataset. The prediction is based on the fitted XGBoost regression model in step 2 and the TF-IDF features of these datasets. We call the combination of these two datasets (remaining 50% of the Labeled dataset and the Unlabeled

Dataset) as *Full Dataset*;

4. **Estimate the Effect θ_o:** Use the Full Dataset to conduct Two-step and GMM DML estimations. Meanwhile, we also conduct classical DML estimations with the remaining 50% (not used in the training) of the Labeled dataset and the full dataset separately as benchmark estimations.

5. Repeat step 1 to step 4 300 times and calculate the means and standard deviations of the results from all estimation methods.

We present the results in Tables B.1. By running the DML on the original dataset with true labels, we obtain the true effect θ_o and we calculate the estimation bias as $(\hat{\theta}_o - \theta_o)$, where we use $\hat{\theta}_o$ to denote the estimation for θ_o. DML(Full) estimation uses the full dataset and the predicted sentiment variable without accessing the true sentiment variable. The DML(Labeled) estimation method only uses the labeled dataset so the changes in the number of estimators in XGBoost do not affect the DML(Labeled) estimation result. As mentioned earlier, the sample size of the labeled dataset is much smaller than the sample size of datasets used by other estimation methods. As a result, the standard deviation of the DML(Labeled) estimation is highest among all estimation results. TwoStep DML uses the full dataset and conducts the Estimator 5. GMM DML uses the full dataset and conducts the Estimator 6. The number of random partition folds is 5 in both TwoStep DML and GMM DML. Overall, we can see that the GMM DML estimation achieves the lowest magnitude of average bias as well as standard deviation. As we increase the number of estimators in the XGBoost algorithm, the sentiment prediction performs better and all estimations leveraging ML-predicted sentiment variables improve.

4.4.2 Correcting GPT AI-Predicted Variable Bias

In this section, we leverage the recent emerging generative AI ChatGPT to predict sentiment labels and then estimate the effect of the sentiment variable on the review helpfulness through our

proposed estimator and other potential estimators. Rather than training a complex machine learning model that requires substantial labeled data and time-intensive feature engineering, leveraging the ChatGPT would allow us to efficiently access exceptionally accurate sentiment predictions. ChatGPT has been trained on a vast dataset, and has state-of-the-art natural language processing capabilities built-in. By making API calls, we tap directly into these powerful capabilities to predict user sentiment scores for review content.

Specifically, we construct a dataset of 10,000 reviews by randomly drawing samples from the original dataset without replacement. Then we provide the dataset and a prompt for ChatGPT API to generate sentiment ratings for all reviews in the dataset. We list our prompt in the following.

> *"As an AI with expertise in language and emotion analysis, your task is to give a sentiment rating of the following text without explanation. Please consider the overall tone of the discussion, the emotion conveyed by the language used, and the context in which words and phrases are used. Indicate a rating within the range from 1 to 5, where 1 means extremely negative and 5 means extremely positive. A higher rating means less negative and more positive. A rating of 3 means neutral."*

We show the joint probability density distribution of the predicted rating from the ChatGPT API and the true rating labels in Figure B.5. The ChatGPT API predictions exhibit a slight conservative bias relative to the true ratings, as illustrated in the figure. When the true rating indicates an extremely negative sentiment (a rating of 1), ChatGPT's predicted rating tends to be slightly higher. Conversely, for reviews with a very positive true rating of 5, ChatGPT's predictions tend to be modestly lower. Given that 39.15% of the true ratings are at the lowest end of the scale (indicating extremely negative sentiment), ChatGPT's more centered, conservative predictions appear higher on average compared with the true ratings.

We describe the afterward whole estimation process as follows.

1. **Construct Dataset:** Randomly draw 2,000 samples from the dataset of 10,000 reviews with replacement as the *Full Dataset*. Shuffle the 2,000 samples and split them with a proportion

ratio of 10%:90%. We use the 10% dataset with true user ratings (and with ChatGPT AI-predicted sentiment) as the *Labeled Dataset* and use the 90% dataset without true user ratings (but with ChatGPT AI-predicted sentiment) as the *Unlabeled Dataset*.

2. **Estimate the Effect θ_o:** Use the Full Dataset to conduct Two-step and GMM DML estimations. Meanwhile, we also conduct classical DML estimations with the Labeled dataset and the full dataset separately as benchmark estimations.

3. Repeat step 1 and step 2 100 times and calculate the means and standard deviations of the results from all estimation methods.

Notice that in the above process, we set the proportion of the labeled dataset as 10%. To examine the estimation performance under different proportions of labeled datasets, we also conduct estimations with proportions 20% and 30%. We present the empirical distributions of estimation results in Figures B.6. We follow the notations in figures in Section 4.3.2. By running the DML on the dataset with true labels, we obtain the true effect of sentiment, -0.023, which is presented as a black vertical line in the figures.

As shown in Figure B.6, estimations from all estimators perform quite well with little bias and low variance. The direct DML estimation using the entire dataset (with GPT predictions but without true labels) exhibits a slight upward bias, especially as presented in Figure B.6 (c). This upward bias can be caused by the conservative predictions from ChatGPT especially on reviews of extremely negative sentiment. In contrast, Estimators 4 (denoted as dml_labelonly in figures), 5 (denoted as 2step dml_full in figures), and 6 (denoted as gmm dml_full in figures) demonstrate lower bias by utilizing the labeled dataset to correct the bias. For the estimation variance, our proposed Estimators 5 and 6 consistently exhibit lower variance than other estimators. The outperformance of Estimator 6 is more significant when we increase the proportion of the labeled dataset in the dataset. In summary, the sentiment predictions from the ChatGPT API demonstrate impressively high accuracy. As a result, using these predictions directly for the DML estimation introduces minimal bias. However, by applying our proposed estimators, especially the GMM

DML estimator, we can further refine the estimates reducing bias and improving efficiency. Therefore, coupling ChatGPT with our proposed estimators (Estimators 5 and 6) enables efficient and accurate estimations of θ_o.

4.5 Discussion

With the significant advancements in ML and AI (including LLMs) and the widespread availability of various APIs, IS researchers are presented with promising opportunities to utilize ML/AI-generated regressors for deriving new insights using econometric analysis (Shin et al., 2020; Zhang et al., 2022; Yu et al., 2023, 2024). Meanwhile, empirical researchers are increasingly turning to partially linear models for econometric analysis due to their ability to integrate various ML algorithms and their strengths in high-dimensional causal inference (Chernozhukov et al., 2018; Yu et al., 2020; Wang et al., 2021). Consequently, it is timely and crucial to address the debiasing of ML/AI-generated regressors in both linear and partially linear models.

This essay proposes two-step estimators and GMM estimators for debiasing ML/AI-generated regressors in partially linear models while keeping linear models as special cases. Therefore, our work advances the IS and management science literature on ML/AI-generated regressors and linear models (Yang et al., 2018; Qiao and Huang, 2021; Fong and Tyler, 2021; Allon et al., 2023). In addition, we contribute to the econometrics literature by extending the classical DML framework (Chernozhukov et al., 2018) to accommodate the case when we have generated regressors in the full sample and their true labels in a small subsample. Further, we theoretically prove the asymptotic consistency and normality of our proposed two-step estimators and GMM estimators, which advances the extant studies focusing on algorithmic development and simulations (Yang et al., 2018; Wei and Malik, 2022). We provide consistent variance estimators, which are instrumental for statistical inference. Further, we synthetically simulate cases of measurement errors in linear regression and partially linear regression models. Through the comparison of other methods and our proposed estimators, we show the outperformance of our two-step estimators and GMM estimators in both unbiasedness and low variance in the numerical experiments. We also provide empirical applications of our proposed estimators on an app review dataset demonstrating the per-

formance of our estimators versus other estimators. In the empirical applications, we show how our proposed estimators can correct the estimation bias from ML-predicted and ChatGPT AI-predicted sentiment variables as well as improve estimation efficiency.

Our work offers important practical implications. First, experimental systems are commonly integrated into large digital platforms to make high-dimensional causal inferences and generate valuable practical insights. Partially linear models are especially useful for high-dimensional causal inferences and are commonly adopted in large-scale experimental systems (e.g., Tencent Fast Causal Inference, Amazon AWS DoubleML services, Microsoft EconML tool). Our framework can enhance the current experimental systems by debiasing ML/AI-generated regressors in partially linear models. Second, ML/AI predictions are biased, which has become a key concern for the use of ML/AI in causal inference and high-stake decision-making in recent years. Our work does not restrict the sources of errors generated by ML/AI, and therefore, has the potential generalizability to address ML/AI biasedness and promote ML/AI fairness. We show that ML/AI-generated predictions can be asymptotically reliable in understanding causal effects, given a small human-annotated subsample. The debiased causal estimation can then be leveraged for unbiased ML/AI-driven decision-making. Third, our work is compatible with leveraging LLMs for estimation and inference. LLMs have the capability of zero- or few-shot predictions. For example, they can predict textual sentiments with zero or few examples. However, due to the randomness of LLMs, it is questionable whether we may obtain reliable predictions and econometric estimation. Our work can unleash the power of zero- or few-shot predictions by LLMs, particularly for partially linear models. Fourth, although our primary motivation stems from ML/AI-driven IS practices, our framework possesses the versatility to be applicable in a variety of other domains and disciplines.

Our work opens new opportunities for future researchers. First, we keep the parameters of interest linear for the interpretation of the "average effect" and model the control variables as non-parametric, resulting in a semi-parametric model. However, our framework may be extended to a fully non-parametric specification in future research. Second, we focus on the estimation and inference for the regression coefficients of interest. Future researchers may extend our work to debias decision-making with ML/AI-generated variables. Third, we debias the ML/AI-generated

regressors using a small labeled dataset. However, sometimes labeled datasets may be difficult or costly to obtain. Without using labeled datasets, future researchers may leverage partial identification approaches to establish identification sets or utilize distributional robust optimization methods to provide estimators that are robust to potentially unknown distributions of measurement errors.

Chapter 5

WHAT AND WHEN TO SELL IN LIVESTREAMING: A STRUCTURAL ANALYSIS OF ONLINE DECISION-MAKING

5.1 Research Context and Data

5.1.1 Research Context

Our research context is primarily focused on the dynamic environment of livestream sales, specifically within one of China's largest video platforms. Livestream shopping has emerged as a powerful e-commerce tool, combining real time video content with interactive shopping experiences. Each livestream operates within a limited timeframe, typically lasting several hours. Prior to the commencement of a livestream, hosts are allocated a specific product pool. This pool comprises the items they are authorized to present and sell during the broadcast. The selection and timing of product presentations within this allocated pool become critical decisions for the hosts.

The livestream environment is characterized by its highly dynamic and interactive nature. Hosts actively showcase products, emphasizing their features and benefits to attract potential buyers. This real-time presentation allows for immediate engagement with the audience, creating an experience that closely mimics in-person shopping. Throughout the duration of a livestream, the audience composition is in constant flux. New viewers may join at any point, while others may leave. This continuous ebb and flow of consumers creates a dynamic atmosphere that hosts must navigate skillfully. Throughout the livestream, hosts have access to real-time metrics such as viewer count and purchase amounts. This immediate feedback allows them to gauge the effectiveness of their presentation strategies and product selections. Based on this information, hosts can dynamically adjust their approach to maximize total sales.

Compared to Online Commerce: In livestreams, products are demonstrated in real-time, allowing hosts to showcase features, answer questions, and address concerns immediately. Online

shopping relies on static images, pre-recorded videos, and written descriptions, which, while informative, lack the dynamic and responsive nature of livestream presentations.

In livestream settings, recommendations are geared towards a crowd of viewers simultaneously, based on real-time audience engagement and feedback. These recommendations are often influenced by the host's ability to read the audience and adapt, creating a "group shopping" experience that feels more organic and spontaneous. Livestream recommendations may leverage trending items or time-limited offers to create urgency, catering to the collective interests of the viewing audience. In contrast, online shopping platforms typically employ personalized recommendation systems tailored to individual users. These systems utilize sophisticated algorithms and machine learning to analyze a user's browsing history, purchase behavior, and demographics, predicting individual preferences with often high accuracy. While these personalized recommendations can be highly effective, they may sometimes feel impersonal or overly predictive. Online shopping platforms often present these as "You might also like" or "Frequently bought together" suggestions. The dynamic, crowd-oriented nature of livestream recommendations contrasts sharply with the data-driven, individualized approach of traditional e-commerce, each offering unique advantages in engaging and converting customers.

Compared to Offline (In-Store) Sales: While livestream commerce brings many elements of the in-store experience online, it differs from traditional offline sales in several key aspects related to our work. Livestream commerce can reach a much larger audience simultaneously, unrestricted by physical store capacity or geographic location. This allows for potentially greater sales volume and wider market reach compared to individual physical stores.

Livestream commerce holds a significant advantage in its ability to harness real-time data for rapid optimization of sales strategies. During a livestream, hosts and platform algorithms can continuously track viewer engagement metrics, such as view counts and purchase behaviors. This wealth of instantaneous data allows for dynamic adjustments to product selections, presentation styles, and promotional strategies within the same streaming session. For instance, if a particular product or feature generates high viewer interest, hosts can allocate more time to it or introduce similar items. In contrast, offline commerce faces considerable challenges in collecting and utiliz-

ing such granular, real-time consumer data. Traditional brick-and-mortar stores typically rely on historical sales data, periodic customer surveys, and general market trends to inform their strategies. The lag between data collection and strategy implementation in offline settings makes it difficult to respond swiftly to changing consumer preferences or to optimize product presentations in real-time. While some offline retailers are adopting technologies like IoT sensors and mobile apps to bridge this gap, they still cannot match the immediacy and depth of data available in livestream commerce. This disparity in real-time data utilization capabilities gives livestream commerce a distinct edge in rapidly understanding and catering to evolving consumer preferences, potentially leading to more effective sales optimization and higher conversion rates.

5.1.2 Data Description

Our dataset contains 5,178 livestreams by 114 livestream hosts during the time window from April 5, 2023 to April 28, 2023 on the platform. The data has 238,985 observations where each observation represents a product selected by a host for presentation at a specific time point during a livestream. On average, each livestream contains 13 sessions (product sets), with approximately 67 products sold and generating revenue of 47,256 USD. Within each session, hosts typically present 5 products as a product set, spending an average of 4 minutes per session. We provide plots of the distribution of the number of sessions per livestream in Figure C.1 and the distribution of the session duration in Figure C.2 in Appendix. This rich dataset allows us to examine the hosts' product selection and presentation timing decisions based on various aspects of livestream commerce, including sales outcomes, environmental factors, product characteristics, and host attributes. We detail the collection and construction of variables of an observation in the following.

First, to model the livestream hosts' demand learning behavior, we use the sales of each product presented as the outcome variable within the demand learning process. Specifically, we calculate this as the total sales generated after the product's presentation. We also collect several key livestream environment features and product features to model the demand learning. To simplify the analysis, we denote each presentation (including all interactions between the livestream hosts and audience) for a product set as a session and record the session numbers in time order. For each

session within a livestream, we record the number of viewers present at the time of introduction. This viewer count serves as a crucial environmental variable, acting as a proxy for the potential audience size and engagement level. We collect product price, brand status, and category as the main product features that livestream hosts consider in demand learning. Product categories are encoded using one-hot encoding. Furthermore, we encode the brand status as a binary variable, where 1 represents a branded product and 0 indicates a non-branded product.

Second, to model the hosts' presentation timing behavior, we collect time length of each session (presentation time window for each product set selected) and construct feature variables for each productset selected in the data. We begin by measuring the diversity of product categories within each set. The variable SetCategoryDiversity represents the number of unique categories present in a productset, providing insight into the range of products hosts choose to present together. To capture the overall brand presence in a productset, we calculate SetIsBrand as the average brand status across all products in the set. We also compute several aggregate measures to characterize the economic potential of each productset. SetSales and SetRevenue represent the average sales and revenue generated by the products in the set, respectively. These variables allow us to assess the effect of overall performance of product sets on hosts' presentation timing decisions. We also calculate the average price of products in the set as the variable SetPrice. Further, we include SetSize, which simply counts the number of products in a productset.

Lastly, to model the heterogeneity in hosts' decisions, we use the comprehensive set of variables capturing host characteristics in the dataset. We use the total number of videos posted by each host to serve as an indicator of their experience and content production volume. To measure host's popularity and engagement within the platform community, we use two variables: the number of favorites they have given to others' content and the number of favorites their content has received on the platform. We also capture the host's influence and connectivity within the platform ecosystem by using the number of followers they have and the number of users they follow. Host demographics are also included in our dataset. Gender is encoded using a trichotomous scheme: 0 for teams including both male and female hosts, 1 for male hosts, and 2 for female hosts. We also include a binary variable for education, where 1 indicates the host attended university and 0

indicates they did not.

Summary statistics for key variables are presented in Table 5.1.

Table 5.1: Descriptive Statistics

Variable	Min	Max	Mean	Median	Std. dev.
Product Features					
Log(*Sales*)	0.0000	11.8501	1.1761	0.6931	1.6345
Log(*Price*)	-4.6052	12.2051	5.7350	5.5759	1.9131
Log(*Viewers*)	0.0000	12.7032	5.5985	5.3613	1.8941
IsBrand	0.0000	1.0000	0.6435	1.0000	0.4790
Product Set Features					
SetCategoryDiversity	1.0000	22.0000	1.7320	1.0000	1.6895
SetIsBrand	0.0000	1.0000	0.6435	1.0000	0.4634
Log(*SetSales*)	0.0000	11.8501	2.4345	1.9459	2.2243
Log(*SetRevenue*)	0.0000	16.4180	6.5830	7.6000	3.8779
Log(*SetPrice*)	0.0100	12.2051	5.9139	5.6553	1.8048
Log(*SetSize*)	0.0000	4.5850	1.3310	1.3860	1.1608
Host Features					
Log(*TotalVideoCount*)	1.0986	9.4957	6.5212	6.5432	1.3142
Log(*FavoritingCount*)	0.0000	10.3796	0.2349	0.0000	1.3393
Log(*TotalFavorited*)	8.9582	20.3765	14.2266	14.1434	2.0177
Log(*FollowerCount*)	11.1727	17.3720	14.3003	14.3662	1.0923
Log(*FollowingCount*)	0.0000	9.1041	3.4914	3.1772	1.8989
Gender	0.0000	2.0000	0.7826	0.0000	0.9251
Education	0.0000	1.0000	0.1043	0.0000	0.3070

5.2 Model

In this section, we propose a learning and decision-making model to describe the dynamics of livestream hosts' product selection and presentation scheduling behaviors. First, we present the sequence of decisions a livestream host conducts in Section 5.2.1. Second, we model how livestream hosts learn about the demand through a process of exploration and exploitation over time and make product selection decisions in Section 5.2.2. For this learning and decision-making process, we propose several strategies hosts can potentially take in Sections 5.2.2 and 5.2.2. We also discuss benchmark strategies, especially non-learning strategies, in Section 5.2.2. Furthermore, given

the product selection decisions, we model how livestream hosts utilize product features, dynamic livestream engagement, and the time to schedule the product presentations in Section 5.2.3.

5.2.1 Sequence of Decisions

Each host can hold multiple livestreams sequentially (see Figure 5.1), with each livestream consisting of distinct selling sessions where the host presents products and interacts with viewers in real-time. The success of these sequential livestreams can be influenced by various factors including the host's experience and the strategic selection of products and presentation timing. We begin by summarizing the sequence of decisions by the host in each session within a livestream. We provide an illustration of the decision sequence within a livestream in Figure 5.2.

The host's objective is to maximize the total sales throughout a livestream l by deciding the product assortment $S_{l,t}$ and the presentation duration $D_{l,t}$ over time. In the first time period of a livestream l ($t = 1$),

1. The host selects one or multiple products from a given product pool $\mathcal{A}_l$. We denote the selected products (assortment) as $S_{l,t}$.

2. The host presents the products selected $S_{l,t}$ and observes the number of staying viewers V_l in real time until the host decides to stop the current presentation, leading to the duration of the presentation of $S_{l,t}$ as $D_{l,t}$.

3. The host observes the sales $\pi_{l,it}$ of each product $i \in S_{l,t}$.

4. The decisions above are repeated in the following periods ($t = 2, 3, \cdots$).

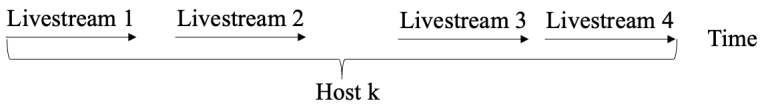


Figure 5.1: Multiple livestreams of a host k

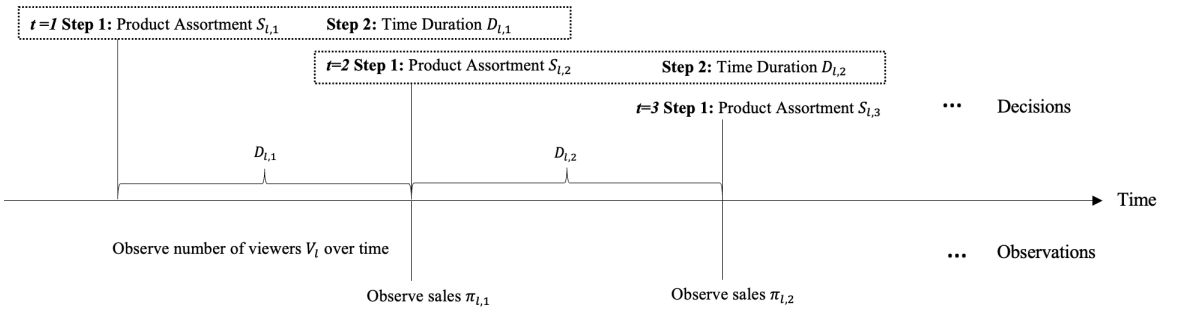


Figure 5.2: Timeline of hosts' decisions in livestream l

5.2.2 Step 1: Product Selection

We discuss how hosts conduct product selection decisions in this section. Before a livestream l begins, the host is assigned a pool of products $\mathcal{A}_l$. The host needs to optimize the product assortment decisions throughout the whole livestream to maximize the total sales at the end of the livestream. The sales of a product i at the time period t of the livestream l, denoted as $\pi_{l,it}$, is a function of the product features (including price, category, and brand status), the livestream engagement (measured by number of viewers), and the number of sessions elapsed. We provide the function in Equation (5.1)):

$$\pi_{l,it} = f(\mathbf{x}_i, V_{l,t}, t) + \zeta_{l,it}, \tag{5.1}$$

where the variable $\mathbf{x}_i \in \mathbb{R}^d$ denotes the feature vector of product i, the variable $V_{l,t}$ denotes the number of viewers at session t in livestream l, and $\zeta_{l,it}$ represents a random error term.

Given the real-time nature of livestream commerce and the need for quick decision-making, hosts typically favor simple, interpretable models over complex ones. While nonparametric functions might offer greater flexibility in capturing complex relationships, they require substantially more data and computational resources, which may not be practical during live sessions. Moreover, linear functions provide hosts with intuitive interpretations of the impact of various factors on

sales performance. Following this practical consideration, we formalize the host's approach as the following. Heuristically, the host assumes that the function f is linear with unknown parameters $\boldsymbol{\beta}_o \in \mathbb{R}^d, \boldsymbol{\gamma}_o \in \mathbb{R}^4$ and the error term $\zeta_{l,it}$ follows a normal distribution $\mathcal{N}(0, \sigma_\zeta^2)$. We provide the linear function in Equation (5.2), where we use superscript (k) to represent the kth element of a vector.

$$f(\mathbf{x}_i, V_{l,t}, t) = \mathbf{x}_i' \boldsymbol{\beta}_o + \boldsymbol{\gamma}_o^{(1)} + \boldsymbol{\gamma}_o^{(2)} log(t) + \boldsymbol{\gamma}_o^{(3)} log(V_{l,t}). \tag{5.2}$$

Now the host's objective is to maximize the total sales of a livestream, which is

$$\max_{[S_t]_{t=1}^T} \sum_{t=1,\cdots,T} \sum_{i \in S_t} \pi_{l,it}, \tag{5.3}$$

where we use T to denote the total number of sessions in the livestream l, and $[S_t]_{t=1}^T$ to denote the sequence of product sets selected throughout the livestream l. To simplify later analysis, let $\tilde{\boldsymbol{x}}_{l,it} = \begin{bmatrix} \boldsymbol{x}_i' & 1 & log(t) & log(V_{l,t}) \end{bmatrix}'$ and $\boldsymbol{\theta}_o = \begin{bmatrix} \boldsymbol{\beta}_o' & \boldsymbol{\gamma}_o' \end{bmatrix}'$.

To achieve this objective, the host can potentially learn consumer preferences based on sales observations and utilize the learned results to improve later product selection decisions. During this process, the host needs to balance learning consumer preferences while simultaneously maximizing immediate sales. This creates an inherent tension between exploration and exploitation - hosts must decide whether to experiment with new or unfamiliar products to gather information about viewer preferences (exploration) or to rely on their existing knowledge to select products that have historically performed well (exploitation). Too much exploration may lead to missed sales opportunities, while excessive exploitation might prevent the discovery of potentially high-performing products. Given this fundamental trade-off, we model the host's learning and decision-making process through the Linear Multi-Armed Bandit framework. This framework is particularly well-suited for capturing the hosts' heuristic decision-making in product assortment selection, as it explicitly addresses the exploration-exploitation dilemma.

In the subsequent sections, we develop two Linear Multi-Armed Bandit strategies tailored to

the livestream commerce context. For each proposed strategy, we derive rigorous likelihood formulations that facilitate empirical validation and comparative analysis of hosts' decision-making patterns.

Learning in ϵ-Greedy Style

Livestream hosts can employ an ϵ-Greedy style strategy to select products. This approach balances exploitation of known high-performing products with exploration of potentially undiscovered gems through the random probability (parameter) ϵ. A higher ϵ encourages more exploration, which can be beneficial in rapidly changing markets or when introducing new products. A lower ϵ favors exploitation, which can maximize short-term sales but may miss out on identifying new trends or successful products.

For a first-time host with no prior livestream experience, product selection in the initial session can be random since no historical data exists. Thus in our model, we assume a host conducts random decisions in the first session of the first livestream. For experienced hosts, the selection decisions become data-driven: they can leverage sales data from their previous livestream to forecast potential first-session sales for each product in the available pool. We assume that hosts conduct ridge regressions to estimate the linear coefficients $\boldsymbol{\theta}_o$ based on historical observations. Then with probability $1-\epsilon$, the host selects products with positive sales predictions; with probability ϵ, the host selects products randomly.

After the first session of a livestream l, the host can utilize the observation data generated from the previous sessions to conduct the estimation of $\boldsymbol{\theta}_o$ as the following,

$$B_{l,t-1} = I_{d+4} + \sum_{\tau=1}^{t-1} \sum_{i,j \in S_{l,\tau}} \tilde{\boldsymbol{x}}_{l,i\tau} \tilde{\boldsymbol{x}}'_{l,j\tau}, \tag{5.4}$$

$$\hat{\boldsymbol{\theta}}_{l,t-1} = B_{l,t-1}^{-1} \left(\sum_{\tau=1}^{t-1} \sum_{i,j \in S_{\tau}} \tilde{\boldsymbol{x}}_{l,i\tau} \pi_{l,j\tau} \right), \tag{5.5}$$

where I_{d+4} is an identity matrix of dimension $(d+4)$. Then, for each product in the given pool $\mathcal{A}_l$,

host use the estimation $\hat{\boldsymbol{\theta}}_{l,t-1}$ to predict the potential sales in the current session t,

$$\hat{\pi}_{l,it} = \hat{\boldsymbol{\theta}}'_{l,t-1}\tilde{\boldsymbol{x}}_{l,it}. \tag{5.6}$$

Recall that the objective is the sum of sales generated from all products selected, a natural greedy strategy is to choose products that have positive predicted sales. Here, with probability (1 - ϵ), hosts select products with positive sales predictions. This exploits knowledge gained from past performance, focusing on products likely to sell well. With probability ϵ, hosts perform a random product set selection from the entire product pool within the current livestream. This allows for exploration, giving a chance to products that might not have shown promise based on limited past data or changing market conditions. We summarize this ϵ-Greedy style product selection strategy in Algorithm 3.

The value of ϵ can be adjusted over time, with two main approaches considered: (1) Constant ϵ: The value of ϵ remains constant throughout each livestream, maintaining a fixed balance between exploration and exploitation. (2) Diminishing ϵ: The value of ϵ decreases over time within each livestream, gradually shifting from more exploration at the beginning to more exploitation towards the end. Both strategies have their merits: a constant ϵ ensures consistent exploration, while a diminishing ϵ allows for more targeted product selection as more data is gathered during the stream. By implementing this ϵ-Greedy strategy, livestream hosts can effectively balance the need to promote consistently well-performing products while also discovering new potential bestsellers. This approach helps in adapting to changing consumer preferences and market dynamics, ultimately optimizing the overall sales performance of the livestream.

Assuming the hosts conduct ϵ-Greedy style product selection decision, The parameters to be estimated are $\{[\epsilon_l]\}_{l=1}^{5,178}$. We provide the calculation formula for the likelihood of the observed product selection decisions in our dataset in the following and then estimate the parameters $\epsilon_{l,t}$ through Maximum Likelihood Estimation (MLE). First, when a host randomly selects products, two random decisions are involved: (i) Choosing how many products to select (n_p), (ii) Selecting specific products given this chosen number. Thus, for random selection, the probability of selecting

Algorithm 3 ϵ-Greedy Product Selection Strategy

Require: Product pool $\mathcal{A}_l$, exploration rate $\epsilon_{l,t}$
Ensure: Selected product set $S_{l,t}$

Initialize empty product set $S_{l,t} \leftarrow \emptyset$
if $l = 1$ and $t = 1$ **then** ▷ First session of first livestream
    $S_{l,t} \leftarrow$ Random subset of products from $\mathcal{A}_l$

else
    if $t = 1$ **then** ▷ First session of later livestreams
        Compute $\hat{\boldsymbol{\theta}}_{l-1,T_{l-1}}$ using previous livestream data ▷ T_{l-1} is last session of previous livestream
    else ▷ Later sessions within livestream
        Compute $B_{l,t-1} = I_{d+4} + \sum_{\tau=1}^{t-1} \sum_{i,j \in S_{l,\tau}} \tilde{\boldsymbol{x}}_{l,i\tau} \tilde{\boldsymbol{x}}'_{l,j\tau}$
        Compute $\hat{\boldsymbol{\theta}}_{l,t-1} = B_{l,t-1}^{-1} (\sum_{\tau=1}^{t-1} \sum_{i,j \in S_{l,\tau}} \tilde{\boldsymbol{x}}_{l,i\tau} \pi_{l,j\tau})$
    end if
    Predict sales $\hat{\pi}_{l,it} = \hat{\boldsymbol{\theta}}'_{l,t-1} \tilde{\boldsymbol{x}}_{l,it}$ for all $i \in \mathcal{A}_l$
    Generate random number $r \sim \text{Uniform}(0, 1)$
    if $r > \epsilon_{l,t}$ **then** ▷ Exploitation phase
        $S_{l,t} \leftarrow \{i \in \mathcal{A}_l : \hat{\pi}_{l,it} > 0\}$ ▷ Select products with positive predicted sales
    else ▷ Exploration phase
        $S_{l,t} \leftarrow$ Random subset of products from $\mathcal{A}_l$
    end if
end if
return $S_{l,t}$

any single product can be calculated as the following,

$$\begin{aligned}\text{Prob(a product is selected)} &= \sum_{n_p=1}^{|\mathcal{A}_l|} \text{Prob(the number of products is chosen to be } n_p)\times \\ &\quad \text{Prob(a product is selected}|\text{the number of products is chosen to be } n_p) \\ &= \sum_{n_p=1}^{|\mathcal{A}_l|} \frac{1}{|\mathcal{A}_l|} \times \frac{n_p}{|\mathcal{A}_l|} \\ &= \frac{1+|\mathcal{A}_l|}{2|\mathcal{A}_l|}. \end{aligned} \tag{5.7}$$

Here, $\frac{1}{|\mathcal{A}_l|}$ represents the uniform probability of choosing n_p products, and $\frac{n_p}{|\mathcal{A}_l|}$ is the probability of selecting a specific product given n_p products are chosen. Hence, the complete likelihood of the ϵ-Greedy style strategy for a livestream l is

$$\begin{aligned}\prod_{t=1}^{T} L(S_{l,t}|\epsilon_l, \text{Data}) = \prod_{t=1}^{T} &\left(\prod_{i_{in}\in S_{l,t}} (1-\epsilon_l)\mathbf{1}\{\tilde{\boldsymbol{x}}'_{i_{in}}\hat{\boldsymbol{\theta}}_{l,t-1} > 0\} + \epsilon_l \frac{1+|\mathcal{A}_l|}{2|\mathcal{A}_l|} \right) \times \\ &\left(\prod_{i_{out}\in S_{l,t}} (1-\epsilon_l)\mathbf{1}\{\tilde{\boldsymbol{x}}'_{i_{out}}\hat{\boldsymbol{\theta}}_{l,t-1} \leq 0\} + \epsilon_l(1-\frac{1+|\mathcal{A}_l|}{2|\mathcal{A}_l|}) \right). \end{aligned} \tag{5.8}$$

This likelihood function accounts for (i) Selected products ($i_{in} \in S_{l,t}$). They could be chosen either through exploitation ($(1-\epsilon_l)$ when predicted sales are positive) or exploration (ϵ_l times the random selection probability). (ii) Non-selected products ($i_{out} \notin S_{l,t}$). They could be excluded either through exploitation ($(1-\epsilon_l)$ when predicted sales are negative) or not chosen during random exploration. By maximizing this likelihood function, we can estimate the exploration rate ϵ_l that best explains the observed product selection decisions in our dataset.

Learning in Thompson Sampling Style

Livestream hosts can also employ a Thompson Sampling style strategy to select products. This approach strikes a balance between featuring proven bestsellers and testing promising new items

by creating a probability distribution based on the host's estimates of consumer preference $\boldsymbol{\theta}_o$, with the variance parameter σ reflecting their exploration degree.

The estimation process is the same as the process in the ϵ-Greedy style decisions in Section 5.2.2. The host begins with a random product selection decision in the first session in the first livestream. After the first livestrem, the host utilizes the estimation of $\boldsymbol{\theta}_o$ from the last livestream to predict sales in the first session. After the first session, the host utilizes the observed sales within the current livestream to conduct ridge regression estimation of $\boldsymbol{\theta}_o$, which is denoted by $\hat{\boldsymbol{\theta}}_{l,t-1}$.

Instead of directly using the estimation $\hat{\boldsymbol{\theta}}_{l,t-1}$ to predict the sales, the host construct a normal distribution around $\hat{\boldsymbol{\theta}}_{l,t-1}$ and randomly draw a $\tilde{\boldsymbol{\theta}}$ from this normal distribution to make sales prediction in the Thompson Sampling style strategy. Specifically, the host holds a prior for θ_o at the beginning of the time period t in livestream l as given by $\mathcal{N}(\hat{\boldsymbol{\theta}}_{l,t-1}, \sigma_l B_{l,t-1}^{-1})$. Upon observing the sales at the end of time period t, the host updates the posterior distribution of the parameters as,

$$\begin{aligned}
\text{Prob}(\tilde{\boldsymbol{\theta}}|\pi_{l,t}) &\propto \text{Prob}(\pi_{l,t}|\tilde{\boldsymbol{\theta}})\text{Prob}(\tilde{\boldsymbol{\theta}}) \\
&\propto \exp\left(-\frac{1}{2\sigma_l^2}\left((\pi_{l,t} - \sum_{i\in S_{l,t}} \tilde{\boldsymbol{\theta}}'\tilde{\boldsymbol{x}}_{l,it})^2 + (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_t)' B_{l,t-1}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_t)\right)\right) \\
&\propto \mathcal{N}(\hat{\boldsymbol{\theta}}_t, \sigma_l^2 B_{l,t}^{-1}).
\end{aligned} \tag{5.9}$$

Note that the matrix $B_{l,t}$ is always positive definite and increases with t, which leads to a weaker effect of σ_l on the distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}_{l,t}, \sigma_l^2 B_{l,t}^{-1})$ as t increases. Subsequently, the host of livestream l samples a $\boldsymbol{\theta}$ from the posterior distribution $\mathcal{N}(\hat{\theta}_{l,t}, \sigma_l^2 B_{l,t}^{-1})$ and selects the products with positive sales predictions based on $\boldsymbol{\theta}$. We summarize this strategy in Algorithm 4.

The fundamental logic behind these selections is based on predicted sales. Hosts make their choices by drawing random guesses around their best estimates, with the amount of randomness controlled by their exploration parameter σ_l. A host who explores more will have more variation in their guesses, leading to more diverse product selections. This randomness in their predictions creates uncertainty, which is essential for balancing exploration of new products with exploitation of known successful ones.

The exploration dynamic evolves naturally over time under this Thompson Sampling style strategy. As $B_{l,t}$ grows with more observations, its inverse shrinks, reducing the variance of the posterior distribution. This creates an automatic transition from exploration to exploitation. In early sessions, the higher variance encourages exploration; while in later sessions, reduced variance leads to more focused exploitation. The parameter σ_l acts as a tuning knob for this exploration-exploitation trade-off.

Algorithm 4 Thompson Sampling Product Selection Strategy

Require: Product pool $\mathcal{A}_l$, exploration parameter σ_l
Ensure: Selected product set $S_{l,t}$

Initialize empty product set $S_{l,t} \leftarrow \emptyset$
if $l = 1$ **and** $t = 1$ **then** ▷ First session of first livestream
    $S_{l,t} \leftarrow$ Random subset of products from $\mathcal{A}_l$
else
    if $t = 1$ **then** ▷ First session of later livestreams
        Use $\hat{\boldsymbol{\theta}}_{l-1,T}$ from previous livestream ▷ T is last session
        $B_{l,0} \leftarrow B_{l-1,T}$ ▷ Initialize precision matrix
    else ▷ Later sessions within livestream
        Compute $B_{l,t-1} = I_{d+4} + \sum_{\tau=1}^{t-1} \sum_{i,j \in S_{l,\tau}} \tilde{\boldsymbol{x}}_{l,i\tau} \tilde{\boldsymbol{x}}'_{l,j\tau}$
        Compute $\hat{\boldsymbol{\theta}}_{l,t-1}$ via ridge regression
    end if
    Draw $\boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_{l,t-1}, \sigma_l^2 B_{l,t-1}^{-1})$
    Predict sales $\hat{\pi}_{l,it} = \boldsymbol{\theta}' \tilde{\boldsymbol{x}}_{l,it}$ for all $i \in \mathcal{A}_l$
    $S_{l,t} \leftarrow \{i \in \mathcal{A}_l : \hat{\pi}_{l,it} > 0\}$
end if
return $S_{l,t}$

The parameters to be estimated in the Thompson Sampling style strategy are $\{[\sigma_l]\}_{l=1}^{5,178}$. To estimate these parameters, note that a host always chooses products that achieve $\tilde{\boldsymbol{x}}'_{l,it+1}\boldsymbol{\theta}_{l,t} > 0$ and $\boldsymbol{\theta}_{l,t} \in \mathcal{N}(\hat{\boldsymbol{\theta}}_{l,t}, \sigma_l^2 B_{l,t}^{-1})$ so that the host can maximize the total predicted sales $\hat{\pi}_{l,t+1}$. When calculating likelihoods, we consider two distinct scenarios. For products that were selected, we calculate how likely it was for the host to predict positive sales. The higher their confidence in positive predictions, the more likely they were to select those products. However, more exploration makes these predictions more uncertain, potentially leading to selections that might seem counterintuitive

based on past performance alone. Conversely, for products that weren't selected, we calculate how likely it was for the host to predict negative sales, with similar principles applying to the role of uncertainty in these decisions.

Hence, the likelihood of observed product assortment decisions over time within a livestream l by a host is given in Equation (5.10).

$$\begin{aligned}
\prod_{t=1}^{T} L(S_{l,t}|\sigma_l, \text{Data}) &= \prod_{t=1}^{T} \prod_{i_{in}\in S_{l,t}} \text{Prob}(i_{in} \in S_{l,t}|i_{in} \in \mathcal{A}_l) \prod_{i_{out}\in \mathcal{A}_l/S_{l,t}} \text{Prob}(i_{out} \notin S_{l,t}|i_{out} \in \mathcal{A}_l) \\
&= \prod_{t=1}^{T} \prod_{i_{in}\in S_{l,t}} \text{Prob}(\tilde{\boldsymbol{x}}'_{i_{in}}\boldsymbol{\theta}_{l,t-1} > 0|i_{in} \in \mathcal{A}_l) \times \prod_{i_{out}\in \mathcal{A}_l/S_{l,t}} \text{Prob}(\tilde{\boldsymbol{x}}'_{i_{out}}\boldsymbol{\theta}_{l,t-1} \leq 0|i_{out} \in \mathcal{A}_l) \\
&= \prod_{t=1}^{T} \prod_{i_{in}\in S_{l,t}} (1 - \Phi(-\frac{\tilde{\boldsymbol{x}}'_{i_{in}}\hat{\boldsymbol{\theta}}_{l,t-1}}{\tilde{\boldsymbol{x}}'_{i_{in}}\sigma_l B_{l,t}^{-1}\tilde{\boldsymbol{x}}_{i_{in}}})) \times \prod_{i_{out}\in S_{l,t}} \Phi(-\frac{\tilde{\boldsymbol{x}}'_{i_{out}}\hat{\boldsymbol{\theta}}_{l,t-1}}{\tilde{\boldsymbol{x}}'_{i_{out}}\sigma_l B_{l,t}^{-1}\tilde{\boldsymbol{x}}_{i_{out}}}).
\end{aligned} \tag{5.10}$$

By finding the exploration parameter σ_l that maximizes this likelihood, we can estimate each host's natural tendency to balance exploring new options versus exploiting known successes. This statistical approach provides insights into how hosts learn from past performance and adapt their strategies over time, helping us understand their decision-making process in the dynamic environment of livestream commerce.

Exploration and Exploitation: Both the constructions of $\boldsymbol{\theta}$ in the Thompson Sampling strategy and the ϵ_l-greedy mechanism capture how hosts balance the exploration and exploitation during a livestream. The host explores by sampling $\boldsymbol{\theta}$ from the posterior distribution, which allows the host to consider a range of plausible parameter values and gather information about the choices' potential rewards (sales). The uncertainty $\sigma_l^2 B_{l,t}^{-1}$ in the posterior distribution encourages exploration, as the host may sample parameters that lead to selecting less explored choices. With probability ϵ_l, the host explores by randomly selecting a product assortment, regardless of predicted sales. This also allows the host to gather information about all choices, even those that may not have been selected based on current reward estimates. Thus, larger σ_l and ϵ_l imply a higher uncertainty of the product selection decisions allowing the host to learn more about unexplored product choices'

sales distribution.

Benchmark Models

To better understand host product selection behaviors reflected in the data, we propose three other alternative models of product selection: Random selection decisions, Fixed Group selection decisions, and Greedy selection decisions. Through a comparative analysis using likelihood-based metrics, we aim to determine which model best fits the observed host behaviors based on the data.

Random Decisions In the random product selection model, hosts do not rely on past performance data or predictions. Instead, the number of products to present in each session is randomly determined, and products are uniformly randomly selected from the available pool within each livestream. This approach provides maximum exploration, potentially discovering unexpected trends or popular products. However, it ignores valuable historical data and may lead to inconsistent performance.

Fixed Group Decisions Another benchmark product selection strategy the livestream hosts can potentially adopt is a mixture of random decisions and fixed sequence of decisions. Hosts can potentially present lower-priced and branded products to attract consumers at the beginning of a livestream and then turn to higher-priced and non-branded products to gain profit. To capture such a strategy, we model that a livestream host sort the products in the given pool by price and branded variables at the very beginning of each livestream. Then the host divided the whole pool of products into two groups based on the sorting: first group with lower-priced and branded products, second group with higher-priced and non-branded products. During the first half of the livestream, the host always randomly draw product sets from the first group; during the second half of the livestream, the host always randomly draw product set from the second group.

Greedy Decisions The greedy selection model represents the opposite extreme of random selection. In this approach, hosts strictly adhere to estimations based on past observations of sales and $\tilde{x}$, selecting only products with positive sales predictions. The process involves estimating linear coefficients θ using historical (sales, $\tilde{x}$) data, predicting sales for all available products, and then selecting only products where positive sales predictions. This method maximizes short-term per-

formance based on historical data but lacks exploration, potentially missing new trends or changes in consumer behavior. The greedy model assumes hosts are purely profit-driven and have high confidence in their predictive models.

The estimation results are presented in Table 5.2, which compares the log likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) for each model.

Table 5.2: Likelihood Comparison of Models without Host Features Embedded

	Without Learning		With Learning			
	Random	Fixed Group	Greedy	TS	ϵ-Greedy(2)	ϵ-Greedy(1)
Log Likelihood	-3,957,941	-103,624,589	-79,671,028	-3,417,737	-11,696,789	-1,227,680
AIC	7,915,882	207,249,178	159,342,056	6,845,830	23,403,934	2,465,716
BIC	7,915,882	207,249,178	159,342,056	6,899,599	23,457,703	2,519,485

Notes. For Random and Greedy models, we do not estimate parameters so AIC is equal to BIC.
ϵ-Greedy(2) refers to the ϵ-Greedy model with a diminishing ϵ over time within each livestream.
ϵ-Greedy(1) refers to the ϵ-Greedy model with a constant ϵ over time within each livestream.

The ϵ-Greedy (2) model, which maintains a constant ϵ within each livestream, demonstrates the best fit to the observed data. It exhibits the highest log likelihood (-1,227,680) and the lowest AIC (2,465,716) and BIC (2,519,485) values among all models. This finding suggests that hosts tend to maintain a consistent balance between exploration and exploitation throughout their livestreams, rather than adjusting their strategy as the stream progresses. The Thompson Sampling (TS) model emerges as the second-best performer, with the next highest log likelihood and next lowest AIC and BIC values. This indicates that some hosts may be employing a more sophisticated probabilistic approach to balance exploration and exploitation, though not to the same extent as the ϵ-Greedy strategy.

Comparing the two ϵ-Greedy models, we observe that the constant ϵ model (ϵ-Greedy (2)) significantly outperforms the diminishing ϵ model (ϵ-Greedy (1)) in fitting the data. This result further supports the notion that hosts do not tend to reduce their exploration rate as the livestream progresses, instead maintaining a consistent approach throughout. Both the Random and Greedy models perform poorly compared to the other strategies, with the Greedy model showing the worst fit. This indicates that hosts neither rely on purely random selection nor depend solely on past per-

formance when choosing products to showcase. The particularly poor performance of the Greedy model underscores the importance of ongoing exploration in the dynamic livestream environment.

It is noteworthy that the models incorporating learning (ϵ-Greedy and Thompson Sampling) generally outperform the non-learning models (Random and Greedy). This supports the hypothesis that hosts dynamically adjust their strategies based on ongoing performance during the livestream, rather than adhering to a fixed approach.

5.2.3 Step 2: Presentation Timing

Now we model how livestream hosts decide the presentation schedule using a survival analysis. The survival analysis estimates the instantaneous probability (hazard rate) of a livestream host stopping the presentation of the current product set and turning to another set of products (termed as an event) conditional on the observable features (including product set features, current sales performance, current livestream engagement level, etc.) given that the host presented the current product set so far.

Specifically, we use the stratified Cox Proportional Hazards model (Cox, 1972) through the survival package in R (Therneau et al., 2015) for the estimation. The stratified Cox Proportional Hazards model allows the hosts' scheduling decisions dependent on the dynamic of the livestream through different baseline hazards over time within a livestream. In addition, the model uses non-parametric estimation for the baseline hazard function, making it more flexible and less restrictive compared to parametric survival models. The hazard rate is formulated in Equation (5.11).

$$\begin{aligned} h_t(D_{l,t}) =& h_{0k}(D_{l,t}) \times \\ & \exp\left(\boldsymbol{z}'_{l,S_{l,t}}\boldsymbol{\beta}_c + \gamma_c^{(1)}(N_l - \sum_{\tau=1}^{t} S_{l,\tau}) + \gamma_c^{(2)} V_{l,t} + \gamma_c^{(3)} IsWeekend_l + \gamma_c^{(4)} IsDaytime_l\right), \end{aligned} \tag{5.11}$$

where $h_t(D_{l,t})$ is the hazard rate for the tth time period, i.e., the instantaneous probability of stopping the current presentation at $D_{l,t}$th second in the tth time period of a livestream. The hazard

rate is the product of a baseline hazard rate $h_{0k}(D_{l,t})$ and an exponential function of a linear combination of covariates. The subscript k represents a host. The baseline hazard rate varies with different hosts and is estimated non-parametrically. The linear component includes features of the current product set $S_{l,t}$, denoted by $\boldsymbol{z}_{S_{l,t}}$, as well as the number of products which have not been selected to present until the current time period, denoted by $N_l - \sum_{\tau=1}^{t} S_{l,\tau}$ (*RemainingProducts*), the number of viewers V_t. The variable $\boldsymbol{z}_{S_{l,t}}$ includes number of products in $S_{l,t}$ (*SetSize*), total revenue achieved for $S_{l,t}$ (*SetRevenue*), the average price of products in $S_{l,t}$ (*SetPrice*), total sales achieved for $S_{l,t}$ (*SetSales*), the probability that a product is branded in $S_{l,t}$ (*SetIsBrand*, i.e., number of branded products divided by the total number of products in $S_{l,t}$), number of distinct product categories in $S_{l,t}$ (*SetCategoryDiversity*). Additionally, we use variables *IsWeekend* and *IsDaytime* to control for the effects of the day of a week and the time of a day on the hazard rate.

Estimation of Timing Strategy

We present the results of a stratified Cox proportional hazards ratio model in Table 5.3. The model identifies key factors that influence a host's presentation-stopping decision during livestream sales. The product set characteristics (size, revenue, sales, branding, category diversity), viewership, remaining products, and timing of the livestream all play significant roles in shaping the instantaneous probability of stopping the current presentation.

Specifically, a one-unit increase in the logged set size multiplies the hazard rate by exp(0.9349) = 2.55, holding other variables constant. In other words, larger product sets are associated with a substantially higher instantaneous probability of the host stopping the current presentation and switching to a new product set. Similarly, branded product sets have exp(0.3295) = 1.39 times the hazard rate of non-branded sets, all else being equal. Higher product set sales and a greater number of remaining products also increase the hazard rate, suggesting that hosts are more likely to stop current presentation when sales are high or when there are many products left to present. On the other hand, factors such as higher set revenues, more diverse set categories, a larger number of viewers, and daytime livestreams are associated with a lower hazard rate. For instance, a one-unit increase in the logged number of viewers multiplies the hazard rate by exp(-0.0996) = 0.91,

indicating a 9% decrease in the instantaneous probability of stopping the current presentation for each unit increase in the logged viewer count. In addition, the model's concordance of 0.708 indicates a good predictive power, suggesting that the model can accurately discriminate between high-probability and low-probability cases. The likelihood ratio test with 10 degrees of freedom is highly significant ($p < 0.01$), indicating that the model as a whole is statistically significant.

5.3 How Host Features Affect Exploration

To gain deeper insights into the factors influencing product selection strategies, we extended our ϵ-Greedy model to incorporate host features. This extension allows us to examine how various characteristics of the livestream hosts affect their degree of exploration versus exploitation. We model the exploration parameter ϵ_l as a function of host features, using a logistic transformation to ensure ϵ_l remains between 0 and 1.

$$\epsilon_l = \frac{1}{1 + \exp\left(\boldsymbol{\alpha}_\epsilon \boldsymbol{z_l}\right)}, \tag{5.12}$$

where $\boldsymbol{z_l}$ denotes the vector of host features for livestream l. These features include host gender, number of videos posted, number of likes received, number of followers, number of accounts followed, education level, and livestreaming experience.

We formulate the likelihood function for the observed product assortments in Equation (5.13), capturing the probability of selecting each product in the assortment $S_{l,t}$ for livestream l at time t. The likelihood incorporates both the exploitation component, governed by the estimated utility $\tilde{\boldsymbol{x}}_i' \boldsymbol{\theta}_{l,t-1}$, and the exploration component, represented by the uniform probability $1/2^{|\mathcal{A}_l|}$. The ϵ_l parameter, modeled as a function of host features $\boldsymbol{z_l}$, determines the balance between these two components. This formulation allows us to estimate the impact of host characteristics on the exploration-exploitation trade-off while accounting for the evolving livestream environment over

Table 5.3: Cox Proportional Hazards Ratio Model Estimation Results

Dependent Var.: Hazard Rate	Coef.
Log(*SetSize*)	0.9349 ***
	(0.0234)
Log(*SetRevenue*)	-0.0304 ***
	(0.0033)
Log(*SetPrice*)	0.0096
	(0.0067)
Log(*SetSales*)	0.1769 ***
	(0.0079)
SetIsBrand	0.3295 ***
	(0.0294)
SetCategoryDiversity	-0.0477 ***
	(0.0145)
Log(*Viewers*)	-0.0996 ***
	(0.0091)
Log(*RemainingProducts*)	0.0512 ***
	(0.0083)
IsWeekend	-0.0513
	(0.0419)
IsDaytime	-0.0805 **
	(0.0329)
Observations	226,786
Concordance	0.708
	(0.003)
Likelihood Ratio Test (10 Degrees of Freedom)	98798***

Notes. Standard errors are presented in parentheses.
*p<0.1, **p<0.05, ***p<0.01

time.

$$\prod_{t=1}^{T} L(S_{l,t}|\boldsymbol{\alpha}_\epsilon, \text{Data}) = \prod_{t=1}^{T} \left(\prod_{i_{in} \in S_{l,t}} (1 - \frac{1}{1+\exp(\boldsymbol{\alpha}_\epsilon \boldsymbol{z_l})}) \mathbf{1}\{\tilde{\boldsymbol{x}}'_{i_{in}} \boldsymbol{\theta}_{l,t-1} > 0\} + \frac{1}{1+\exp(\boldsymbol{\alpha}_\epsilon \boldsymbol{z_l})} \frac{1+|\mathcal{A}_l|}{2|\mathcal{A}_l|} \right) \times \left(\prod_{i_{out} \in S_{l,t}} (1 - \frac{1}{1+\exp(\boldsymbol{\alpha}_\epsilon \boldsymbol{z_l})}) \mathbf{1}\{\tilde{\boldsymbol{x}}'_{i_{out}} \boldsymbol{\theta}_{l,t-1} \leq 0\} + \frac{1}{1+\exp(\boldsymbol{\alpha}_\epsilon \boldsymbol{z_l})} (1 - \frac{1+|\mathcal{A}_l|}{2|\mathcal{A}_l|}) \right). \tag{5.13}$$

Table 5.4 presents the maximum likelihood estimation results for the ϵ-Greedy model incorporating host features. The estimation results provide several interesting insights into how host characteristics influence the balance between exploration and exploitation in product selection. The negative coefficient (-0.3327) for gender suggests that female hosts tend to have a lower ϵ, indicating a higher propensity for exploitation compared to male hosts. Hosts who post more videos tend to have a lower ϵ, suggesting that more active content creators rely more on exploitation strategies. Regarding popularity metrics, hosts with more likes and followers are more inclined towards exploration. Interestingly, the number of accounts a host follows has the strongest positive impact on exploration, suggesting that hosts who are more engaged with the platform community tend to be more exploratory in their product selections. Hosts with higher education show a slightly lower tendency for exploration, although the effect is relatively small. The positive coefficient for live experience indicates that more experienced hosts tend to engage in more exploration, possibly due to increased confidence in their ability to handle diverse products.

These findings have several implications for understanding host behavior and platform dynamics. More experienced hosts and those with larger networks (followers and following) tend to be more exploratory. This could be due to their confidence in handling diverse products and their ability to leverage their network for successful promotion of new items. The negative relationship between video posting activity and exploration suggests that hosts who are more focused on content creation might prefer a more stable, exploitation-focused product selection strategy. The gender effect indicates potential differences in risk-taking or product knowledge between male

and female hosts, which could be further explored in future research. The slight negative effect of higher education on exploration could indicate that more educated hosts might rely more on analytical approaches to product selection, favoring exploitation of known successful products.

Table 5.4: Effect of Host Features on ϵ: MLE Estimation Results

	α_ϵ
Gender	-0.3327 ***
	(0.0036)
Log(*VideoPosted*)	-0.2182 ***
	(0.0016)
Log(*Favorited*)	0.0854 ***
	(0.0016)
Log(*Follower*)	0.1182 ***
	(0.0028)
Log(*Following*)	2.2338 ***
	(0.0168)
HighEducation	-0.0192 **
	(0.0092)
Log(*LiveExperience*)	0.8820 ***
	(0.0009)
Observations	238,985
Log Likelihood	-1,400,481
AIC	2,800,976
BIC	2,801,049

Notes. Standard errors are presented in parentheses.
*p<0.1, **p<0.05, ***p<0.01

5.4 Simulations

Our prior analysis demonstrates how livestream hosts learn and explore to make product assortment decisions, how host features affect their exploration behaviors, and how host features affect their timing behaviors. While our model structure captures the indirect effect of host features on sales, the precise magnitude of this effect is not directly quantifiable through estimated parameters alone.

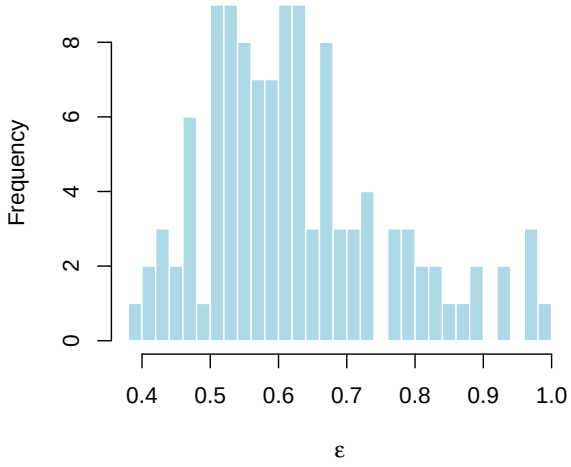


Figure 5.3: Distribution of Exploration Degree ϵ among Hosts

To address this limitation and provide a more comprehensive understanding, we employ simulations in this section. These simulations explicitly demonstrate how host characteristics ultimately influence sales performance through their impact on learning, exploration, product assortment and presentation timing.

We use our parameter estimates α_ϵ, $\boldsymbol{\beta}_c$ and $\boldsymbol{\gamma}_c$ to simulate livestream hosts' product assortment and timing decisions, and sales generated from the viewers. To begin with, we assume that the viewer arrivals in each livestream follow Poisson distributions, where the intensity of the Poisson arrivals can vary with hosts (see Equation (5.14)).

$$\text{Prob}(V_{l,t} = v) = \frac{\lambda_k^v \exp(-\lambda_k)}{v!}. \tag{5.14}$$

We estimate the Poisson arrival rate for each host from the data by Maximum Likelihood Estimation. Specifically, viewers can arrive as well leave. Thus, to estimate the arrival rate, we first fix a departure rate at 0.01 and assume the number of viewers can decrease ($0.01\times$ *Time Duration*)

if there are no new arrivals. Further, we simulate sales as a linear function of product features and livestream environment features as given in Equation (5.2). We estimate the linear coefficients from the original data.

Given the products of each livestream in the original dataset, we simulate that each host conducts product assortment decisions following the ϵ-Greedy framework with the estimate of α_ϵ. Then, based on the product set selected, the host decides the presentation timing following a Cox proportional hazards framework with the estimates of $\boldsymbol{\beta}_c$ and $\boldsymbol{\gamma}_c$. To simulate the presentation timing, we use the estimated cumulative baseline hazard function $H_0(t)$ and calculate the inverse function $H_0^{-1}(t)$ by linear interpolation. Specifically, by Leemis (1987) and Bender et al. (2005), we know that the presentation timing can be generated through (5.15),

$$H_0^{-1}\left(-\log(u)\exp\left(\boldsymbol{z}'_{l,S_{l,t}}\boldsymbol{\beta}_c+\gamma_c^{(1)}(N_l-\sum_{\tau=1}^{t}S_{l,\tau})+\gamma_c^{(2)}V_{l,t}+\gamma_c^{(3)}IsWeekend_l+\gamma_c^{(4)}IsDaytime_l\right)\right), \tag{5.15}$$

where u is randomly drawn from a uniform distribution ranging from $[0,1]$. Since we assume that the hazard function is nonparametric, we can only get step functions for $H_0(t)$ from the estimation. Thus, we use linear interpolation technique to estimate $H_0^{-1}(t)$. Based on this set up, we conduct simulations to see how to improve sales heuristically in the following section.

5.4.1 Improving Sales Heuristically

In this section, we apply our model estimation results from Section 5.3 to demonstrate directive heuristic policy-making methods to improve livestream sales. Our analysis compares the performance of several potential product selection strategies, leveraging the insights gained from our earlier modeling efforts. Specifically, we evaluate five strategies: the original estimated ϵ-Greedy approach (denoted as ϵ-Greedy(1)), a strictly Greedy strategy, an ϵ-Greedy strategy with diminishing ϵ over time within each livestream (denoted as ϵ-Greedy(2)), an ϵ-Greedy strategy with constant ϵ over time within each livestream (denoted as ϵ-Greedy(3)) where ϵ is lower than original estimated from the data, and the Thompson Sampling (TS) strategy.

Table 5.5 presents the mean sales and standard deviation for each strategy. The Thompson Sampling strategy achieves the highest mean sales at 1757.8, followed closely by ϵ-Greedy(3) at 1751.2 and ϵ-Greedy(2) at 1735.0. All three strategies substantially outperform both the strictly Greedy approach (1674.0) and the original strategy (1482.6). The improvement rates range from 12.91% for the Greedy strategy to 18.56% for Thompson Sampling, with all improvements being statistically significant (p-values < 0.01).

To better understand the effectiveness of these strategies across different host characteristics, we conducted a heterogeneity analysis presented in Table 5.6. The results reveal several distinctive patterns in strategy effectiveness across different host segments. Our analysis of host activity levels shows that while all strategies demonstrate significant improvements for both more and less active hosts, less active hosts particularly benefit from Thompson Sampling, achieving a 23.56% increase compared to a 12.36% increase for more active hosts.

The impact of these strategies varies dramatically based on the host's follower count. Hosts with more followers experience substantial improvements ranging from 25% to 33% across different strategies, while those with fewer followers see performance decreases between 17% and 19%. This stark contrast suggests that these optimization strategies are most effective when applied to hosts with established audiences.

Examining social network engagement reveals that hosts with more followings consistently show stronger improvements, ranging from 16% to 25% across strategies, compared to those with fewer followings, where the improvements are either minimal or negative. This pattern indicates that hosts who are more actively engaged in following others on the platform may be better positioned to benefit from these optimization strategies.

The analysis of education levels yields an unexpected finding: hosts with lower education levels demonstrate remarkably higher improvements, ranging from 81% to 91% across strategies, compared to those with higher education levels, where the improvements are not statistically significant. This suggests that these optimization strategies might serve as particularly valuable tools for hosts with less formal education, potentially helping to level the playing field in livestream commerce.

In terms of platform engagement, hosts with fewer favorites received show stronger improvements (17-26% across strategies) compared to those with more favorites. This pattern suggests that these strategies might be particularly beneficial for hosts who are still building their reputation on the platform, offering them tools to improve their performance despite having less established content popularity.

These findings illustrate that while all proposed strategies can significantly improve upon the original approach, their effectiveness varies substantially across different host segments. The Thompson Sampling strategy consistently shows strong performance across different host characteristics, while the ϵ-Greedy variants demonstrate particular strength in specific contexts. This heterogeneity in strategy effectiveness highlights the importance of considering host characteristics when implementing optimization strategies in livestream commerce platforms, suggesting that a more nuanced, targeted approach to strategy deployment might yield the best results.

Table 5.5: Simulated Sales Comparison over Different Strategies

Sales	Original	Greedy	ϵ-Greedy(2)	ϵ-Greedy(3)	TS
Mean	1482.627	1674.004	1735.031	1751.247	1757.842
Std.	9913.010	9193.994	9936.557	11347.735	11165.471
Mean Increase Rate	-	12.91%	17.02%	18.12%	18.56%
p-value	-	0.0057	0.0013	0.0094	0.0050

Notes. All p-values are based on paired t-tests that compare each strategy to the Original. ϵ-Greedy(2) refers to the ϵ-Greedy model with a diminishing ϵ over time within each livestream. ϵ-Greedy(3) refers to the ϵ-Greedy model with a constant ϵ over time within each livestream where ϵ is lower than the original strategy estimated.

5.5 Discussion

Livestream sales have become increasingly popular in recent years. Hosts, aiming to maximize total sales throughout livestreams, face crucial decisions regarding product assortment and presentation timing. Despite the importance of these decisions, few studies have explored this area. Our work contributes to the literature by investigating how livestream hosts make product assortment and presentation timing choices using real-world data and structural modeling techniques.

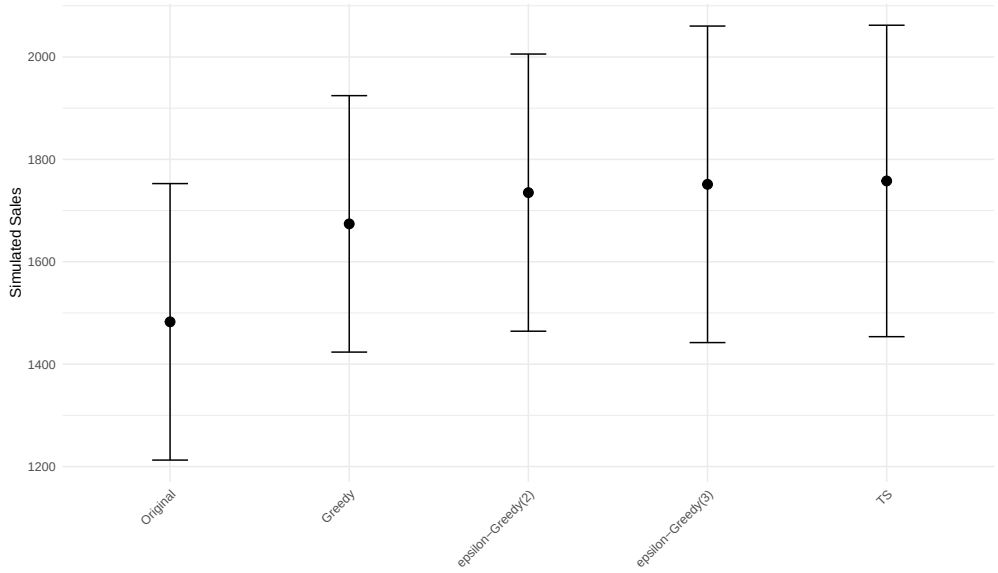


Figure 5.4: Comparison of 95% Confidence Intervals of Simulated Sales

Table 5.6: Host Heterogeneity in Simulated Sales Improvement over Different Strategies

Mean Increase Rate	Greedy	ϵ-Greedy(2)	ϵ-Greedy(3)	TS
Less Active	13.49%***	17.22%***	18.87%***	23.56%***
More Active	12.18%***	16.78%***	17.19%***	12.36%***
Less Followers	-17.50%***	-18.81%***	-19.16%***	-18.58%***
More Followers	25.03%***	31.31%***	32.98%***	33.37%***
Less Followings	2.36%	-5.75%*	-3.50%	2.76%
More Followings	16.37%***	24.50%***	25.21%***	23.75%***
High Educated	-8.00%	-0.44%	1.26%	0.46%
Low Educated	91.57%***	82.70%***	81.52%***	86.65%***
Less Favorited	-0.17%	-7.44%	-5.05%	-1.36%
More Favorited	17.50%***	25.61%***	26.24%***	25.55%***

Notes. *p<0.1, **p<0.05, ***p<0.01. ϵ-Greedy(2) refers to the ϵ-Greedy model with a diminishing ϵ over time within each livestream. ϵ-Greedy(3) refers to the ϵ-Greedy model with a constant ϵ over time within each livestream where ϵ is lower than the original strategy estimated.

To study product assortment decisions, we focus on how hosts learn about viewer demand over time and exploit this knowledge to optimize their choices. We fit the livestream sales data into Multi-Armed Bandit (MAB) models, which allows us to examine the learning and exploitation process. Additionally, we identify how hosts' characteristics influence their exploration and exploitation behaviors in product assortment decisions. The maximum likelihood estimation results reveal that a host's exploration behavior, as measured by the parameters ϵ_l, is significantly influenced by their activeness and social network on the platform, past livestream experience, and personal demographic characteristics. This finding highlights the dynamic nature of hosts' decision-making processes and the importance of considering both historical and contextual factors.

Furthermore, we investigate product presentation timing decisions using a stratified Cox proportional hazards model. This model allows us to examine the factors that influence a host's decision to stop the current product presentation and move on to the next one. Our analysis reveals that product set characteristics, viewership, remaining products, and the timing of the livestream play significant roles in shaping the instantaneous probability of stopping the current presentation. Specifically, larger product set sizes, branded products, higher sales, and a greater number of remaining products are more likely to lead to a presentation switch. Conversely, higher revenues, category diversity, viewership, and daytime livestreams are associated with a lower probability of stopping the current presentation. These findings provide valuable insights into the complex interplay of factors that drive hosts' presentation timing decisions.

Building on these empirical findings, we conducted simulation experiments to evaluate various product selection strategies and their potential for improving sales performance. Our results demonstrate that alternative strategies, particularly Thompson Sampling and modified ϵ-Greedy approaches, can significantly outperform the current practices, with improvements ranging from 12.91% to 18.56% in overall sales. However, the effectiveness of these strategies varies substantially across different host segments. Hosts with larger follower bases show remarkable improvements of 25-33%, while those with fewer followers may experience decreased performance. Similarly, hosts with lower education levels demonstrate particularly strong improvements of 81-91%, suggesting these strategies could serve as valuable tools for democratizing success on the plat-

form. The heterogeneous effects extend to other host characteristics, with less active hosts, those with more followings, and those with fewer favorites generally showing stronger improvements from these optimization strategies.

Our work offers several important implications for livestream sales practitioners and platform providers. By understanding the factors that influence hosts' product assortment and presentation timing decisions, platforms can develop targeted tools and features to support different host segments in making informed choices. For instance, platforms could provide customized real-time analytics and recommendations based on viewer engagement, sales performance, and product set characteristics, with particular attention to supporting hosts who might benefit most from optimization strategies, such as those with fewer followers or less platform experience.

Moreover, our research lays the groundwork for future AI and ML-assisted livestream sales. The insights gained from our empirical analysis and simulation experiments can inform the development of intelligent systems that support hosts in real-time decision-making. These systems could be tailored to different host segments, recognizing that the effectiveness of optimization strategies varies significantly across host characteristics. AI or ML models can be trained on historical data to predict viewer demand, recommend optimal product assortments, and suggest presentation timing based on real-time livestream performance, with customizations for different host profiles. Such targeted support could significantly enhance hosts' abilities to adapt to dynamic market conditions and make data-driven decisions, ultimately improving the overall effectiveness of livestream sales while helping to level the playing field across different host segments.

Chapter 6

CONCLUDING REMARKS

As digital platforms continue to evolve and AI systems become more prevalent in decision-making contexts, the methodologies and insights developed in this dissertation provide valuable tools for ensuring that these systems make optimal decisions based on accurate causal understanding. By addressing fundamental challenges such as endogeneity, measurement error, and dynamic learning, this research contributes to the development of more reliable, effective, and equitable data-driven decision systems.

This concluding chapter synthesizes the key findings from the three essays, highlights their interconnections, and outlines promising directions for future research that build upon this foundation.

The first essay addressed the critical problem of endogeneity in online decision-making contexts. By formulating the decision-making process as a contextual linear bandit model, we demonstrated how many economic settings characterized by endogeneity can lead to biased estimation and suboptimal decisions. Our proposed ϵ-BanditIV algorithm integrates instrumental variables with existing linear bandit algorithms to correct this bias and optimize online decisions. The theoretical properties of our algorithm include an upper bound for the total expected regret and the consistency and asymptotic normality of the estimator. Through extensive simulations on synthetic data, we demonstrated that the ϵ-BanditIV algorithm outperforms several benchmark linear bandit algorithms in endogenous settings. Empirical validation using datasets from mobile app recommendation and Real-Time Bidding (RTB) systems confirmed the algorithm's effectiveness in estimating causal impacts while maintaining near-oracle regret performance.

The second essay addressed the growing prevalence of ML and AI-generated regressors (including those from Large Language Models) in econometric analysis. With researchers increas-

ingly turning to partially linear models for their ability to integrate various ML algorithms and their strengths in high-dimensional causal inference, we developed novel two-step estimators and GMM estimators to debias ML/AI-generated regressors in these models. Our work extended the classical Double Machine Learning (DML) framework to accommodate scenarios with generated regressors in a full sample alongside true labels in a small subsample. Theoretical proofs established the asymptotic consistency and normality of the proposed estimators, and we provided consistent variance estimators for statistical inference. Synthetic simulations demonstrated the superior performance of our methods in terms of unbiasedness and low variance compared to existing approaches. Empirical applications using app review data showed how our estimators effectively correct bias from ML-predicted and ChatGPT-predicted sentiment variables while improving estimation efficiency.

The third essay examined how livestream hosts make product assortment and presentation timing decisions to maximize sales. Using real-world data and structural modeling techniques, we investigated how hosts learn about viewer demand and optimize their choices over time. For product assortment decisions, we employed Multi-Armed Bandit (MAB) models to examine hosts' learning and exploitation processes. Maximum likelihood estimation revealed that a host's exploration behavior is significantly influenced by their activeness, social network presence, past livestream experience, and demographic characteristics. For presentation timing decisions, a stratified Cox proportional hazards model identified factors influencing when hosts switch products, including product set characteristics, viewership, remaining products, and livestream timing. Simulation experiments evaluating alternative product selection strategies found that Thompson Sampling and modified ϵ-Greedy approaches can improve overall sales by 12.91% to 18.56%, with effectiveness varying across host segments. Particularly notable were the improvements for hosts with larger follower bases (25-33%) and those with lower education levels (81-91%), suggesting these strategies could serve as valuable tools for democratizing success on livestream platforms.

Building upon the foundation established by these three essays, several promising directions for future research emerge. First, future research could develop bandit algorithms that can handle both endogeneity (from the first essay) and noisy ML/AI-generated features (from the second essay) in

online decision-making contexts. This would extend both theoretical frameworks while providing practical tools for applications such as personalized recommendation systems or adaptive pricing strategies. Second, building on insights from all three essays, future work could also develop AI-assisted decision support systems for livestream hosts that correct for both endogeneity and measurement error while optimizing product assortment and timing decisions. By incorporating the theoretical advances from the first two essays into the applied context of the third, such systems could provide real-time, theoretically grounded recommendations to hosts. Third, all three essays employ linear or partially linear models for their theoretical guarantees and interpretability. Future research could extend these frameworks to fully non-parametric specifications that can capture more complex relationships while maintaining statistical guarantees. This would be particularly valuable for the debiasing framework in the second essay and could enhance the performance of the bandit algorithms in the first and third essays.

BIBLIOGRAPHY

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

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd international conference on World Wide Web*, pages 13–24, 2013.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Gad Allon, Daniel Chen, Zhenling Jiang, and Dennis Zhang. Machine learning and prediction errors in causal inference. *Available at SSRN 4480696*, 2023.

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

Asim Ansari, Skander Essegaier, and Rajeev Kohli. Internet recommendation systems, 2000.

Nicolás Aramayo, Mario Schiappacasse, and Marcel Goic. A multi-armed bandit approach for house ads recommendations. *Available at SSRN 4107976*, 2022.

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

Edvard Bakhitov and Amandeep Singh. Causal gradient boosting: Boosted instrumental variable regression. *arXiv preprint arXiv:2101.06078*, 2021.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881, 2022.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

Neeraj Bharadwaj, Michel Ballings, Prasad A Naik, Miller Moore, and Mustafa Murat Arat. A new livestream retail analytics framework to assess the sales impact of emotional displays. *Journal of Marketing*, 86(1):27–47, 2022.

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

Gordon Burtch, Edward McFowland III, Mochen Yang, and Gediminas Adomavicius. Ensembleiv: Creating instrumental variables from ensemble learners for robust statistical inference. *arXiv preprint arXiv:2303.02820*, 2023.

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

Raymond J Carroll, David Ruppert, and Leonard A Stefanski. *Measurement error in nonlinear models*, volume 105. CRC press, 1995.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.

Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Matthew T Clements and Hiroshi Ohashi. Indirect network effects and the product cycle: video games in the us, 1994–2002. *The Journal of Industrial Economics*, 53(4):515–542, 2005.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

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

Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453, 2019.

Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951, 2021.

Arindrajit Dube, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. Monopsony in online labor markets. *American Economic Review: Insights*, 2(1):33–46, 2020.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

Daria Dzyabura and John R Hauser. Active machine learning for consideration heuristics. *Marketing Science*, 30(5):801–819, 2011.

Daria Dzyabura and Alex Tuzhilin. Not by search alone: How recommendations complement search results. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 371–374, 2013.

Vivek Farias, Ciamac Moallemi, Tianyi Peng, and Andrew Zheng. Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*, 2022.

Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

Nitai Fingerhut, Matteo Sesia, and Yaniv Romano. Coordinated double machine learning. In *International Conference on Machine Learning*, pages 6499–6513. PMLR, 2022.

Christian Fong and Matthew Tyler. Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484, 2021.

David Godes and Dina Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4):545–560, 2004.

Brett R Gordon, Kinshuk Jerath, Zsolt Katona, Sridhar Narayanan, Jiwoong Shin, and Kenneth C Wilbur. Inefficiencies in digital advertising markets. *Journal of Marketing*, 85(1):7–25, 2021.

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

Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22, 1977.

Xian Gu, Xiaoxi Zhang, and PK Kannan. Influencer mix strategies in livestream commerce: Impact on product sales. *Journal of Marketing*, page 00222429231213581, 2024a.

Zheyin Jane Gu, Xuying Zhao, and DJ Wu. A model of shoppertainment live streaming. *Management Science, Forthcoming*, 2024b.

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

Bruce Hansen. *Econometrics*. Princeton University Press, 2022.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Jerry A Hausman. Specification and estimation of simultaneous equation models. *Handbook of econometrics*, 1:391–448, 1983.

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

Lalit Jain, Zhaoqi Li, Erfan Loghmani, Blake Mason, and Hema Yoganarasimhan. Effective adaptive exploration of prices and promotions in choice-based demand models. *Marketing Science*, 43(5):1002–1030, 2024.

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

Michael C Knaus. A double machine learning approach to estimate the effects of musical practice on student's skills. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184 (1):282–300, 2021.

Michael C Knaus. Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3):602–627, 2022.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Tomás Lagos, Ramón Auad, and Felipe Lagos. The online shortest path problem: Learning travel times using a multiarmed bandit framework. *Transportation Science*, 2024.

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

Lawrence M Leemis. Variate generation for accelerated life and proportional hazards models. *Operations research*, 35(6):892–894, 1987.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Hairen Liao, Lingxiao Peng, Zhenchuan Liu, and Xuehua Shen. ipinyou global rtb bidding algorithm competition dataset. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–6, 2014.

Xi Lin, Luyi Gui, and Yixin Lu. Managing sales via livestream commerce: Implications of price negotiation and consumer price search. *Production and Operations Management*, page 10591478231224930, 2022.

Yan Lin, Dai Yao, and Xingyu Chen. Happiness begets money: Emotion and engagement in live streaming. *Journal of Marketing Research*, 58(3):417–438, 2021.

Xiao Liu. Dynamic coupon targeting using batch deep reinforcement learning: An application to livestream shopping. *Marketing Science*, 42(4):637–658, 2023.

Pei-San Lo, Yogesh K Dwivedi, Garry Wei-Han Tan, Keng-Boon Ooi, Eugene Cheng-Xi Aw, and Bhimaraya Metri. Why do consumers buy impulsively during live streaming? a deep learning-based dual-stage sem-ann analysis. *Journal of Business Research*, 147:325–337, 2022.

Jeffrey M Wooldridge. Introductory econometrics, 2014.

Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Dung Daniel T Ngo, Logan Stapleton, Vasilis Syrgkanis, and Steven Wu. Incentivizing compliance with algorithmic instruments. In *International Conference on Machine Learning*, pages 8045–8055. PMLR, 2021.

Hai Thanh Nguyen, Jérémie Mary, and Philippe Preux. Cold-start problems in recommendation systems via contextual-bandit algorithms. *arXiv preprint arXiv:1405.7544*, 2014.

Alessandro Nuara, Francesco Trovo, Nicola Gatti, and Marcello Restelli. A combinatorial-bandit algorithm for the online joint bid/budget optimization of pay-per-click advertising campaigns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. 2013.

Wei Qian and Yuhong Yang. Randomized allocation with arm elimination in a bandit problem with covariates. 2016.

Mengke Qiao and Ke-Wei Huang. Correcting misclassification bias in regression models with variables generated via data mining. *Information Systems Research*, 32(2):462–480, 2021.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

Omid Rafieian and Hema Yoganarasimhan. Targeting and privacy in mobile advertising. *Marketing Science*, 40(2):193–218, 2021.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Donghyuk Shin, Shu He, Gene Moo Lee, Andrew B Whinston, Suleyman Cetintas, and Kuang-Chih Lee. Enhancing social media analysis with visual data analytics: A deep learning approach. *MIS Quarterly*, 44(4):1459–1492, 2020.

David Simchi-Levi and Chonghuan Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics*, pages 3086–3097. PMLR, 2023.

Michael Sinkinson and Amanda Starc. Ask your doctor? direct-to-consumer advertising of pharmaceuticals. *The Review of Economic Studies*, 86(2):836–881, 2019.

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 323–332, 2015.

Terry Therneau et al. A package for survival analysis in s. *R package version*, 2(7):2014, 2015.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

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

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. *arXiv preprint arXiv:2112.03432*, 2021.

Huazheng Wang, Qingyun Wu, and Hongning Wang. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1633–1642, 2016.

Juan Wang, Yifan Yu, Wendao Xue, and Yong Tan. Covid-19, urban transportation, and air pollution. *SSRN 3859071*, 2021.

Yingfei Wang, Hua Ouyang, Chu Wang, Jianhui Chen, Tsvetan Asamov, and Yi Chang. Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Yingfei Wang, Inbal Yahav, and Balaji Padmanabhan. Smart testing with vaccination: A bandit algorithm for active sampling for managing covid-19. *Information Systems Research*, 35(1): 120–144, 2024.

Yanhao'Max' Wei and Nikhil Malik. Machine learning, econometric models, and estimation bias. *Econometric Models, and Estimation Bias (May 18, 2022)*, 2022.

Si Xie, Amit Mehra, and Siddhartha Sharma. Designing e-commerce livestreams: How quicker product presentation affects sales? *Available at SSRN 4114942*, 2022.

Mochen Yang, Gediminas Adomavicius, Gordon Burtch, and Yuqing Ren. Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, 29(1):4–24, 2018.

Mochen Yang, Edward McFowland III, Gordon Burtch, and Gediminas Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science*, 1(2):138–155, 2022.

Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

Zikun Ye, Zhiqi Zhang, Dennis Zhang, Heng Zhang, and Renyu Philip Zhang. Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *Available at SSRN 4375327*, 2023.

Dezhi Yin, Samuel D Bond, and Han Zhang. Keep your cool or let it out: Nonlinear effects of expressed arousal on perceptions of consumer reviews. *Journal of Marketing Research*, 54(3): 447–463, 2017.

Yifan Yu, Shan Huang, Yuchen Liu, and Yong Tan. Emotions in online content diffusion. *arXiv preprint arXiv:2011.09003*, 2020.

Yifan Yu, Yang Yang, Jinghua Huang, and Yong Tan. Unifying algorithmic and theoretical perspectives: Emotions in online reviews and sales. *MIS Quarterly*, 47(1):127–160, 2023.

Yifan Yu, Xue Tan, and Yong Tan. Understanding volunteer crowdsourcing from a multiplex perspective. *Information Systems Research*, 2024.

Daniel Zantedeschi, Eleanor McDonnell Feit, and Eric T Bradlow. Measuring multichannel advertising response. *Management Science*, 63(8):2706–2728, 2017.

Shunyuan Zhang, Dokyun Lee, Param Vir Singh, and Kannan Srinivasan. What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science*, 68(8):5644–5666, 2022.

Xiaoxue Zhao, Weinan Zhang, and Jun Wang. Interactive collaborative filtering. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1411–1420, 2013.

Tongxin Zhou, Yingfei Wang, Lu Yan, and Yong Tan. Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach. *Information Systems Research*, 34(4):1493–1512, 2023.

Appendix A

APPENDIX TO CAUSAL BANDITS

A.0.1 Illustration Figure for Endogeneity and Diagram for Literature Review

We provide a figure to illustrate the endogeneity problem in product recommendation in Figure A.1 and a diagram to illustrate the relationship between related literature and this essay in Figure A.2.

A.0.2 Synthetic Simulation Results with Noisier Instrumental Variables

To investigate whether our proposed algorithms are robust to noisier instrumental variables, we vary the variances of errors $\xi_{1,t}$ and $\xi_{3,t}$ in Equation (3.5) and report the comparison results in Figure A.3. Here are our main observations:

- When the variance of the error term which includes the exogenous error $\xi_{3,t}$ and the endogenous error $\xi_{1,t}$ increases, the finite sample estimation bias of all methods in our comparison would increase and all the regrets would increase;

- Among all cases, the *BanditIV* algorithm and ϵ-*BanditIV* algorithm persistently outperform benchmark algorithms.

A.0.3 Extra Simulation Results on Mobile App Recommendation

The flexibility of our proposed framework allows it to capture a wide range of market conduct behaviors, including strategic interactions between firms, without requiring the researcher to explicitly model these relationships. Specifically, the error component $\xi_{3,t,a}$ in Equation (3.7) can

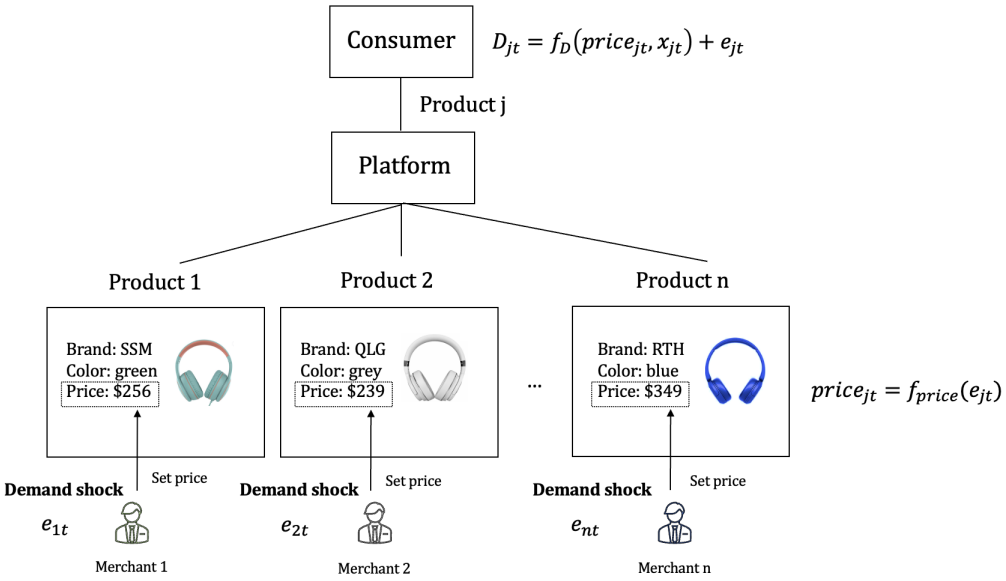


Figure A.1: Illustration for endogeneity in product recommendation

In this figure, we illustrate how endogeneity can occur in the context of a platform operator recommending products to customers in an online fashion. Every period, the platform has n products to choose from and aims to make recommendations that maximize some cumulative reward (say demand). The realized reward for any chosen arm (D_{jt}) depends on that product's features $\{price_{jt}, x_{jt}\}$ and unobserved demand shocks e_{jt}, that are unknown to the platform operator. These product features are determined by third-party sellers or merchants. Specifically, merchant j can adjust the $price_{jt}$ of product j in response to the realization of demand shocks e_{jt}, meaning that $price_{jt}$ is a function (f_{price}) of e_{jt}. This potential correlation, if ignored, can lead to sub-optimal recommendation decisions.

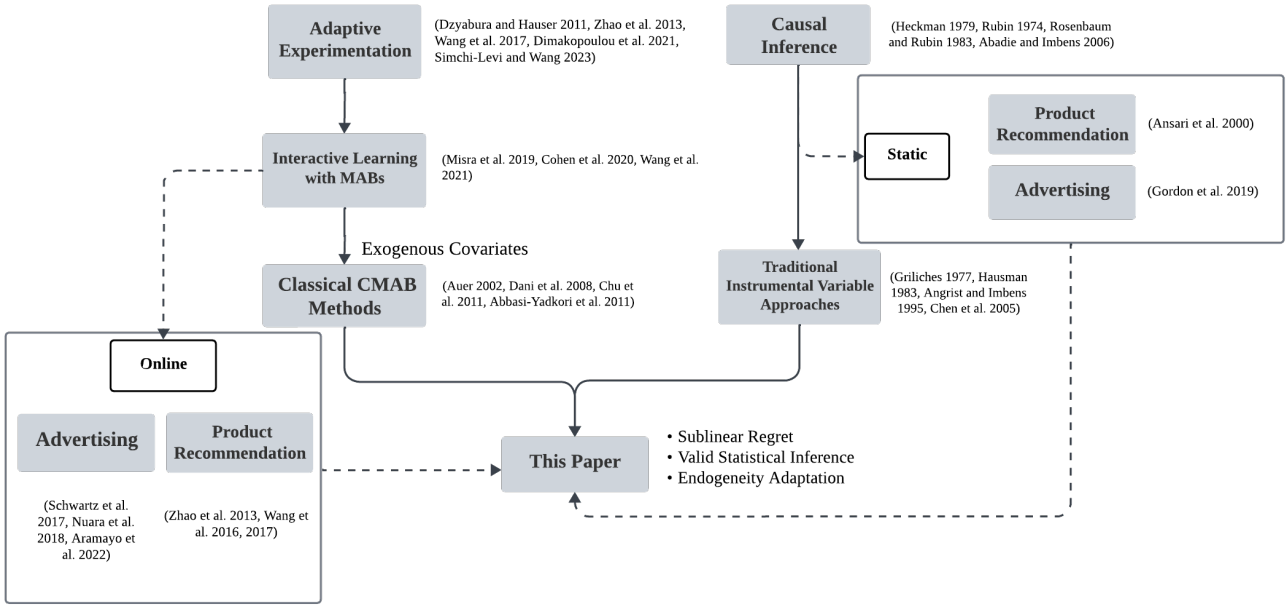


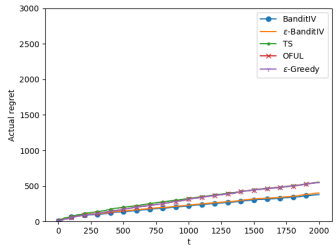
Figure A.2: Literature Overview Diagram

incorporate strategic pricing responses to competitors. For instance, $\xi_{3,t,a}$ could represent a function of competitors' prices: $\xi_{3,t,a} = f(Price_{-a,t'})$, where we use $-a$ to denote competitors of the focal app and t' to denote a specific time. This allows firms to engage in strategic price-setting based on competitor behavior without requiring us to explicitly model the precise form of competition.

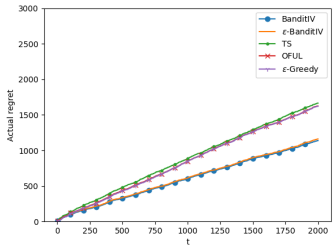
To demonstrate the robustness of our proposed approach, we generate data where firms explicitly respond to competitor pricing through the following data generating process:

$$\xi_{3,t,a} \sim \mathcal{N}(\text{Mean}(Price_{-a,t-1}), \text{Var}(Price_{t-1})).$$

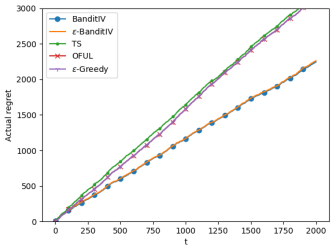
Our results, presented in Figure A.4, show that our methodology maintains its effectiveness even when firms engage in strategic pricing behavior.



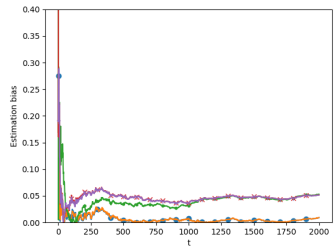
(a) $\sigma_{\xi_{3,t}} = 0.1$, $\sigma_{\xi_{1,t}} = 0.1$, $\rho = 5$



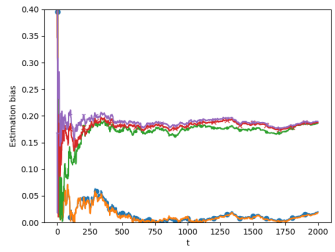
(b) $\sigma_{\xi_{3,t}} = 0.2$, $\sigma_{\xi_{1,t}} = 0.2$, $\rho = 5$



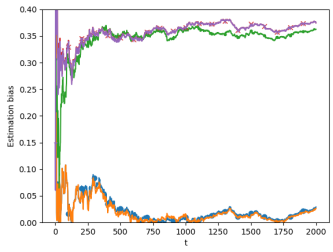
(c) $\sigma_{\xi_{3,t}} = 0.3$, $\sigma_{\xi_{1,t}} = 0.3$, $\rho = 5$



(d) $\sigma_{\xi_{3,t}} = 0.1$, $\sigma_{\xi_{1,t}} = 0.1$, $\rho = 5$



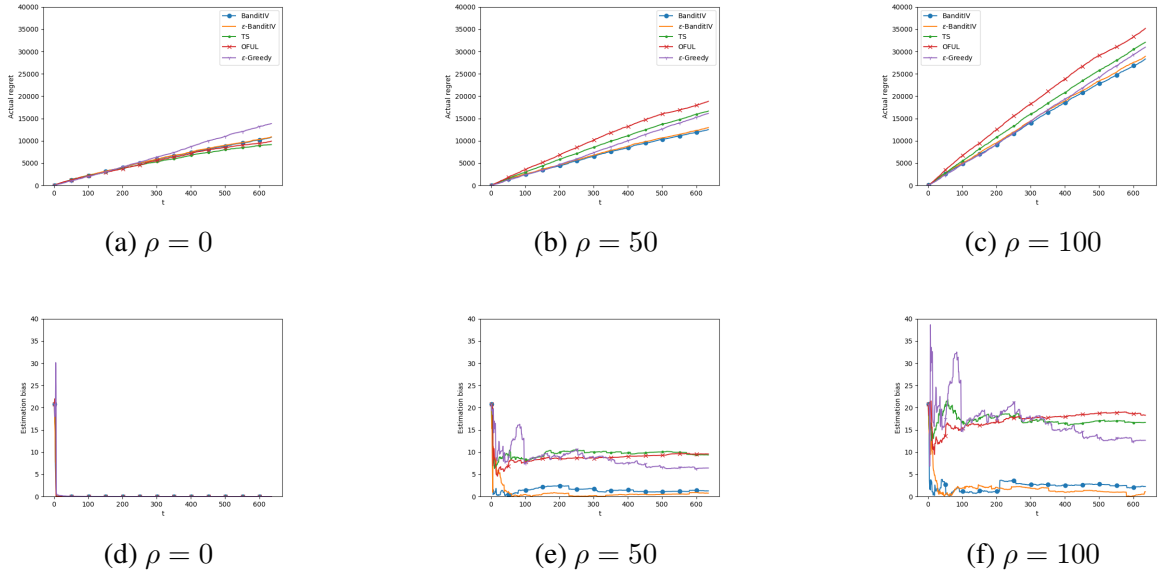
(e) $\sigma_{\xi_{3,t}} = 0.2$, $\sigma_{\xi_{1,t}} = 0.2$, $\rho = 5$



(f) $\sigma_{\xi_{3,t}} = 0.3$, $\sigma_{\xi_{1,t}} = 0.3$, $\rho = 5$

Figure A.3: Noisier instrumental variable: Simulation results on synthetic data with different noise levels on IV

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the actual regret. We use the absolute difference between the estimated coefficient of p_t and the true value of the coefficient -1 to measure the estimation bias and the actual cumulative regret to measure the regret. As mentioned before in this essay, we denote the *ϵ-BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm.



(a) $\rho = 0$ (b) $\rho = 50$ (c) $\rho = 100$

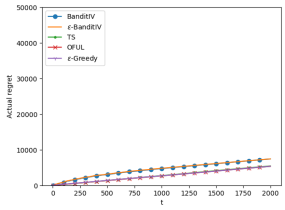
(d) $\rho = 0$ (e) $\rho = 50$ (f) $\rho = 100$

Figure A.4: Simulation results on mobile app data with $\xi_{3,t,a}$ drawn from a normal distribution $\mathcal{N}(\text{Mean}(\textit{Price}_{-a,t-1}), \text{Var}(\textit{Price}_{t-1}))$

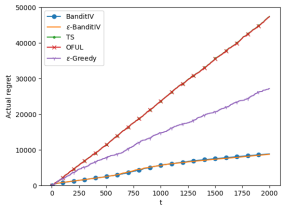
The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the regret. We use actual cumulative regret to measure the difference between the numbers of downloads from the true optimal at each time and from the focal algorithm. We use the absolute difference between the true value of the price effect and the estimated price effect by the focal algorithm to measure the estimation bias. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm.

A.0.4 Extra Simulation Results on RTB

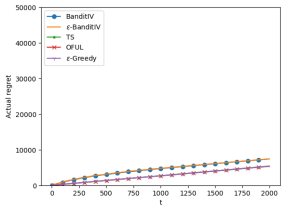
We provide results with $\beta^{(0)} \in [1000, 10000]$ in Figure A.5. The results are similar with the result of $\beta^{(0)} = 100$ reported in Section 3.5.2. Our proposed *BanditIV* and ϵ-*BanditIV* significantly outperform the other three classic algorithms both on regret and estimation bias, especially under higher endogeneity cases. When there are no endogeneity issues, i.e., $\rho = 0$, all of the algorithms have close-to-zero estimation bias while the *BanditIV* and the ϵ-*BanditIV* algorithms have slightly higher regrets. When endogeneity increases ($\rho = 100$), the estimation bias from classic algorithms increases dramatically while *BanditIV* algorithms still keep extremely low estimation bias and regret.



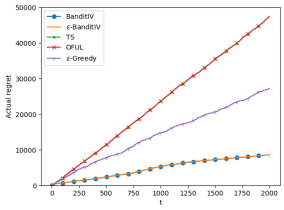
(a) $\rho = 0$, $\beta^{(0)} = 1,000$



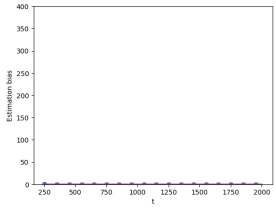
(b) $\rho = 100$, $\beta^{(0)} = 1,000$



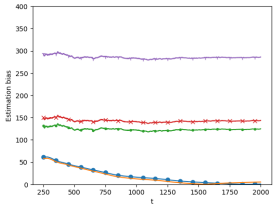
(c) $\rho = 0$, $\beta^{(0)} = 10,000$



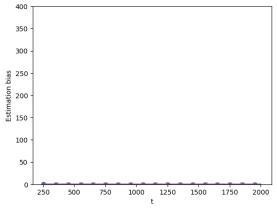
(d) $\rho = 100$, $\beta^{(0)} = 10,000$



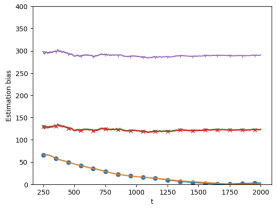
(e) $\rho = 0$, $\beta^{(0)} = 1,000$



(f) $\rho = 100$, $\beta^{(0)} = 1,000$



(g) $\rho = 0$, $\beta^{(0)} = 10,000$



(h) $\rho = 100$, $\beta^{(0)} = 10,000$

Figure A.5: Simulation results on RTB data with different endogeneity degrees and different $\beta^{(0)}$

The x-axis shows the number of time steps and the y-axis shows the performance indicator which is either estimation bias or the regret. We use actual cumulative regret to measure the difference between the revenue from the true optimal at each time and from the focal algorithm. We use the absolute difference between the true value of the ad effect and the estimated ad effect by the focal algorithm to measure the estimation bias. As mentioned before in this essay, we denote the ϵ-*BanditIV* algorithm with $\epsilon_t = 0$ for $t = 1, 2, \cdots, T$ as the *BanditIV* algorithm.

A.0.5 Hyperparameter Search

To ensure fair comparison between bandit algorithms, we perform hyperparameter tuning for each method. While this typically requires a validation set, bandit algorithms present unique challenges since they operate in online learning environments rather than offline supervised settings. Our hyperparameter tuning process follows three distinct steps. First, we generate a validation simulator whose data generating process precisely mimics that of the final test environment, but with a time horizon set to 30%[1] of the test dataset's duration. Second, we run each bandit algorithm with various hyperparameter configurations in this validation simulator. Third, we select the specific hyperparameter setting for each algorithm that minimizes its total cumulative regret on the validation data. This approach maintains statistical consistency between validation and testing environments while providing sufficient data for hyperparameter optimization without compromising test data integrity. Specifically, using a separate validation dataset prevents data leakage from the test set, avoids overfitting hyperparameters to test data patterns, and eliminates selection bias that could occur if algorithms were optimized directly on the evaluation benchmark.

Note that theoretically, if we know the upper bound (denoted by S in Abbasi-Yadkori et al. (2011)) of the l_2 norm of β_0, upper bound (denoted by R in Agrawal et al. (2013) and Abbasi-Yadkori et al. (2011)) of errors, we can fix the value of v by fixing the values of ε and δ, and fix the value of C_t by fixing the value of λ and δ. Similarly, if we know the true value of L_z and the upper bound of $\|\beta_0\|_2$, $\|\Gamma_0\|_F^2$, $\max_{i\in\{1,\cdots,d\}} \|\Gamma_0^{(i)}\|_2$, we can fix the values of B_t and G_t by fixing the values of γ_x, γ_z, and δ. However, in simulation or practical settings, the true values of S, R, L_z, and the upper bound of $\|\beta_0\|_2$, $\|\Gamma_0\|_F^2$, $\max_{i\in\{1,\cdots,d\}} \|\Gamma_0^{(i)}\|_2$ are hard to be specified. Thus we fine tune the hyperparameters v, C_t, B_t, G_t and ε by introducing multipliers that scale these variance, confidence bounds, or random probability appropriately. This approach allows us to control the level of exploration without requiring precise knowledge of the theoretical upper bounds, while

[1]For synthetic simulations in Section 3.4, the final test task has 2,000 rounds and we set the number of rounds in validation simulator as 600 which is 30% of the time horizon in the test task. For the mobile app application in Section 3.5.1, the total rounds in the final test task is 637 which is short and we set the number of rounds in the validation simulator as 600 to ensure sufficient learning in the validation process.

maintaining the structural form of variance, confidence bounds, or random probability.

A.0.6 Proof of Main Results

We provide proofs for main results in Appendix A.0.6 and auxiliary lemmas in Appendix A.0.7.

Theorem 1 (Theorem 3.2.1.)**.** *The expected cumulative regret of the ϵ-BanditIV at time T, with probability at least $1-\delta$, is upper-bounded by*

$$R_T \leq 3B_T\sqrt{2Td\log(\frac{T+d\gamma_x}{d\gamma_x})} + \tilde{G}_T\sqrt{2Tk\log(\frac{T+k\gamma_z}{k\gamma_z})} + \\ + 2B_T\frac{4\|\Gamma_0\|_2 L_z}{\sqrt{\lambda}(\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z}G_T)}(\sqrt{T} - \sqrt{T^\dagger - 1}) + 2L_y((1-\epsilon_1)T^\dagger + \epsilon_1 T)$$

where $B_T = \sqrt{\gamma_x}\|\beta_0\|_2 + \sqrt{2\log(\frac{6T}{\delta}) + d\log(\frac{2TkL_z^2}{d\gamma_x}\|\Gamma_0\|_F^2 + \frac{4kT}{\gamma_x}\log(\frac{6Td}{\delta})\log(\frac{T+k\gamma_z}{k\gamma_z}) + 1)}$,
$\tilde{G}_T = \sqrt{\gamma_z}\|\Gamma_0\beta_0\|_2 + \|\beta_0\|_1\sqrt{2\log(\frac{3T}{\delta}) + 2\log(\|\beta_0\|_1) + k\log(\frac{TL_z^2}{\gamma_z} + 1)}$,
$G_T = \sqrt{\gamma_z}\max_{i\in\{1,\cdots,d\}}\|\Gamma_0^{(i)}\|_2 + \sqrt{2\log(\frac{6Td}{\delta}) + k\log\left(\frac{TL_z^2}{\gamma_z} + 1\right)}$,
and $T^\dagger = \frac{6L_z^2}{\lambda}\log(\frac{6k}{\delta})$.

Proof. Proof. This regret can be decomposed into

$$\begin{aligned} R_T &= \sum_{t=1}^{T^\dagger}\mathbb{E}_{u_t}[\langle x_t^*, \beta_0\rangle - \langle x_t, \beta_0\rangle] + \sum_{t=T^\dagger+1}^{T}\mathbb{E}_{u_t}[\langle x_t^*, \beta_0\rangle - \langle x_t, \beta_0\rangle] \\ &\leq 2T^\dagger L_y + \sum_{t=T^\dagger+1}^{T}(1-\epsilon_t)\mathbb{E}_{u_t}[\langle x_t^*, \beta_0\rangle - \langle x_t, \beta_0\rangle]\mathbf{1}[x_t = \Gamma_0' z_t] + \\ &\quad + \sum_{t=T_\dagger+1}^{T}\epsilon_t\mathbb{E}_{u_t}[\langle x_t^*, \beta_0\rangle - \langle x_t, \beta_0\rangle]\mathbf{1}[x_t \neq \Gamma_0' z_t] \\ &\leq 2T^\dagger L_y + \sum_{t=T^\dagger+1}^{T}(\langle \Gamma_0' z_t^*, \beta_0\rangle - \langle \Gamma_0' z_t, \beta_0\rangle) + 2\epsilon_1(T - T^\dagger)L_y \end{aligned}$$

Next, we will upper bound the second term. Let $(\tilde{\Gamma}_t, \tilde{\beta}_t) = \arg\max_{\Gamma\in C_{1,t}, \beta\in C_{2,t}}\langle \Gamma' z_t, \beta\rangle$. By

using the *two-stage* optimism we can decompose our regret as follows,

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

where the first inequality comes from the Lemma 3.2.1, 3.2.2 that $\Gamma_0 \in C_{1,t}, \beta_0 \in C_{2,t}$ for all t, the second inequality comes from the definition of z_t, the first equality comes from the definition of $\hat{x}_t$ and some careful decomposition, and the third inequality comes from Cauchy-Schwartz inequality.

Now, our goal is to upper bound these three terms, which mainly comes from the estimation error of Γ_0 in the first stage and the estimation error of β_0 in the second stage. For the first term, by applying Cauchy-Schwartz inequality and the standard elliptical potential lemma in Carpentier et al. (2020), we have

$$
\begin{aligned}
\sum_{t=T^\dagger+1}^{T} \|\hat{x}_t\|_{W_t^{-1}} \|\tilde{\beta}_t - \beta_0\|_{W_t} &\leq \max_{t \in \{1,2,\cdots,T\}} \|\tilde{\beta}_t - \beta_0\|_{W_t} \sum_{t=T^\dagger+1}^{T} \|\hat{x}_t\|_{W_t^{-1}} \\
&\leq \max_{t \in \{1,2,\cdots,T\}} \|\tilde{\beta}_t - \beta_0\|_{W_t} \sqrt{2Td \log(\frac{T + d\gamma_x}{d\gamma_x})}
\end{aligned}
$$

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

$$\sum_{t=T^\dagger+1}^{T} |\beta_0'(\tilde{\Gamma}_t - \Gamma_0)' z_t| \leq \sum_{t=T^\dagger+1}^{T} \|\beta_0'(\tilde{\Gamma}_t - \Gamma_0)'\|_{U_t} \|z_t\|_{U_t^{-1}}$$
$$\leq \max_{t\in\{1,2,\cdots,T\}} \|\beta_0'(\tilde{\Gamma}_t - \Gamma_0)'\|_{U_t} \sqrt{2Tk\log(\frac{T+k\gamma_z}{k\gamma_z})}$$

Finally, for the third term,

$$2\sum_{t=T^\dagger+1}^{T} |\langle(\Gamma_0 - \hat{\Gamma}_t)' z_t, \tilde{\beta}_t - \beta_0\rangle| \leq 2\sum_{t=T^\dagger+1}^{T} (|\langle\Gamma_0' z_t, \tilde{\beta}_t - \beta_0\rangle| + |\langle\hat{\Gamma}_t' z_t, \tilde{\beta}_t - \beta_0\rangle|)$$
$$\leq 2\sum_{t=T^\dagger+1}^{T} \|\Gamma_0' z_t\|_{W_t^{-1}} \|\tilde{\beta}_t - \beta_0\|_{W_t} + 2\sum_{t=T^\dagger+1}^{T} \|\hat{x}_t\|_{W_t^{-1}} \|\tilde{\beta}_t - \beta_0\|_{W_t}$$
$$\leq 2\max_t \|\tilde{\beta}_t - \beta_0\|_{W_t} \sum_{t=T^\dagger+1}^{T} \|\Gamma_0' z_t\|_{W_t^{-1}} + 2\max_{t\in\{1,2,\cdots,T\}} \|\tilde{\beta}_t - \beta_0\|_{W_t} \sqrt{2Td\log(\frac{T+d\gamma_x}{d\gamma_x})}$$

Therefore, by combing these three terms and Lemmas 3.2.1, A.0.1, and A.0.2, we complete the proof.

□

Lemma 1 (Lemma 3.2.1.). *With high prob* $1-\delta/3$*, for all* $t \in \{1, 2, \cdots, T\}$*,*

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

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

least squares estimator as follows,

$$\begin{aligned}\hat{\beta}_t &= W_t^{-1} Q_t \\ &= W_t^{-1} \hat{X}'_{t-1}(X_{t-1}\beta_0 + \mathbf{e}_{t-1}) \\ &= W_t^{-1} X'_{t-1} P_{Z_{t-1}}(X_{t-1}\beta_0 + \mathbf{e}_{t-1}) \\ &= W_t^{-1}(W_t - \gamma_x I)\beta_0 + W_t^{-1}\hat{X}'_{t-1}\mathbf{e}_{t-1}\end{aligned}$$

where we denote $Z_{t-1}(Z'_{t-1}Z_{t-1})^{-1}Z'_{t-1}$ as $P_{Z_{t-1}}$. Therefore, we can write the following

$$\begin{aligned}\|\hat{\beta}_t - \beta_0\|_{W_t} &= \|W_t^{-1}\hat{X}'_{t-1}\mathbf{e}_{t-1} - \gamma_x W_t^{-1}\beta_0\|_{W_t} \\ &= \|\hat{X}'_{t-1}\mathbf{e}_{t-1} - \gamma_x\beta_0\|_{W_t^{-1}} \\ &\leq \|\hat{X}'_{t-1}\mathbf{e}_{t-1}\|_{W_t^{-1}} + \gamma_x\|\beta_0\|_{W_t^{-1}} \\ &\leq \|\hat{X}'_{t-1}\mathbf{e}_{t-1}\|_{W_t^{-1}} + \gamma_x\|\beta_0\|_{(\gamma_x I)^{-1}} \\ &= \|\hat{X}'_{t-1}\mathbf{e}_{t-1}\|_{W_t^{-1}} + \sqrt{\gamma_x}\|\beta_0\|_2\end{aligned}$$

Finally, by noticing that each element in $\mathbf{e}_t$ is 1-subgaussian, and according to Lemma 20.2 and Theorem 20.4 in Lattimore and Szepesvári (2020), that $\exp(\langle q, \hat{X}'_{t-1}\mathbf{e}_{t-1}\rangle - \frac{\|q\|^2_{W_t}}{2})$ for all $q \in \mathbb{R}^d$ is a supermartingale, we have that, with probability at least $1 - \delta/(6T)$,

$$\|\hat{X}'_{t-1}\mathbf{e}_{t-1}\|_{W_t^{-1}} \leq \sqrt{2\log(\frac{6T}{\delta}) + \log(\frac{\det(W_t)}{\gamma_x^d})}$$

By combining the above result, the explicit calculation of W_t detailed in Lemma A.0.7 and the Boole's inequality, we complete the proof. □

Lemma 2 (Lemma 3.2.2.)**.** *With high prob* $1 - \delta/6$*, for all* $t \in \{1, 2, \cdots, T\}$ *and all* $i \in \{1, 2, \cdots, d\}$*,*

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

Proof. Proof. The proof steps are similar to the previous lemma. At any fixed time t and dimension $i \in \{1, 2, \cdots, d\}$, we denote $\mathbf{u}_t^{(i)} \in \mathbb{R}^t$ as a collection of $\{u_s^{(i)}\}_{s \le t}$ and $\mathbf{u}_t \in \mathbb{R}^{t \times d}$ as a collection of $\{u_s\}_{s \le t}$. We again get the closed-form of $\hat{\Gamma}_t$ as

$$\hat{\Gamma}_t = U_t^{-1} V_t = U_t^{-1}(U_t - \gamma_z I)\Gamma_0 + U_t^{-1} Z'_{t-1}\mathbf{u}_{t-1} = \Gamma_0 - \gamma_z U_t^{-1}\Gamma_0 + U_t^{-1} Z'_{t-1}\mathbf{u}_{t-1}$$

And therefore, $\|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_{U_t} \le \|Z'_{t-1}\mathbf{u}_{t-1}^{(i)}\|_{U_t^{-1}} + \sqrt{\gamma_z}\|\Gamma_0^{(i)}\|_2$.

Finally, again by noticing that each element in $\mathbf{u}_t$ is 1-subgaussian and according to Lemma 20.2 and Theorem 20.4 in Lattimore and Szepesvári (2020), that $\exp(\langle q, Z'_{t-1}\mathbf{u}_{t-1}^{(i)}\rangle - \frac{\|q\|_{U_t}^2}{2})$ for all $q \in \mathbb{R}^k$ is a supermartingale, we have that, with probability at least $1 - \delta/(6Td)$,

$$\|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_{U_t} \le \sqrt{\gamma_z}\|\Gamma_0^{(i)}\|_2 + \sqrt{2\log(\frac{6Td}{\delta}) + \log(\frac{\det(U_t)}{\gamma_z^k})}$$

where the last inequality comes from the explicit calculation of U_t detailed in Lemma A.0.7. We complete the proof by the Boole's inequality. □

Lemma A.0.1. With high prob $1 - \delta/3$, for all $t \in \{1, 2, \cdots, T\}$,

$$\|(\hat{\Gamma}_t - \Gamma_0)\beta_0\|_{U_t} \le \sqrt{\gamma_z}\|\Gamma_0\beta_0\|_2 + \|\beta_0\|_1\sqrt{2\log(\frac{3T}{\delta}) + 2\log(\|\beta_0\|_1) + k\log(\frac{TL_z^2}{\gamma_z} + 1)} \equiv \tilde{G}_T$$

Proof. Proof. Following the proof of Lemma 3.2.2, we have $\|(\hat{\Gamma}_t - \Gamma_0)\beta_0\|_{U_t} \le \|Z'_{t-1}\mathbf{u}_{t-1}\beta_0\|_{U_t^{-1}} + \sqrt{\gamma_z}\|\Gamma_0\beta_0\|_2$.

Again by noticing that each element in $\mathbf{u}_t\beta_0$ is $\|\beta_0\|_1$-subgaussian and according to Lemma 20.2 and Theorem 20.4 in Lattimore and Szepesvári (2020), that $\exp(\langle q, Z'_{t-1}\mathbf{u}_{t-1}\beta_0\rangle - \frac{\|\beta_0\|_1^2\|q\|_{U_t}^2}{2})$

for all $q \in \mathbb{R}^k$ is a supermartingale, we have that, with probability at least $1 - \delta/(3T)$,

$$\begin{aligned}\|(\hat{\Gamma}_t - \Gamma_0)\beta_0\|_{U_t} &\leq \sqrt{\gamma_z}\|\Gamma_0\beta_0\|_2 + \|\beta_0\|_1\sqrt{2\log(\frac{3T}{\delta}) + \log(\frac{\|\beta_0\|_1^2 \det(U_t)}{\gamma_z^k})} \\ &\leq \sqrt{\gamma_z}\|\Gamma_0\beta_0\|_2 + \|\beta_0\|_1\sqrt{2\log(\frac{3T}{\delta}) + 2\log(\|\beta_0\|_1) + k\log(\frac{TL_z^2}{\gamma_z} + 1)}\end{aligned}$$

where the last inequality comes from the explicit calculation of U_t detailed in Lemma A.0.7.

□

Lemma A.0.2.

$$\sum_{t=T^\dagger+1}^{T} \|\Gamma_0' z_t\|_{W_t^{-1}} \leq \frac{4\|\Gamma_0\|_2 L_z}{\sqrt{\lambda}(\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z}G_T)}(\sqrt{T} - \sqrt{T^\dagger - 1}), \quad w.p. \geq 1 - \frac{\delta}{3}$$

where $T^\dagger = \frac{6L_z^2}{\lambda}\log(\frac{6k}{\delta})$.

Proof. Proof. On the one hand, by Weyl's inequality and Lemma A.0.4, we have the following with a high probability at least $1 - \delta/6$,

$$\lambda_{\min}(Z_t'Z_t) \geq t\lambda_{\min}(\hat{\Sigma}_t) \geq \frac{t}{4}\lambda, \quad \text{for all } t \geq T^\dagger$$

On the other hand, by Lemmas 3.2.2 and A.0.6, we have the following with a high probability at least $1 - \delta/6$,

$$\|\hat{\Gamma}_t - \Gamma_0\|_1 = \max_{i\in\{1,\ldots,d\}} \|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_1 \leq \max_{i\in\{1,\ldots,d\}} \sqrt{2}\|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_2 \leq \max_{i\in\{1,\ldots,d\}} \frac{\sqrt{2}}{\gamma_z}\|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_{U_t} \leq \frac{\sqrt{2}}{\gamma_z}G_t$$

and hence,

$$\lambda_{min}(\hat{\Gamma}_t) \geq \lambda_{min}(\Gamma_0) - \frac{\sqrt{2}}{\gamma_z}G_t\sqrt{d}$$

By Min-max theorem,

$$\begin{aligned}
\lambda_{\min}(\hat{X}'_{t-1}\hat{X}_{t-1}) &= \lambda_{\min}(\hat{\Gamma}'_t Z'_{t-1} Z_{t-1} \hat{\Gamma}_t) \\
&= \min_{\|a\|_2=1} a'\hat{\Gamma}'_t Z'_{t-1} Z_{t-1} \hat{\Gamma}_t a \frac{(\hat{\Gamma}_t a)'(\hat{\Gamma}_t a)}{(\hat{\Gamma}_t a)'(\hat{\Gamma}_t a)} \\
&\geq \lambda_{\min}(Z'_{t-1} Z_{t-1}) \min_{\|a\|_2=1} (\hat{\Gamma}_t a)'(\hat{\Gamma}_t a) \\
&= \lambda_{\min}(Z'_{t-1} Z_{t-1}) \lambda_{\min}(\hat{\Gamma}'_t \hat{\Gamma}_t) \\
&= \lambda_{\min}(Z'_{t-1} Z_{t-1}) (\lambda_{\min}(\hat{\Gamma}_t))^2
\end{aligned}$$

$$\lambda_{\min}(\hat{X}'_{t-1}\hat{X}_{t-1}) \geq \frac{t-1}{4}\lambda \left(\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z} G_t \right)^2, \quad \text{for all } t > T^\dagger$$

By Cauchy–Schwarz inequality,

$$\begin{aligned}
\sum_{t=T^\dagger+1}^{T} \|\Gamma'_0 z_t\|_{W_t^{-1}} &\leq \sum_{t=T^\dagger+1}^{T} \|\Gamma'_0 z_t\|_2 \sqrt{\lambda_{\max}(W_t^{-1})} \\
&\leq \|\Gamma_0\|_2 L_z \sum_{t=T^\dagger+1}^{T} \sqrt{\lambda_{\max}(W_t^{-1})} \\
&= \|\Gamma_0\|_2 L_z \sum_{t=T^\dagger+1}^{T} \sqrt{\frac{1}{\lambda_{\min}(W_t)}} \\
&\leq \|\Gamma_0\|_2 L_z \sum_{t=T^\dagger+1}^{T} \sqrt{\frac{1}{\lambda_{\min}(\hat{X}'_{t-1}\hat{X}_{t-1})}} \\
&\leq \|\Gamma_0\|_2 L_z \sum_{t=T^\dagger+1}^{T} \sqrt{\frac{4}{(t-1)\lambda \left(\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z} G_t \right)^2}} \\
&\leq \frac{4\|\Gamma_0\|_2 L_z}{\lambda_{min}(\Gamma_0) - \frac{\sqrt{2d}}{\gamma_z} G_T} (\sqrt{T} - \sqrt{T^\dagger - 1})
\end{aligned}$$

where the last inequality is from the fact that $\sum_{t=T^\dagger+1}^{T} \frac{1}{\sqrt{t-1}} = \sum_{t=T^\dagger+1}^{T} \frac{\int_{t-1}^{t} ds}{\sqrt{t-1}} \leq \sum_{t=T^\dagger+1}^{T} \int_{t-1}^{t} \frac{ds}{\sqrt{s-1}} = \int_{T^\dagger}^{T} \frac{ds}{\sqrt{s-1}} \leq 2\sqrt{T-1} - 2\sqrt{T^\dagger - 1}$. □

Lemma A.0.3. (Dependent OLS Tail Inequality). For the online decision-making model, if all realizations of z_t satisfy $\|z_t\|_\infty \leq L_z$ for all t, and $\hat{\Sigma} = \frac{1}{t}\sum_{s=1}^{t} z_s z_s'$ has minimum eigenvalue $\lambda_{min}(\hat{\Sigma}) > \lambda$ for some $\lambda > 0$ almost surely. Then for any $\eta > 0$,

$$P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta) \geq 1 - 2k \exp\left(-\frac{t\lambda^2\eta^2}{2k^2\sigma_v^2 L_z^2}\right)$$

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

$$\|\hat{\vartheta}_t - \vartheta_0\|_2 = \|\hat{\Sigma}^{-1}(\frac{1}{t}\sum_{s=1}^{t} z_s v_s)\|_2 \leq \frac{1}{t}\|\hat{\Sigma}^{-1}\|_2 \|\sum_{s=1}^{t} z_s v_s\|_2 \leq \frac{1}{t\lambda}\|\sum_{s=1}^{t} z_s v_s\|_2$$

Hence, we have

$$\begin{aligned}
P(\|\hat{\vartheta}_t - \vartheta_0\|_2 \leq \eta) &\geq P(\|\sum_{s=1}^{t} z_s v_s\|_2 \leq t\lambda\eta) \\
&\geq P(|\sum_{s=1}^{t} z_s^{(1)} v_s| \leq \frac{t\lambda\eta}{\sqrt{k}}, \ldots, |\sum_{s=1}^{t} z_s^{(k)} v_s| \leq \frac{t\lambda\eta}{\sqrt{k}}) \\
&= 1 - P(\bigcup_{j=1}^{k} \{|\sum_{s=1}^{t} z_s^{(j)} v_s| > \frac{t\lambda\eta}{\sqrt{k}}\}) \\
&\geq 1 - \sum_{j=1}^{k} P(|\sum_{s=1}^{t} z_s^{(j)} v_s| > \frac{t\lambda\eta}{\sqrt{k}})
\end{aligned}$$

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

Lemma 1 of Chen et al. (2021) and have the following

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

Thus,

$$P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta) \geq P(\|\hat{\vartheta}_t - \vartheta_0\|_2 \leq \frac{\eta}{\sqrt{k}}) \geq 1 - 2k\exp(-\frac{t\lambda^2\eta^2}{2k^2\sigma_v^2 L_z^2})$$

□

Lemma A.0.4. Let $\{z_t : t = 1, \ldots, T\}$ be a sequence of i.i.d. k-dimension random vectors such that all realizations of z_t satisfy $\|z_t\|_2 \leq L_z$ for all t. Denote $\hat{\Sigma}_t = \frac{1}{t}\sum_{s=1}^{t} z_s z_s'$. If $\Sigma = \mathbb{E}[z_t z_t']$ has minimum eigenvalue $\lambda_{min}(\Sigma) > \lambda$ for some $\lambda > 0$, then for a fixed t,

$$P(\lambda_{\min}(\hat{\Sigma}_t) \leq \frac{\lambda}{4}) \leq k\exp(-\frac{t\lambda}{6L_z^2})$$

Proof. Proof. This proof is based on the proof of Lemma 3 in Chen et al. (2021) with minor changes. First, we have

$$\lambda_{max}(\frac{z_s z_s'}{t}) = \max_{\|a\|_2=1} a'(\frac{z_s z_s'}{T})a = \frac{1}{t}\max_{\|a\|_2=1}(a'z_s)^2 \leq \frac{2}{t}\max_{\|a\|_2=1}\sum_{i=1}^{k}(a^{(i)}z_s^{(i)})^2 \leq \frac{2}{t}\sum_{i=1}^{k}(z_s^{(i)})^2 \leq \frac{2L_z^2}{t}$$

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

Then, notice that $4^{1/8} = 2^{\frac{7}{5}/(4\cdot\frac{7}{5})} < \exp(1/(4\cdot\frac{7}{5})) = \exp(\frac{5}{28})$, using the Matrix Chernoff bound in Theorem 5.1.1 in Tropp et al. (2015),

$$P(\lambda_{\min}(\hat{\Sigma}_t) \leq \frac{\mu_{min}}{4}) \leq k\left(\frac{\exp(-\frac{3}{4})}{(1/4)^{1/4}}\right)^{\frac{\mu_{\min}t}{2L_z^2}} = k\exp(-\frac{3\mu_{\min}t}{8L_z^2})4^{\frac{\mu_{\min}t}{8L_z^2}} \leq k\exp(-\frac{\mu_{\min}t}{L_z^2}\frac{11}{56}) \leq k\exp(-\frac{\mu_{\min}t}{L_z^2}\frac{1}{6})$$

$$P(\lambda_{min}(\hat{\Sigma}_t) \leq \frac{\lambda}{4}) \leq P(\lambda_{\min}(\hat{\Sigma}) \leq \frac{\mu_{min}}{4}) \leq k\exp(-\frac{t\mu_{min}}{6L_z^2}) \leq k\exp(-\frac{t\lambda}{6L_z^2})$$

□

Lemma A.0.5.

$$\|\hat{\Gamma}(\hat{\beta}-\beta_0)\|_1 \geq \frac{\lambda_{min}(\hat{\Gamma})}{\sqrt{2}}\|\hat{\beta}-\beta_0\|_1$$

where $\lambda_{min}(\hat{\Gamma}) = \sqrt{\lambda_{min}(\hat{\Gamma}'\hat{\Gamma})}$, i.e., the smallest nonzero singular value of $\hat{\Gamma}$.

Proof. Proof. By doing Singular Value Decomposition for $\hat{\Gamma}$, we have the following,

$$\|\hat{\Gamma}(\hat{\beta}-\beta_0)\|_1 \geq \|\hat{\Gamma}(\hat{\beta}-\beta_0)\|_2 = \|U\Sigma V'(\hat{\beta}-\beta_0)\|_2 = \|\Sigma V'(\hat{\beta}-\beta_0)\|_2$$

where the last equality is due to the orthonormality of U. WLOG, we assume $rank(\hat{\Gamma}) = min\{k,d\} = d$.

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

where the last equality is due to that V is orthonormal. Lastly, because

$$\|(\hat{\beta}-\beta_0)\|_2^2 = \sum_{i=1}^{d}(\hat{\beta}^{(i)}-\beta_0^{(i)})^2 \geq \frac{1}{2}(\sum_{i=1}^{d}|\hat{\beta}^{(i)}-\beta_0^{(i)}|)^2 = \frac{1}{2}\|(\hat{\beta}-\beta_0)\|_1^2, \qquad \text{(A.2)}$$

combing inequalities (A.1) and (A.2), we complete the proof. □

Lemma A.0.6. If $\|\hat{\Gamma}-\Gamma_0\|_1 \leq \eta_2$, then we have

$$\lambda_{min}(\hat{\Gamma}) \geq \lambda_{min}(\Gamma_0) - \eta_2\sqrt{d} \qquad \text{(A.3)}$$

Proof. Proof. On the event that $\|\hat{\Gamma} - \Gamma_0\|_1 \leq \eta_2$, we can rewrite the equation for each element in $\hat{\Gamma}$ using a newly defined matrix $A \in \mathbb{R}^{k\times d}, s.t. \|A\|_1 \leq 1$ and

$$\hat{\Gamma} = \Gamma_0 + \eta_2 A \tag{A.4}$$

We know that all singular values are non-negative. Here, due to the requirement of the instrumental variable method, we assume the singular values of Γ_0 are strictly positive. We can write the singular values of Γ_0 and $\hat{\Gamma}$ as two increasing sequences, $0 < \sigma_\Gamma^{(1)} \leq \sigma_\Gamma^{(2)} \leq \cdots \leq \sigma_\Gamma^{(d)}$ and $\sigma^{(1)} \leq \sigma^{(2)} \leq \cdots \leq \sigma^{(d)}$, respectively. Using the min-max principle for singular values, we have

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

where the first equality is directly derived by applying the min-max theorem. The second equality is from the two facts. First, a should be of order $d \times 1$ to fit the order of Γ_0. Second, there are only two unique vectors fitting the constraint $\|a\|_2 = 1$ in the given space S and they generate the same objective values $\|\Gamma_0 a\|_2$. Similarly, we can rewrite the smallest singular value of $\hat{\Gamma}$ as $\sigma^{(1)} = \min_{\substack{a\in\mathbb{R}^d \\ \|a\|_2=1}} \|\hat{\Gamma} a\|_2$.

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

where the first inequality is derived from the equation (A.4) and triangle inequality. The third

equality is from the constraint that $\|a\|_2 = 1$. The fourth inequality is from the fact that $\|A\|_2 \leq \sqrt{d}\|A\|_1$

□

Proposition 1 (Proposition 3.3.1.)**.** *In the online decision-making model with the ϵ-greedy policy, if Assumptions 3.1.1 and 3.1.2 are satisfied, and ϵ_t is non-increasing, then for any $\eta_1, \eta_2 > 0$, any $i \in \{1, \ldots, d\}$,*

$$P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta_1) \geq 1 - \exp(-\frac{t\epsilon_t}{8}) - k\exp(-\frac{t\epsilon_t\lambda}{12L_z^2}) - 2k\exp(-\frac{t\epsilon_t^2\lambda^2\eta_1^2}{128k^2\sigma_v^2L_z^2})$$

$$P(\|\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)}\|_1 \leq \eta_2) \geq 1 - \exp(-\frac{t\epsilon_t}{8}) - k\exp(-\frac{t\epsilon_t\lambda}{12L_z^2}) - 2k\exp(-\frac{t\epsilon_t^2\lambda^2\eta_2^2}{128k^2\sigma_u^2L_z^2})$$

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

$$\hat{\vartheta}_t - \vartheta_0 = (\frac{1}{t}\sum_{s=1}^{t} z_s z_s')^{-1}\frac{1}{t}\sum_{s=1}^{t} z_s v_s = \{\hat{\Sigma}(\mathcal{S}_t)\}^{-1}\frac{1}{t}\sum_{s=1}^{t} z_s v_s$$

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

We denote the following event,

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

Applying Lemma A.0.4, if $\lambda_{min}(\Sigma) > \lambda$, then we have

$$P(E) \geq 1 - k \exp(-\frac{|\mathcal{R}_t|\lambda}{6L_z^2})$$

Meanwhile, we know that

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

By Weyl's inequality, on event E,

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

Applying Lemma A.0.3, we have

$$P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta | E) \geq 1 - 2k \exp\left(-\frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2}\right)$$

Hence,

$$\begin{aligned}
P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta) &\geq P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \leq \eta | E) P(E) \\
&\geq 1 - 2k \exp\left(-\frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2}\right) - k \exp(-\frac{|\mathcal{R}_t|\lambda}{6L_z^2}) + 2k^2 \exp(-\frac{|\mathcal{R}_t|\lambda}{6L_z^2} - \frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2}) \\
&\geq 1 - 2k \exp\left(-\frac{|\mathcal{R}_t|^2\lambda^2\eta^2}{32tk^2\sigma_v^2 L_z^2}\right) - k \exp(-\frac{|\mathcal{R}_t|\lambda}{6L_z^2}) \\
&\equiv P(\tilde{f}(|\mathcal{R}_t|))
\end{aligned}$$

The checking step for that $|\mathcal{R}_t|$ is large enough is exactly the same as that in Chen et al. (2021) and we obtain $P(|\mathcal{R}_t| \leq \frac{t\epsilon_t}{2}) \leq \exp(-\frac{t\epsilon_t}{8})$. Let event $\tilde{E} := \{|\mathcal{R}_t| > \frac{t\epsilon_t}{2}\}$ and we know this event is with a probability at least $1 - \exp(-\frac{t\epsilon_t}{8})$. We complete the proof by combining all the results

above,

$$
\begin{aligned}
P(\|\hat{\vartheta}_t - \vartheta_0\|_1 \le \eta) &\ge P(\tilde{f}(|\mathcal{R}_t|)|\tilde{E})P(\tilde{E}) \ge \inf P(\tilde{f}(|\mathcal{R}_t|)|\tilde{E})P(\tilde{E}) \\
&\ge \left(1 - 2k\exp\left(-\frac{t\epsilon_t^2\lambda^2\eta^2}{128k^2\sigma_v^2 L_z^2}\right) - k\exp\left(-\frac{t\epsilon_t\lambda}{12L_z^2}\right)\right)\left(1 - \exp\left(-\frac{t\epsilon_t}{8}\right)\right) \\
&\ge 1 - \exp\left(-\frac{t\epsilon_t}{8}\right) - k\exp\left(-\frac{t\epsilon_t\lambda}{12L_z^2}\right) - 2k\exp\left(-\frac{t\epsilon_t^2\lambda^2\eta^2}{128k^2\sigma_v^2 L_z^2}\right)
\end{aligned}
$$

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

Proposition 2 (Proposition 3.3.2.)**.** *In the online decision-making model with ϵ-greedy policy, if the Assumptions 3.1.1 and 3.1.2 are satisfied, and ϵ_t is non-increasing, then for any $\eta_1, \eta_2 > 0$, $\eta_2 \ne \frac{\lambda_{min}(\Gamma_0)}{\sqrt{d}}$,*

$$P(\|\hat{\beta}_t - \beta_0\|_1 \le C_\beta) \ge 1 - (P_1 + dP_2)$$

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

$$
\begin{aligned}
P_1 &= \exp\left(-\frac{t\epsilon_t}{8}\right) + k\exp\left(-\frac{t\epsilon_t\lambda}{12L_z^2}\right) + 2k\exp\left(-\frac{t\epsilon_t^2\lambda^2\eta_1^2}{128k^2\sigma_v^2 L_z^2}\right), \\
P_2 &= \exp\left(-\frac{t\epsilon_t}{8}\right) + k\exp\left(-\frac{t\epsilon_t\lambda}{12L_z^2}\right) + 2k\exp\left(-\frac{t\epsilon_t^2\lambda^2\eta_2^2}{128k^2\sigma_u^2 L_z^2}\right)
\end{aligned}
$$

Proof. Proof. On the events that $\|\hat{\Gamma} - \Gamma_0\|_1 \le \eta_2$ and $\|\hat{\vartheta} - \vartheta_0\|_1 \le \eta_1$, we can prove the upper bound of $\|\hat{\beta} - \beta_0\|_1$ as the following by utilizing triangle inequalities.

$$\|\hat{\Gamma}(\hat{\beta} - \beta_0)\|_1 - \|(\hat{\Gamma} - \Gamma_0)\beta_0\|_1 \le \|\hat{\Gamma}(\hat{\beta} - \beta_0) + (\hat{\Gamma} - \Gamma_0)\beta_0\|_1 = \|\hat{\Gamma}\hat{\beta} - \Gamma_0\beta_0\|_1 = \|\hat{\vartheta} - \vartheta_0\|_1 \le \eta_1$$

Hence,

$$\|\hat{\Gamma}(\hat{\beta} - \beta_0)\|_1 \le \eta_1 + \|(\hat{\Gamma} - \Gamma_0)\beta_0\|_1 \le \eta_1 + \|\hat{\Gamma} - \Gamma_0\|_1\|\beta_0\|_1 \le \eta_1 + \eta_2\|\beta_0\|_1 \tag{A.5}$$

Combining inequality (A.5), Lemma A.0.5 and Lemma A.0.6, we complete the proof by the

Boole's inequality(or the union bound).

□

Theorem 2 (Theorem 3.3.1)**.** *If Assumptions 3.1.1 and 3.1.2 are satisfied, then*

$$\sqrt{t}(\hat{\vartheta}_t - \vartheta_0) \xrightarrow{d} \mathcal{N}_k(0, \mathbb{E}[v_t^2](\int zz' d\mathcal{P}_z)^{-1})$$

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

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

$$\sqrt{t}(\hat{\vartheta}_t - \vartheta_0) = (\frac{1}{t}\sum_{s=1}^{t} z_s z_s')^{-1}(\frac{1}{\sqrt{t}}\sum_{s=1}^{t} z_s v_s)$$

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

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

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

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

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

where v_s, $s = 1, \cdots, t$ are i.i.d. from $\mathcal{P}_v$. Notice that $v_1^2 I\{v_1^2 > \frac{\zeta^2 t}{\|q\|_2^2 L_z^2}\}$ is dominated by v_1^2 with $\mathbb{E}[v_1^2] \leq (\|\beta_0\|_1 + 1)^2$. Also, $v_1^2 I\{v_1^2 > \frac{\zeta^2 t}{\|q\|_2^2 L_z^2}\}$ converges to 0 almost surely when $t \to \infty$. Thus,

applying dominated convergence theorem, we can get, as $t \to \infty$,

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

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

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

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

$$\begin{aligned}
&\sum_{s=1}^{t} \mathbb{E}[\frac{1}{t}(q'z_s)^2 v_s^2|\mathcal{F}_{t,s-1}] \\
&= \frac{\mathbb{E}[v_1^2]}{t} \sum_{s=1}^{t} \mathbb{E}[(q'z_s)^2|\mathcal{F}_{s-1}] \\
&\xrightarrow{p} \mathbb{E}[v_1^2] q'(\int z_s z_s' d\mathcal{P}_z) q
\end{aligned} \tag{A.6}$$

Applying Martingale Limit Theorem, we have

$$\frac{1}{\sqrt{t}} \sum_{s=1}^{t} z_s v_s \xrightarrow{d} \mathcal{N}_k(0, \mathbb{E}[v_1^2] \int z_s z_s' d\mathcal{P}_z) \tag{A.7}$$

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

$$\frac{1}{t} \sum_{s=1}^{t} (q'z_s z_s' q - \mathbb{E}[q'z_s z_s' q|\mathcal{F}_{s-1}]) \xrightarrow{p} 0 \tag{A.8}$$

From the result in (A.6) and (A.8), we have $\frac{1}{t}\sum_{s=1}^{t} q'z_s z_s' q \xrightarrow{p} q'(\int z_s z_s' d\mathcal{P}_z)q$. Applying

Lemma 6 in Chen et al. (2021) and Continuous Mapping Theorem, we obtain

$$(\frac{1}{t}\sum_{s=1}^{t} z_s z_s')^{-1} \xrightarrow{p} (\int z_s z_s' d\mathcal{P}_z)^{-1} \tag{A.9}$$

Combining the results in (A.7) and (A.9), we complete the proof.

□

Proof of "consistency of the variance estimator" in Theorem 3.3.2

Proof. Proof. We want to show that

$$\sum_{s=1}^{t} \hat{v}_s^2 (\hat{\Gamma}_t' \sum_{s=1}^{t} z_s z_s' \hat{\Gamma}_t)^{-1} \xrightarrow{p} S \tag{A.10}$$

By rewriting the squared error term in (A.10), we have

$$\frac{1}{t}\sum_{s=1}^{t} \hat{v}_s^2 = \frac{1}{t}\sum_{s=1}^{t}((\vartheta_0 - \hat{\vartheta}_t)' z_s + v_s)^2 = \frac{1}{t}\sum_{s=1}^{t}((\vartheta_0 - \hat{\vartheta}_t)' z_s)^2 + \frac{2}{t}\sum_{s=1}^{t}(\vartheta_0 - \hat{\vartheta}_t)' z_s v_s + \frac{1}{t}\sum_{s=1}^{t} v_s^2 \tag{A.11}$$

Now we analyze the asymptotic properties for the three terms in (A.11). Notice that by Proposition 3.3.1, the first term in (A.11) can be written as

$$\frac{1}{t}\sum_{s=1}^{t}((\vartheta_0 - \hat{\vartheta}_t)' z_s)^2 \leq \frac{1}{t}\sum_{s=1}^{t} \|\vartheta_0 - \hat{\vartheta}_t\|_2^2 \|z_s\|_2^2 \leq L_z^2 \|\vartheta_0 - \hat{\vartheta}_t\|_1^2 \xrightarrow{p} 0$$

For the second term, by Proposition 3.3.1 and Lemma A.0.9,

$$(\vartheta_0 - \hat{\vartheta}_t)' \frac{2}{t}\sum_{s=1}^{t} z_s v_s \xrightarrow{p} 0$$

For the third term, we apply the weak law of large numbers and obtain

$$\frac{1}{t}\sum_{s=1}^{t} v_s^2 \xrightarrow{p} \mathbb{E}[v_s^2]$$

Combining the results above and Proposition 3.3.1 with continuous mapping theorem, we complete the proof. □

A.0.7 Auxiliary Lemmas

Lemma A.0.7. For each t, we can explicitly upper bound the following terms,

$$W_t \leq X_t' Z_t (Z_t' Z_t)^{-1} Z_t' X_t + \gamma_x I, \quad \hat{X}_t \leq P_{Z_t} X_t, \quad P_{Z_t} = Z_t (Z_t' Z_t)^{-1} Z_t'$$
$$\log(\det(W_t)) \leq d \log(\frac{2TkL_z^2}{d}\|\Gamma_0\|_F^2 + 4kT\log(\frac{6Td}{\delta})\log(\frac{T+k\gamma_z}{k\gamma_z}) + \gamma_x), \text{w.p } \geq (1-\delta/(6T))$$
$$\log(\det(U_t)) \leq k \log\left(TL_z^2 + \gamma_z\right)$$

Proof. Proof. First, we have

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

Following the above result, we obtain the upper bound of W_t by $W_t = \hat{X}_t' \hat{X}_t + \gamma_x I$. Second, we have

$$\begin{aligned}
\log(\det(W_t)) &\leq d \log\left(\frac{1}{d}\text{trace}(W_t)\right) \\
&= d \log(\frac{1}{d}(\text{trace}(\hat{X}_t' \hat{X}_t) + \text{trace}(\gamma_x I))) \\
&= d \log(\frac{1}{d}(\sum_{s=1}^{t} \text{trace}(\hat{x}_s \hat{x}_s') + \gamma_x d)) \\
&= d \log(\frac{1}{d}\sum_{s=1}^{t} \|\hat{x}_s\|_2^2 + \gamma_x)
\end{aligned}$$

Now from the triangle inequality,

$$\|\hat{x}_t\|_2 \leq \|\Gamma_0' z_t\|_2 + \|(\hat{\Gamma}_t - \Gamma_0)' z_t\|_2$$

We first prove the upper bound for $\|\Gamma_0' z_t\|_2$ and then for $\|(\hat{\Gamma}_t - \Gamma_0)' z_t\|_2$ in the following.

$$\|\Gamma_0' z_t\|_2 = \sqrt{\sum_{j=1}^{d}(\sum_{i=1}^{k}\Gamma_0^{(i)(j)} z_t^{(i)})^2} \leq \sqrt{\sum_{j=1}^{d}(\sum_{i=1}^{k}\Gamma_0^{(i)(j)})^2 L_z^2} \leq \sqrt{\sum_{j=1}^{d}\sum_{i=1}^{k}(\Gamma_0^{(i)(j)})^2 k L_z^2} = \sqrt{k} L_z \|\Gamma_0\|_F$$

where the second inequality is from the Cauchy–Schwarz inequality. To prove an upper bound for $\|(\hat{\Gamma}_t - \Gamma_0)' z_t\|_2$, we first prove an upper bound of $q'(\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)})$ for any $q \in \mathbb{R}^k$ and then let $q = z_t$ as follows.

$$\begin{aligned}
\hat{\Gamma}_t - \Gamma_0 &= (\gamma_z I + Z_t' Z_t)^{-1} Z_t' X_t - \Gamma_0 \\
&= (\gamma_z I + Z_t' Z_t)^{-1} Z_t' (Z_t \Gamma_0 + \tilde{U}_t) - \Gamma_0 \\
&= ((\gamma_z I + Z_t' Z_t)^{-1} Z_t' Z_t - I)\Gamma_0 + (\gamma_z I + Z_t' Z_t)^{-1} Z_t' \tilde{U}_t \\
&\leq (\gamma_z I + Z_t' Z_t)^{-1} Z_t' \tilde{U}_t = M_t \tilde{U}_t
\end{aligned}$$

where we denote $M_t = (\gamma_z I + Z_t' Z_t)^{-1} Z_t' \in \mathbb{R}^{k \times t}$ and $\tilde{U}_t = X_t - Z_t \Gamma_0$. Hence, we have $\hat{\Gamma}_t^{(i)} - \Gamma_0^{(i)} \leq M_t \tilde{U}_t^{(i)}$.

Let $\tilde{r} = M_t' q \in \mathbb{R}^t$, $\tilde{\lambda} = \frac{\delta}{\|\tilde{r}\|_2^2}$. We have

$$\begin{aligned}
\mathbb{P}(q' M_t \tilde{U}_t^{(i)} > \delta) &\leq \exp(-\tilde{\lambda}\delta)\mathbb{E}[\exp(\tilde{\lambda}\sum_{s=1}^{t} \tilde{r}^{(s)} \tilde{U}_t^{(i)(s)})] \\
&= \exp(-\tilde{\lambda}\delta) \prod_{s=1}^{t} \mathbb{E}[\exp(\tilde{\lambda}\tilde{r}^{(s)} \tilde{U}_t^{(i)(s)})] \\
&\leq \exp(-\tilde{\lambda}\delta) \prod_{s=1}^{t} \exp(\frac{\tilde{\lambda}^2 (\tilde{r}^{(s)})^2}{2}) \\
&= \exp(-\tilde{\lambda}\delta) \exp(\frac{\tilde{\lambda}^2}{2}\|\tilde{r}\|_2^2) \\
&= \exp(-\frac{\delta^2}{2\|\tilde{r}\|_2^2})
\end{aligned}$$

where the first inequality is due to the Chernoff Bound for any $\tilde{\lambda} \in \mathbb{R}$; the second equality and inequality are due to the independency between $\tilde{U}_t^{(i)(s)}$ and the subgausianity of $\tilde{U}_t^{(i)(s)}$ for $s \in \{1, 2, \cdots, t\}$, respectively. Meanwhile, we have

$$\|\tilde{r}\|_2^2 = \tilde{r}'\tilde{r} = q' M_t M_t' q \leq q'(\gamma_z I + Z_t' Z_t)^{-1} q = \|q\|^2_{(\gamma_z I + Z_t' Z_t)^{-1}}$$

Letting $q = z_t$, we have the following with a high probability at least $1 - \delta$,

$$z_t'(\hat{\Gamma}^{(i)} - \Gamma_0^{(i)}) \leq z_t' M_t \tilde{U}_t^{(i)} \leq \sqrt{2\log(\frac{1}{\delta})}\|\tilde{r}\|_2 \leq \sqrt{2\log(\frac{1}{\delta})}\|z_t\|_{(\gamma_z I + Z_t' Z_t)^{-1}}$$

Therefore, with a high probability at least $1 - \delta$,

$$\sum_{s=1}^{t} \|(\hat{\Gamma}_s - \Gamma_0)' z_s\|_2 = \sum_{s=1}^{t} \sqrt{\sum_{i=1}^{d} (z_s'(\hat{\Gamma}_s^{(i)} - \Gamma_0^{(i)}))^2} \leq \sum_{s=1}^{t} \sqrt{2d\log(\frac{d}{\delta})}\|z_s\|_{U_{s+1}^{-1}} \leq \sqrt{2d\log(\frac{d}{\delta}) tk\log(\frac{t + k\gamma_z}{k\gamma_z})}$$

where the last inequality is from the elliptical potential lemma (Carpentier et al., 2020).

Combining the results above, we can obtain, with a high probability at least $1 - \delta/(6T)$,

$$\sum_{s=1}^{t} \|\hat{x}_t\|_2^2 \leq 2\sum_{s=1}^{t} \|\Gamma_0' z_s\|_2^2 + 2\sum_{s=1}^{t} \|(\hat{\Gamma}_t - \Gamma_0)' z_t\|_2^2 \leq 2T(\sqrt{k}L_z\|\Gamma_0\|_F)^2 + 2(\sum_{s=1}^{t} \|(\hat{\Gamma}_t - \Gamma_0)' z_t\|_2)^2$$
$$\leq 2TkL_z^2\|\Gamma_0\|_F^2 + 4dkT\log(\frac{6Td}{\delta})\log(\frac{T + k\gamma_z}{k\gamma_z})$$

which completes the proof for the upper bound of $\log(\det(W_t))$.

Lastly, we have

$$\begin{aligned}
\log(\det(U_t)) &\leq k\log\left(\frac{1}{k}\text{trace}(U_t)\right) \\
&= k\log\left(\frac{1}{k}\text{trace}(Z_t'Z_t + \gamma_z I)\right) \\
&= k\log\left(\frac{1}{k}\text{trace}(Z_t'Z_t) + \gamma_z\right) \\
&= k\log\left(\frac{1}{k}\|Z_t\|_F^2 + \gamma_z\right) \\
&\leq k\log\left(tL_z^2 + \gamma_z\right)
\end{aligned}$$

□

Lemma A.0.8. Consider a sequence of vectors $(x_t)_{t=1}^T$, $x_t \in \mathbb{R}^d$, and assume that $\|x_t\|_2 \leq a$ for all t. Let $V_t = \lambda I + \sum_{s=1}^t x_s x_s'$ for some $\lambda > 0$. Then, we will have that $\|x_t\|_{V_{t-1}^{-1}} > b$ at most

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

times.

Lemma A.0.8 is directly from lemma 6.2 in Wagenmaker et al. (2021).

Proof of "$v_{t,a}$ is $(\|\beta_0\|_1 + 1)$-subgaussian"

Proof. Proof. First we prove that $\beta_0^{(i)} u_{t,a}^{(i)}$ is $|\beta_0^{(i)}|$-subgaussian as follows. Because each element in

$u_{t,a}$ is 1-subgaussian, for any $i \in \{1, \ldots, d\}$, $\lambda \in \mathbb{R}$,

$$\mathbb{E}[exp(\lambda(\beta_0^{(i)} u_{t,a}^{(i)}))] \leq exp(\frac{\lambda^2 |\beta_0^{(i)}|^2}{2})$$

Then we prove that $\sum_{i=1}^{d} \beta_0^{(i)} u_{t,a}^{(i)} + e_{t,a}$ is $(\|\beta_0\|_1 + 1)$-subgaussian as follows. Using Holder's inequality, for any sequence $p_i > 1, i \in \{1, ..., d+1\}$, s.t., $\sum_{i=1}^{d+1} \frac{1}{p_i} = 1$,

$$\begin{aligned}
\mathbb{E}[exp(\lambda v_{t,a}] &= \mathbb{E}[exp(\lambda(\sum_{i=1}^{d} u_{t,a}^{(i)} \beta_0^{(i)} + e_{t,a}))] \\
&\leq \prod_{i=1}^{d} \mathbb{E}[exp(\lambda p_i u^{(i)} \beta_0^{(i)})]^{\frac{1}{p_i}} \mathbb{E}[exp(\lambda p_{d+1} e_{t,a})]^{\frac{1}{p_{d+1}}} \\
&\leq \prod_{i=1}^{d} exp(\frac{\lambda^2 p_i \beta_0^{(i)^2}}{2}) exp(\frac{\lambda^2 p_{d+1}}{2})
\end{aligned} \tag{A.12}$$

where the last inequality is because that $u_{t,a}^{(i)}$ and $e_{t,a}$ are 1-subgaussian with mean zero.

By letting $p_i = \frac{\|\beta_0\|_1 + 1}{|\beta_0^{(i)}|}, \forall i \in \{1, ..., d\}$ and $p_{d+1} = \|\beta_0\|_1 + 1$, we have

$$\prod_{i=1}^{d} exp(\frac{\lambda^2 p_i \beta_0^{(i)^2}}{2}) exp(\frac{\lambda^2 p_{d+1}}{2}) = exp(\frac{\lambda^2 (\|\beta_0\|_1 + 1)^2}{2}) \tag{A.13}$$

Combining (A.12) and (A.13), we complete the proof.

□

Proof of "$\hat{\vartheta} = \hat{\Gamma}_{Z,X} \hat{\beta}_{\hat{X},Y}$"

Proof. Proof. We want to prove the following equation

$$\hat{\Gamma}_{Z,X} \hat{\beta}_{\hat{X},Y} = \hat{\vartheta} \tag{A.14}$$

Multiplying $((Z\hat{\Gamma}_{Z,X})' Z\hat{\Gamma}_{Z,X})^{-1} (Z\hat{\Gamma}_{Z,X})' Z$ to both sides of Equation (A.14), we can obtain the exact form of two-stage least squares estimator which we defined before. Therefore, we can com-

plete the proof.

□

Lemma A.0.9. Suppose $\{\mathcal{F}_t : t = 1, \ldots, T\}$ is an increasing filtration of σ-fields. Let $\{\tilde{W}_t : t = 1, \ldots, T\}$ be a sequence of random variables such that $\tilde{W}_t$ is $\mathcal{F}_{t-1}$ measurable and $|\tilde{W}_t| \leq L_{\tilde{w}}$ almost surely for all t. Let $\{\tilde{e}_t : t = 1, \ldots, T\}$ be independent $\sigma_{\tilde{e}}$-subgaussian, and $\tilde{e}_t \perp \mathcal{F}_{t-1}$ for all t. Let $\tilde{S} = \{s_1, \ldots, s_{|\tilde{S}|}\} \subseteq \{1, \ldots, T\}$ be an index set where $|\tilde{S}|$ is the number of elements in $\tilde{S}$. Then for $\kappa > 0$,

$$P(\sum_{s \in \tilde{S}} \tilde{W}_s \tilde{e}_s \geq \kappa) \leq \exp(-\frac{\kappa^2}{2|\tilde{S}|\sigma_{\tilde{e}}^2 L_{\tilde{w}}^2}) \tag{A.15}$$

Lemma A.0.9 is from Lemma 1 in Chen et al. (2021).

Appendix B

APPENDIX TO DEBIASING ML/AI GENERATED REGRESSORS

B.1 Tables and Figures

Table B.1: Estimation Results Comparison with ML-predicted Ratings

	DML(Full)	DML(Labeled)	TwoStep DML	GMM DML
		No. Estimators in XGBoost = 4		
Avg(Bias)	-0.0089	0.0012	-0.0011	-0.0008
Std	0.0040	0.0043	0.0031	0.0028
		No. Estimators in XGBoost = 5		
Avg(Bias)	-0.0068	0.0012	-0.0008	-0.0006
Std	0.0034	0.0043	0.0029	0.0026
		No. Estimators in XGBoost = 6		
Avg(Bias)	-0.0053	0.0012	-0.0006	-0.0004
Std	0.0030	0.0043	0.0026	0.0024
		No. Estimators in XGBoost = 7		
Avg(Bias)	-0.0043	0.0012	-0.0006	-0.0004
Std	0.0028	0.0043	0.0025	0.0023

B.2 Literature Comparison

Comparison between this essay and existing literature:

1. Solutions for continuous independent variable
2. Solutions for binary independent variable
3. Solutions for binary dependent variable
4. Closed-form solutions
5. Can solutions correct the estimation inconsistency?

6. Can solutions be directly applied to semi-parametric models?

7. Do solutions include inference for semi-parametric models?

8. Do solutions utilize cross-fitting methods to prevent overfitting?

	Yang et al. (2018)	Qiao and Huang (2021)	Wei and Malik (2022)	Fong and Tyler (2021)	Allon et al. (2023)	This essay
1	Yes	No	No	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes	Yes	Yes
3	No	Yes	No	No	Yes	No
4	No(simulation)	Yes	No	Yes	Yes	Yes
5	Partially	Completely	Partially	Completely	Completely	Completely
6	Yes	No	Models estimable by MLE or GMM	No	No	Yes
7	No	No	No	No	No	Yes
8	No	No	No	No	No	Yes

B.3 Additional Assumptions

We consolidate the standard regularization assumptions required for Theorems 1-4 below. These assumptions are readily satisfied in typical empirical applications.

First, for Theorems 1 and 2, we have the following additional assumptions: 1(c) $E[X^*(X^*)']$ and $E[XX']$ are positive definite; 1(d) $E[Y^2] < \infty$, $E[\|X^*\|^2] < \infty$, $E[\|X\|^2] < \infty$, $E[X^*X'] < \infty$; 1(e) $E[Y^4] < \infty$, $E[\|X^*\|^4] < \infty$; 1(f) $E[X^*(X^*)'\varepsilon^2] < \infty$, $E[(\xi'\beta_o)^2XX'] < \infty$, $E[\varepsilon\xi'\beta_oXX'] < \infty$; 1(g) $\beta_o \in$ interior$(\mathcal{B})$ and $\mathcal{B} \subset \mathbb{R}^{d_D+d_Z}$ is compact.

Assumptions 1(c)-1(g) are commonly adopted regularity assumptions for consistency and asymptotic normality of the TSLS estimator and GMM estimator. Similar assumptions are used in Chernozhukov et al. (2018).

Second, for Theorems 3 and 4, we have the following extra assumptions: For all probability laws $P \in \mathcal{P}$ for $W = (Y, D^*, D, Z)$, $q > 2$, the following conditions hold:2(c) $\|Y\|_{P,q} + \|D^*\|_{P,q} + \|D\|_{P,q} \leq C$; 2(d) $\|\varepsilon u\|_{P,2} \geq c^2$, $\|\varepsilon e\|_{P,2} \geq c^2$, $\|ue\|_{P,2} \geq c^2$, $E_P[u^2] \geq c$, and $E_P[e^2] \geq c$; 2(e) $\|E_P[\varepsilon^2|Z]\|_{P,\infty} \leq C$, $\|E_P[u^2|Z]\|_{P,\infty} \leq C$, $\|E_P[v^2|Z]\|_{P,\infty} \leq C$, and $\|E_P[e^2|Z]\|_{P,\infty} \leq C$; and 2(f) given a random subset I of $[n]$ of size $N_I = \frac{n}{k}$, the nuisance parameter estimators $\hat{\eta} := (\hat{l}, \hat{r}, \hat{m}) = \hat{\eta}(W_i)_{i\in I^c}$ obeys the following conditions for all $N_I \geq 1$. Let $\{\Delta_n\}_{n\geq 1}$ and $\{\delta_n\}_{n\geq 1}$ be some sequences of positive constants converging to zero. With P-probability no less than $1 - \Delta_n$,

$\|\hat{\eta} - \eta_o\|_{P,q} \leq C$, $\|\hat{\eta} - \eta_o\|_{P,2} \leq \delta_n$. For the score ψ in (4.14), $\|\hat{r} - r_o\|_{P,2} \times \|\hat{s} - s_o\|_{P,2} \leq \delta_n n^{-1/2}$, and $\|\hat{r} - r_o\|_{P,2} \times (\|\hat{r} - r_o\|_{P,2} + \|\hat{l} - l_o\|_{P,2}) \leq \delta_n n^{-1/2}$. For the score ψ_1 in (4.17), $\|\hat{m} - m_o\|_{P,2} \times (\|\hat{m} - m_o\|_{P,2} + \|\hat{r} - r_o\|_{P,2} + \|\hat{l} - l_o\|_{P,2}) \leq \delta_n n^{-1/2}$; 2(g) $\theta_o \in$ interior(Θ), and $\Theta \subset \mathbb{R}$ is compact.

Assumptions 2(c)-2(g) represent the standard regularity assumptions necessary to guarantee a finite asymptotic variance. Specifically, Assumption 2(f) imposes an assumption on the convergence rate in estimating the function using the ML algorithm. Unlike the typical root-n convergence required for establishing asymptotic variance, this assumption allows the ML algorithm to converge at a rate slower than root-n.

Now we provide a detailed explanation for how Assumption 2 is satisfied in empirical application, especially in the context of Section 4.4.

Assumption 2(a) states that the data generating process follows a standard partially linear model. This assumption is valid in our context, where we aim to study the linear impact of review sentiment on review helpfulness while controlling for all other possible confounding factors. The partially linear structure allows us to isolate the direct effect of sentiment while accounting for the high-dimensional features of text data, including higher-order terms and interaction effects, without imposing restrictive functional form assumptions on these complex relationships.

Assumption 2(b) states that the L_q norms of the random variables Y_i, D_i^*, D_i are bounded by a constant C. This assumption is satisfied in our dataset since these variables have finite ranges. These empirical constraints ensure that the moment conditions required by the assumption are met.

Assumption 2(c) ensures that the variances and covariances of the error terms are bounded away from zero. In our review helpfulness context, this is naturally satisfied because there will always be unexplained variation in both review helpfulness and sentiment that cannot be fully captured by observable features. The presence of these unobservable factors—such as reader preferences and idiosyncratic elements of writing style—guarantees that the error terms maintain sufficient variation for identification, making this assumption realistic in our empirical setting.

Assumption 2(d) places uniform bounds on the conditional second moments of the error terms. In our context, this condition is naturally satisfied because the variables Y_i, D_i, and D_i^* all have

finite ranges. Since these variables are bounded, their corresponding error terms necessarily have bounded conditional variances regardless of the conditioning set Z.

Assumption 2(e) specifies the accuracy requirements for estimating the nuisance parameters $\hat{\eta} := (\hat{l}, \hat{r}, \hat{m})$. This assumption requires that with high probability $(1 - \Delta_n)$, these estimators are bounded and converge to the true parameters at specific rates. Particularly, the product of estimation errors must be smaller than $\delta_n n^{-1/2}$. These conditions are indeed mild and can be satisfied by many modern machine learning methods, as established in Chernozhukov et al. (2018). For instance, lasso, random forests, neural networks, and boosting algorithms can achieve these convergence rates under standard regularity conditions. The beauty of these requirements is that they allow for individual nuisance components to be estimated at slower rates (e.g., $n^{-1/4}$) as long as their product achieves the faster $n^{-1/2}$ rate. This "product rate" condition is central to the double/debiased machine learning framework, enabling us to handle high-dimensional controls while maintaining $\sqrt{n}$-consistency and asymptotic normality of our target parameter. In our implementation, we employ random forest and gradient boosting that have been shown to satisfy these theoretical guarantees.

For Assumption 2(f), this condition is naturally satisfied in our empirical context. Our parameter of interest represents the effect of review sentiment on review helpfulness, which is expected to be finite and bounded. For estimation purposes, we work with a parameter space that contains the true parameter value well within its boundaries, not at the edges. This is a standard regularity assumption in statistical estimation that ensures our estimator exists and is uniquely defined. Such an assumption is routinely satisfied in applications like ours and poses no practical restrictions on our estimation procedure.

B.4 Theoretical Results of Linear Estimators

Theorem B.4.1 (Consistency)**.** *For the linear model described in (4.3), and under Assumption 4.2.1, any estimator $\hat{\beta} \in \{\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}\}$ is a consistent estimator for β_o, i.e.,*

$$\hat{\beta} \xrightarrow{p} \beta_o, \quad \text{as } n \to \infty. \tag{B.1}$$

Theorem B.4.1 shows that the provided estimators converge in probability to the true estimator as the sample size increases and ensures the asymptotic accuracy for the provided estimators. This is important in business as consistent estimators enable better downstream decision-making. We provide the proof of Theorem B.4.1 in Appendix (B.7).

Theorem B.4.2 (Asymptotic Normality)**.** *Under Assumption 4.2.1, any estimator* $\hat{\beta} \in \{\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}\}$ *is asymptotically normal, i.e.,*

$$\sqrt{n_l}(\hat{\beta} - \beta_o) \xrightarrow{d} \mathcal{N}(0, V), \tag{B.2}$$

where $X_i^* \coloneqq ((D_i^*)', Z_i')'$ *is an* $\mathbb{R}^{d_D+d_Z}$*-valued random variable. where we denote the variance* V *for* $\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}$ *as* $V_{ols}, V_{tsls}, V_{gmm}$*, and the expressions are given as below.*

$$V_{ols} = (E[X^*(X^*)'])^{-1}\Xi_{ols}(E[X^*(X^*)'])^{-1}, \text{ where } \Xi_{ols} = E[X^*(X^*)'\varepsilon^2].$$

$$V_{tsls} = ([E[XX']\Gamma_o]^{-1}\Xi_{tsls}[\Gamma_o'E[XX']]^{-1}, \text{ where}$$

$$\Xi_{tsls} = \lambda_l E[(\varepsilon + \xi'\beta_o)^2 XX'] + E[(\xi'\beta_o)^2 XX'] - 2\lambda_l E[(\varepsilon + \xi'\beta_o)\xi'\beta_o XX'].$$

$$V_{gmm} = (\triangle'\Xi^{-1}\triangle)^{-1}, \text{ where } \triangle = \begin{pmatrix} -E[X^*(X^*)'] \\ -E[X(X^*)'] \end{pmatrix}, \text{ and } \Xi = \begin{pmatrix} \Xi_{ols} & \Xi_{ols,tsls} \\ \Xi_{ols,tsls} & \Xi_{tsls} \end{pmatrix} \text{ with}$$

$$\Xi_{ols,tsls} = \lambda_l E[\varepsilon\,(\varepsilon + \xi'\beta_o)\, X^*X'] - E[\varepsilon\xi'\beta_o X^*X'].$$

Theorem B.4.2 facilitates statistical inference by utilizing the properties of the normal distribution. The following remark provides further details on conducting inference using confidence intervals. The proof of the Theorem B.4.2 is provided in Appendix (B.8).

Remark B.4.1 (Confidence Intervals). Based on Theorem B.4.2, we are able to construct the confidence intervals for the scalar parameter $t'\beta_o$ for some $(d_D + d_Z) \times 1$ vector t. The confidence interval for each estimator $\hat{\beta} \in \{\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}\}$:

$$CI := \left[t'\hat{\beta} \pm \Phi^{-1}(1 - \alpha/2)\left(t'\hat{V}t\right)^{-1/2}/\sqrt{n_l}\right]$$

obeys $\lim_{n_l\to\infty}\Pr\left(\beta_o\in CI\right)=1-\alpha$, where $\hat{V}$ represents a consistent estimator of V, Φ^{-1} represents the inverse of the CDF of a standard normal distribution. Potential choices for consistent estimators of V_{ols}, V_{tsls}, and V_{gmm} are provided by

$$\hat{V}_{ols}=\left(\frac{1}{n_l}\sum_{i=1}^{n}\iota_i\left[X_i^*\left(X_i^*\right)'\right]\right)^{-1}\frac{1}{n_l}\sum_{i=1}^{n}\iota_i\left[X_i^*\left(X_i^*\right)'\left(Y_i-\hat{\beta}_{ols}'X_i^*\right)^2\right]\left(\frac{1}{n_l}\sum_{i=1}^{n}\iota_i\left[X_i^*\left(X_i^*\right)'\right]\right)^{-1}.$$

$$\hat{V}_{tsls}=\left(\frac{1}{n}\sum_{i=1}^{n}\left[X_iX_i'\right]\hat{\Gamma}\right)^{-1}\hat{\Xi}_{tsls}\left(\hat{\Gamma}'\frac{1}{n}\sum_{i=1}^{n}\left[X_iX_i'\right]\right)^{-1},\text{ where}$$

$$\hat{\Xi}_{tsls}=\frac{n_l}{n}\sum_{i=1}^{n}\iota_i\left[\left(Y_i-\hat{\beta}_{tsls}'X_i^*+\left(X_i^*-\hat{\Gamma}'X_i\right)'\beta_o\right)^2X_iX_i'\right]+\frac{1}{n}\sum_{i=1}^{n}\left[\left(\left(X_i^*-\hat{\Gamma}'X_i\right)'\beta_o\right)^2X_iX_i'\right]$$
$$-2\frac{n_l}{n}\sum_{i=1}^{n}\iota_i\left[\left(Y_i-\hat{\beta}_{tsls}'X_i^*+\left(X_i^*-\hat{\Gamma}'X_i\right)'\beta_o\right)\left(X_i^*-\hat{\Gamma}'X_i\right)'\beta_oX_iX_i'\right].$$

$$\hat{V}_{gmm}=\left(\hat{\triangle}'\hat{\Xi}^{-1}\hat{\triangle}\right)^{-1},\text{ where }\hat{\triangle}=\begin{pmatrix}-\frac{1}{n_l}\sum_{i=1}^{n}\iota_i\left[X_i^*\left(X_i^*\right)'\right]\\-\frac{1}{n_l}\sum_{i=1}^{n}\iota_i\left[X_i\left(X_i^*\right)'\right]\end{pmatrix},$$

and a consistent estimator for Ξ is given by

$$\hat{\Xi}=\frac{1}{n_l}\sum_{i=1}^{n}h_ih_i',\quad\text{with }h_i=\begin{pmatrix}\iota_iX_i^*(Y_i-(X_i^*)'\tilde{\beta})\\\hat{\lambda}_lX_i(Y_i-(\hat{\Gamma}'X_i)'\tilde{\beta})-\iota_iX_i(X_i^*-\hat{\Gamma}'X_i)'\tilde{\beta}\end{pmatrix}.\tag{B.3}$$

Here $\hat{\lambda}_l=\frac{n_l}{n}$ and $\tilde{\beta}$ is some consistent estimator of β_o (such as the OLS estimator on the labeled set).

B.5 Confidence Intervals for DML Estimators

The confidence interval for each estimator $\hat{\theta}\in\{\hat{\theta}_{dml},\hat{\theta}_{tsdml},\hat{\theta}_{gdml}\}$:

$$CI:=\left[\hat{\theta}\pm\Phi^{-1}(1-\alpha/2)\left(\hat{V}_{plr}\right)^{-1/2}/\sqrt{n_l}\right]$$

obeys $\lim_{n_l\to\infty}\Pr\left(\theta_o\in CI\right)=1-\alpha$, where $\hat{V}_{plr}$ represents a consistent estimator of V_{plr}, Φ^{-1} represents the inverse of the CDF of a standard normal distribution. Potential choices for consistent

estimators of V_{dml}, V_{tsdml}, and V_{gdml} are provided by

$$\hat{V}_{dml} = (\frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}[(D^* - \hat{r}_k(Z))^2])^{-1}\frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}[(Y - \hat{l}_k(Z) - \hat{\theta}_{dml}(D^* - \hat{r}_k(Z)))^2(D^* - \hat{r}_k(Z))^2]$$
$$\times (\frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}[(D^* - \hat{r}_k(Z))^2])^{-1}.$$

$\hat{V}_{tsdml} = \hat{J}^{-1}\hat{\Xi}_{tsdml}\hat{J}^{-1}$, where $\hat{J} = \frac{1}{K}\sum_{k=1}^{K}\left[\hat{\Gamma}\left(D - \hat{m}_k\left(Z\right)\right)^2\right]$, and

$$\hat{\Xi}_{tsdml} = \lambda_l\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[(Y - \hat{l}_k(Z) - \hat{\theta}\hat{\Gamma}(D - \hat{m}_k(Z)))^2(D - \hat{m}_k(Z))^2]$$
$$+ \frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}[(D^* - \hat{r}_k(Z) - \hat{\Gamma}(D - \hat{m}_k(Z)))^2(D - \hat{m}_k(Z))^2\hat{\theta}^2]$$
$$-2\lambda_l\frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}[(Y - \hat{l}_k(Z) - \hat{\theta}\hat{\Gamma}(D - \hat{m}_k(Z)))(D^* - \hat{r}_k(Z) - \hat{\Gamma}(D - \hat{m}_k(Z)))(D - \hat{m}_k(Z))^2\hat{\theta}].$$

$\hat{V}_{gdml} = (\hat{\triangle}'_{dml}W_{dml}\hat{\triangle}_{dml})^{-1}$, and $\hat{\triangle}_{dml} = \begin{pmatrix} -\frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}(D^* - \hat{r}_k(Z))^2 \\ -\frac{1}{K}\sum_{k=1}^{K}\hat{E}_{k,l}(D^* - \hat{r}_k(Z))(D - \hat{m}_k(Z)) \end{pmatrix}$.

B.6 Proof of the equivalence between Estimator 2 and the estimator in the first example in Qiao and Huang (2021)

Proof. Proof. We abuse the notation of X, X^* in this proof to be the matrices of the observations. Let Y be a vector of the observations. The TSLS estimator:

$$\hat{X^*} = P_X X^*, P_X = X(X'X)^{-1}X'$$
$$(\hat{\theta}, \hat{\beta}')' = (\hat{X^*}'\hat{X^*})^{-1}(\hat{X^*}'Y) = (X'X^*)^{-1}X'Y$$

Their estimator:

$$\begin{aligned}(\hat{\theta}, \hat{\beta}')' &= (I - (X'X)^{-1}X'(X - X^*))^{-1}(X'X)^{-1}X'Y \\ &= ((X'X)^{-1}X'X^*)^{-1}(X'X)^{-1}X'Y \\ &= (X'X^*)^{-1}X'Y\end{aligned}$$

where we assume $X'X^*$ is invertible. □

B.7 Proof of Theorem B.4.1.

Theorem 3 (Theorem 1.)**.** *[Consistency] For the linear model described in (4.3), and under Assumption 4.2.1, any estimator $\hat{\beta} \in \{\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}\}$ is a consistent estimator for β_o, i.e.,*

$$\hat{\beta} \xrightarrow{p} \beta_o, \quad \text{as } n \to \infty. \tag{B.4}$$

Proof. Proof. The consistency of the OLS estimator, GMM estimator is directly from Theorem 6.2.1 in Hansen (2022), Theorem 2.6 in Newey and McFadden (1994), respectively. We provide the proof of the consistency of the TSLS estimator as the following.

By Weak Law of Large Number (WLLN) and Assumption 4.2.1.3, 4.2.1.5,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{X}_i^*(\hat{X}_i^*)' &= \frac{1}{n}\sum_{i=1}^{n}((\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j^* X_j')(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j X_j')^{-1}X_i X_i'(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j X_j')^{-1}(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j (X_j^*)') \\ &= (\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j^* X_j')(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j X_j')^{-1}\frac{1}{n}\sum_{i=1}^{n}X_i X_i'(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j X_j')^{-1}(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j (X_j^*)') \\ &\xrightarrow{p} E[X^*X']E[XX']^{-1}E[X(X^*)'] \end{aligned} \tag{B.5}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{X}_i^* Y_i &= \frac{1}{n}\sum_{i=1}^{n}((\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j^* X_j')(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j X_j')^{-1}X_i(\beta_o' X_i^* + \varepsilon_i)) \\ &= (\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j^* X_j')(\frac{1}{n_l}\sum_{j=1}^{n}\iota_i X_j X_j')^{-1}((\frac{1}{n}\sum_{i=1}^{n}X_i(X_i^*)'\beta_o) + \frac{1}{n}\sum_{i=1}^{n}X_i\varepsilon_i) \\ &\xrightarrow{p} E[X^*X']E[XX']^{-1}E[X(X^*)']\beta_o \end{aligned} \tag{B.6}$$

By Slutsky Theorem,

$$\hat{\beta}_{tsls} \xrightarrow{p} (E[X^*X']E[XX']^{-1}E[X(X^*)'])^{-1}E[X^*X']E[XX']^{-1}E[X(X^*)']\beta_o = \beta_o \tag{B.7}$$

□

B.8 Proof of Theorem B.4.2.

Theorem 4 (Theorem 2.)**.** *[Asymptotic Normality] Under Assumption 4.2.1, any estimator* $\hat{\beta} \in \{\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}\}$ *is asymptotically normal, i.e.,*

$$\sqrt{n_l}(\hat{\beta} - \beta_o) \xrightarrow{d} \mathcal{N}(0, V) \tag{B.8}$$

where we denote the variance V for $\hat{\beta}_{ols}, \hat{\beta}_{tsls}, \hat{\beta}_{gmm}$ as $V_{ols}, V_{tsls}, V_{gmm}$, and the expressions are given as below.

$$
\begin{aligned}
V_{ols} &= (E[X^*(X^*)'])^{-1}\Xi_{ols}(E[X^*(X^*)'])^{-1}, \\
\Xi_{ols} &= E[X^*(X^*)'\varepsilon^2], \\
V_{tsls} &= ([E[XX']\Gamma_o]^{-1}\Xi_{tsls}[\Gamma_o' E[XX']]^{-1}, \\
\Xi_{tsls} &= \lambda_l E[(\varepsilon + \xi'\beta_o)^2 XX'] + E[(\xi'\beta_o)^2 XX'] - 2\lambda_l E[(\varepsilon + \xi'\beta_o)\xi'\beta_o XX'], \\
V_{gmm} &= (\triangle'\Xi^{-1}\triangle)^{-1}, \\
\Xi &= \begin{pmatrix} \Xi_{ols} & \lambda_l E[\varepsilon\,(\varepsilon + \xi'\beta_o)\, X^*X'] \\ & -E[\varepsilon\xi'\beta_o X^*X'] \\ \lambda_l E[\varepsilon\,(\varepsilon + \xi'\beta_o)\, X(X^*)'] & \Xi_{tsls} \\ -E[\varepsilon\xi'\beta_o X(X^*)'] & \end{pmatrix}, \\
\triangle &= \begin{pmatrix} -E[X^*(X^*)'] \\ -E[X(X^*)'] \end{pmatrix}.
\end{aligned}
$$

Proof. Proof. The asymptotic normality of $\hat{\beta}_{ols}$ is directly from Theorem 6.3.2 in Hansen (2022). We provide the proof of the asymptotic normality of $\hat{\beta}_{tsls}$ as the following.

The moment condition for the TSLS estimator is $E[X(Y - (X^*)'\beta_o)] = 0$. $\hat{\beta}$ satisfies $\hat{g}_2(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^n X_i(Y_i - \hat{X}_i^{*'}\hat{\beta}) = \frac{1}{n}\sum_{i=1}^n X_i(Y_i - (\hat{\Gamma}'X_i)'\hat{\beta}) = 0$. This can be further rewritten as the following,

$$
\begin{aligned}
0 &= \hat{g}_2(\hat{\beta}) \\
&= \frac{1}{n}\sum_{i=1}^n X_i(Y_i - (\hat{\Gamma}'X_i)'\beta_o) + \frac{\partial}{\partial\beta}(\frac{1}{n}\sum_{i=1}^n X_i(Y_i - (\hat{\Gamma}'X_i)'\beta))|_{\beta=\beta_o}(\hat{\beta} - \beta_o) \qquad \text{(B.9)}
\end{aligned}
$$

By rearranging Equation (B.9), we have that

$$\begin{aligned}&\sqrt{n_l}(\hat{\beta}-\beta_o)\\&=\left(\frac{\partial}{\partial\beta}(\frac{1}{n}\sum_{i=1}^{n}X_i(Y_i-(\hat{\Gamma}'X_i)'\beta))|_{\beta=\beta_o}\right)^{-1}\left(-\frac{\sqrt{n_l}}{n}\sum_{i=1}^{n}X_i(Y_i-(\hat{\Gamma}'X_i)'\beta_o\right)\end{aligned}\tag{B.10}$$

In the meantime, by Assumptions 1.1, 1.3, $E[\xi X']=0$ and Hansen (2022), we have $\hat{\Gamma}\xrightarrow{\text{p}}\Gamma_o$, and

$$\sqrt{n_l}(\hat{\Gamma}-\Gamma_o)=\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\iota_i(X_i^*-\Gamma_o'X_i)X_i'E[XX']^{-1}+o_P(1)\tag{B.11}$$

From the law of large number and the continuous mapping theorem, we have

$$(\frac{\partial}{\partial\beta}(\frac{1}{n}\sum_{i=1}^{n}X_i(Y_i-(\hat{\Gamma}'X_i)'\beta))|_{\beta=\beta_o})^{-1}=-(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\hat{\Gamma})^{-1}\xrightarrow{\text{p}}(-E[XX']\Gamma_o)^{-1}\tag{B.12}$$

Further,

$$\begin{aligned}\sqrt{n_l}\hat{g}_2(\beta_o)&=\frac{\sqrt{n_l}}{n}\sum_{i=1}^{n}X_i(Y_i-(\hat{\Gamma}'X_i)'\beta_o)\\&=\frac{\sqrt{n_l}}{n}\sum_{i=1}^{n}[X_i(Y_i-(\Gamma_o'X_i)'\beta_o)-X_iX_i'(\hat{\Gamma}-\Gamma_o)\beta_o]\\&=\frac{\sqrt{n_l}}{n}\sum_{i=1}^{n}[X_i(Y_i-(\Gamma_o'X_i)'\beta_o)]-E[XX']\sqrt{n_l}(\hat{\Gamma}-\Gamma_o)\beta_o+o_P(1)\\&=\frac{\sqrt{n_l}}{n}\sum_{i=1}^{n}X_i(Y_i-(\Gamma_o'X_i)'\beta_o)-\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\iota_iX_i(X_i^*-\Gamma_o'X_i)'\beta_o+o_P(1)\\&=\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}[\frac{n_l}{n}X_i(Y_i-(\Gamma_o'X_i)'\beta_o)-\iota_iX_i(X_i^*-\Gamma_o'X_i)'\beta_o]+o_P(1)\\&=\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}[\frac{n_l}{n}X_i(\varepsilon_i+\xi_i'\beta_o)-\iota_iX_i\xi_i'\beta_o]+o_P(1)\\&=\frac{\sqrt{n_l}}{\sqrt{n}}\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_2(\beta_o)+o_P(1)\end{aligned}\tag{B.13}$$

where we denote $(\varepsilon_i + \xi_i'\beta_o)X_i - \frac{n}{n_l}\iota_i\xi_i'\beta_o X_i \triangleq \tilde{g}_2(\beta_o)$ to simplify the notation.

Notice that

$$\lim_{n_l\to\infty}\frac{n_l}{n}E[\tilde{g}_2(\beta_o)\tilde{g}_2'(\beta_o)] = \lambda_l E[(\varepsilon+\xi'\beta_o)^2XX'] + E[(\xi'\beta_o)^2XX'] - 2\lambda_l E[(\varepsilon+\xi'\beta_o)\xi'\beta_o XX'] \triangleq \Xi_{tsls} \tag{B.14}$$

Hence, combining Equations (B.10), (B.12) and (B.14), we conclude that

$$\sqrt{n_l}(\hat{\beta}_{tsls} - \beta_o) \xrightarrow{\text{d}} \mathcal{N}(0, ([E[XX']\Gamma_o]^{-1}\Xi_{tsls}[\Gamma_o'E[XX']]^{-1}) \tag{B.15}$$

The proof of the asymptotic normality of $\hat{\beta}_{gmm}$ is based on Theorem 3.4 in Newey and McFadden (1994). We can construct the weighting matrix W as the following,

$$W^{-1} = \frac{1}{n_l}\sum_{i=1}^{n} h_i h_i' \tag{B.16}$$

$$h_i = \begin{pmatrix} \iota_i X_i^*(Y_i - (X_i^*)'\tilde{\beta}) \\ \lambda_l X_i(Y_i - (\hat{\Gamma}'X_i)'\tilde{\beta}) - \iota_i X_i(X_i^* - \hat{\Gamma}'X_i)'\tilde{\beta} \end{pmatrix} \tag{B.17}$$

where $\lambda_l = \lim_{n_l\to\infty}\frac{n_l}{n}$ and $\tilde{\beta}$ is some $\sqrt{n_l}$ consistent estimator of β_o (such as the OLS estimator on the labeled set).

We denote $\triangle = \text{plim}_{n_l\to\infty}\frac{\partial}{\partial b}g(b)$, and have

$$\triangle = \begin{pmatrix} -E[X^*(X^*)'] \\ -E[X(X^*)'] \end{pmatrix} \tag{B.18}$$

Hence,

$$\sqrt{n_l}(\hat{\beta}_{gmm} - \beta_o) \xrightarrow{\text{d}} \mathcal{N}(0, (\triangle'\Xi^{-1}\triangle)^{-1}) \tag{B.19}$$

where Ξ is the first-order asymptotic variance of $\sqrt{n_l}g(\beta_o)$ and W^{-1} consistently estimates Ξ.

Now we calculate Ξ through the following. We rewrite $g_1(\beta_o)$ and $g_2(\beta_o)$ as the following,

$$g_1(\beta_o) = E[X^*\varepsilon],$$
$$g_2(\beta_o) = E[X(Y - (X^*)'\beta_o)]$$

To simplify the notations, we denote $\frac{n}{n_l}\iota_i X_i^*\varepsilon_i \triangleq \tilde{g}_1(\beta_o)$, and recall that we have denoted $X_i(Y_i - (\Gamma' X_i)'\beta_o) + \frac{n}{n_l}\iota_i X_i(X_i^* - \Gamma' X_i)'\beta_o = (\varepsilon_i + \xi_i'\beta_o)X_i + \frac{n}{n_l}\iota_i\xi_i'\beta_o X_i \triangleq \tilde{g}_2(\beta_o)$.

Meanwhile, notice that the labeled set is randomly assigned and is independent of $X_i, X_i^*, \varepsilon_i, \xi_i$, i.e., $E[\iota f(X, X^*, \varepsilon, \xi)] = E[\iota] \cdot E[f(X, X^*, \varepsilon, \xi)]$, where $f(\cdot)$ can be any functions and

$$E[\iota] = E[\iota^2] = Prob(\iota = 1) = \frac{n_l}{n}$$

Hence, we have

$$\tilde{g}_1(\beta_o)\tilde{g}_1'(\beta_o) = \frac{n^2}{n_l^2}\iota_i^2\varepsilon_i^2 X_i^*(X_i^*)'$$
$$\tilde{g}_2(\beta_o)\tilde{g}_1'(\beta_o) = \frac{n}{n_l}\iota_i\varepsilon_i(\varepsilon_i + \xi_i'\beta_o)X_i(X_i^*)' - \frac{n^2}{n_l^2}\iota_i^2\varepsilon_i\xi_i'\beta_o X_i(X_i^*)'$$
$$\lim_{n_l\to\infty}\frac{n_l}{n}E[\tilde{g}_1(\beta_o)\tilde{g}_1'(\beta_o)] = E[\varepsilon^2 X^*(X^*)']$$
$$\lim_{n_l\to\infty}\frac{n_l}{n}E[\tilde{g}_2(\beta_o)\tilde{g}_1'(\beta_o)] = \lambda_l E[\varepsilon(\varepsilon + \xi'\beta_o)X(X^*)'] - E[\varepsilon\xi'\beta_o X(X^*)']$$

combining the result in Equation (B.14), which gives that

$$\Xi = \begin{pmatrix} E[\varepsilon^2 X^*(X^*)'] & \lambda_l E[\varepsilon\left(\varepsilon + \xi'\beta_o\right) X^* X'] \\ & -E[\varepsilon\xi'\beta_o X^* X'] \\ \lambda_l E[\varepsilon\left(\varepsilon + \xi'\beta_o\right) X(X^*)'] & \lambda_l E[\left(\varepsilon + \xi'\beta_o\right)^2 XX'] \\ -E[\varepsilon\xi'\beta_o X(X^*)'] & +E[((\xi'\beta_o)^2 XX')] \\ & -2\lambda_l E[\left(\varepsilon + \xi'\beta_o\right)\xi'\beta_o\left(XX'\right)] \end{pmatrix}.$$

□

B.9 Proof of Neyman Orthogonality

Let t be a scalar in this proof.

$$\begin{aligned}
&\partial_\eta E\left[\psi_1\left(W; \Gamma_o^{plr}, \eta_o\right)\right][\eta - \eta_0] \\
=&\partial_t E\left[\psi_1\left(W; \Gamma_o^{plr}, \eta_o + t(\eta - \eta_o)\right)\right]|_{t=0} \\
=&\partial_t E[(D_i^* - (r_o(Z_i) + t\left(r(Z_i) - r_o(Z_i)\right)) - \Gamma_o^{plr}(D_i - (m_o(Z_i) + t(m(Z_i) - m_o(Z_i)))\times \\
&\quad (D_i - (m_o(Z_i) + t(m(Z_i) - m_o(Z_i))))]|_{t=0}, \\
=&E[(r_o(Z_i) - r(Z_i) + \Gamma_o^{plr}(m(Z_i) - m_o(Z_i))) \times (D_i - (m_o(Z_i) + t(m(Z_i) - m_o(Z_i))))+ \\
&\quad (D_i^* - (r_o(Z_i) + t\left(r(Z_i) - r_o(Z_i)\right)) - \Gamma_o^{plr}(D_i - (m_o(Z_i) + t(m(Z_i) - m_o(Z_i))) \times (m_o(Z_i) - m(Z_i))]|_{t=0} \\
=&E[(r_o(Z_i) - r(Z_i) + \Gamma_o^{plr}(m(Z_i) - m_o(Z_i))) \times (D_i - m_o(Z_i))+ \\
&\quad \left(D_i^* - r_o(Z_i) - \Gamma_o^{plr}(D_i - m_o(Z_i)\right) \times (m_o(Z_i) - m(Z_i))].
\end{aligned}$$

Similarly,

$$
\begin{aligned}
&\partial_\eta E\left[\psi_2\left(W;\theta_o,\Gamma_o^{plr},\eta_o\right)\right][\eta-\eta_0] \\
=&\partial_t E[(Y_i-(l_o(Z_i)+t(l(Z_i)-l_o(Z_i)))-\theta_o\Gamma_o^{plr}(D_i-(m_o(Z_i)+t(m(Z_i)-m_o(Z_i)))))\times \\
&(D_i-(m_o(Z_i)+t(m(Z_i)-m_o(Z_i))))]|_{t=0} \\
=&E[(l_o(Z_i)-l(Z_i))+\theta_o\Gamma_o^{plr}(m(Z_i)-m_o(Z_i))\times(D_i-(m_o(Z_i)+t(m(Z_i)-m_o(Z_i))))+ \\
&(Y_i-(l_o(Z_i)+t(l(Z_i)-l_o(Z_i)))-\theta_o\Gamma_o^{plr}(D_i-(m_o(Z_i)+t(m(Z_i)-m_o(Z_i)))))\times \\
&(m_o(Z_i)-m(Z_i))]|_{t=0} \\
=&E[(l_o(Z_i)-l(Z_i)+\theta_o\Gamma_o^{plr}(m(Z_i)-m_o(Z_i)))\times(D_i-m_o(Z_i))+ \\
&(Y_i-l_o(Z_i)-\theta_o\Gamma_o^{plr}(D_i-m_o(Z_i)))\times(m_o(Z_i)-m(Z_i))].
\end{aligned}
$$

From $E[D-m_o(Z)]=E[Y-l_o(Z)]=E[D^*-r_o(Z)]=0$ and Assumption 4.2.2, we complete the proof.

B.10 Proof of Theorems 4.2.1 and 4.2.2.

Theorem 5 (Theorem 3.)**.** *[Consistency] Under Assumptions 4.2.2, any estimator* $\hat{\theta}\in\{\hat{\theta}_{dml},\hat{\theta}_{tsdml},\hat{\theta}_{gdml}\}$ *is a consistent estimator for* θ_o*, i.e.,*

$$\hat{\theta}\xrightarrow{p}\theta_o,\quad \text{as } n\to\infty. \tag{B.20}$$

Theorem 6 (Theorem 4.)**.** *[Asymptotic Normality] Under Assumptions 4.2.2, any estimator* $\hat{\theta}\in\{\hat{\theta}_{dml},\hat{\theta}_{tsdml},\hat{\theta}_{gdml}\}$ *follows asymptotically normality, i.e.,*

$$\sqrt{n_l}(\hat{\theta}-\theta_o)\xrightarrow{d}\mathcal{N}(0,V_{plr}), \tag{B.21}$$

where we denote the variance V_{plr} *for* $\hat{\theta}_{dml}, \hat{\theta}_{tsdml}, \hat{\theta}_{gdml}$ *as* $V_{dml}, V_{tsdml}, V_{gdml}$ *as follows:*

$$V_{dml} = (E[u^2])^{-1}\Xi_{dml}(E[u^2])^{-1},$$
$$V_{tsdml} = J_o^{-1}\Xi_{tsdml}J_o^{-1},$$
$$V_{gdml} = (\triangle'_{dml}\Xi^{-1}_{gdml}\triangle_{dml})^{-1},$$

and

$$\Xi_{dml} = E[\varepsilon^2 u^2], \quad J_o = E\left[\Gamma_o^{plr} e^2\right], \quad \triangle_{dml} = \begin{pmatrix} -E[u^2] & -E[ue] \end{pmatrix}',$$
$$\Xi_{tsdml} = \lambda_l E[(\nu - \theta_o\Gamma_o^{plr}e)^2 e^2] + E[(u - \Gamma_o^{plr}e)^2 e^2\theta_o^2] - 2\lambda_l E[(\nu - \theta_o\Gamma_o^{plr}e)(u - \Gamma_o^{plr}e)e^2\theta_o],$$
$$\Xi_{gdml} = \begin{pmatrix} \Xi_{dml} & \begin{matrix} \lambda_l E[(\nu - \theta_o u)(\nu - \theta_o\Gamma_o^{plr}e)eu] \\ -E[(\nu - \theta_o u)(u - \Gamma_o^{plr}e)ue\theta_o] \end{matrix} \\ \begin{matrix} \lambda_l E[(\nu - \theta_o u)(\nu - \theta_o\Gamma_o^{plr}e)eu] \\ -E[(\nu - \theta_o u)(u - \Gamma_o^{plr}e)ue\theta_o] \end{matrix} & \Xi_{tsdml} \end{pmatrix}.$$

Proof. Proof. We omit the proof of the asymptotic consistency and normality of the DML estimator $\hat{\theta}_{dml}$ as it is provided or can be trivially derived based on the proofs in Chernozhukov et al. (2018). In the following, we mainly provide the proofs of the consistency and asymptotic normality of the proposed $\hat{\theta}_{tsdml}$ and $\hat{\theta}_{gdml}$.

(i) Asymptotic consistency and normality of $\hat{\theta}_{tsdml}$**:** Let Γ_o^{plr} denote the projection of residual $D^* - r_o(Z)$ on residual $D - m_o(Z)$, that is

$$\Gamma_o^{plr} = E[(D - m_o(Z))(D^* - r_o(Z))]E[(D - m_o(Z))(D - m_o(Z))]^{-1} \tag{B.22}$$
$$\zeta_i = D_i^* - r_o(Z_i) - \Gamma_o^{plr}(D_i - m_o(Z_i)) \tag{B.23}$$
$$\theta_o = \left(E\left[\left(\Gamma_o^{plr}(D - m_o(Z))\right)\left(\Gamma_o^{plr}(D - m_o(Z))\right)'\right]\right)^{-1} E\left[(Y - l_o(Z))\Gamma_o^{plr}(D - m_o(z))\right] \tag{B.24}$$

Notice that $E(\zeta|Z) = 0$ by definition of (l_o, r_o, m_o), and $E(\zeta(D - m_o(Z))) = 0$ is naturally

satisfied by the construction of Γ_o^{plr}.

The estimator $\hat{\theta}_{tsdml}$ can be written as

$$\hat{\theta}_{tsdml} = \left(\frac{1}{K} \sum_{k=1}^{K} \hat{E}_k \left[\left(\hat{\Gamma}_{tsdml} \left(D - \hat{m}_k \left(Z \right) \right) \right) \left(\hat{\Gamma}_{tsdml} \left(D - \hat{m}_k \left(Z \right) \right) \right)' \right] \right)^{-1} \cdot \left(\frac{1}{K} \sum_{k=1}^{K} \hat{E}_k \left[\left(Y - \hat{l}_k \left(Z \right) \right) \hat{\Gamma}_{tsdml} \left(D - \hat{m}_k \left(Z \right) \right) \right] \right),$$

where

$$\hat{\Gamma}_{tsdml} = \left(\frac{1}{K} \sum_{k=1}^{K} \hat{E}_{k,l} \left[\left(D^* - \hat{r}_{k,l} \left(Z \right) \right) \left(D - \hat{m}_k \left(Z \right) \right)' \right] \right) \left(\frac{1}{K} \sum_{k=1}^{K} \hat{E}_{k,l} \left[\left(D - \hat{m}_k \left(Z \right) \right) \left(D - \hat{m}_k \left(Z \right) \right)' \right] \right)^{-1}.$$

In the following proof, notice that we have assumed $Y, D^*, D, s_o(\cdot), m_o(\cdot) \in \mathbb{R}$.

Step 1:

We employ the score function in (4.17) to obtain $\hat{\Gamma}$. Following the proof of Theorem 4.1 in Chernozhukov et al. (2018), we can easily obtain

$$\begin{aligned} &\sqrt{n_l} \left(\hat{\Gamma}_{tsdml} - \Gamma_o^{plr} \right) \\ =&\left(\frac{1}{\sqrt{n_l}} \sum_{i=1}^{n} \iota_i (D_i^* - r_o \left(Z_i \right) - \Gamma_o^{plr} \left(D_i - m_o \left(Z_i \right) \right)) (D_i - m_o(Z_i)) \right) \left(E \left[(D - m_o(Z))^2 \right] \right)^{-1} + o_P \left(1 \right). \end{aligned} \tag{B.25}$$

which we can rewrite as $\left(\hat{\Gamma}_{tsdml} - \Gamma_o^{plr} \right) = O_P(n_l^{-1/2})$ or $\left(\hat{\Gamma} - \Gamma_o^{plr} \right) = o_P(1)$.

Step 2: We employ the score function in (4.18) to obtain $\hat{\theta}$, where we set $\Gamma = \hat{\Gamma}_{tsdml}$ from the first step in Equation (4.15).

Notice that $\psi_2(W; \theta, \Gamma, \eta) = \psi^a(W; \Gamma, \eta)\theta + \psi^b(W; \eta)$, where

$$\psi^a(W; \Gamma, \eta) = -\Gamma(D - m(Z))^2 \tag{B.26}$$

$$\psi^b(W; \eta) = (Y - l(Z))(D - m(Z)) \tag{B.27}$$

We write the negative expectation of $\psi^a(W;\Gamma,\eta)$ as J_o and its sample analog as $\hat{J}$,

$$J_o := E_P\left[\Gamma_o^{plr}\left(D - m_o\left(Z\right)\right)^2\right],$$
$$\hat{J} := \frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\hat{\Gamma}\left(D - \hat{m}_k\left(Z\right)\right)^2\right],$$

In the following, we first show that

$$\hat{J} = J_o + o_p(1) \tag{B.28}$$
$$\sqrt{n_l}\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi_2(W;\theta_o,\Gamma_o^{plr},\hat{\eta}_k)] = \frac{\sqrt{n_l}}{n}\sum_{i=1}^{n}\psi_2(W_i;\theta_o,\Gamma_o^{plr},\eta_o) + o_P\left(1\right) \tag{B.29}$$
$$\sqrt{n_l}\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\frac{\partial}{\partial\Gamma}\psi^a(W;\Gamma_o^{plr},\hat{\eta}_k) + o_P(1)\right]\right\}\theta_o\left(\hat{\Gamma}_{tsdml} - \Gamma_o^{plr}\right)$$
$$= -\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\iota_i\psi_1(W_i;\Gamma_o^{plr},\eta_o)\theta_o + o_P(1) \tag{B.30}$$

To show (B.28), note that by equation (A.5) in Chernozhukov et al. (2018) and Assumptions 4.2.2.(a), 4.2.2.(e), we have

$$\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\psi^a\left(W;\Gamma_o^{plr},\hat{\eta}_k\right)\right] - E\left[\psi^a\left(W;\Gamma_o^{plr},\eta_o\right)\right] = o_P(1). \tag{B.31}$$

By continuous mapping theorem and $\hat{\Gamma}_{tsdml} - \Gamma_o^{plr} = o_P(1)$ from equation (B.25), we have

$$\hat{J} := \frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\psi^a\left(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k\right)\right] = \frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\psi^a\left(W;\Gamma_o^{plr},\hat{\eta}_k\right)\right] + o_P(1) = J_o + o_P(1).$$

To show (B.29), by (A.6) in Chernozhukov et al. (2018), we have

$$\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi_2(W;\theta_o,\Gamma_o^{plr},\hat{\eta}_k)] = \frac{1}{n}\sum_{i=1}^{n}\psi_2(W_i;\theta_o,\Gamma_o^{plr},\eta_o) + o_P\left(n^{-1/2}\right),$$

Hence (B.29) holds.

To show (B.30), by Assumptions 4.2.2.(d), 4.2.2.(e) and the definition of $\psi^a(\cdot)$ in (B.26), taking the derivative with respect to Γ_o^{plr} in (B.31) gives us,

$$\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\frac{\partial}{\partial\Gamma}\psi^a(W;\Gamma_o^{plr},\hat{\eta}_k)+o_P(1)\right]=-E\left[(D-m_o(Z))^2\right]+o_P(1).$$

Moreover, note that by (B.25), Assumption 4.2.2.(d), and the definition of $\psi_1(\cdot)$, we have

$$\begin{aligned}
&\sqrt{n_l}\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\frac{\partial}{\partial\Gamma}\psi^a(W;\Gamma_o^{plr},\hat{\eta}_k)+o_P(1)\right]\right\}\theta_o\left(\hat{\Gamma}_{tsdml}-\Gamma_o^{plr}\right)\\
&=\left\{-E\left[(D-m_o(Z))^2\right]+o_P(1)\right\}\theta_o\left\{\begin{array}{c}\left(\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\iota_i(D_i^*-r_o(Z_i)-\Gamma_o^{plr}(D_i-m_o(Z_i)))(D_i-m_o(Z_i))\right)\\(E\left[(D-m_o(Z))^2\right])^{-1}+o_P(1)\end{array}\right\}\\
&=-\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\iota_i\left[D_i^*-r_o(Z_i)-\Gamma_o^{plr}(D_i-m_o(Z_i))\right](D_i-m_o(Z_i))\theta_o+o_P(1)\\
&=-\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\iota_i\psi_1(W_i;\Gamma_o^{plr},\eta_o)\theta_o+o_P(1).
\end{aligned}$$

Hence (B.30) holds.

Now let's derive the asymptotic linear representation for $\hat{\theta}-\theta_o$. Note that, $\hat{g}_{2,dml}(\hat{\theta}_{tsdml})=0$, which can be rewritten as

$$\begin{aligned}
\hat{g}_{2,dml}(\hat{\theta}_{tsdml})&=\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]\hat{\theta}+\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^b(W;\hat{\eta}_k)]\\
&=\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]\theta_o+\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^b(W;\hat{\eta}_k)]\\
&\quad+\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]+o_p(1)\right\}\left(\hat{\theta}-\theta_o\right),
\end{aligned}$$

Then, by (B.28), (B.29) and (B.30), we can write

$$
\begin{aligned}
&\sqrt{n_l}\left(\hat{\theta}-\theta_o\right) \\
&= -\sqrt{n_l}\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]+o_p(1)\right\}^{-1} \\
&\quad\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]\theta_o+\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^b(W;\hat{\eta}_k)]\right\} \\
&= -\left\{\hat{J}+o_P(1)\right\}^{-1}\sqrt{n_l}\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]\theta_o+\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^b(W;\hat{\eta}_k)]\right\} \\
&= -\left\{\hat{J}+o_P(1)\right\}^{-1}\sqrt{n_l}\left\{\begin{array}{c}\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^a(W;\Gamma_o^{plr},\hat{\eta}_k)]\theta_o+\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k[\psi^b(W;\hat{\eta}_k)] \\ +\frac{1}{K}\sum_{k=1}^{K}\hat{E}_k\left[\frac{\partial}{\partial\Gamma}\psi^a(W;\Gamma_o^{plr},\hat{\eta}_k)+o_P(1)\right]\theta_o\left(\hat{\Gamma}_{tsdml}-\Gamma_o^{plr}\right)\end{array}\right\} \\
&= -\left\{J_o+o_P(1)\right\}^{-1}\left\{\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n}\left[\frac{n_l}{n}\psi_2(W_i;\theta_o,\Gamma_o^{plr},\eta_o)-\iota_i\psi_1\left(W_i;\Gamma_o^{plr},\eta_o\right)\theta_o\right]\right\}+o_P(1) \\
&= -\left\{J_o+o_P(1)\right\}^{-1}\left\{\frac{\sqrt{n_l}}{\sqrt{n}}\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left[\psi_2(W_i;\theta_o,\Gamma_o^{plr},\eta_o)-\frac{n}{n_l}\iota_i\psi_1\left(W_i;\Gamma_o^{plr},\eta_o\right)\theta_o\right]\right\}+o_P(1)
\end{aligned}
$$

To simplify the notation further, we let

$$
\tilde{g}_{2,dml}(\theta_o) := \psi_2(W_i;\theta_o,\Gamma_o^{plr},\eta_o)-\frac{n}{n_l}\iota_i\psi_1\left(W_i;\Gamma_o^{plr},\eta_o\right)\theta_o
$$

Notice that

$$
\begin{aligned}
\lim_{n_l\to\infty}\frac{n_l}{n}E[\tilde{g}_{2,dml}(\beta_o)\tilde{g}'_{2,dml}(\beta_o)] =&\lambda_l E[(\nu-\theta_o\Gamma_o^{plr}e)^2e^2]+E[(u-\Gamma_o^{plr}e)^2e^2\theta^2] \\
&-2\lambda_l E[(\nu-\theta_o\Gamma_o^{plr}e)(u-\Gamma_o^{plr}e)e^2\theta_o] \qquad \text{(B.32)}
\end{aligned}
$$

This is from the facts below. Notice that ι_i follows a Bernoulli distribution with probability λ_l when n, n_l approach infinity and ι_i is independent with W_i. As a result, we have $\lim_{n,n_l\to\infty}E[\frac{n}{n_l}\iota]=1$

and $\lim_{n,n_l\to\infty} E[\iota^2] = 1 \cdot \lambda_l + 0 = \lambda_l$. Hence, we have

$$\lim_{n,n_l\to\infty} E[\frac{n^2}{n_l^2}\iota^2] = \lim_{n,n_l\to\infty} \frac{n^2}{n_l^2} E[\iota^2] = \frac{1}{\lambda_l}$$

By central limit theorem and Slutsky Theorem, we have

$$\sqrt{n_l}\left(\hat{\theta} - \theta_o\right) \xrightarrow{d} N\left(0, V_{tsdml}\right),$$

where $V_{tsdml} = J_o^{-1}\Xi_{tsdml}J_o^{-1}$, and

$$\begin{aligned}\Xi_{tsdml} = \lambda_l E[(\nu - \theta_o\Gamma_o^{plr}e)^2e^2] + E[(u - \Gamma_o^{plr}e)^2e^2\theta^2] \\ -2\lambda_l E[(\nu - \theta_o\Gamma_o^{plr}e)(u - \Gamma_o^{plr}e)e^2\theta_o]\end{aligned}$$

(ii) Asymptotic consistency and normality of $\hat{\theta}_{gdml}$: Lastly, we prove the asymptotic normality of the estimator $\hat{\theta}_{gdml}$ by Lemmas 5.2-5.4 in Newey (1994) and consistency can be trivially

derived. We write the moment conditions $g_{1,dml}(\theta_o)$ and $g_{2,dml}(\theta_o)$ as the following,

$$\begin{aligned}
g_{1,dml}(\theta_o) &= E[\psi(W;\theta_o,\eta_o)] \\
&= E[(Y - l_o(Z) - \theta_o(D^* - r_o(Z)))(D^* - r_o(Z))] \\
&= E[(\nu - \theta u)u], \\
g_{2,dml}(\theta_o) &= E[\psi_2(W;\theta_o,\Gamma_o^{plr},\eta_o)] \\
&= E[(Y - l_o(Z) - \theta_o(D^* - r_o(Z))(D - m(Z))] \\
&= E[(\nu - \theta_o ue] \\
\hat{g}_{1,dml}(\theta_o) &= \frac{1}{K}\sum_{k=1}^{K} \hat{E}_{k,l}[\psi(W;\theta_o,\hat{\eta}_k)] \\
\hat{g}_{2,dml}(\theta_o) &= \frac{1}{K}\sum_{k=1}^{K} \hat{E}_k[\psi_2(W;\theta_o,\hat{\Gamma}_{tsdml},\hat{\eta}_k)] \\
&= \frac{1}{K}\sum_{k=1}^{K} \hat{E}_k[\psi^a(W;\hat{\Gamma}_{tsdml},\hat{\eta}_k)]\theta_o + \frac{1}{K}\sum_{k=1}^{K} \hat{E}_k[\psi^b(W;\hat{\eta}_k)]
\end{aligned}$$

Let $\tilde{g}_{1,dml}(\theta_o) := \frac{n}{n_l}\iota_i(\nu_i - \theta_o u_i)u_i$. Through the proofs of Theorems 3.1 and 4.1 in Chernozhukov et al. (2018), we know that

$$\begin{aligned}
\|\sqrt{n_l}\hat{g}_{1,dml}(\theta_o)\| &= O_P(1) \\
\sqrt{n_l}\hat{g}_{1,dml}(\theta_o) &\xrightarrow{d} N(0, E[((\nu - \theta u)u)^2])
\end{aligned}$$

From Equation (B.32), we know that

$$\sqrt{n_l}\hat{g}_{2,dml}(\theta_o) \xrightarrow{d} N(0, \Xi_{tsdml})$$

We can construct the weighting matrix W_{dml} as the following,

$$W_{dml}^{-1} = \frac{1}{n_l}\sum_{i=1}^{n} h_{i,dml}h_{i,dml}' \tag{B.33}$$

$$h_{i,dml} = \begin{pmatrix} \iota_i\psi(W;\tilde{\theta},\tilde{\eta}) \\ \lambda_l\psi_2(W_i;\tilde{\theta},\tilde{\Gamma}^{plr},\tilde{\eta}) - \iota_i\psi_1\left(W_i;\tilde{\Gamma}^{plr},\tilde{\eta}\right)\tilde{\theta} \end{pmatrix} \tag{B.34}$$

where $\lambda_l = \lim_{n_l,n\to\infty}\frac{n_l}{n}$, the estimators $\tilde{\theta}$ and $\tilde{\beta}$ are some $\sqrt{n_l}$ consistent estimator of θ_o and Γ_o^{plr} (such as the DML estimators on the labeled set, $\hat{\theta}_{dml}$ and $\hat{\Gamma}_{tsdml}$), the estimator $\tilde{\eta}$ is an estimator for η_o under Assumptions 4.2.2.

We denote $\triangle_{dml} = \text{plim}_{n_l,n\to\infty}\frac{\partial}{\partial\theta}g_{dml}(\theta)$, and have

$$\triangle_{dml} = \begin{pmatrix} -E[(D^* - r_o(Z))^2] \\ -E[(D^* - r_o(Z))(D - m_o(Z))] \end{pmatrix} \tag{B.35}$$

Hence,

$$\sqrt{n_l}(\hat{\theta}_{gdml} - \theta_o) \xrightarrow{\text{d}} \mathcal{N}(0, (\triangle_{dml}'\Xi_{gdml}^{-1}\triangle_{dml})^{-1}) \tag{B.36}$$

where Ξ_{gdml} is the first-order asymptotic variance of $\sqrt{n_l}g_{dml}(\theta_o)$ and W_{dml}^{-1} consistently estimates Ξ_{gdml}.

Meanwhile, notice that

$$\tilde{g}_{1,dml}(\theta_o)\tilde{g}_{1,dml}'(\theta_o) = \frac{n^2}{n_l^2}\iota_i^2(\nu_i - \theta_o u_i)^2 u_i^2$$

$$\tilde{g}_{2,dml}(\theta_o)\tilde{g}_{1,dml}'(\theta_o) = \frac{n}{n_l}\iota_i(\nu_i - \theta_o u_i)(\nu_i - \theta_o\Gamma_o^{plr}e_i)e_i u_i - \frac{n^2}{n_l^2}\iota_i^2(\nu_i - \theta_o u_i)(u_i - \Gamma_o^{plr}e_i)u_i e_i\theta_o$$

$$\lim_{n_l\to\infty}\frac{n_l}{n}E[\tilde{g}_{1,dml}(\theta_o)\tilde{g}_{1,dml}'(\theta_o)] = E[(\nu - \theta_o u)^2 u^2]$$

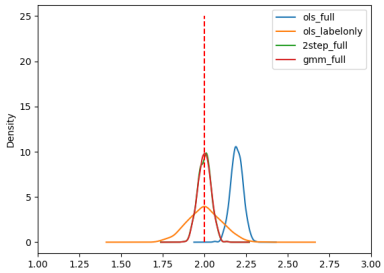
$$\lim_{n_l\to\infty}\frac{n_l}{n}E[\tilde{g}_{2,dml}(\theta_o)\tilde{g}_{1,dml}'(\theta_o)] = \lambda_l E[(\nu - \theta_o u)(\nu - \theta_o\Gamma_o^{plr}e)eu] - E[(\nu - \theta_o u)(u - \Gamma_o^{plr}e)ue\theta_o]$$

combining the result above, which gives that

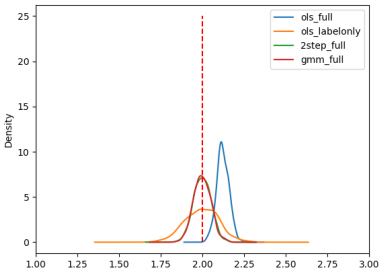
$$\Xi_{gdml} = \begin{pmatrix} E[(\nu - \theta u)^2 u^2] & \lambda_l E[(\nu - \theta_o u)(\nu - \theta_o \Gamma_o^{plr} e) e u] \\ & -E[(\nu - \theta_o u)(u - \Gamma_o^{plr} e) u e \theta_o] \\ \lambda_l E[(\nu - \theta_o u)(\nu - \theta_o \Gamma_o^{plr} e) e u] & \\ -E[(\nu - \theta_o u)(u - \Gamma_o^{plr} e) u e \theta_o] & \Xi_{tsdml} \end{pmatrix}.$$

As $\nu_i = \theta_o u_i + \varepsilon_i$, we complete the proof.

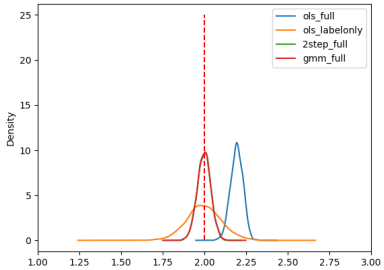
□



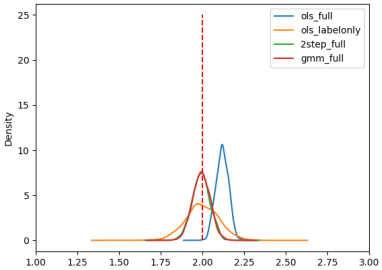
(a) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,0,0.1)



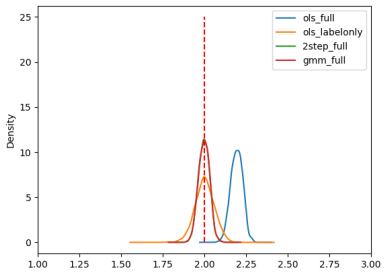
(b) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,0,0.2)



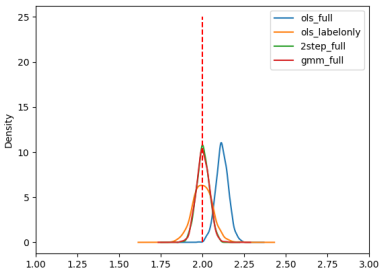
(c) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,1,0.1)



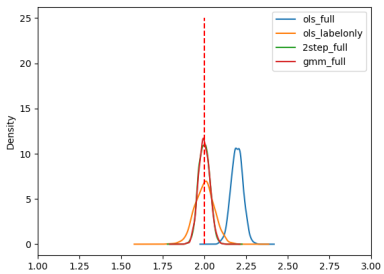
(d) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,1,0.2)



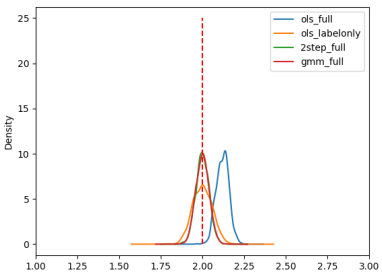
(e) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,0,0.1)



(f) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,0,0.2)

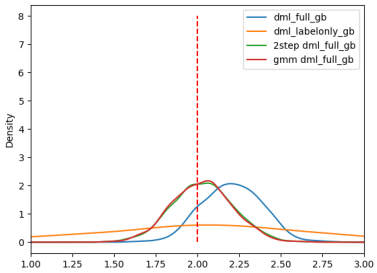


(g) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,1,0.1)

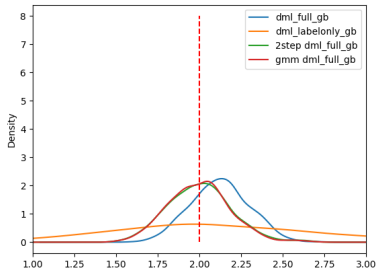


(h) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,1,0.2)

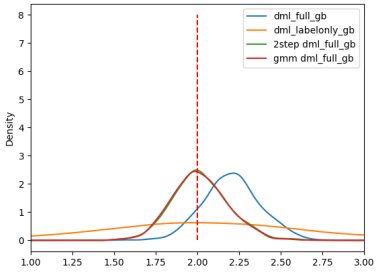
Figure B.1: Linear case results



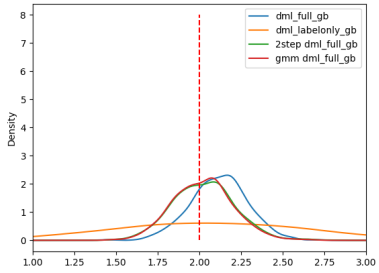
(a) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,0,0.1)



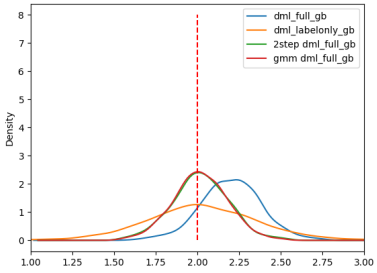
(b) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,0,0.2)



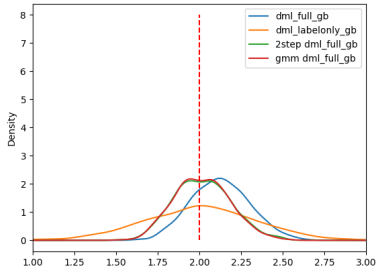
(c) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,1,0.1)



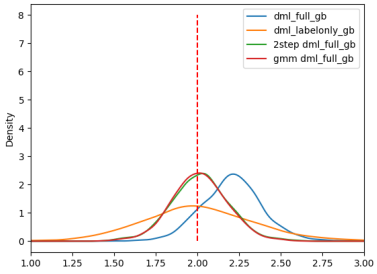
(d) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,1,0.2)



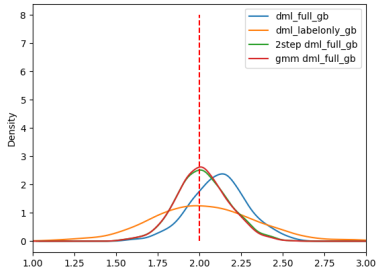
(e) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,0,0.1)



(f) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,0,0.2)

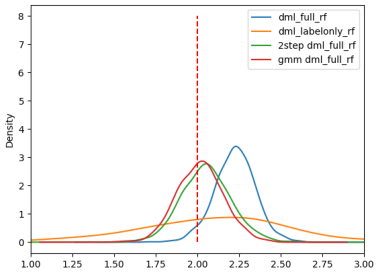


(g) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,1,0.1)

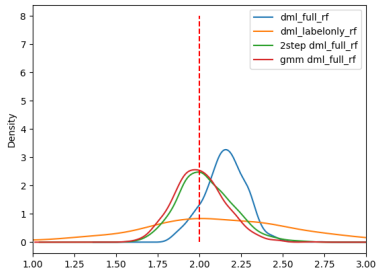


(h) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,1,0.2)

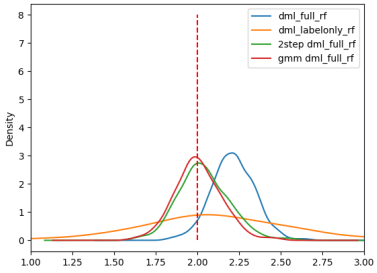
Figure B.2: Partially linear case results using gradient boosting



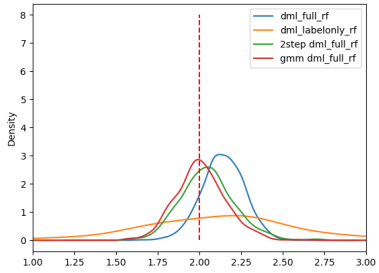
(a) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,0,0.1)



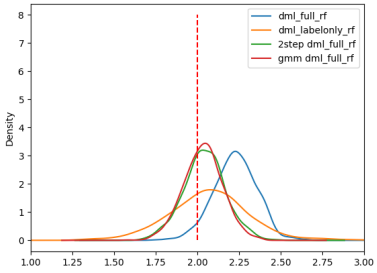
(b) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,0,0.2)



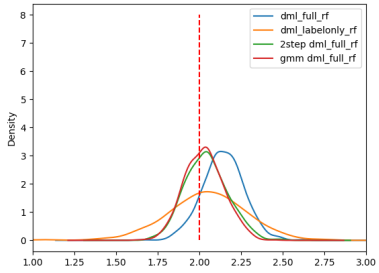
(c) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,1,0.1)



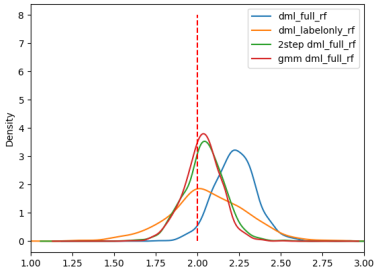
(d) ($\frac{n_l}{n}$,μ_1,σ_1):(0.1,1,0.2)



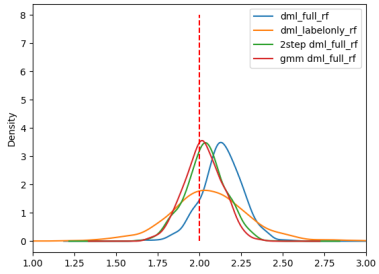
(e) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,0,0.1)



(f) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,0,0.2)



(g) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,1,0.1)



(h) ($\frac{n_l}{n}$,μ_1,σ_1):(0.3,1,0.2)

Figure B.3: Partially linear case results using random forest

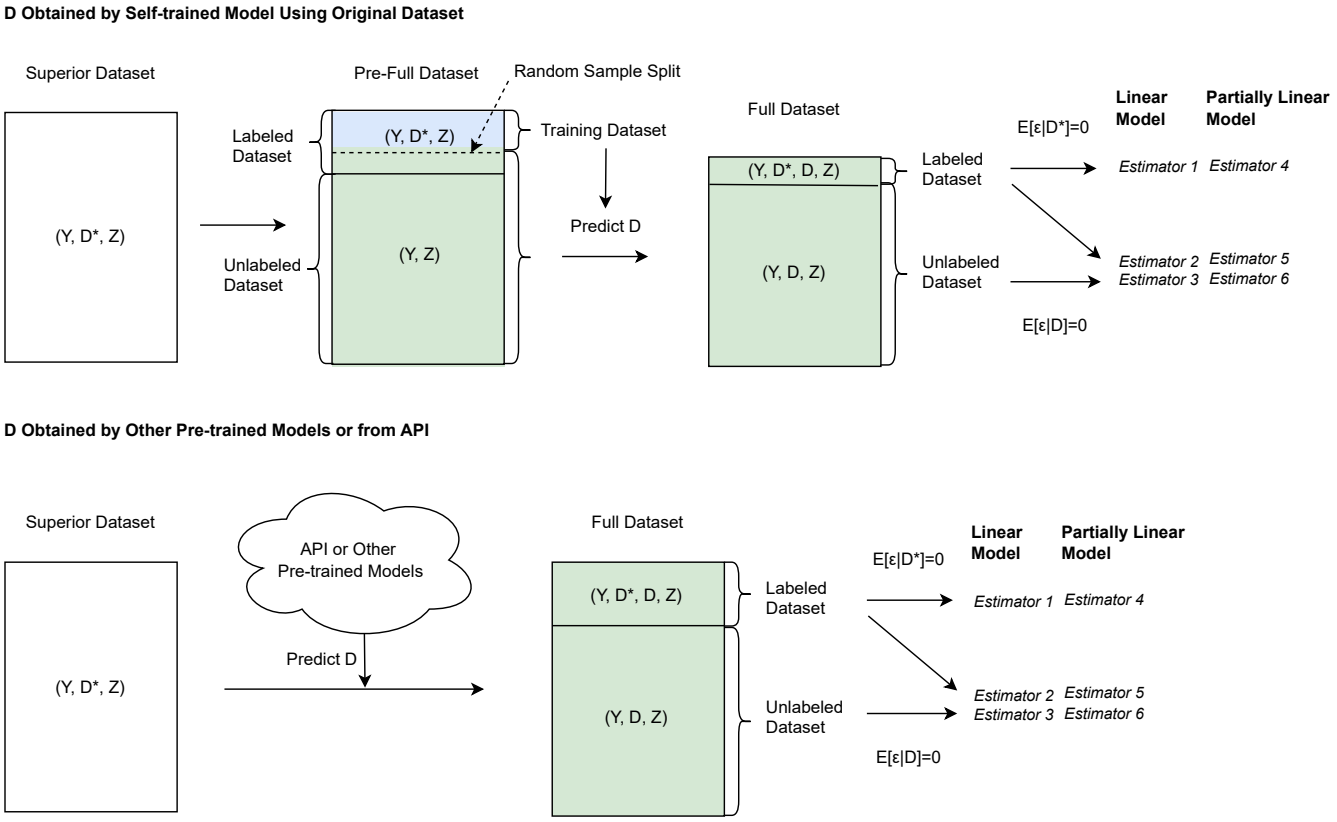


Figure B.4: Illustration of Datasets and Sample Splitting

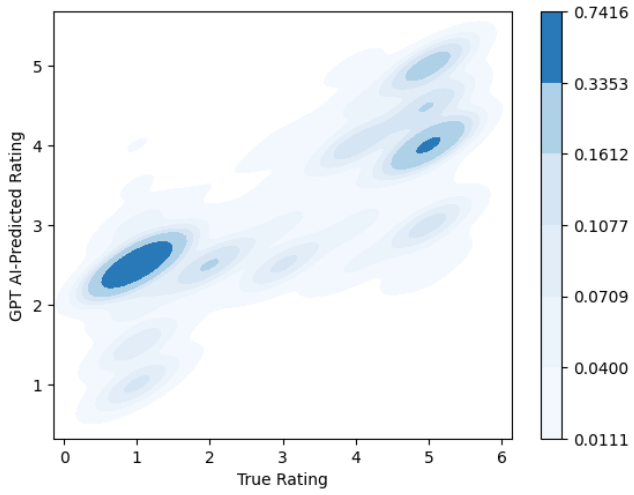
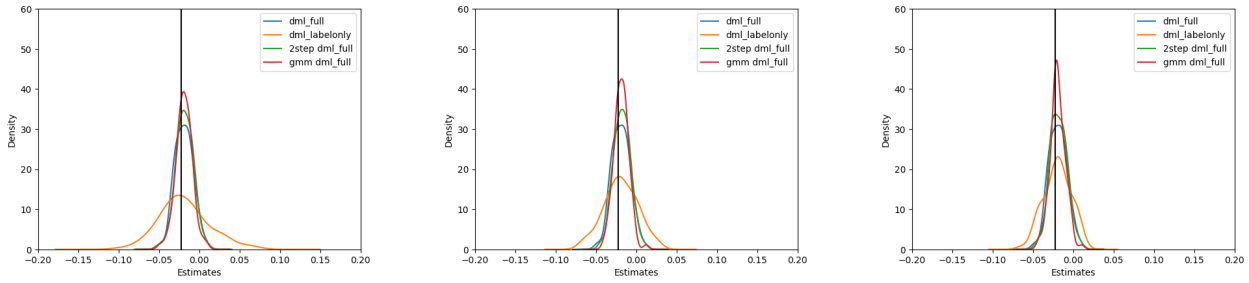


Figure B.5: Joint Distribution of GPT AI-Predicted Rating and True Rating



(a) $\frac{n_l}{n} = 0.1$ (b) $\frac{n_l}{n} = 0.2$ (c) $\frac{n_l}{n} = 0.3$

Figure B.6: Estimation Results Comparison with GPT AI-predicted Ratings

Appendix C

APPENDIX TO WHAT AND WHEN TO SELL IN LIVESTREAMING

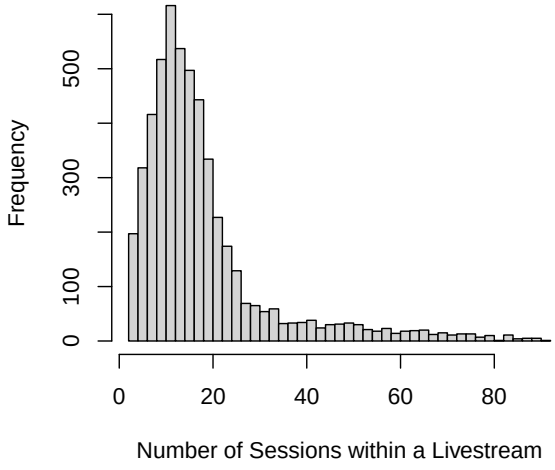


Figure C.1: Distribution of Number of Sessions within Each Livestream

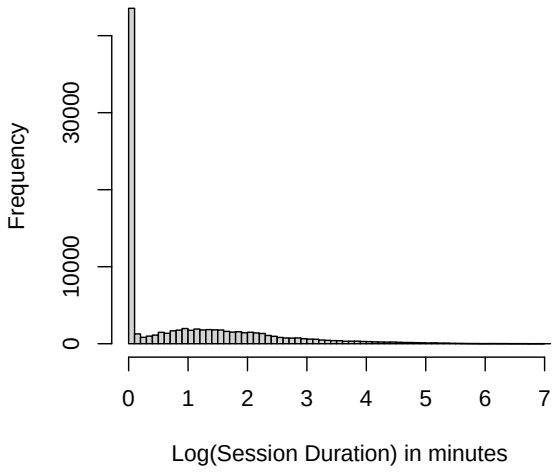


Figure C.2: Session Duration Distribution