

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Topics in Causal Inference: Data Fusion, Time-Series Experiments, and Matching Estimators

Permalink

<https://escholarship.org/uc/item/27t6p17q>

ISBN

9798189270352

Author

Lin, Zhexiao

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Topics in Causal Inference: Data Fusion, Time-Series Experiments,
and Matching Estimators

by

Zhexiao Lin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Peng Ding, Co-chair

Professor Peter J. Bickel , Co-chair

Professor Fang Han

Associate Professor Avi Feller

Professor Adityanand Guntuboyina

Spring 2026

Topics in Causal Inference: Data Fusion, Time-Series Experiments,
and Matching Estimators

Copyright 2026
by
Zhexiao Lin

Abstract

Topics in Causal Inference: Data Fusion, Time-Series Experiments,
and Matching Estimators

by

Zhexiao Lin

Doctor of Philosophy in Statistics

University of California, Berkeley

Associate Professor Peng Ding, Co-chair

Professor Peter J. Bickel , Co-chair

This dissertation studies problems in causal inference arising in modern empirical research. The three chapters in this dissertation develop novel theory and methodology aimed at improving the validity, robustness, and efficiency of causal inference in experimental and observational studies.

Chapter 1 discusses my joint work with Professor Peng Ding and Professor Peter J. Bickel on data fusion, adapted from [Lin et al. \(2026\)](#). In empirical research, when we have multiple estimators for the same parameter of interest, a central question arises: how do we combine unbiased but less precise estimators with biased but more precise ones to improve the inference? Under this setting, the point estimation problem has attracted considerable attention. In this paper, we focus on a less studied inference question: how can we conduct valid statistical inference in such settings with unknown bias? We propose a strategy to combine unbiased and biased estimators from a sensitivity analysis perspective. We derive a sequence of confidence intervals indexed by the magnitude of the bias, which enable researchers to assess how conclusions vary with the bias levels. Importantly, we introduce the notion of the b-value, a critical value of the unknown maximum relative bias at which combining estimators does not yield a significant result. We apply this strategy to three canonical combined estimators: the precision-weighted estimator, the pretest estimator, and the soft-thresholding estimator. For each estimator, we characterize the sequence of confidence intervals and determine the bias threshold at which the conclusion changes. Based on the theory, we recommend reporting the b-value based on the soft-thresholding estimator and its associated confidence intervals, which are robust to unknown bias and achieve the lowest worst-case risk among the alternatives.

Chapter 2 discusses my joint work with Professor Peng Ding on time-series experiments,

adapted from [Lin and Ding \(2025\)](#). Time-series experiments, also called switchback experiments or N-of-1 trials, play increasingly important roles in modern applications in medical and industrial areas. Under the potential outcomes framework, recent research has studied time-series experiments from the design-based perspective, relying solely on the randomness in the design to drive the statistical inference. Focusing on simpler statistical methods, we examine the design-based properties of regression-based methods for estimating treatment effects in time-series experiments. We demonstrate that the treatment effects of interest can be consistently estimated using ordinary least squares with an appropriately specified working model and transformed regressors. Our analysis allows for estimating a diverging number of treatment effects simultaneously, and establishes the consistency and asymptotic normality of the regression-based estimators. Additionally, we show that asymptotically, the heteroskedasticity and autocorrelation consistent variance estimators provide conservative estimates of the true, design-based variances. Importantly, although our approach relies on regression, our design-based framework allows for misspecification of the regression model.

Chapter 3 discusses my joint work with Professor Fang Han and Professor Peng Ding on matching estimators, adapted from [Lin et al. \(2023\)](#). Nearest neighbor (NN) matching is widely used in observational studies for causal effects. [Abadie and Imbens \(2006\)](#) provided the first large-sample analysis of NN matching. Their theory focuses on the case with the number of NNs, M fixed. We reveal something new out of their study and show that, once allowing M to diverge with the sample size, an intrinsic statistic in their analysis constitutes a consistent estimator of the density ratio with regard to covariates across the treated and control groups. Consequently, with a diverging M , the NN matching with [Abadie and Imbens \(2011\)](#)'s bias correction yields a doubly robust estimator of the average treatment effect and is semiparametrically efficient if the density functions are sufficiently smooth and the outcome model is consistently estimated. It can thus be viewed as a precursor of the double machine learning estimators. Follow-up work to Chapter 3 includes [Lin and Han \(2025\)](#), coauthored with Professor Fang Han, and [Cattaneo et al. \(2025\)](#), coauthored with Professor Fang Han and Professor Matias Cattaneo.

To my parents

Contents

Contents	ii
List of Figures	vi
List of Tables	vii
1 Introducing the b-value: combining unbiased and biased estimators from a sensitivity analysis perspective	1
1.1 Introduction	1
1.2 Problem setup and a review of point estimation	3
1.2.1 Problem setup	3
1.2.2 Point estimation: a review	7
1.3 Confidence intervals, hypothesis testing, and the b-value	8
1.3.1 Confidence interval based on the precision-weighted estimator	9
1.3.2 Confidence interval based on the pretest estimator	9
1.3.3 Confidence interval based on the soft-thresholding estimator	11
1.3.4 Comparison of the point estimators and confidence intervals	12
1.3.5 Generalization to the dependent case	14
1.4 Generalization to the multivariate case	14
1.4.1 Setup	15
1.4.2 Point estimation	17
1.4.3 Confidence intervals	18
1.5 Generalization to multiple estimators	21
1.5.1 Setup	21
1.5.2 Point estimation	22
1.5.3 Confidence intervals	23
1.6 Empirical studies	25
1.7 Discussion	27
2 Unifying regression-based and design-based causal inference in time-series experiments	28
2.1 Introduction to time-series experiments	28

2.2	Setup, estimand and estimator	30
2.2.1	Setup	30
2.2.2	Causal estimand	32
2.3	Theory	35
2.3.1	Consistency and asymptotic normality	35
2.3.2	HAC variance estimation	40
2.3.3	Comparisons with the existing estimands	41
2.4	Extension to continuous treatment	42
2.5	Empirical studies	45
2.5.1	Simulations	45
2.5.2	Real data analysis	47
2.6	Extensions	48
3	Estimation based on nearest neighbor matching: from density ratio to average treatment effect	50
3.1	Introduction	50
3.2	Density ratio estimation via NN-matching	52
3.3	Revisiting the bias-corrected matching estimator of the ATE	54
3.4	Asymptotic analysis with diverging M	56
3.4.1	The original matching-based estimator	56
3.4.2	A double machine learning version of the matching	59
3.5	Simulation	61
3.6	Some final remarks	62
3.7	Density ratio estimation I: computation	65
3.8	Density ratio estimation II: theory	68
3.8.1	Consistency	68
3.8.2	Rates of convergence	70
3.9	Proofs of the results in Sections 3.3 and 3.4	73
3.9.1	Proof of Lemma 3.3.1	73
3.9.2	Proof of Theorem 3.4.1	73
3.9.3	Proof of Theorem 3.4.2	77
	Bibliography	80
A	Appendix for Chapter 1	93
A.1	More discussions	93
A.1.1	Computing the b-value efficiently	93
A.1.2	One-sided confidence bounds	95
A.1.3	Bias-dependent point estimator	96
A.1.4	More details for the dependence case	98
A.1.5	Relationship between the generic soft-thresholding estimator and the estimator in Berger (1981); Bickel (1984)	99

A.2	Proofs of the main results	99
A.2.1	Proof of Theorem 1.3.1	99
A.2.2	Proof of Theorem 1.3.2	100
A.2.3	Proof of Theorem 1.3.3	102
A.2.4	Proof of Theorem 1.4.1	104
A.2.5	Proof of Theorem 1.4.2	105
A.2.6	Proof of Theorem 1.4.3	107
A.2.7	Proof of Theorem 1.5.1	109
A.2.8	Proof of Theorem 1.5.2	110
A.2.9	Proof of Theorem 1.5.3	112
A.3	Proofs of additional results	113
A.3.1	Proof of Lemma 1.3.1 and Lemma 1.3.2	113
A.3.2	Proof of Theorem A.1.1	115
A.3.3	Proof of Theorem A.1.2	116
A.3.4	Proof of Theorem A.1.3	117
A.3.5	Proof of Theorem A.1.4	118
A.3.6	Proof of Theorem A.1.5	118
A.3.7	Proof of Theorem A.1.6	119
B	Appendix for Chapter 2	121
B.1	Additional results	121
B.1.1	Relationship between full OLS, marginal OLS and weighted least squares	121
B.1.1.1	Alternative OLS	123
B.1.1.2	Full OLS and marginal OLS	123
B.1.1.3	Marginal OLS and WLS	124
B.1.2	Regression with interaction terms	125
B.1.3	General exposure mapping	128
B.2	Lemmas for the main results	131
B.2.1	Proof of Lemma B.2.1	131
B.2.2	Proof of Lemma B.2.2	133
B.2.3	Proof of Lemma B.2.3	136
B.3	Proofs	139
B.3.1	Proof of Theorem 2.3.1	139
B.3.2	Proof of Theorem 2.3.2	140
B.3.3	Proof of Theorem 2.3.3	140
B.3.4	Proofs of Theorem 2.4.1 and Theorem 2.4.2	146
B.3.5	Proof of Theorem 2.4.3	146
B.3.6	Proofs of Theorem B.1.1 and Theorem B.1.2	147
B.3.7	Proofs of Theorem B.1.3 and Theorem B.1.4	147
B.3.8	Proof of Proposition B.1.1	148
B.3.9	Proof of Proposition B.1.2	149

C	Appendix for Chapter 3	153
C.1	Proofs of the results in Sections 3.3 and 3.4	153
C.2	Proofs of the results in Section 3.7	153
C.2.1	Proof of Theorem 3.7.1	153
C.3	Proofs of the results in Section 3.8	154
C.3.1	Proof of Lemma 3.8.1	154
C.3.2	Proof of Theorem 3.8.1	161
C.3.3	Proof of Theorem 3.8.2	162
C.3.4	Proof of Corollary 3.8.1	165
C.3.5	Proof of Theorem 3.8.3	165
C.3.6	Proof of Proposition 3.8.1	171
C.3.7	Proof of Theorem 3.8.4	171
C.3.8	Proof of Proposition 3.8.2	172
C.4	Proofs of the results in Section 3.9	173
C.4.1	Proof of Lemma 3.9.1	173
C.4.2	Proof of Lemma 3.9.2	175
C.4.3	Proof of Lemma 3.9.3	176
C.5	Proofs of the results in Appendix	177
C.5.1	Proof of Lemma C.3.1	177
C.5.2	Proof of Lemma C.3.3	178
C.5.3	Proof of Lemma C.3.4	179

List of Figures

1.1	Confidence intervals against the maximum relative bias $ \Delta/\sigma_0 \leq b$	13
1.2	Confidence intervals against the maximum relative bias $ \Delta/\sigma_0 \leq b$	26
2.1	Histograms of $\hat{\tau}_k - \tau_k$ for $k = 0, \dots, 5$ under $T = 1000$	46
2.2	$\ \hat{\tau} - \tau\ _2$ versus T . The center bar is the median.	46
2.3	Normalized Frobenius error $\ \hat{\mathbf{V}} - \mathbf{V}\ _F/\ \mathbf{V}\ _F$ versus T for different bandwidths L	47
2.4	Lag-specific estimates $\hat{\tau}_k$ with 95% confidence intervals for $k = 0, \dots, 4$ under different markets. The number in parentheses following each market label reports the joint p-value for $H_0 : \tau = \mathbf{0}$	49

List of Tables

2.1	Empirical 95% coverage for $\hat{\tau}_k$ using $\hat{\mathbf{V}}$ with $L = \lfloor T^{1/4} \rfloor$	47
3.1	Simulation results, LaLonde (1986), $\sigma^* = 9.55$	63
3.2	Simulation results, Shadish et al. (2008), $\sigma^* = 3.85$	64

Acknowledgments

First and foremost, I would like to thank my three fantastic advisors, Professor Peng Ding, Professor Peter Bickel, and Professor Fang Han, for their generous support and guidance throughout my PhD journey. It has been a great privilege and fortune to work with them, and I am deeply grateful for their patience, encouragement, and invaluable advice, all of which have shaped both my research and my growth as a scholar.

Professor Peng Ding was the first person to introduce me to the fascinating world of causal inference. Before working with him, I knew very little about the field. Through our discussions, I learned many foundational ideas and gained many deep insights that could never be acquired from textbooks alone. These insights have guided me in formulating research questions that are not only intellectually interesting but also meaningful and useful to the broader community. His knowledge spans virtually all areas of causal inference, giving me the freedom to explore the topics that interest me most. No matter what problem I am working on, he always provides valuable guidance and suggestions. He also taught me to understand causal inference through the lens of empirical research, which has shaped my sense of what makes a research problem truly important and has encouraged me to think more carefully about the broader audience for my work. I am especially grateful for his detailed feedback on my research and for the time and care he devoted to revising my papers. His mentorship has been indispensable in helping me grow into an independent researcher.

Professor Peter Bickel was someone I had already known of through my study and research in statistics before coming to Berkeley. I cannot fully express how excited I was when he agreed to advise and work with me. He is one of the most brilliant statisticians I have ever met. At the same time, he is extraordinarily kind, patient, and always willing to help. I have long been inspired by his enthusiasm and passion for statistics, which have motivated me to pursue a deeper understanding of statistical theory and to strive to become a better researcher. He has remarkable intuition, deep insight, and an extraordinary wealth of ideas. What I especially appreciate is his patience in writing down his ideas carefully and clearly, which has been invaluable in helping me understand them, especially when they initially felt abstract or difficult to grasp. It has been a great pleasure to learn from him, to follow the development of his ideas, and to see them ultimately work in our joint projects.

Professor Fang Han was my first advisor in my formal research journey and the person who introduced me to statistical research during my master's study. He was also the one who sparked my deep interest in statistics. Over the past five years, we have written many papers together, all of which were published in top journals, and it has been both an enjoyable and deeply rewarding experience to work with him. He is supportive, creative, and generous. Even though I had no prior experience in statistical research before working with him, he was willing to guide me through every step of doing research and writing papers. We share many research interests and it has always been a pleasure to exchange ideas with him. He has extensive knowledge and many original ideas in theoretical statistics, and I have learned enormously from our discussions. I still remember that, during one period, I proposed meeting with him every day and worried that I might be taking too much of his personal

time. Yet he still agreed, and every single day he gave me detailed comments and thoughtful suggestions on my research. I am truly grateful for the tremendous amount of time and effort he devoted to my work. My research experience with him helped me gain admission to top PhD programs and made it possible for me to continue pursuing statistical research, for which I will always be grateful. He has always encouraged me, introduced me to many of his collaborators and friends in the community, and promoted our work at conferences and seminars, which allowed many researchers to become familiar with my name and my research. Beyond academia, I have learned a great deal from him as well. He has been not only a wonderful academic advisor, but also a life mentor and a dear friend. He taught me not only how to do better research, but also how to become a better person. He shared many of his life experiences with me and offered valuable advice on both life and career. Although most of our interactions took place over Zoom, and we met in person only a few times because I was in Seattle for only a few months, his mentorship has been invaluable.

My thanks also go to many other professors and collaborators. I would like to thank Professor Avi Feller and Professor Aditya Guntuboyina for kindly serving on my committee and for providing valuable feedback and suggestions on this dissertation and on my PhD research more broadly. I would like to thank Professor Wei Sun, who offered me a research associate position during my master's study, introduced me to the world of biostatistics, and worked with me on many interesting applied projects. I would like to thank Professor Matias Cattaneo, who collaborated with me on a theoretical project and offered many insightful suggestions. I would like to thank my collaborators in the United States, including Weitong Li, Tianhao Zhang, Zhecheng Sheng, Chen Jiang, Yuanyuan Gao, and Yilong Hou. I would like to thank all the people and institutions who invited me to present my research, including Professor Rong Chen, Professor Lihua Lei, Professor Qizhai Li, Rui Liu, Yi Liu, Professor Adrian Röllin, Yilin Song, Professor Fang Yao, Professor Xiaoke Zhang, Professor Yichuan Zhao, among many others. Without their support, I would not have had the opportunity to give more than 30 presentations across many cities and countries during my PhD journey.

Next, I would like to thank the Berkeley Statistics community, which I am very proud to be part of. I would like to thank all of Peng's students and the members of his weekly group meeting for exchanging ideas in an environment that was both rigorous and relaxed, especially Kevin Benac, Mengsi Gao, Yiyi Huo, Sizhu Lu, Lei Shi, Fangzhou Su, Xinyi Wang, Mingrui Zhang, as well as all other visiting faculty members and students; the faculty members who organized the Casual Causal Group and all of its members for many interesting talks and discussions; many other outstanding fellow students in my PhD program for their conversations and friendship, in particular Tianyu Guo, Yaxuan Huang, Dohyeong Ki, Licong Lin, Yichen Pan, Dan Soriano, Xuelin Yang, Ruiqi Zhang, and many others.

I would also like to thank many people outside Berkeley: my internship managers Fan Zhang and Hari Adireddy for their patient and generous mentorship and support, and Can Wang for discussions at Two Sigma; Yuanyuan Li, Neeraj Sarna, and Michael von Gablenz at Munich Re; Pablo Crespo and Julie Beckley at Etsy. My four years in the PhD program have also been greatly enriched by friendship. I would like to thank all my friends in China, the Bay Area, New York City, and around the world.

Finally, I owe my deepest gratitude to my parents, who have always supported me and encouraged me to pursue my goals. I am grateful for their unconditional love and unwavering support throughout my life. This achievement would not have been possible without them.

Chapter 1

Introducing the b-value: combining unbiased and biased estimators from a sensitivity analysis perspective

1.1 Introduction

In empirical research, it is common for researchers to employ different methods to estimate the same parameter of interest. These differences may arise from the use of distinct datasets or from imposing different model assumptions on the same dataset. We motivate our paper with the following two examples of combining estimators.

Example 1.1.1. Randomized controlled trials (RCTs) are the gold standard for estimating treatment effects due to their ability to eliminate unmeasured confounding. However, RCTs often suffer from limited sample sizes, as large-scale experiments can be costly or infeasible. In contrast, observational data are more readily available from the target population of interest. However, estimates using observational data may be biased in estimating treatment effects due to unmeasured confounding, raising concerns about the internal validity. See [Brantner et al. \(2023\)](#) and [Colnet et al. \(2024\)](#) for recent reviews on motivations and methods for combining RCTs and observational studies.

Example 1.1.2. The ordinary least squares (OLS) estimator is biased in estimating the unknown parameters under the linear model when the error term is correlated with the regressor. In contrast, the instrumental variables (IV) estimator can provide unbiased estimates for the parameters of interest with a valid instrumental variable that is uncorrelated with the error term but correlated with the regressor. However, the IV estimator is usually much less precise than the OLS estimator, especially when the IV is weakly correlated with the regressor ([Bound et al., 1995](#)). In empirical studies, e.g., [Angrist and Krueger \(1991\)](#), researchers often report results from both OLS and IV estimators. How to combine OLS and IV estimators is gaining increasing interest ([Armstrong et al., 2025b](#)).

When the estimators are from different datasets, e.g., Example 1.1.1, the estimators are independent as long as the datasets are independent. When the estimators are from the same dataset but with different model assumptions, e.g., Example 1.1.2, the estimators are dependent in general. Given access to multiple potentially dependent estimators, some unbiased but less precise and others biased but more precise, a natural question is: How can we combine the unbiased and potentially biased estimators to improve the inference with unknown bias? From the point estimation perspective, this problem has been extensively studied (Bickel, 1984; Green and Strawderman, 1991; Giles and Giles, 1993; Chen et al., 2015; Athey et al., 2020; de Chaisemartin and D’Haultfœuille, 2020; Rosenman et al., 2023a; Gao and Yang, 2023; Yang et al., 2025). Many methods have been proposed for constructing combined estimators that perform well when the bias is small and have bounded risks when the bias is large. From the statistical inference perspective, this problem is less studied. In this paper, we answer the following question: how can we conduct valid statistical inference after combining the estimators?

The inference question has received considerably less attention. The primary difficulty lies in the impossibility of characterizing the distribution of the combined estimator with unknown bias (Noack and Rothe, 2024; Armstrong et al., 2025b; Chen et al., 2025). Once the information about the bias is introduced, e.g., an upper bound on its magnitude, confidence intervals for the parameter of interest become possible. In the absence of such information, we focus on the following question:

How can we construct a sequence of confidence intervals for the parameter of interest across bias levels?

The sequence of confidence intervals provides a way to quantify how the level of bias impacts the uncertainty in point estimation. One important application of this sequence is in hypothesis testing. Suppose the null hypothesis is not rejected based on the unbiased but less precise estimator. Now consider a scenario where the statistical test based on the biased but more precise estimator rejects the null hypothesis. If we have prior knowledge suggesting the bias is small, incorporating the biased estimator may yield a more precise estimator to reject the null hypothesis. In such cases, the sequence of confidence intervals enables us to address the following question:

How large must the bias be to change the conclusion of a hypothesis test—from rejection to non-rejection?

The idea of constructing the sequence of confidence intervals and examining how conclusions change as the assumed level of bias varies is related to sensitivity analysis in causal inference with unmeasured confounding, e.g., Cornfield et al. (1959); Rosenbaum and Rubin (1983a) and VanderWeele and Ding (2017). In observational studies, sensitivity analysis assesses how the causal conclusions change with respect to different degrees of unmeasured confounding by varying the sensitivity parameter (Rosenbaum, 2002; Ding and VanderWeele,

2016). Our proposed framework has a similar flavor: by indexing inference results over a continuum of bias levels, we can assess the robustness of statistical inference.

We formalize two statistical inference questions, confidence interval and hypothesis testing, in the context of combining unbiased and biased estimators. Under regularity conditions, the estimators satisfy a joint central limit theorem. Consequently, we present our formulation in a finite-sample Gaussian setting, assuming exact normality for both the unbiased and biased estimators. This reduction to a Gaussian model is motivated by Le Cam’s classical asymptotic argument (Le Cam, 1956), and because of this, our Gaussian formulation should be viewed as an asymptotic idealization rather than a restrictive finite-sample assumption.

We develop a general framework that applies to any point estimator formed by combining such estimators. Within this framework, we construct a sequence of confidence intervals indexed by the bias level, and importantly, we introduce the notion of the b-value, a critical value of the unknown maximum relative bias at which combining estimators does not yield a significant result. We examine three canonical combined estimators: the precision-weighted estimator, the pretest estimator, and the soft-thresholding estimator. For each estimator, we derive either analytically or numerically the sequence of confidence intervals and the b-value. Among the three, we advocate for the soft-thresholding estimator, as it offers robustness to unknown bias compared with the precision-weighted estimator, and exhibits lower worst-case risk and more desirable properties for confidence interval construction than the pretest estimator. We provide a Python package for the proposed methods, available at <https://github.com/zhexiaolin/b-value>.

Notation. For a vector $a = (a_1, \dots, a_d)^\top \in \mathbb{R}^d$, let $\|a\|_1 = \sum_{i=1}^d |a_i|$, $\|a\|_2 = (\sum_{i=1}^d a_i^2)^{1/2}$, and $|a| = (|a_1|, \dots, |a_d|)^\top$. For vectors $a = (a_1, \dots, a_d)^\top$ and $b = (b_1, \dots, b_d)^\top$, let $a \odot b = (a_1 b_1, \dots, a_d b_d)^\top$ be the Hadamard (element-wise) product, and $a \leq b$ denote $a_i \leq b_i$ for all $i = 1, 2, \dots, d$. For a scalar $b \in \mathbb{R}$ and a set $A \subset \mathbb{R}$, we write $b - A = \{b - a : a \in A\}$, which extends naturally to vectors and sets in $\mathbb{R}^d$. We use $\Phi(\cdot)$ and $\phi(\cdot)$ to denote the cumulative distribution function and density function of the standard normal distribution, respectively. We use $\phi_{\mu, \Sigma}(\cdot)$ to denote the density function of the multivariate normal distribution with mean μ and covariance matrix Σ . We use $\Psi_d(\cdot; \lambda)$ to denote the cumulative distribution function of the noncentral chi-squared distribution with noncentrality parameter λ and degrees of freedom d . We use c_α to denote the $(1 - \alpha)$ upper quantile of the standard normal distribution.

1.2 Problem setup and a review of point estimation

1.2.1 Problem setup

We consider the following setting:

Assumption 1.2.1. Suppose we observe two independent random variables: one unbiased estimator $\hat{\tau}_0 \sim N(\tau, \sigma_0^2)$ and one biased estimator $\hat{\tau}_1 \sim N(\tau + \Delta, \sigma_1^2)$. Here, $\tau \in \mathbb{R}$ is

the unknown parameter of interest. We assume that σ_0^2 and σ_1^2 are known, whereas Δ is unknown.

In practice, both the unbiased and biased estimators for the same parameter of interest are constructed from data. Under regularity conditions, these estimators are jointly asymptotically normal with an unknown covariance matrix. As long as this covariance matrix can be consistently estimated, the problem of combining estimators reduces to an exact normality framework under Le Cam's asymptotic framework (Bickel, 1984). Le Cam (1956) showed that a wide class of estimators and test statistics can be approximated in large samples by Gaussian experiments with known covariances. Within this framework, we treat the estimators as arising from Gaussian experiments in which the variances are replaced by their consistent estimates, and the validity of the inference procedures is preserved asymptotically. For this reason, we present our analysis in the exact normality setting. Nevertheless, our results represent the asymptotic limits of a broad class of more general and practically relevant inference problems. The analysis can be generalized to dependent (Section 1.3.5), multivariate (Section 1.4) and multiple estimators (Section 1.5) cases.

Let $\gamma = \sigma_0^2/\sigma_1^2$ be the variance ratio between $\hat{\tau}_0$ and $\hat{\tau}_1$. In general, the unbiased estimator is less precise but the biased estimator is more precise. Therefore, we focus on the regime in which σ_0^2 is large and σ_1^2 is small, which indicates that γ is large.

Under Assumption 1.2.1, we address the following central question:

How can we construct a sequence of two-sided confidence intervals for τ across different levels of bias (Δ)?

We focus on constructing two-sided symmetric confidence intervals centered at prespecified point estimators. In other words, we do not let the point estimator itself depend on the bias level Δ . An alternative approach would be to use a bias-dependent point estimator, where the estimator incorporates Δ . We discuss using bias-dependent point estimators in Section 1.7.

Since $\hat{\tau}_0$ is an unbiased estimator of τ , one natural approach is to construct a confidence interval based solely on $\hat{\tau}_0$, which is invariant to the bias Δ . Given a significance level $\zeta \in (0, 1)$, we can construct a standard two-sided confidence interval

$$[\hat{\tau}_0 - \sigma_0 c_{\zeta/2}, \hat{\tau}_0 + \sigma_0 c_{\zeta/2}],$$

where $c_{\zeta/2}$ is the $(1 - \zeta/2)$ upper quantile of the standard normal distribution. While being valid regardless of the bias Δ , this confidence interval may be too wide when $\hat{\tau}_0$ is not precise, whose length $2\sigma_0 c_{\zeta/2}$ scales proportionally with σ_0 . As we have access to the additional biased but more precise estimator $\hat{\tau}_1$, a natural question arises: can we shorten the confidence interval by incorporating information from $\hat{\tau}_1$?

This motivates us to combine the two estimators $\hat{\tau}_0$ and $\hat{\tau}_1$ to construct a two-sided confidence interval with shorter length. Consider a generic combined estimator

$$\hat{\tau} = \hat{\tau}(\hat{\tau}_0, \hat{\tau}_1, \sigma_0^2, \sigma_1^2)$$

of τ , which depends only on observed data $(\hat{\tau}_0, \hat{\tau}_1)$ and known variances (σ_0^2, σ_1^2) , but not on the unknown bias Δ . If Δ were known, constructing an exact two-sided confidence interval would be straightforward, as the distributions of both estimators are known, which leads to the known distribution of $\hat{\tau}$. However, since Δ is unknown, the exact confidence interval depends on the magnitude of Δ .

To analyze the problem, we impose bounds on the bias Δ and construct two-sided confidence intervals for different levels of bias. We assume that $|\Delta/\sigma_0| \leq b$ for some $b \geq 0$, and study how the confidence interval based on $\hat{\tau}$ changes as a function of the bias bound b . Here we focus on the relative bias Δ/σ_0 instead of the absolute bias Δ to ensure that the parameterization is invariant to the scale of the estimators. For a given b , we aim to construct a two-sided confidence interval that achieves correct coverage uniformly over all Δ satisfying $|\Delta/\sigma_0| \leq b$. We thus define the confidence interval below.

Definition 1.2.1. Given a significance level $\zeta \in (0, 1)$ and a maximum relative bias $b \geq 0$, we want to construct an interval $\mathcal{I}(b, \zeta) = \mathcal{I}(b, \zeta, \hat{\tau}_0, \hat{\tau}_1, \sigma_0^2, \sigma_1^2)$ such that

$$\inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(\tau \in \hat{\tau} - \mathcal{I}(b, \zeta)) = \inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(\hat{\tau} - \tau \in \mathcal{I}(b, \zeta)) \geq 1 - \zeta, \quad (1.2.1)$$

where P_{Δ} explicitly denotes the dependence of the distribution of $\hat{\tau}$ on Δ . The confidence interval for τ based on $\hat{\tau}$ is then given by $\hat{\tau} - \mathcal{I}(b, \zeta)$.

We next impose two natural monotonicity conditions on the interval $\mathcal{I}(b, \zeta)$ below. We want the family of confidence intervals $\{\hat{\tau} - \mathcal{I}(b, \zeta)\}_{b \geq 0}$ to evolve in a coherent way as b changes. In particular, although Definition 1.2.1 concerns validity at a fixed bias bound b , for sensitivity analysis it is useful to require that larger bias bounds lead to larger confidence intervals. This monotone structure will also ensure that the b-value defined in Definition 1.2.2 later has a natural threshold interpretation.

Assumption 1.2.2. *We assume:*

1. For fixed $(\hat{\tau}_0, \hat{\tau}_1, \sigma_0^2, \sigma_1^2)$ and ζ , we have $\mathcal{I}(b, \zeta) \subset \mathcal{I}(b', \zeta)$ whenever $b \leq b'$.
2. For fixed $(\hat{\tau}_0, \hat{\tau}_1, \sigma_0^2, \sigma_1^2)$ and b , we have $\mathcal{I}(b, \zeta) \subset \mathcal{I}(b, \zeta')$ whenever $\zeta \geq \zeta'$.

The first condition in Assumption 1.2.2 requires that as we allow greater bias in $\hat{\tau}_1$, the confidence interval becomes wider and contains the previous intervals. The second condition in Assumption 1.2.2 requires that to guarantee higher coverage rate, the confidence interval must widen. A common class of intervals satisfying both conditions is the class of symmetric fixed-length centered intervals, whose length does not depend on $(\hat{\tau}_0, \hat{\tau}_1)$, i.e., $\mathcal{I}(b, \zeta) = [-c(b, \zeta, \sigma_0^2, \sigma_1^2), c(b, \zeta, \sigma_0^2, \sigma_1^2)]$ for some $c(b, \zeta, \sigma_0^2, \sigma_1^2) \geq 0$ depending only on b , ζ , σ_0^2 , and σ_1^2 . In this case, the confidence interval $\hat{\tau} - \mathcal{I}(b, \zeta)$ in Definition 1.2.1 is given by $[\hat{\tau} - c(b, \zeta, \sigma_0^2, \sigma_1^2), \hat{\tau} + c(b, \zeta, \sigma_0^2, \sigma_1^2)]$.

Therefore, for given $(\hat{\tau}_0, \hat{\tau}_1, \sigma_0^2, \sigma_1^2)$ and ζ , we can regard the confidence interval $\hat{\tau} - \mathcal{I}(b, \zeta)$ in Definition 1.2.1 as a function of the bias bound b . Thus, to address the central question

of constructing a sequence of two-sided confidence intervals for τ across different levels of bias, we compute this confidence interval for a range of values of b : $\{\hat{\tau} - \mathcal{I}(b, \zeta)\}_{b \geq 0}$.

Once we construct the sequence of confidence intervals, a natural application is hypothesis testing about the parameter τ . We focus on the two-sided test:

$$H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau \neq 0. \quad (1.2.2)$$

We generalize the discussion to one-sided tests, such as testing $\tau = 0$ versus $\tau > 0$ or testing $\tau \leq 0$ versus $\tau > 0$, in Section A.1.2. It is straightforward to extend the two-sided test to hypotheses of the form $\tau = \tau^*$ versus $\tau \neq \tau^*$ for any given $\tau^* \in \mathbb{R}$. A common practice for the two-sided test involves constructing confidence intervals for τ and checking whether the null value 0 lies within these intervals.

Under Assumption 1.2.2, the width of the confidence interval increases with the bias bound b . This leads to a central question when using the combined estimator $\hat{\tau}$ with the confidence interval $\hat{\tau} - \mathcal{I}(b, \zeta)$:

How large must the bias bound b be to change the conclusion of the hypothesis test (1.2.2)?

By the monotonicity of $\mathcal{I}(b, \zeta)$ in b (holding $(\hat{\tau}_0, \hat{\tau}_1)$ and ζ fixed) from Assumption 1.2.2, we define the limiting interval as $\mathcal{I}(\infty, \zeta) = \lim_{b \rightarrow \infty} \mathcal{I}(b, \zeta)$. Then there may exist a critical value of b at which the confidence interval $\hat{\tau} - \mathcal{I}(b, \zeta)$ contains the null value tested in (1.2.2). We focus on cases where the critical value exists. Specifically, we consider the case when $0 \notin \hat{\tau} - \mathcal{I}(0, \zeta)$ but $0 \in \hat{\tau} - \mathcal{I}(\infty, \zeta)$. In this scenario, we define the b-value below.

Definition 1.2.2. Define the b-value as the critical value b^* of testing $\tau = 0$ versus $\tau \neq 0$:

$$b^*(\zeta) = b^*(\zeta, \hat{\tau}_0, \hat{\tau}_1, \sigma_0^2, \sigma_1^2) = \inf \{b \geq 0 : 0 \in \hat{\tau} - \mathcal{I}(b, \zeta)\}. \quad (1.2.3)$$

Although we can define the quantity in (1.2.3) without Assumption 1.2.2, the monotonicity conditions in Assumption 1.2.2 ensure that the b-value is the critical bias level at which the testing conclusion changes. In particular, by the monotonicity conditions in Assumption 1.2.2, we have $0 \notin \hat{\tau} - \mathcal{I}(b, \zeta)$ for $0 \leq b < b^*$ and $0 \in \hat{\tau} - \mathcal{I}(b, \zeta)$ for $b > b^*$. Thus, we reject the null hypothesis $\tau = 0$ when $b < b^*$ and fail to reject it for $b > b^*$. The b-value b^* thus represents the maximum relative bias beyond which the null hypothesis can no longer be rejected. We propose to report b^* defined in (1.2.3) for given estimators $(\hat{\tau}_0, \hat{\tau}_1)$ and significance level ζ .

Compared with the sensitivity analysis literature, the bias bound b in our framework plays a role analogous to the sensitivity parameter. The corresponding confidence interval $\hat{\tau} - \mathcal{I}(b, \zeta)$ serves as a sensitivity curve, illustrating how inference changes as the bias varies. In this context, the b-value b^* serves a role similar to key robustness metrics in prior work: it parallels the E-value by VanderWeele and Ding (2017), the robustness value by Cinelli and Hazlett (2020), the breakdown point and breakdown frontier by Horowitz and Manski (1995) and Masten and Poirier (2020). The b-value is also analogous to the minimal level

of selection necessary to undermine the conclusion in [Kline and Santos \(2013\)](#) for assessing the sensitivity to departures from the missing at random assumption.

The b-value b^* provides a criterion for comparing competing strategies of confidence interval construction. Different choices of the combined estimator $\hat{\tau}$ and different formulations of $\mathcal{I}(b, \zeta)$ can lead to different values of b^* . We prefer procedures that yield a larger b-value b^* , since this indicates that the resulting confidence interval is more robust to potential bias in the biased estimator. This role of b^* is analogous to that use of the design sensitivity in [Rosenbaum \(2004\)](#), where it serves as a criterion for comparing different test statistics and matched designs in observational studies.

Remark 1.2.1. Besides the scenario of primary interest: $0 \notin \hat{\tau} - \mathcal{I}(0, \zeta)$ and $0 \in \hat{\tau} - \mathcal{I}(\infty, \zeta)$, two other less interesting scenarios may occur. First, if $0 \in \hat{\tau} - \mathcal{I}(0, \zeta)$, then $0 \in \hat{\tau} - \mathcal{I}(b, \zeta)$ for all $b \geq 0$, and we always fail to reject the null hypothesis regardless of the bias magnitude. Second, if $0 \notin \hat{\tau} - \mathcal{I}(\infty, \zeta)$, then $0 \notin \hat{\tau} - \mathcal{I}(b, \zeta)$ for all $b \geq 0$, and we always reject the null hypothesis regardless of the bias magnitude. In these two scenarios, combining the estimators or not does not change the statistical result qualitatively. Therefore, we ignore them in our discussion. Depending on the choice of $\hat{\tau}$ and construction of $\mathcal{I}(b, \zeta)$, not all three scenarios may arise. We focus on the case when the b-value b^* defined in (1.2.3) is in the interval $(0, \infty)$.

1.2.2 Point estimation: a review

Before diving into the details of the method for constructing confidence intervals and obtaining the b-value, we first review the existing approaches to point estimation. Specifically, we review three point estimators: the precision-weighted estimator, the pretest estimator, and the soft-thresholding estimator.

First, we recall the precision-weighted estimator:

$$\hat{\tau}_{\text{PW}} := \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2} \hat{\tau}_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2} \hat{\tau}_1 = \hat{\tau}_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2} (\hat{\tau}_1 - \hat{\tau}_0) = \hat{\tau}_0 + \frac{\gamma}{1 + \gamma} (\hat{\tau}_1 - \hat{\tau}_0).$$

When Δ is known to be 0, the precision-weighted estimator is the maximum likelihood estimator and best linear unbiased estimator of τ . Moreover, by the classical Le Cam asymptotic decision theory ([Le Cam, 1956](#)) and classical results for the Gaussian shift model, it is asymptotically admissible and minimax under L_2 risk in general, i.e., $E[(\hat{\tau} - \tau)^2]$, among all regular estimators under standard regularity conditions.

However, $\hat{\tau}_{\text{PW}}$ is not robust to bias: its risk is large when $|\Delta|$ is large. This motivates us to consider a combined estimator that performs nearly as well as $\hat{\tau}_{\text{PW}}$ when the bias is small, and is robust to unknown bias Δ , ensuring that the worst-case risk $\sup_{\Delta \in \mathbb{R}} E_{\Delta}[(\hat{\tau} - \tau)^2]$ remains bounded. This motivates the following two estimators.

Second, we recall the pretest estimator, which involves incorporating a pretest for $\Delta = 0$ versus $\Delta \neq 0$, a procedure that is commonly used ([Bancroft, 1944](#); [Wallace, 1977](#); [Bancroft and Han, 1977](#); [Giles and Giles, 1993](#)). Under $\Delta = 0$, given the independence between $\hat{\tau}_1$ and

$\hat{\tau}_0$, their difference follows: $\hat{\tau}_1 - \hat{\tau}_0 \sim N(0, \sigma^2)$, where $\sigma^2 = \sigma_0^2 + \sigma_1^2$. For a fixed significance level $\alpha \in (0, 1)$, we consider the test statistic $(\hat{\tau}_1 - \hat{\tau}_0)/\sigma$. Let

$$\mathcal{A} = \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\}$$

denote the event that the pretest fails to reject the null hypothesis ($\Delta = 0$). If the pretest fails to reject the null hypothesis, i.e., $|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}$, the pretest estimator uses the precision-weighted estimator:

$$\hat{\tau}_0 + \frac{\gamma}{1 + \gamma}(\hat{\tau}_1 - \hat{\tau}_0).$$

If the pretest rejects the null hypothesis, i.e., $|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}$, the pretest estimator employs hard-thresholding by reverting to the unbiased estimator $\hat{\tau}_0$. Combining the two cases, the pretest estimator is:

$$\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + \frac{\gamma}{1 + \gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbb{1}(\mathcal{A}).$$

Third, we recall the soft-thresholding estimator, which ensures continuity at the pretest boundary ($|\hat{\tau}_1 - \hat{\tau}_0| = \sigma c_{\alpha/2}$). If the pretest fails to reject the null hypothesis, the soft-thresholding estimator also uses the precision-weighted estimator. If the pretest rejects the null hypothesis, the soft-thresholding estimator employs soft-thresholding by:

$$\hat{\tau}_0 + \frac{\gamma}{1 + \gamma}\sigma c_{\alpha/2} \text{sign}(\hat{\tau}_1 - \hat{\tau}_0).$$

Combining the two cases, the soft-thresholding estimator is:

$$\hat{\tau}_{\text{ST}} = \hat{\tau}_0 + \frac{\gamma}{1 + \gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbb{1}(\mathcal{A}) + \frac{\gamma}{1 + \gamma}\sigma c_{\alpha/2} \text{sign}(\hat{\tau}_1 - \hat{\tau}_0)\mathbb{1}(\mathcal{A}^c).$$

While $\hat{\tau}_{\text{PT}}$ reduces to $\hat{\tau}_0$ with probability nearly one when the bias is large, the risk of $\hat{\tau}_{\text{PT}}$ is much higher than that of $\hat{\tau}_{\text{ST}}$ when the bias is moderate. See [Bickel \(1983\)](#) and [Armstrong et al. \(2025b\)](#) for a comparison of worst-case risk between $\hat{\tau}_{\text{PT}}$ and $\hat{\tau}_{\text{ST}}$ and the superior performance of $\hat{\tau}_{\text{ST}}$. See also [Chen et al. \(2025, Section 3.2\)](#) for optimality results for $\hat{\tau}_{\text{ST}}$.

In both $\hat{\tau}_{\text{PT}}$ and $\hat{\tau}_{\text{ST}}$, the role of α is different from its usual interpretation in hypothesis testing. Rather than a Type I error rate, α here acts as a tuning parameter that balances the bias and variance of the point estimator. In this sense, α simply indexes a family of estimators, much like the penalty level in penalized regression methods (e.g., the Lasso). Choosing α optimally depends on Δ , which is unknown in practice. In our empirical studies, we set $\alpha = 0.05$ as a default, following the standard convention, and examine how the estimator's performance varies across different bias levels.

1.3 Confidence intervals, hypothesis testing, and the b-value

Given the canonical point estimators introduced in [Section 1.2.2](#), we now discuss the problem of constructing confidence intervals, hypothesis testing, and the b-value.

1.3.1 Confidence interval based on the precision-weighted estimator

As a warm-up, we use the precision-weighted estimator $\hat{\tau}_{\text{PW}}$ to construct the sequence of confidence intervals and to illustrate how the confidence interval changes with respect to b . Based on the theory below, we do not recommend using the precision-weighted estimator and its corresponding confidence intervals in practice. Under Assumption 1.2.1, we have $\hat{\tau}_{\text{PW}} \sim N(\tau + (1 + \gamma)^{-1}\gamma\Delta, (1 + \gamma)^{-1}\sigma_0^2)$. The following theorem provides the sequence of confidence intervals based on $\hat{\tau}_{\text{PW}}$ depending on b .

Theorem 1.3.1. *Let $\hat{L}_{\text{PW}} = \hat{L}_{\text{PW}}(b, \zeta, \gamma) \geq 0$ denote the solution to the equation of L :*

$$\Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}b\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1+\gamma}}b\right) = 1 - \zeta.$$

The $\hat{L}_{\text{PW}}$ always exists and is unique. The shortest length symmetric centered confidence interval based on $\hat{\tau}_{\text{PW}}$ for τ satisfying (1.2.1) is given by $[\hat{\tau}_{\text{PW}} - \hat{L}_{\text{PW}}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{PW}} + \hat{L}_{\text{PW}}(1 + \gamma)^{-1/2}\sigma_0]$.

The $\hat{L}_{\text{PW}}$ in Theorem 1.3.1 corresponds to the $(1 - \zeta)$ quantile of the folded normal distribution $|N(\frac{\gamma}{\sqrt{1+\gamma}}b, 1)|$. This distribution also arises in the regression literature, where worst-case bias is incorporated into the construction of confidence intervals (Armstrong et al., 2020, 2025a).

In the special case when $b = 0$, we have $\hat{L}_{\text{PW}}(0, \zeta, \gamma) = c_{\zeta/2}$, which yields the length of the confidence interval based on $\hat{\tau}_{\text{PW}}$ to be $2c_{\zeta/2}(1 + \gamma)^{-1/2}\sigma_0$. For comparison, recall that the length of the confidence interval solely based on $\hat{\tau}_0$ is $2c_{\zeta/2}\sigma_0$. Thus the length of the confidence interval reduces by a factor of $(1 + \gamma)^{-1/2}$ when using $\hat{\tau}_{\text{PW}}$ instead of $\hat{\tau}_0$, which is small when γ is large, i.e., σ_0^2 is much larger than σ_1^2 . This indicates that by using a more precise but potentially biased estimator, if we have prior knowledge that the bias is small, then we can achieve a much shorter confidence interval. Given ζ and γ , the $\hat{L}_{\text{PW}}(b, \zeta, \gamma)$ in Theorem 1.3.1 increases as b increases, and goes to infinity as $b \rightarrow \infty$. Therefore, the precision-weighted estimator is not robust to unknown bias since the confidence interval is not bounded as the bias diverges.

One interesting observation is that $\hat{L}_{\text{PW}}$ depends on σ_0^2 and σ_1^2 only through γ , the variance ratio. This explains why the bias Δ is scaled by σ_0 in the definition of b , and the confidence interval is represented as $[\hat{\tau}_{\text{PW}} - \hat{L}_{\text{PW}}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{PW}} + \hat{L}_{\text{PW}}(1 + \gamma)^{-1/2}\sigma_0]$.

1.3.2 Confidence interval based on the pretest estimator

We now construct the confidence interval for the pretest estimator $\hat{\tau}_{\text{PT}}$. In the following lemma, we present the distribution of $\hat{\tau}_{\text{PT}} - \tau$.

Lemma 1.3.1. *Let Z_1, Z_2 be two independent standard normal random variables. Then $\hat{\tau}_{PT} - \tau$ is distributed as*

$$\frac{1}{\sqrt{1+\gamma}}\sigma_0 Z_1 + \frac{\gamma}{1+\gamma}\Delta \mathbf{1} \left(\left| Z_2 + \sqrt{\frac{\gamma}{1+\gamma}} \frac{\Delta}{\sigma_0} \right| \leq c_{\alpha/2} \right) - \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0 Z_2 \mathbf{1} \left(\left| Z_2 + \sqrt{\frac{\gamma}{1+\gamma}} \frac{\Delta}{\sigma_0} \right| > c_{\alpha/2} \right).$$

By Lemma 1.3.1, the distribution of $\hat{\tau}_{PT}$ is a mixture of a normal distribution and a truncated normal distribution. We define $\hat{L}_{PT} = \hat{L}_{PT}(b, \zeta, \sigma_0^2, \sigma_1^2, \alpha)$ as the smallest length such that the confidence interval $[\hat{\tau}_{PT} - \hat{L}_{PT}(1+\gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1+\gamma)^{-1/2}\sigma_0]$ achieves correct coverage for all Δ satisfying $|\Delta/\sigma_0| \leq b$, where we choose the same scale as in Theorem 1.3.1 for comparison. By Definition 1.2.1, we can formulate $\hat{L}_{PT}$ as the optimization problem

$$\hat{L}_{PT} = \hat{L}_{PT}(b, \zeta, \sigma_0^2, \sigma_1^2, \alpha) = \inf \left\{ L \geq 0 : \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbf{P}_{\Delta}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) \geq 1 - \zeta \right\}. \quad (1.3.1)$$

However, $\hat{L}_{PT}$ generally does not admit a closed-form expression. Moreover, direct computation of $\hat{L}_{PT}$ based on (1.3.1) is computationally challenging since for each $L \geq 0$, the optimization problem (1.3.1) involves finding the infimum of $\mathbf{P}_{\Delta}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0)$ over all Δ satisfying $|\Delta/\sigma_0| \leq b$. Nevertheless, we show that computing $\hat{L}_{PT}$ is tractable due to the following theorem:

Theorem 1.3.2. *For any $L > 0$, $\mathbf{P}_{\Delta}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0)$ as a function of Δ is symmetric about $\Delta = 0$. Then $\hat{L}_{PT} = \hat{L}_{PT}(b, \zeta, \gamma, \alpha)$ in (1.3.1) is the solution to the following equation of L :*

$$\min_{0 \leq t \leq b} \mathbf{P}_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) = 1 - \zeta,$$

where

$$\begin{aligned} & \mathbf{P}_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) \right] \left[\Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \right] \\ & \quad + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \left[\Phi\left(L + \sqrt{\gamma}u\right) - \Phi\left(-L + \sqrt{\gamma}u\right) \right] \phi(u) du \\ & \quad + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \left[\Phi\left(L + \sqrt{\gamma}u\right) - \Phi\left(-L + \sqrt{\gamma}u\right) \right] \phi(u) du. \end{aligned} \quad (1.3.2)$$

$[\hat{\tau}_{PT} - \hat{L}_{PT}(1+\gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1+\gamma)^{-1/2}\sigma_0]$ is the shortest length symmetric centered confidence interval based on $\hat{\tau}_{PT}$ for τ satisfying (1.2.1).

We have two main observations from Theorem 1.3.2. First, for any $L > 0$, the coverage probability $P_\Delta(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ is symmetric about $\Delta = 0$. As a result, instead of minimizing over all Δ satisfying $|\Delta/\sigma_0| \leq b$ as in (1.3.1), we can restrict attention to the interval $0 \leq \Delta/\sigma_0 \leq b$. Second, Theorem 1.3.2 provides an explicit expression for the coverage probability $P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ as a function of t and L . The first term in (1.3.2) corresponds to the event that the pretest fails to reject the null hypothesis ($\Delta = 0$), in which case $\hat{\tau}_{PT}$ reduces to the precision-weighted estimator. The second and third terms in (1.3.2) integrate over the regions where the pretest rejects the null hypothesis, in which case $\hat{\tau}_{PT}$ reduces to the unbiased estimator.

Although the probability $P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ admits a closed form for any $L > 0$, the minimum of $P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ over $0 \leq t \leq b$ is not explicitly available in closed form, and thus must be computed numerically. Since the function $P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ is monotonically increasing with respect to L for $L \geq 0$, L_{PT}^* can be computed numerically using, for example, the bisection method, combined with numerical computation of the minimum of $P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ with respect to t for each candidate L .

1.3.3 Confidence interval based on the soft-thresholding estimator

We now construct the confidence interval for the soft-thresholding estimator $\hat{\tau}_{ST}$. In the following lemma, we present the distribution of $\hat{\tau}_{ST} - \tau$.

Lemma 1.3.2. *Let Z_1, Z_2 be two independent standard normal random variables. Then $\hat{\tau}_{ST} - \tau$ is distributed as*

$$\begin{aligned} & \frac{1}{\sqrt{1+\gamma}}\sigma_0 Z_1 + \frac{\gamma}{1+\gamma}\Delta \mathbb{1} \left(\left| Z_2 + \sqrt{\frac{\gamma}{1+\gamma}} \frac{\Delta}{\sigma_0} \right| \leq c_{\alpha/2} \right) \\ & - \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0 \left(Z_2 - c_{\alpha/2} \text{sign} \left(Z_2 + \sqrt{\frac{\gamma}{1+\gamma}} \frac{\Delta}{\sigma_0} \right) \right) \mathbb{1} \left(\left| Z_2 + \sqrt{\frac{\gamma}{1+\gamma}} \frac{\Delta}{\sigma_0} \right| > c_{\alpha/2} \right). \end{aligned}$$

By Lemma 1.3.2, the distribution of $\hat{\tau}_{ST}$ is a mixture of a normal distribution and a truncated normal distribution. Suppose $|\Delta/\sigma_0| \leq b$ for some $b > 0$. We seek the shortest length $\hat{L}_{ST} = \hat{L}_{ST}(b, \zeta, \sigma_0^2, \sigma_1^2, \alpha)$ such that the confidence interval $[\hat{\tau}_{ST} - \hat{L}_{ST}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{ST} + \hat{L}_{ST}(1 + \gamma)^{-1/2}\sigma_0]$ achieves correct coverage uniformly over all Δ with $|\Delta/\sigma_0| \leq b$. By Definition 1.2.1, we can formulate $\hat{L}_{ST}$ as the optimization problem:

$$\hat{L}_{ST} = \inf \left\{ L \geq 0 : \inf_{\Delta: |\Delta/\sigma_0| \leq b} P_\Delta(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) \geq 1 - \zeta \right\}. \quad (1.3.3)$$

However, $\hat{L}_{ST}$ generally does not admit a closed-form expression. Moreover, direct computation of $\hat{L}_{ST}$ based on (1.3.3) is computationally challenging since for each $L \geq 0$, the optimization problem (1.3.3) involves finding the infimum of $P_\Delta(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ over all Δ satisfying $|\Delta/\sigma_0| \leq b$. Nevertheless, we show that $\hat{L}_{ST}$ can be computed efficiently due to the following theorem:

Theorem 1.3.3. *For any $L > 0$, $P_{\Delta}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ as a function of Δ is symmetric about $\Delta = 0$ and monotonically decreasing in $|\Delta|$. Then $\hat{L}_{ST} = \hat{L}_{ST}(b, \zeta, \gamma, \alpha)$ in (1.3.3) is the solution to the following equation of L :*

$$P_{\Delta/\sigma_0=b}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta,$$

where

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) \right] \left[\Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \right] \\ &+ \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \left[\Phi\left(L + \sqrt{\gamma}(u + c_{\alpha/2})\right) - \Phi\left(-L + \sqrt{\gamma}(u + c_{\alpha/2})\right) \right] \phi(u) du \\ &+ \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \left[\Phi\left(L + \sqrt{\gamma}(u - c_{\alpha/2})\right) - \Phi\left(-L + \sqrt{\gamma}(u - c_{\alpha/2})\right) \right] \phi(u) du. \end{aligned} \quad (1.3.4)$$

$[\hat{\tau}_{ST} - \hat{L}_{ST}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{ST} + \hat{L}_{ST}(1 + \gamma)^{-1/2}\sigma_0]$ is the shortest length symmetric centered confidence interval based on $\hat{\tau}_{ST}$ for τ satisfying (1.2.1).

We have two main observations from Theorem 1.3.3. First, for any $L > 0$, the coverage probability $P_{\Delta}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ is symmetric about $\Delta = 0$ and, unlike the pretest estimator, decreases monotonically in $|\Delta|$. This monotonicity implies that the worst-case coverage over the bias Δ satisfying $|\Delta/\sigma_0| \leq b$ as in (1.3.3) is always attained at the boundary $\Delta/\sigma_0 = b$. The monotonicity of the coverage probability $P_{\Delta}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ makes the computation of $\hat{L}_{ST}$ more efficient than that of $\hat{L}_{PT}$. Second, Theorem 1.3.3 provides an explicit expression for the coverage probability $P_{\Delta/\sigma_0=t}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ as a function of t and L . The first term in (1.3.4) corresponds to the event that the pretest fails to reject the null hypothesis ($\Delta = 0$), in which case $\hat{\tau}_{ST}$ reduces to the precision-weighted estimator. The second and third terms in (1.3.4) integrate over the regions where the pretest rejects the null hypothesis, in which case $\hat{\tau}_{ST}$ reduces to the unbiased estimator with a constant shift to ensure continuity at the pretest boundary.

Since $P_{\Delta/\sigma_0=b}(|\hat{\tau}_{ST} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0)$ is monotonically increasing in L , $\hat{L}_{ST}$ can be efficiently computed by, for example, the bisection method.

1.3.4 Comparison of the point estimators and confidence intervals

Before comparing the point estimators and confidence intervals, we note that the point estimators considered in this paper are all anchored at the unbiased estimator $\hat{\tau}_0$ and incorporate the biased but more precise estimator $\hat{\tau}_1$ only through shrinkage toward it. Compared with $\hat{\tau}_{PW}$, $\hat{\tau}_{ST}$ has bounded worst-case risk. The mean squared error of $\hat{\tau}_{ST}$ remains bounded regardless of the magnitude of the bias, whereas the mean squared error of $\hat{\tau}_{PW}$ grows without bound as the bias increases. Thus, $\hat{\tau}_{ST}$ offers a more balanced compromise between

efficiency and robustness: it retains efficiency comparable to $\hat{\tau}_{PW}$ when the bias is small, yet its performance is comparable to the unbiased estimator when the bias is large.

Compared with $\hat{\tau}_{PT}$, $\hat{\tau}_{ST}$ enjoys a desirable monotonicity property: for any $L > 0$, the probability $P_{\Delta}(|\hat{\tau}_{ST} - \tau| \leq L)$ decreases monotonically in $|\Delta|$ (Theorem 1.3.3). This behavior parallels the result in Bickel (1983), which shows that the mean squared error of $\hat{\tau}_{ST}$ increases monotonically in $|\Delta|$. This monotonicity of $\hat{\tau}_{ST}$ allows us to compute the confidence interval based on $\hat{\tau}_{ST}$ more efficiently than $\hat{\tau}_{PT}$, as shown in Theorem 1.3.3.

Figure 1.1 compares the confidence intervals based on $\hat{\tau}_{ST}$, $\hat{\tau}_{PT}$, $\hat{\tau}_{PW}$, and the unbiased estimator $\hat{\tau}_0$, as the bias bound b varies. We fix $\sigma_0^2 = 1$ and $\hat{\tau}_0 = 1$. Let $\zeta = 0.05$ and $\alpha = 0.05$. We set $\hat{\tau}_1 = 2$ and examine two scenarios: $\gamma = 10$ and $\gamma = 100$. We observe that when the bias bound b is small, the confidence intervals based on $\hat{\tau}_{ST}$, $\hat{\tau}_{PT}$, and $\hat{\tau}_{PW}$ are all shorter than the one based on $\hat{\tau}_0$. Importantly, when the bias bound b is small, the confidence interval based on $\hat{\tau}_{ST}$ is shorter than $\hat{\tau}_{PT}$ and is comparable to $\hat{\tau}_{PW}$. Furthermore, the confidence interval length based on $\hat{\tau}_{ST}$ remains bounded, whereas that based on $\hat{\tau}_{PW}$ grows without bound as b increases. Therefore, $\hat{\tau}_{ST}$ combines the efficiency gains of precision-weighted methods when bias is small with robustness to large biases, making it a superior choice for inference in practice.

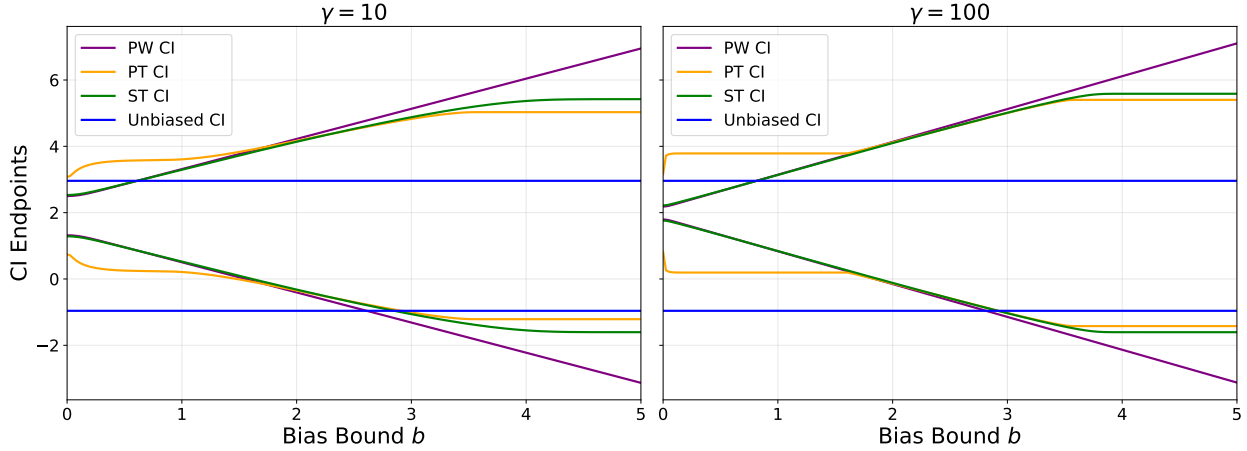


Figure 1.1: Confidence intervals against the maximum relative bias $|\Delta/\sigma_0| \leq b$

Remark 1.3.1 (Computing the b-value). After plotting the estimators and their confidence intervals as in Figure 1.1, we are ready to read the b-value for each estimator given the significance level ζ . Since the sequence of confidence intervals is different for different estimators, the b-values are also different. Alternatively, we can also compute the b-value directly. A naive way is to use the bisection method by the monotonicity of the confidence interval, which involves computing the confidence interval for each possible bias level b , and see whether the confidence interval contains the null value. However, this procedure is computationally heavy since the computation of the confidence interval for a given b also involves

the bisection method when based on $\hat{\tau}_{PT}$ and $\hat{\tau}_{ST}$. In Appendix A.1.1, we propose a method to compute the b-value efficiently.

1.3.5 Generalization to the dependent case

So far we assume independent unbiased and biased estimators. In this section, we generalize the above discussion to the case where $\hat{\tau}_0$ and $\hat{\tau}_1$ are jointly normal with known correlation ρ . We assume $\rho\sigma_1 \neq \sigma_0$, which trivially holds when $\sigma_0^2 > \sigma_1^2$, or equivalently $\gamma > 1$. The key is to construct a biased estimator which is independent of $\hat{\tau}_0$ without losing information of $\hat{\tau}_1$. To achieve that, we define the reparametrization

$$\hat{\tau}'_1 = \frac{\hat{\tau}_1 - (\rho\sigma_1/\sigma_0)\hat{\tau}_0}{1 - \rho\sigma_1/\sigma_0}. \quad (1.3.5)$$

Then, $\hat{\tau}_0$ and $\hat{\tau}'_1$ are independent, with $\hat{\tau}'_1 \sim N(\tau + \Delta', \sigma_1'^2)$, where

$$\Delta' = \frac{\Delta}{1 - \rho\sigma_1/\sigma_0}, \quad \sigma_1'^2 = \frac{(1 - \rho^2)\sigma_1^2}{(1 - \rho\sigma_1/\sigma_0)^2}.$$

Thus, the problem reduces to the independent case considered above.

In the reparametrization in (1.3.5), we need to know how the transformation rescales the bias and how to interpret the relative bias. We can compute the confidence intervals and the b-value based on the reparametrization with independent unbiased estimator $\hat{\tau}_0$ and transformed biased estimator $\hat{\tau}'_1$. Let the confidence intervals and the b-value in the transformed problem be $\hat{\tau} - \mathcal{I}'(b, \zeta)$ and b^* , respectively. Then the confidence intervals for the original problem are $\hat{\tau} - \mathcal{I}(b, \zeta) = \hat{\tau} - \mathcal{I}'(b/|1 - \rho\sigma_1/\sigma_0|, \zeta)$ and the b-value is $b^* = |1 - \rho\sigma_1/\sigma_0|b^*$, respectively. We relegate the technical details to Appendix A.1.4.

1.4 Generalization to the multivariate case

In this section, we extend our framework to the multivariate case. Many applications involve vector-valued parameters rather than scalars. Consider two leading examples in causal inference. First, when a treatment has multiple levels, the parameter of interest is a vector of treatment effects across those levels. This setting also includes factorial designs, where researchers aim to estimate factorial effects jointly. Second, when the population is partitioned into subgroups, the focus is often on subgroup treatment effects, yielding a vector of conditional average treatment effects indexed by the subgroup variable (Schwartz et al., 2026). Beyond causal inference, similar issues arise in regression settings. For instance, when combining OLS and IV estimators, the target parameter can be a vector of regression coefficients. Therefore, generalization to the multivariate case is essential for applications where researchers must make inferences about multiple parameters.

1.4.1 Setup

We consider the following multivariate setting:

Assumption 1.4.1. *Suppose we observe two independent random vectors: one unbiased estimator $\hat{\tau}_0 \sim N(\tau, \Sigma_0)$ and one biased estimator $\hat{\tau}_1 \sim N(\tau + \Delta, \Sigma_1)$. Here $\tau \in \mathbb{R}^d$ is the unknown parameter of interest. We assume that Σ_0 and Σ_1 are known but Δ is unknown.*

As in the univariate case, we present our analysis in the exact normality setting which asymptotically captures the essential structure of inference as in [Le Cam \(1956\)](#). Our goal is to construct confidence regions for τ at different bias levels. Although we assume independent unbiased and biased estimators, the extension to the dependent case is straightforward; see [Section 1.3.5](#). Since $\hat{\tau}_0$ is unbiased for τ , given significance level $\zeta \in (0, 1)$, a natural confidence region based solely on $\hat{\tau}_0$ is the ellipsoid:

$$\{\tau \in \mathbb{R}^d : (\hat{\tau}_0 - \tau)^\top \Sigma_0^{-1} (\hat{\tau}_0 - \tau) \leq \chi_{d,1-\zeta}^2\}, \quad (1.4.1)$$

where $\chi_{d,1-\zeta}^2$ is the $(1 - \zeta)$ upper quantile of the chi-squared distribution with d degrees of freedom.

Then we consider combining the two estimators $\hat{\tau}_0$ and $\hat{\tau}_1$, and we consider a generic combined estimator $\hat{\tau} = \hat{\tau}(\hat{\tau}_0, \hat{\tau}_1, \Sigma_0, \Sigma_1)$. We assume that $|[\Sigma^{-1/2}\Delta]_j| \leq b_j$ for some $b_j \geq 0$ for all $j = 1, 2, \dots, d$ and study how the confidence region changes with the maximum relative bias vector $\mathbf{b} = (b_1, b_2, \dots, b_d)^\top$. Here $\Sigma = \Sigma(\Sigma_0, \Sigma_1) \in \mathbb{R}^{d \times d}$ can be any fixed positive definite scaling matrix that depends on Σ_0 and Σ_1 , which corresponds to σ_0^2 we used in the univariate case. For a given bias vector $\mathbf{b}$, different choices of Σ correspond to different regions for the bias Δ . In this paper, we do not discuss how to choose the scaling matrix Σ optimally. For simplicity, one may take $\Sigma = \mathbf{I}_d$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix, or set $\Sigma = \Sigma_0$ as in the univariate case. Analogous to [Definition 1.2.1](#) for the univariate case, the confidence region is defined below.

Definition 1.4.1. Given a significance level $\zeta \in (0, 1)$ and the maximum relative bias vector $\mathbf{b}$ with $b_j \geq 0$ for all $j = 1, 2, \dots, d$, we seek a region $\mathcal{I}(\mathbf{b}, \zeta) = \mathcal{I}(\mathbf{b}, \zeta, \hat{\tau}_0, \hat{\tau}_1, \Sigma_0, \Sigma_1)$ such that

$$\inf_{\Delta: |\Sigma^{-1/2}\Delta| \leq \mathbf{b}} P_{\Delta}(\tau \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)) = \inf_{\Delta: |\Sigma^{-1/2}\Delta| \leq \mathbf{b}} P_{\Delta}(\hat{\tau} - \tau \in \mathcal{I}(\mathbf{b}, \zeta)) \geq 1 - \zeta. \quad (1.4.2)$$

The confidence region based on $\hat{\tau}$ is then given by $\hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)$.

Then we introduce the monotonicity conditions on the region $\mathcal{I}(\mathbf{b}, \zeta)$ below, which generalizes [Assumption 1.2.2](#) to the multivariate case.

Assumption 1.4.2. *We assume:*

1. For fixed $(\hat{\tau}_0, \hat{\tau}_1, \Sigma_0, \Sigma_1)$ and ζ , we have $\mathcal{I}(\mathbf{b}, \zeta) \subset \mathcal{I}(\mathbf{b}', \zeta)$ whenever $\mathbf{b} \leq \mathbf{b}'$, i.e., $b_j \leq b'_j$ for all $j = 1, 2, \dots, d$.

2. For fixed $(\hat{\tau}_0, \hat{\tau}_1, \Sigma_0, \Sigma_1)$ and $\mathbf{b}$, we have $\mathcal{I}(\mathbf{b}, \zeta) \subset \mathcal{I}(\mathbf{b}, \zeta')$ whenever $\zeta \geq \zeta'$.

Assume $\mathcal{I}(\mathbf{b}, \zeta)$ satisfies the monotonicity conditions of Assumption 1.4.2. A common family of such regions is fixed-length centered ellipsoids: $\mathcal{I}(\mathbf{b}, \zeta) = \{\mathbf{h} \in \mathbb{R}^d : \mathbf{h}^\top \mathbf{A}^{-1} \mathbf{h} \leq c(\mathbf{b}, \zeta, \Sigma_0, \Sigma_1)\}$ with some constant $c(\mathbf{b}, \zeta, \Sigma_0, \Sigma_1) \geq 0$ and a positive definite matrix $\mathbf{A}$. In this case, the confidence region $\hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)$ in Definition 1.4.1 is given by $\{\boldsymbol{\tau} \in \mathbb{R}^d : (\hat{\tau} - \boldsymbol{\tau})^\top \mathbf{A}^{-1} (\hat{\tau} - \boldsymbol{\tau}) \leq c(\mathbf{b}, \zeta, \Sigma_0, \Sigma_1)\}$.

Remark 1.4.1. Unlike the univariate case, the construction of confidence regions in the multivariate setting is more complicated. First, in one dimension, the only convex confidence set is an interval, whereas in higher dimensions, there exists a wide variety of admissible convex confidence regions. Although ellipsoidal regions enjoy optimality properties under specific conditions (Stein, 1962; Wald, 1949), the appropriate geometry depends on the family of contrasts and the norm used to measure uncertainty. Ellipsoidal regions based on the Mahalanobis distance (Hotelling, 1931) arise naturally under the Gaussian model, but alternative geometries have been studied in the literature (Tukey, 1949; Scheffé, 1953; Šidák, 1967). Second, the optimal choice of the center of the confidence region is not straightforward. For $d \geq 3$, Stein's phenomenon implies that recentering at shrinkage estimators such as the James–Stein estimator can yield confidence regions with smaller volume while maintaining nominal coverage (Stein, 1962; Hwang and Casella, 1982; Berger, 1985). In this section we follow one specific track: constructing ellipsoidal confidence regions under the Mahalanobis distance, centered at prespecified estimators, in parallel with our discussion for the univariate case in Section 1.3. We do not focus on the optimality among all possible shapes or centers, but rather focus on this formulation for its analytical tractability and interpretability within our framework.

Then we define the multivariate b-value below.

Definition 1.4.2. Define the b-value as the critical boundary $\mathbf{b}^*$ of testing $\boldsymbol{\tau} = \mathbf{0}$ versus $\boldsymbol{\tau} \neq \mathbf{0}$ as

$$\mathbf{b}^*(\zeta) = \mathbf{b}^*(\zeta, \hat{\tau}_0, \hat{\tau}_1, \Sigma_0, \Sigma_1) = \partial \{\mathbf{b} \geq \mathbf{0} : \mathbf{0} \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)\}. \quad (1.4.3)$$

Definition 1.4.2 extends the univariate notion of the b-value (Definition 1.2.2) to the multivariate setting. In higher dimensions $d > 1$, the set $\{\mathbf{b} \geq \mathbf{0} : \mathbf{0} \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)\}$ is a convex region in the first quadrant. The b-value is defined as the boundary of this region, $\partial \{\mathbf{b} \geq \mathbf{0} : \mathbf{0} \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)\}$, which is a $(d - 1)$ -dimensional surface, e.g., a curve when $d = 2$, in the first quadrant. When $d = 1$, this boundary reduces to the single left endpoint $\inf \{b \geq 0 : 0 \in \hat{\tau} - \mathcal{I}(b, \zeta)\}$, which is exactly Definition 1.2.2. By the monotonicity conditions in Assumption 1.4.2, the geometry of this boundary yields a natural decision rule: we reject the null hypothesis $\boldsymbol{\tau} = \mathbf{0}$ if the bias bound vector $\mathbf{b}$ lies to the left of the b-value surface, and fail to reject it if $\mathbf{b}$ lies on or to the right of the b-value surface. To compute and visualize the multivariate b-value, we can plot the estimators and their confidence regions, and then we are ready to read the b-value for each estimator given the significance level ζ .

1.4.2 Point estimation

First, we recall the precision-weighted estimator defined as

$$\hat{\tau}_{\text{PW}} = (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}(\Sigma_0^{-1}\hat{\tau}_0 + \Sigma_1^{-1}\hat{\tau}_1) = \hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0).$$

The precision-weighted estimator is the maximum likelihood estimator of τ if Δ is known to be all zeros. However, its risk is large when $\|\Delta\|_2$ is large. Thus, we seek a combined estimator that performs nearly as well as $\hat{\tau}_{\text{PW}}$ when the bias is small, and is robust to unknown bias Δ , ensuring that the maximum risk $\sup_{\Delta \in \mathbb{R}^d} \mathbb{E}_{\Delta}[\|\hat{\tau}_{\text{PW}} - \tau\|_2^2]$ remains bounded. To develop estimators that perform well under zero bias yet remain robust to unknown bias, various estimators have been proposed in the empirical Bayes and shrinkage estimation literature (Berger, 1981; Bickel, 1984; Green and Strawderman, 1991; Green et al., 2005; Rosenman et al., 2023a,b). In this paper, we focus on the generic pretest estimator and the generic soft-thresholding estimator, as discussed below.

Second, we recall the pretest estimator, which involves incorporating a pretest for $\Delta = \mathbf{0}$ versus $\Delta \neq \mathbf{0}$. Under $\Delta = \mathbf{0}$, given the independence between $\hat{\tau}_1$ and $\hat{\tau}_0$, their difference follows: $\hat{\tau}_1 - \hat{\tau}_0 \sim N(\mathbf{0}, \Sigma_0 + \Sigma_1)$. We consider the test statistic $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2$ and the critical value $q \geq 0$. The critical value q here plays a similar role as the significance level α in Section 1.2.2. Let

$$\mathcal{A} = \{\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q\}$$

denote the event that the pretest fails to reject the null hypothesis ($\Delta = \mathbf{0}$). If the pretest fails to reject the null hypothesis, i.e., $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q$, the pretest estimator uses the precision-weighted estimator:

$$\hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0).$$

If the pretest rejects the null hypothesis, i.e., $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q$, the pretest estimator employs hard-thresholding by reverting to the unbiased estimator $\hat{\tau}_0$. Combining the two cases, the pretest estimator is:

$$\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(\mathcal{A}).$$

Third, we recall the soft-thresholding estimator, which ensures continuity at the pretest boundary ($\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 = q$). If the pretest fails to reject the null hypothesis, the soft-thresholding estimator also uses the precision-weighted estimator. If the pretest rejects the null hypothesis, the soft-thresholding estimator employs soft-thresholding by:

$$\hat{\tau}_0 + h_q^*(\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2)(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0),$$

where $h_q^*(\cdot) : [q, \infty) \rightarrow [0, 1]$ is a non-increasing function with $h_q^*(q) = 1$. Combining the two cases, the soft-thresholding estimator is:

$$\hat{\tau}_{\text{ST}} = \hat{\tau}_0 + h_q(\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2)(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0),$$

where $h_q(\cdot) : [0, \infty) \rightarrow [0, 1]$ is defined as $h_q(r) = \mathbb{1}(0 \leq r \leq q) + h_q^*(r)\mathbb{1}(r > q)$. Here both q and $h_q^*(\cdot)$ can depend on Σ_0 and Σ_1 .

The generic soft-thresholding estimator $\hat{\tau}_{\text{ST}}$ contains many classical estimators as special cases, such as the estimator studied in Berger (1981, Theorem 3) and Bickel (1984, Section 4), which generalizes the univariate soft-thresholding estimator $\hat{\tau}_{\text{ST}}$ to the multivariate setting. We provide a detailed discussion of the relationship between the generic soft-thresholding estimator and the estimator in Berger (1981) and Bickel (1984) in Appendix A.1.5. In this paper, we focus on general Σ_0 , Σ_1 , and $h_q^*(\cdot)$.

1.4.3 Confidence intervals

First, we construct the confidence regions based on the precision-weighted estimator $\hat{\tau}_{\text{PW}}$. Under Assumption 1.4.1, we have $\hat{\tau}_{\text{PW}} \sim N(\tau + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}\Delta, (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1})$. By using the ellipsoidal confidence region under the Mahalanobis distance and the covariance matrix of $\hat{\tau}_{\text{PW}}$, we consider the confidence region based on $\hat{\tau}_{\text{PW}}$ with the following form:

$$\{\tau \in \mathbb{R}^d : (\hat{\tau}_{\text{PW}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{\text{PW}} - \tau) \leq M\},$$

for some $M \geq 0$. The following theorem explicitly characterizes the confidence region based on $\hat{\tau}_{\text{PW}}$ as a function of bias bound $\mathbf{b}$.

Theorem 1.4.1. *Let $\widehat{M}_{\text{PW}} = \widehat{M}_{\text{PW}}(\mathbf{b}, \zeta, \Sigma_0, \Sigma_1, \Sigma) \geq 0$ be the $(1 - \zeta)$ upper quantile of the noncentral chi-squared distribution with d degrees of freedom and noncentrality parameter*

$$\sup_{\mathbf{s} \in \{\pm 1\}^d} \|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} \Sigma^{1/2} \mathbf{b} \odot \mathbf{s}\|_2^2.$$

The confidence region for τ satisfying (1.4.2) is given by $\{\tau \in \mathbb{R}^d : (\hat{\tau}_{\text{PW}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{\text{PW}} - \tau) \leq \widehat{M}_{\text{PW}}\}$.

Theorem 1.4.1 extends the univariate result in Theorem 1.3.1 to the multivariate setting. In the univariate case, taking $\Sigma = \sigma_0^2$ and $\mathbf{b} = b$ reduces the multivariate confidence region in Theorem 1.4.1 exactly to the confidence interval in Theorem 1.3.1. In the univariate case, the bias constraint $\{\Delta : |\Delta/\sigma_0| \leq b\}$ has only two boundary points, and the absolute bias of $\hat{\tau}_{\text{PW}}$ is the same at both endpoints. By contrast, the multivariate bias constraint $\{\Delta : |\Sigma^{-1/2}\Delta| \leq \mathbf{b}\}$ is a convex hyperrectangle with 2^d vertices, and the maximum L_2 norm of the bias of $\hat{\tau}_{\text{PW}}$ may occur at any of these vertices. Theorem 1.4.1 therefore characterizes the maximum possible bias of $\hat{\tau}_{\text{PW}}$ over the entire bias constraint.

In the special case when $\mathbf{b} = 0$, we have $\widehat{M}_{\text{PW}} = \chi_{d,1-\zeta}^2$. Compared with the ellipsoid confidence region based on $\hat{\tau}_0$, the scaling matrix in the definition of ellipsoid confidence region reduces from Σ_0^{-1} to $\Sigma_0^{-1} + \Sigma_1^{-1} = (\Sigma_1^{-1}\Sigma_0 + \mathbf{I}_d)\Sigma_0^{-1}$, analogous to the efficiency gain in the univariate case, where $\Sigma_1^{-1}\Sigma_0$ can be regarded as the variance ratio γ in the univariate case. As $\mathbf{b}$ increases, $\widehat{M}_{\text{PW}}$ increases, and goes to infinity as $\mathbf{b} \rightarrow \infty$.

Second, we construct the confidence regions based on the generic pretest estimator $\widehat{\boldsymbol{\tau}}_{\text{PT}}$. Similar to the confidence regions based on the precision-weighted estimator, we consider the confidence regions based on $\widehat{\boldsymbol{\tau}}_{\text{PT}}$ with the following form:

$$\{\boldsymbol{\tau} \in \mathbb{R}^d : (\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M\},$$

for some $M \geq 0$. Here, we focus on the ellipsoid confidence region defined using the same scaling matrix, $\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1}$, as that used in the precision-weighted estimator. This choice allows for a direct comparison with the confidence region based on the precision-weighted estimator through the upper bound M in the confidence region. To ensure the confidence regions satisfy Definition 1.4.1, we need to find the minimal value $\widehat{M}_{\text{PT}} = \widehat{M}_{\text{PT}}(\mathbf{b}, \zeta, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma})$ such that the confidence region

$$\{\boldsymbol{\tau} \in \mathbb{R}^d : (\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq \widehat{M}_{\text{PT}}\}$$

achieves correct coverage for all $\boldsymbol{\Delta}$ satisfying $|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}|_j \leq b_j$ for all $j = 1, 2, \dots, d$. We can formulate $\widehat{M}_{\text{PT}}$ as the optimization problem

$$\widehat{M}_{\text{PT}} = \inf \left\{ M \geq 0 : \inf_{\boldsymbol{\Delta}: |\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}| \leq \mathbf{b}} \mathbb{P}_{\boldsymbol{\Delta}}((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M) \geq 1 - \zeta \right\}.$$

We show how to compute $\widehat{M}_{\text{PT}}$ in the following theorem.

Theorem 1.4.2. *$\widehat{M}_{\text{PT}}$ is the solution to the equation of M :*

$$\inf_{\boldsymbol{\Delta}: |\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}| \leq \mathbf{b}} \mathbb{P}_{\boldsymbol{\Delta}}((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M) = 1 - \zeta,$$

with an explicit form given by:

$$\begin{aligned} & \mathbb{P}_{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}=\mathbf{t}}((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M) \\ &= \Psi_d \left(M; \|(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}^{1/2} \mathbf{t}\|_2^2 \right) \Psi_d \left(q; \|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} \boldsymbol{\Sigma}^{1/2} \mathbf{t}\|_2^2 \right) \\ &+ \int_{\|\mathbf{u}\|_2^2 > q} \Psi_d \left(M; \|(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} [\boldsymbol{\Sigma}^{1/2} \mathbf{t} - (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{1/2} \mathbf{u}]\|_2^2 \right) \phi_{(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} \boldsymbol{\Sigma}^{1/2} \mathbf{t}, \mathbf{I}_d}(\mathbf{u}) d\mathbf{u}. \end{aligned} \tag{1.4.4}$$

Theorem 1.4.2 extends the univariate result in Theorem 1.3.2 to the multivariate setting. Theorem 1.4.2 provides an explicit expression for the coverage probability $\mathbb{P}_{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}=\mathbf{t}}((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M)$ as a function of M and $\mathbf{t}$. The first term in (1.4.4) corresponds to the event that the pretest fails to reject the null hypothesis ($\boldsymbol{\Delta} = \mathbf{0}$), in which case $\widehat{\boldsymbol{\tau}}_{\text{PT}}$ reduces to the precision-weighted estimator. The second term in (1.4.4) integrates over the regions where the pretest rejects the null hypothesis, in which case $\widehat{\boldsymbol{\tau}}_{\text{PT}}$ reduces to the unbiased estimator.

Third, we construct the confidence regions based on the generic soft-thresholding estimator $\widehat{\boldsymbol{\tau}}_{\text{ST}}$. Similar to the confidence regions based on the precision-weighted estimator, we consider the confidence regions based on $\widehat{\boldsymbol{\tau}}_{\text{ST}}$ with the following form:

$$\{\boldsymbol{\tau} \in \mathbb{R}^d : (\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \leq M\},$$

for some $M \geq 0$. To ensure the confidence regions satisfy Definition 1.4.1, we need to find the minimal value $\widehat{M}_{\text{ST}} = \widehat{M}_{\text{ST}}(\mathbf{b}, \zeta, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma})$ such that the confidence region

$$\{\boldsymbol{\tau} \in \mathbb{R}^d : (\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \leq \widehat{M}_{\text{ST}}\}$$

achieves correct coverage for all $\boldsymbol{\Delta}$ satisfying $|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}|_j \leq b_j$ for all $j = 1, 2, \dots, d$. We can formulate $\widehat{M}_{\text{ST}}$ as the optimization problem

$$\widehat{M}_{\text{ST}} = \inf \left\{ M \geq 0 : \inf_{\boldsymbol{\Delta} : |\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}| \leq \mathbf{b}} \mathbb{P}_{\boldsymbol{\Delta}}((\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \leq M) \geq 1 - \zeta \right\}.$$

We show that $\widehat{M}_{\text{ST}}$ can be computed efficiently in the following theorem.

Theorem 1.4.3. $\widehat{M}_{\text{ST}}$ is the solution to the equation of M :

$$\inf_{\boldsymbol{\Delta} : \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta} = \mathbf{b} \odot \mathbf{s}, \mathbf{s} \in \{-1, 1\}^d} \mathbb{P}_{\boldsymbol{\Delta}}((\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \leq M) = 1 - \zeta,$$

with an explicit form given by:

$$\begin{aligned} & \mathbb{P}_{\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta} = \mathbf{t}}((\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \leq M) \\ &= \Psi_d \left(M; \|(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}^{1/2} \mathbf{t}\|_2^2 \right) \Psi_d \left(q; \|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} \boldsymbol{\Sigma}^{1/2} \mathbf{t}\|_2^2 \right) \\ &+ \int_{\|\mathbf{u}\|_2^2 > q} \Psi_d \left(M; \|(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} [\boldsymbol{\Sigma}^{1/2} \mathbf{t} - (1 - h_q^*(\|\mathbf{u}\|_2^2))(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{1/2} \mathbf{u}]\|_2^2 \right) \\ &\phi_{(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} \boldsymbol{\Sigma}^{1/2} \mathbf{t}, \mathbf{I}_d}(\mathbf{u}) d\mathbf{u}. \end{aligned} \tag{1.4.5}$$

Theorem 1.4.3 extends the univariate result in Theorem 1.3.3 to the multivariate setting. Theorem 1.4.3 provides an explicit expression for the coverage probability $\mathbb{P}_{\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta} = \mathbf{t}}((\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \leq M)$ as a function of M and $\mathbf{t}$. The first term in (1.4.5) corresponds to the event that the pretest fails to reject the null hypothesis ($\boldsymbol{\Delta} = \mathbf{0}$), in which case $\widehat{\boldsymbol{\tau}}_{\text{ST}}$ reduces to the precision-weighted estimator. The second term in (1.4.5) integrates over the regions where the pretest rejects the null hypothesis, in which case $\widehat{\boldsymbol{\tau}}_{\text{ST}}$ reduces to the unbiased estimator with a nonconstant shift to ensure continuity at the pretest boundary. As in the univariate case, the monotonicity of the coverage probability makes the computation of $\widehat{M}_{\text{ST}}$ more efficient than that of $\widehat{M}_{\text{PT}}$. For any $M > 0$, the coverage probability is minimized at the boundary of the bias region, namely at the vertices $\boldsymbol{\Delta}$ satisfying $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta} = \mathbf{b} \odot \mathbf{s}$ for some $\mathbf{s} \in \{-1, 1\}^d$.

Using Theorems 1.4.1–1.4.3, we can construct confidence regions based on $\widehat{\boldsymbol{\tau}}_{\text{PW}}$, $\widehat{\boldsymbol{\tau}}_{\text{PT}}$, and $\widehat{\boldsymbol{\tau}}_{\text{ST}}$ for various bias bounds $\mathbf{b}$ and significance levels ζ .

1.5 Generalization to multiple estimators

In this section, we extend our framework to the multiple estimators setting. In empirical research, analysts often face the challenge of integrating evidence from datasets of heterogeneous quality. A leading example in causal inference is to combine an RCT with multiple observational studies. While the RCT provides unbiased but often noisy estimates, observational studies can offer much larger sample sizes but are subject to hidden bias due to unmeasured confounding. Beyond causal inference, similar problems occur when synthesizing results from multiple registries, surveys, or administrative databases, where no single source is sufficient on its own. We present our theory below.

1.5.1 Setup

We consider the following multiple estimators setting with one unbiased estimator and K potentially biased estimators. We can generalize to the case where there are multiple unbiased estimators estimating the same parameter. In that case, we can first combine all the unbiased estimators into a single unbiased estimator with smaller variance using the precision-weighted estimator. Then we can apply our current framework to combine this aggregated unbiased estimator with multiple biased estimators.

Assumption 1.5.1. *Suppose we observe $K+1$ independent random variables: one unbiased estimator $\hat{\tau}_0 \sim N(\tau, \sigma_0^2)$ and K potentially biased estimators $\hat{\tau}_j \sim N(\tau + \Delta_j, \sigma_j^2)$ for $j = 1, \dots, K$. Here $\tau \in \mathbb{R}$ is the unknown parameter of interest. We assume that $\sigma_0^2, \sigma_1^2, \dots, \sigma_K^2$ are known, whereas $\Delta_1, \dots, \Delta_K$ are unknown.*

Let $\gamma_j = \sigma_0^2/\sigma_j^2$ be the variance ratio between the unbiased estimator and the j -th biased estimator. Let $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)^\top$ be the vector of variance ratios.

We consider a generic combined estimator $\hat{\tau} = \hat{\tau}(\hat{\tau}_0, \hat{\tau}_1, \dots, \hat{\tau}_K, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2)$. We assume that $|\Delta_j/\sigma_0| \leq b_j$ for some $b_j > 0$ for all $j = 1, \dots, K$ and study how the confidence interval changes with the maximum relative bias vector $\mathbf{b} = (b_1, \dots, b_K)^\top$. Let $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_K)^\top$ be the vector of unknown biases. Analogous to Definition 1.2.1 for the univariate case, the confidence interval is defined below.

Definition 1.5.1. Given a significance level $\zeta \in (0, 1)$ and the maximum relative bias vector $\mathbf{b}$ with $b_j \geq 0$ for all $j = 1, \dots, K$, we want to construct an interval $\mathcal{I}(\mathbf{b}, \zeta) = \mathcal{I}(\mathbf{b}, \zeta, \hat{\tau}_0, \hat{\tau}_1, \dots, \hat{\tau}_K, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2)$ such that

$$\inf_{\boldsymbol{\Delta}: |\boldsymbol{\Delta}/\sigma_0| \leq \mathbf{b}} \mathbb{P}_{\boldsymbol{\Delta}}(\tau \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)) = \inf_{\boldsymbol{\Delta}: |\boldsymbol{\Delta}/\sigma_0| \leq \mathbf{b}} \mathbb{P}_{\boldsymbol{\Delta}}(\hat{\tau} - \tau \in \mathcal{I}(\mathbf{b}, \zeta)) \geq 1 - \zeta, \quad (1.5.1)$$

The confidence interval for τ based on $\hat{\tau}$ is then given by $\hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)$.

Then we introduce the monotonicity conditions on the interval $\mathcal{I}(\mathbf{b}, \zeta)$ below, which generalizes Assumption 1.2.2 to the multiple estimators setting.

Assumption 1.5.2. *We assume:*

1. For fixed $(\hat{\tau}_0, \hat{\tau}_1, \dots, \hat{\tau}_K, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2)$ and ζ , we have $\mathcal{I}(\mathbf{b}, \zeta) \subset \mathcal{I}(\mathbf{b}', \zeta)$ whenever $\mathbf{b} \leq \mathbf{b}'$, i.e., $b_j \leq b'_j$ for all $j = 1, \dots, K$.
2. For fixed $(\hat{\tau}_0, \hat{\tau}_1, \dots, \hat{\tau}_K, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2)$ and $\mathbf{b}$, we have $\mathcal{I}(\mathbf{b}, \zeta) \subset \mathcal{I}(\mathbf{b}, \zeta')$ whenever $\zeta \geq \zeta'$.

Assume $\mathcal{I}(\mathbf{b}, \zeta)$ satisfies the monotonicity conditions of Assumption 1.5.2. A common family of such intervals is fixed-length centered intervals: $\mathcal{I}(\mathbf{b}, \zeta) = [-c(\mathbf{b}, \zeta, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2), c(\mathbf{b}, \zeta, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2)]$ with some constant $c(\mathbf{b}, \zeta, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2) \geq 0$. In this case, the confidence interval $\hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta)$ in Definition 1.5.1 is given by $[\hat{\tau} - c(\mathbf{b}, \zeta, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2), \hat{\tau} + c(\mathbf{b}, \zeta, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2)]$.

Then we define the b-value with multiple estimators below.

Definition 1.5.2. Define the b-value as the critical boundary $\mathbf{b}^*$ of testing $\tau = 0$ versus $\tau \neq 0$ as

$$\mathbf{b}^*(\zeta) = \mathbf{b}^*(\zeta, \hat{\tau}_0, \hat{\tau}_1, \dots, \hat{\tau}_K, \sigma_0^2, \sigma_1^2, \dots, \sigma_K^2) = \partial \{ \mathbf{b} \geq \mathbf{0} : 0 \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta) \}. \quad (1.5.2)$$

Definition 1.5.2 extends the univariate notion of the b-value (Definition 1.2.2) to the multiple estimators setting. As in the multivariate b-value defined in Definition 1.4.2, in higher dimensions $K > 1$, the set $\{ \mathbf{b} \geq \mathbf{0} : 0 \in \hat{\tau} - \mathcal{I}(\mathbf{b}, \zeta) \}$ is a convex region in the first quadrant, and the b-value is the boundary of this region, which is a $(K - 1)$ -dimensional surface. See Definition 1.4.2 and the subsequent discussion for more details.

1.5.2 Point estimation

First, we recall the precision-weighted estimator defined as

$$\hat{\tau}_{\text{PW}} = \frac{\sigma_0^{-2}}{\sigma_0^{-2} + \sum_{\ell=1}^K \sigma_\ell^{-2}} \hat{\tau}_0 + \sum_{j=1}^K \frac{\sigma_j^{-2}}{\sigma_0^{-2} + \sum_{\ell=1}^K \sigma_\ell^{-2}} \hat{\tau}_j = \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0).$$

If $\Delta_j = 0$ for all $j = 1, \dots, K$, the precision-weighted estimator is the maximum likelihood estimator and best linear unbiased estimator of τ . Moreover, by the classical Wald–Le Cam asymptotic decision theory (Wald, 1949; Le Cam, 1956), it is asymptotically admissible and minimax under the L_2 risk, i.e., $E[(\hat{\tau} - \tau)^2]$, among all regular estimators under standard regularity conditions.

Second, we introduce a pretest estimator. We incorporate the pretest for $\Delta_j = 0$ separately using $\hat{\tau}_j$ and the unbiased estimator $\hat{\tau}_0$, and combine those $\hat{\tau}_j$'s that fail to reject the null. Here we assume the pretests share the same significance level α for simplicity. The generalization to different significance levels is straightforward. Since $\hat{\tau}_j - \hat{\tau}_0 \sim N(\Delta_j, \sigma_0^2 + \sigma_j^2)$ and $\sigma_0^2 + \sigma_j^2 = (1 + \gamma_j^{-1})\sigma_0^2$, let

$$\mathcal{A}_j = \{ |\hat{\tau}_j - \hat{\tau}_0| \leq (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2} \}$$

denote the event that the pretest fails to reject the null hypothesis ($\Delta_j = 0$). The pretest estimator is:

$$\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) \mathbb{1}(\mathcal{A}_j).$$

Third, we introduce a soft-thresholding estimator, which ensures continuity at the pretest boundary. Instead of dropping a biased estimator entirely when the pretest rejects, soft-thresholding shifts the unbiased estimator by making the estimator continuous at the pretest boundary. The soft-thresholding estimator is:

$$\hat{\tau}_{\text{ST}} = \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} [(\hat{\tau}_j - \hat{\tau}_0) \mathbb{1}(\mathcal{A}_j) + (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2} \text{sign}(\hat{\tau}_j - \hat{\tau}_0) \mathbb{1}(\mathcal{A}_j^c)].$$

By the definitions of the pretest estimator and the soft-thresholding estimator, when the pretests fail to reject for all $j = 1, \dots, K$, both the pretest estimator and the soft-thresholding estimator reduce to the precision-weighted estimator. When $K = 1$, i.e., there is only one biased estimator, both pretest estimator and soft-thresholding estimator reduce to the estimators introduced in Section 1.2.2. Similar to the discussion in Section 1.2.2, the choice of α is a tuning parameter.

The pretest estimator and the soft-thresholding estimator considered above are not the only way to combine $\hat{\tau}_0$ and $\hat{\tau}_1, \dots, \hat{\tau}_K$. For example, one could apply shrinkage or soft-thresholding directly between the precision-weighted estimator $\hat{\tau}_{\text{PW}}$ and the unbiased estimator $\hat{\tau}_0$, rather than at the level of individual biased components. We focus on thresholding at the component level because it admits a transparent interpretation in terms of testing and controlling each bias component Δ_j separately, which aligns naturally with our sensitivity analysis perspective. Alternative combination strategies have also been considered in the data fusion literature; see, for example, [Duan and Wang \(2023\)](#) and [Chen et al. \(2025\)](#). Our framework can accommodate other shrinkage schemes in principle, but we do not pursue them here.

1.5.3 Confidence intervals

First, we construct the confidence intervals based on the precision-weighted estimator $\hat{\tau}_{\text{PW}}$. Under Assumption 1.5.1, we have $\hat{\tau}_{\text{PW}} \sim N(\tau + (1 + \|\boldsymbol{\gamma}\|_1)^{-1} \langle \boldsymbol{\gamma}, \boldsymbol{\Delta} \rangle, (1 + \|\boldsymbol{\gamma}\|_1)^{-1} \sigma_0^2)$. The following theorem provides the confidence intervals based on $\hat{\tau}_{\text{PW}}$ as a function of bias bound $\mathbf{b}$.

Theorem 1.5.1. *Let $\hat{L}_{\text{PW}} = \hat{L}_{\text{PW}}(\mathbf{b}, \zeta, \boldsymbol{\gamma}) \geq 0$ denote the solution to the equation of L :*

$$\Phi \left(L - \frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) - \Phi \left(-L - \frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) = 1 - \zeta.$$

The $\hat{L}_{PW}$ always exists and is unique. The shortest length symmetric centered confidence interval based on $\hat{\tau}_{PW}$ for τ satisfying (1.5.1) is given by $[\hat{\tau}_{PW} - \hat{L}_{PW}(1 + \|\gamma\|_1)^{-1/2}\sigma_0, \hat{\tau}_{PW} + \hat{L}_{PW}(1 + \|\gamma\|_1)^{-1/2}\sigma_0]$.

Theorem 1.5.1 extends the single estimator result in Theorem 1.3.1 to the multiple estimators setting. In the single estimator case, the bias constraint $\{\Delta : |\Delta/\sigma_0| \leq b\}$ has only two boundary points, and the absolute bias of $\hat{\tau}_{PW}$ is the same at both endpoints. By contrast, the multiple estimators bias constraint $\{\Delta : |\Delta/\sigma_0| \leq \mathbf{b}\}$ is a hyperrectangle with 2^K vertices, and the maximum absolute bias of $\hat{\tau}_{PW}$ occurs only at $\Delta/\sigma_0 = \pm \mathbf{b}$. Theorem 1.5.1 therefore characterizes the maximum possible bias of $\hat{\tau}_{PW}$ over the entire bias constraint.

Second, we construct the confidence intervals based on the pretest estimator $\hat{\tau}_{PT}$. We seek the shortest length $\hat{L}_{PT} = \hat{L}_{PT}(\mathbf{b}, \zeta, \gamma, \alpha)$ such that the confidence interval $[\hat{\tau}_{PT} - \hat{L}_{PT}(1 + \|\gamma\|_1)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1 + \|\gamma\|_1)^{-1/2}\sigma_0]$ achieves correct coverage uniformly over all Δ with $|\Delta_j/\sigma_0| \leq b_j$ for all $j = 1, \dots, K$. We can formulate $\hat{L}_{PT}$ as the optimization problem:

$$\hat{L}_{PT} = \inf \left\{ L \geq 0 : \inf_{\Delta : |\Delta/\sigma_0| \leq \mathbf{b}} P_{\Delta}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \|\gamma\|_1)^{-1/2}\sigma_0) \geq 1 - \zeta \right\}. \quad (1.5.3)$$

We show how to compute $\hat{L}_{PT}$ in the following theorem.

Theorem 1.5.2. $\hat{L}_{PT} = \hat{L}_{PT}(\mathbf{b}, \zeta, \gamma, \alpha)$ in (1.5.3) is the solution to the equation of L :

$$\inf_{\Delta : |\Delta/\sigma_0| \leq \mathbf{b}} P_{\Delta}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \|\gamma\|_1)^{-1/2}\sigma_0) = 1 - \zeta,$$

with an explicit form given by:

$$\begin{aligned} & P_{\Delta/\sigma_0=\mathbf{t}}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \|\gamma\|_1)^{-1/2}\sigma_0) \\ &= \int_{\mathbb{R}^K} \left[\Phi \left(L - \frac{\langle \gamma, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\gamma\|_1}} \right) - \Phi \left(-L - \frac{\langle \gamma, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\gamma\|_1}} \right) \right] \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u}, \end{aligned}$$

where for any $\mathbf{u} \in \mathbb{R}^K$, $\mathbf{u}' \in \mathbb{R}^K$ is defined as

$$u'_j = u_j \mathbb{1}(|u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}) \text{ for all } j = 1, \dots, K,$$

and $\mathbf{V}_{ij} = 1 + \gamma_i^{-1} \mathbb{1}(i = j)$ for $i, j = 1, \dots, K$. $[\hat{\tau}_{PT} - \hat{L}_{PT}(1 + \|\gamma\|_1)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1 + \|\gamma\|_1)^{-1/2}\sigma_0]$ is the shortest length symmetric centered confidence interval based on $\hat{\tau}_{PT}$ for τ satisfying (1.5.1).

Theorem 1.5.2 extends the single estimator result in Theorem 1.3.2 to the multiple estimators setting. Theorem 1.5.2 provides an explicit expression for the coverage probability $P_{\Delta/\sigma_0=\mathbf{t}}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \|\gamma\|_1)^{-1/2}\sigma_0)$ as a function of $\mathbf{t}$ and L .

Third, we construct the confidence intervals based on the soft-thresholding estimator $\hat{\tau}_{ST}$. We seek the shortest length $\hat{L}_{ST} = \hat{L}_{ST}(\mathbf{b}, \zeta, \gamma, \alpha)$ such that the confidence interval

$[\widehat{\tau}_{\text{ST}} - \widehat{L}_{\text{ST}}(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0, \widehat{\tau}_{\text{ST}} + \widehat{L}_{\text{ST}}(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0]$ achieves correct coverage uniformly over all $\boldsymbol{\Delta}$ with $|\Delta_j/\sigma_0| \leq b_j$ for all $j = 1, \dots, K$. We can formulate $\widehat{L}_{\text{ST}}$ as the optimization problem:

$$\widehat{L}_{\text{ST}} = \inf \left\{ L \geq 0 : \inf_{\boldsymbol{\Delta}: |\boldsymbol{\Delta}/\sigma_0| \leq \mathbf{b}} \mathbb{P}_{\boldsymbol{\Delta}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0) \geq 1 - \zeta \right\}. \quad (1.5.4)$$

We show that $\widehat{L}_{\text{ST}}$ can be computed efficiently in the following theorem.

Theorem 1.5.3. *For any $L > 0$, $\mathbb{P}_{\boldsymbol{\Delta}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0)$ is symmetric about $\boldsymbol{\Delta} = \mathbf{0}$ and monotonically decreasing in $|\Delta_j|$ for all $j = 1, \dots, K$. Then $\widehat{L}_{\text{ST}} = \widehat{L}_{\text{ST}}(\mathbf{b}, \zeta, \boldsymbol{\gamma}, \alpha)$ in (1.5.4) is the solution to the equation of L :*

$$\mathbb{P}_{\boldsymbol{\Delta}/\sigma_0 = \mathbf{b}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0) = 1 - \zeta,$$

with an explicit form given by:

$$\begin{aligned} & \mathbb{P}_{\boldsymbol{\Delta}/\sigma_0 = \mathbf{t}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0) \\ &= \int_{\mathbb{R}^K} \left[\Phi \left(L - \frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) - \Phi \left(-L - \frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) \right] \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u}, \end{aligned}$$

where for any $\mathbf{u} \in \mathbb{R}^K$, $\mathbf{u}' \in \mathbb{R}^K$ is defined as

$$u'_j = [u_j - (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2} \text{sign}(u_j)] \mathbf{1}(|u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}) \text{ for all } j = 1, \dots, K,$$

and $\mathbf{V}_{ij} = 1 + \gamma_i^{-1} \mathbf{1}(i = j)$ for $i, j = 1, \dots, K$. $[\widehat{\tau}_{\text{ST}} - \widehat{L}_{\text{ST}}(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0, \widehat{\tau}_{\text{ST}} + \widehat{L}_{\text{ST}}(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0]$ is the shortest length symmetric centered confidence interval based on $\widehat{\tau}_{\text{ST}}$ for τ satisfying (1.5.1).

Theorem 1.5.3 extends the single estimator result in Theorem 1.3.3 to the multiple estimators setting. Theorem 1.5.3 provides an explicit expression for the coverage probability $\mathbb{P}_{\boldsymbol{\Delta}/\sigma_0 = \mathbf{t}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0)$ as a function of $\mathbf{t}$ and L . As in the single estimator case, the monotonicity makes the computation of $\widehat{L}_{\text{ST}}$ more efficient than that of $\widehat{L}_{\text{PT}}$. For any $L > 0$, the coverage probability is minimized at the vertices $\boldsymbol{\Delta}$ satisfying $\boldsymbol{\Delta}/\sigma_0 = \pm \mathbf{b}$.

1.6 Empirical studies

In this section, we use the example from Angrist and Krueger (1991) to demonstrate how our framework works in practice, where the authors studied the effect of years of schooling on earnings. Angrist and Krueger (1991) used quarter of birth as an instrument to obtain the IV estimate. They also reported the OLS estimate, which would be biased in the presence of endogeneity.

We consider the OLS estimator and the IV estimator. The OLS estimator of the return to schooling is potentially biased because schooling decisions are endogenous: unobserved factors may influence both schooling and earnings. By contrast, the IV estimator uses quarter of birth as a source of exogenous variation, yielding a consistent estimate of the causal effect under standard IV assumptions. We take the IV estimator as the unbiased estimator $\hat{\tau}_0$ and the OLS estimator as the biased estimator $\hat{\tau}_1$. Using these two estimators, we construct the three combined estimators: the precision-weighted estimator $\hat{\tau}_{PW}$, the pretest estimator $\hat{\tau}_{PT}$, and the soft-thresholding estimator $\hat{\tau}_{ST}$.

Figure 1.2 visualizes the confidence intervals of the IV, OLS, and combined estimators against the maximum relative bias $|\Delta/\sigma_0| \leq b$.

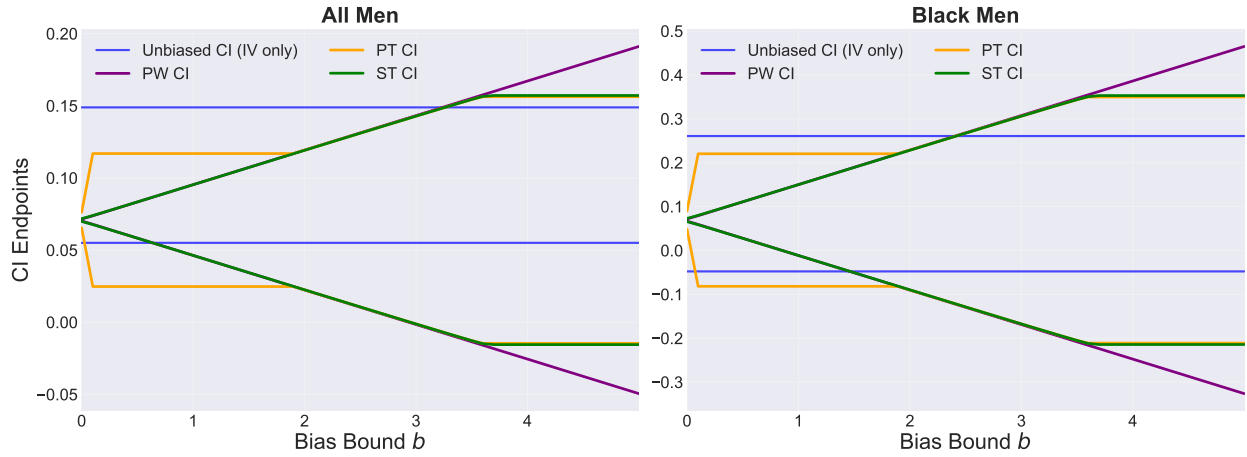


Figure 1.2: Confidence intervals against the maximum relative bias $|\Delta/\sigma_0| \leq b$

In both the full sample (black and white men) and the subsample (black men), we observe that the standard error of the OLS estimator is much smaller than that of the IV estimator. Moreover, although we compute the OLS estimator and the IV estimator using the same sample, their correlation is very low. Therefore, the new biased estimator constructed following Section 1.3.5 is nearly identical to the original biased estimator (OLS). The combined estimators behave as expected. When the bias bound is small, $\hat{\tau}_{PW}$, $\hat{\tau}_{PT}$, and $\hat{\tau}_{ST}$ all yield confidence intervals substantially shorter than that based on the unbiased estimator (IV), reflecting efficiency gains from incorporating the more precise OLS estimator. Notably, the confidence interval of $\hat{\tau}_{ST}$ is nearly identical to that of $\hat{\tau}_{PW}$ and much shorter than that of $\hat{\tau}_{PT}$ when the bias bound is small. Even when the bias bound becomes large, the confidence intervals of $\hat{\tau}_{ST}$ and $\hat{\tau}_{PW}$ remain nearly the same in length. These behaviors arise because the OLS estimator is far more precise than the IV estimator.

However, there are differences between the full sample (black men and white men) and the subsample (black men). In the full sample, the sample size is very large, and the IV estimator is sufficiently precise that the null hypothesis of zero returns to schooling can be rejected using IV alone. In contrast, for the subsample of black men, the sample size is much smaller, leading

to a much larger standard error for the IV estimator. As a result, the IV-based confidence interval is wide enough that the null hypothesis cannot be rejected at the 5% significance level. This contrast illustrates a common empirical challenge: when the unbiased estimator is noisy, inference based solely on it can be inconclusive, even when the overall dataset is large. In such settings, the combined estimators, particularly the precision-weighted estimator $\hat{\tau}_{PW}$ and the soft-thresholding estimator $\hat{\tau}_{ST}$, yield much tighter confidence intervals under small biases. Under small biases, these estimators provide sufficient evidence to reject the null hypothesis of zero returns to schooling, while still allowing researchers to explicitly assess sensitivity to potential bias.

1.7 Discussion

In our framework, the choices of point estimators and confidence intervals are not unique. In this paper, we focus on several convenient estimators and construct the corresponding confidence intervals and b-values. We do not focus on the optimality of confidence intervals and leave it to future research. An alternative approach is to construct confidence intervals based on bias-dependent point estimators. In Appendix A.1.3, we discuss the advantages of using prespecified point estimators rather than subtracting the bias from the point estimator. More generally, one could construct confidence intervals that are optimized separately for each bias bound. For a given bias bound, we can construct the optimal confidence interval by minimizing a criterion, e.g., interval length for two-sided intervals and quantile of the excess length for one-sided intervals, over a class of estimators, subject to valid coverage under the specified bias bound. Such constructions are closely related to confidence intervals in Donoho (1994); Kolesár and Rothe (2018); Armstrong and Kolesár (2018, 2020, 2021b); Noack and Rothe (2024) and Kaplan and Liu (2024). Under this approach, the center of the optimal confidence interval depends on the bias bound. While such bias-dependent constructions can yield shorter confidence intervals for a given bias bound, they may lead to non-monotonic confidence intervals as the bias bound varies. In such cases, the interpretation of the b-value becomes less straightforward. For this reason, we impose the restriction in Definition 1.2.1 and Assumption 1.2.2 that confidence intervals are based on a point estimator that does not depend on the bias bound. The trade-off is that the resulting confidence intervals may be slightly longer than those obtained from bias-dependent constructions.

Chapter 2

Unifying regression-based and design-based causal inference in time-series experiments

2.1 Introduction to time-series experiments

Time-series experiments, also referred to as switchback experiments or N-of-1 trials, have attracted increasing attention across a variety of domains ([Davidson et al., 2022](#)), including clinical trials ([Lillie et al., 2011](#); [Nikles and Mitchell, 2015](#)), online platforms ([Bojinov et al., 2023](#)), and other applied settings ([Gabler et al., 2011](#); [Mirza et al., 2017](#); [Hawksworth et al., 2024](#)). Unlike traditional randomized controlled trials, which assign treatments at the unit level and maintain a fixed assignment throughout the study, time-series experiments repeatedly randomize treatment assignments over time for the same, single unit. Time-series experiments are closely related to longitudinal experiments ([Robins, 1986](#)), but focus on dynamic treatment effects on the same unit rather than average effects over a population. The time-series experimental design allows capturing dynamic causal effects in environments where both treatments and outcomes evolve over time and where treatment effects vary across time periods. Such designs are particularly attractive in modern applications where repeated interventions are feasible, the number of units is limited, and the research interest is to estimate time-varying or lagged treatment effects. [Bojinov and Shephard \(2019\)](#) provided a concrete example of a time-series experiment conducted at AHL Partners LLP, a quantitative hedge fund that randomly assigned large trading orders to either human traders or computer algorithms to compare execution performance. Their example is a time-series experiment with sequential treatment assignments, where the outcome of interest is the relative performance across 10 equity-index futures markets over the course of a year.

Despite the wide applications of time-series experiments, the statistical analysis presents challenges, mainly due to the temporal dependence of outcomes and the risk of model misspecification. In time-series experiments, the treatment effects may persist beyond the period

of assignment, inducing the dependence of the observed outcomes on past treatments. Consequently, the methods for analyzing traditional randomized experiments are not applicable. The design-based perspective offers a principled way to avoid strong assumptions on the outcome model, in contrast to the model-based analysis of time-series experiments (Zucker et al., 2010; Senn, 2017; Schmid and Yang, 2022; Menchetti et al., 2023). By viewing the treatment assignment as the sole source of randomness, we propose treatment effect estimates with desirable properties without strong model assumptions. This perspective builds on a rich literature on design-based analysis for various experimental settings, including regression analysis for completely randomized experiments (Lin, 2013), matched-pairs experiments (Fogarty, 2018), randomized block designs (Koo and Pashley, 2026), factorial experiments (Dasgupta et al., 2015), split-plot designs (Zhao and Ding, 2022; Mukerjee and Dasgupta, 2022), cluster-randomized experiments (Middleton and Aronow, 2015; Su and Ding, 2021), and network experiments (Aronow and Samii, 2017; Leung, 2022; Gao and Ding, 2025). Recent reviews are available in Shi and Li (2024) and Ding (2025). However, the design-based analysis of regression-based estimators in time-series experiments remains underexplored.

In this paper, we develop a unified design-based framework for analyzing regression-based estimators in time-series experiments. We show that regressing the observed outcomes on current and past treatments using an appropriately specified working model and transformed regressors yields estimates of the lagged treatment effects of interest, even when the number of regressors diverges with the length of the time series. A key feature of our working model is the centering and scaling of treatment indicators, which ensures the regression coefficients recover the causal estimands of interest. We establish the consistency and asymptotic normality of the ordinary least squares (OLS) estimator and show that heteroskedasticity and autocorrelation consistent (HAC) variance estimators (Newey and West, 1987) provide conservative variance estimates. Our results do not require the regression model to be correctly specified, reflecting the robustness inherent to the design-based analysis. Our work is related to Bojinov and Shephard (2019), Malenica et al. (2023), and Liang and Recht (2025), who also develop design-based analyses for time-series experiments. However, our paper differs from theirs at two levels. First, at the level of estimands, we focus on treatment effects defined independently of the realized treatment path and outcome model assumptions. Second, at the level of estimators, we focus on simpler, regression-based point and variance estimators. We will provide a detailed comparison in Section 2.3.3.

The remainder of the paper is organized as follows. Section 2.2 introduces the setup of the time-series experiment, defines the causal estimand of interest, and presents our regression-based estimator. Section 2.3 develops the asymptotic theory, establishing consistency, asymptotic normality, and variance estimation. Section 2.4 discusses the extension to continuous treatment. Section 2.5 presents empirical studies, including simulation evidence and an application to the trading experiment data previously analyzed by Bojinov and Shephard (2019). Section 2.6 concludes with a discussion of several extensions. The online supplementary materials contain extensions to the regression with interaction terms and general exposure mapping, as well as the technical details.

Notation. For a vector $a = (a_1, \dots, a_d)^\top \in \mathbb{R}^d$, let $\|a\|_1 = \sum_{i=1}^d |a_i|$, and $\|a\|_2 =$

$(\sum_{i=1}^d a_i^2)^{1/2}$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, let $\|A\|_F = (\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2)^{1/2}$ denote the Frobenius norm, and $\|A\|_2$ denote the spectral norm. Following the R syntax, we use $\mathbf{lm}(Y \sim X)$ to denote the linear regression model with outcome Y and regressors X .

2.2 Setup, estimand and estimator

2.2.1 Setup

Consider a time-series experiment with a binary treatment and a single unit repeatedly measured at T time points. At each time point $t = 1, \dots, T$, we apply $Z_t = 0$ for the control and $Z_t = 1$ for the treatment. For each time point t , define the treatment history up to t as $z_{t:1} = (z_t, z_{t-1}, \dots, z_1)$. Let $\{Y_t(z_T, \dots, z_1) : z_T, \dots, z_1 = 0 \text{ or } 1\}_{t=1}^T$ denote the potential outcomes, which can depend on the whole treatment history. We first simplify the potential outcomes by imposing the no-anticipation assumption below.

Assumption 2.2.1 (No-anticipation). *For every possible treatment path $(z_1, \dots, z_T)$, the potential outcome at time t depends only on the treatment history up to and including time t , but does not depend on any future treatments $z_{t+1}, \dots, z_T$:*

$$Y_t(z_T, \dots, z_{t+1}, z_t, \dots, z_1) = Y_t(z_t, \dots, z_1), \text{ for all } t = 1, \dots, T.$$

Under the no-anticipation assumption (Assumption 2.2.1), the observed outcome simplifies to $Y_t = Y_t(Z_{t:1})$, where $Z_{t:1} = (Z_t, \dots, Z_1)$ represents the treatment history up to and including time t , following the notation in [Bojinov and Shephard \(2019\)](#). The no-anticipation assumption holds when treatments are rapidly randomized (e.g., online experiments), or treatments are administered sequentially (e.g., clinical N-of-1 trials). However, when there exists forward-looking behavior, where the unit can predict future treatments, the no-anticipation assumption would be violated. We do not consider violations of the no-anticipation assumption in this paper.

Time-series experiments and network experiments are closely related in that both extend the classical potential outcomes framework to accommodate structured interference. In network experiments, a unit's potential outcome may depend not only on its own treatment but also on the treatments of its neighbors in a network ([Aronow and Samii, 2017](#)). Similarly, in time-series experiments with a single unit observed over multiple periods, outcomes may depend not only on the current treatment but also on past or future treatments, giving rise to temporal interference. The no-anticipation assumption specifies a form of temporal interference: outcomes at time t may depend on the entire treatment history up to t , but not on future treatments ([Bojinov and Shephard, 2019](#)).

We introduce the following assumption to describe the design of the time-series experiment considered in this paper.

Assumption 2.2.2 (Time-series experiment). *Assume the treatments $Z_1, \dots, Z_T$ are independent. Let $p_t = P(Z_t = 1) \in (0, 1)$ for each $t = 1, \dots, T$. Assume the values of $(p_1, \dots, p_T)$ are known by the design.*

To estimate the lagged treatment effects, which capture how treatments administered in earlier periods continue to influence outcomes at later times, we consider regressing the observed outcomes on past treatments. We motivate the working regression model by the standard causal inference setting with only one time point. Suppose there is a single-time binary treatment Z with known treatment probability $P(Z = 1) = p$, a pair of potential outcomes $(Y(0), Y(1))$, and we observe $Y = Y(Z)$. Then we can construct the standard inverse propensity score weighted estimator for the average treatment effect (ATE) based on the identification formula $E[Y(1) - Y(0)] = E[(ZY/p) - (1 - Z)Y/(1 - p)]$ (Ding, 2024, Chapter 11). The ATE also admits a regression-based formulation. Define the normalized treatment indicator $\tilde{Z} = (Z - p)/[p(1 - p)]$. Consider the population OLS regression of Y on $\tilde{Z}$ without an intercept, and denote the regression coefficient by $\beta = E[\tilde{Z}^2]^{-1}E[\tilde{Z}Y]$. Then we can identify the ATE as

$$E[Y(1) - Y(0)] = E\left[\frac{ZY}{p} - \frac{(1 - Z)Y}{1 - p}\right] = E[\tilde{Z}Y] = E[\tilde{Z}^2]\beta = \frac{1}{\text{Var}[Z]}\beta. \quad (2.2.1)$$

Therefore, we can estimate the ATE by running an OLS regression of Y on $\tilde{Z}$ without intercept, and scaled the coefficient by the variance of Z . Extending the idea to time-series experiments with multiple time points, we normalize the treatment indicator Z_t as

$$\tilde{Z}_t = (Z_t - E[Z_t]) / \text{Var}[Z_t] = (Z_t - p_t) / [p_t(1 - p_t)]. \quad (2.2.2)$$

Suppose we are interested in the lagged treatment effects up to the number of lags K . Now we consider the OLS regression without the intercept:

$$\text{lm}(Y_t \sim \tilde{Z}_t + \tilde{Z}_{t-1} + \cdots + \tilde{Z}_{t-K}) \quad (2.2.3)$$

using observations from $t = K + 1, \dots, T$. Let $\tilde{\tau} = (\tilde{\tau}_0, \dots, \tilde{\tau}_K)^\top \in \mathbb{R}^{K+1}$ denote the OLS coefficient vector from (2.2.3), where $\tilde{\tau}_k$ is expected to be an estimate of the lag- k treatment effect.

Several key points should be noted. First, the centering of Z_t in $\tilde{Z}_t$ by $E[Z_t]$ ensures that the estimand is invariant to location shifts in potential outcomes. Specifically, for any time t , if the potential outcome $Y_t(z_t, \dots, z_1)$ is shifted to $Y_t(z_t, \dots, z_1) + c$ for some constant c and all treatment realizations $z_{t:1} \in \{0, 1\}^t$, the estimand remains unchanged. This transformation invariance property guarantees that the estimand captures causal effects rather than the absolute level of the outcomes. Second, the scaling of Z_t in $\tilde{Z}_t$ by $\text{Var}[Z_t]$ ensures equal weighting across all time points in the estimand. If differential weighting across lags is desired to reflect differences in treatment assignment probabilities, the scaling can be adjusted accordingly. See the discussion about general treatment effects and the connection with Bojinov and Shephard (2019) in Section 2.3.3.

To ensure that the OLS estimator matches the scale of the causal estimand of interest, where the difference in scale arises from dividing Z_t by $\text{Var}[Z_t]$, we apply a linear transformation to $\tilde{\tau}$. This step rescales the regression coefficients back into the natural scale of the

treatment effect. Since $E[\tilde{Z}_t^2] = 1/\text{Var}[Z_t]$ for each $t = 1, \dots, T$, we define

$$\hat{\tau}_k = w_k^{-1} \tilde{\tau}_k \quad \text{for } k = 0, \dots, K, \quad (2.2.4)$$

where w_k is the harmonic mean of the variances of the lag- k treatment indicators:

$$\begin{aligned} w_k &= \left[\frac{1}{T-K} \sum_{t=K+1}^T E[\tilde{Z}_t^2] \right]^{-1} = \left[\frac{1}{T-K} \sum_{t=K+1}^T \frac{1}{\text{Var}[Z_{t-k}]} \right]^{-1} \\ &= \left[\frac{1}{T-K} \sum_{t=K+1}^T \frac{1}{p_{t-k}(1-p_{t-k})} \right]^{-1}. \end{aligned} \quad (2.2.5)$$

Here w_k generalizes the single-time scaling factor $\text{Var}[Z_t]$ in (2.2.1) to the time-series experiments setting by accounting for different treatment probabilities at different time points. By Assumption 2.2.2, the treatment probabilities are fixed by design, and thus w_k is a fixed constant. Let

$$\hat{\tau} = (\hat{\tau}_0, \dots, \hat{\tau}_K)^\top = (w_0^{-1} \tilde{\tau}_0, \dots, w_K^{-1} \tilde{\tau}_K)^\top \in \mathbb{R}^{K+1}$$

denote the OLS estimate from (2.2.3) after the linear transformation (2.2.4). In this paper, we focus on studying the design-based properties of $\hat{\tau}$. Under the design-based perspective, we treat the potential outcomes as fixed and the randomness solely comes from the treatment assignments.

To write $\hat{\tau}$ explicitly, define the outcome vector $\mathbf{Y}_K = (Y_{K+1}, \dots, Y_T)^\top \in \mathbb{R}^{T-K}$ and the regressor matrix $\mathbf{Z}_K = (\tilde{Z}_{K+1:1}^\top, \dots, \tilde{Z}_{T:T-K}^\top)^\top \in \mathbb{R}^{(T-K) \times (K+1)}$ where $\tilde{Z}_{t:t'} = (\tilde{Z}_t, \tilde{Z}_{t-1}, \dots, \tilde{Z}_{t'})^\top$ for $t > t'$. Let $\mathbf{W}_K \in \mathbb{R}^{(K+1) \times (K+1)}$ be the diagonal weight matrix $\mathbf{W}_K = \text{diag}(w_k : k = 0, \dots, K)$. Then, the OLS estimator in (2.2.3) after the linear transformation (2.2.4) can be expressed as

$$\hat{\tau} = \mathbf{W}_K^{-1} \tilde{\tau} = \mathbf{W}_K^{-1} (\mathbf{Z}_K^\top \mathbf{Z}_K)^{-1} \mathbf{Z}_K^\top \mathbf{Y}_K. \quad (2.2.6)$$

The OLS regression is intuitive for estimating lagged treatment effects because it directly links the outcome to both current and past treatments. Each coefficient $\hat{\tau}_k$ captures the average impact of a treatment administered k periods earlier on the current outcome. In this way, the estimator naturally decomposes the outcome into lag-specific contributions, providing a clear measure of how treatments influence present and future outcomes over time. We will show that $\hat{\tau}$ can recover the causal estimand in the next section. We discuss other regression specifications in Section B.1.1 of the supplementary material.

2.2.2 Causal estimand

We define the treatment effect of lag- k treatment on the outcome at time t as:

$$\tau_{t,k} = E[Y_t(Z_{t:t-k+1}, 1, Z_{t-k-1:1}) - Y_t(Z_{t:t-k+1}, 0, Z_{t-k-1:1})], \quad (2.2.7)$$

where the expectation is taken over the joint distribution of $(Z_{t:t-k+1}, Z_{t-k-1:1})$, holding Z_{t-k} at 1 or 0. This estimand $\tau_{t,k}$ represents the ATE on the outcome at time t by setting the treatment at time $t - k$ to 1 rather than 0, averaging over all other treatment assignments. We then define the ATE of lag- k treatment on the outcome over time as:

$$\tau_k = \frac{1}{T - K} \sum_{t=K+1}^T \tau_{t,k}. \quad (2.2.8)$$

Let $\tau = (\tau_0, \dots, \tau_K)^\top \in \mathbb{R}^{K+1}$ be the vector of lagged treatment effects. In this section, we focus on the interpretation of τ , and we establish the asymptotic behavior of $\hat{\tau}$ as an estimator of τ in Section 2.3.

The lag- k treatment effect τ_k in (2.2.8) does not rely on any specific model assumptions about the potential outcomes. However, its interpretation becomes more explicit under additional modeling assumptions. We consider a linear model for the potential outcome to facilitate the interpretation of τ_k .

Example 2.2.1 (Linear model). Consider the linear model for the potential outcome:

$$Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \beta_{t,k} z_{t-k} + \epsilon_t(z_{t:1}), \quad (2.2.9)$$

for all $t = 1, \dots, T$ and all treatment realizations $z_{t:1} \in \{0, 1\}^t$ with non-random coefficients $\{\beta_{t,k}\}_{0 \leq k < t}$. Under the design-based perspective, the noise $\epsilon_t(z_{t:1})$ only depends on the treatment realization $z_{t:1}$. Under the model (2.2.9), the lag- k treatment effect can be expressed as:

$$\tau_k = \frac{1}{T - K} \sum_{t=K+1}^T \beta_{t,k} + \frac{1}{T - K} \sum_{t=K+1}^T \mathbb{E}[\epsilon_t(Z_{t:t-k+1}, 1, Z_{t-k-1:1}) - \epsilon_t(Z_{t:t-k+1}, 0, Z_{t-k-1:1})].$$

If the treatment switch at time $t - k$ does not affect the error term ϵ_t , for example, if ϵ_t does not depend on $z_{t:1}$, then $\tau_k = (T - K)^{-1} \sum_{t=K+1}^T \beta_{t,k}$. Further assuming homogeneity of lag- k treatment effects over time, i.e., $\beta_{t,k} = \beta_k$ for all $t > k$, we obtain $\tau_k = \beta_k$. On the other hand, under heterogeneous lag- k treatment effects over time, the estimand τ_k captures the average treatment effect. This implies that the estimand τ_k remains robust to time-varying treatment effects.

The linear model in Example 2.2.1 contains several important time-series models, e.g., the autoregressive model and the moving-average model. In the following two examples, we illustrate the interpretation of τ_k under these two special models.

Example 2.2.2 (Autoregressive model). Consider the autoregressive model of order p for the potential outcome:

$$Y_t(z_{t:1}) = \mu_t(z_{t:1}) + \sum_{k=1}^p \phi_k Y_{t-k}(z_{t-k:1}) + \epsilon_t(z_{t:1}),$$

for all $t = 1, \dots, T$ and all treatment realizations $z_{t:1} \in \{0, 1\}^t$. Further assume $\mu_t(z_{t:1}) = \mu_t(z_t)$ and $\epsilon_t(z_{t:1}) = \epsilon_t(z_t)$ for all $t = 1, \dots, T$. Then we can write the potential outcome as

$$Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \Psi_k \mu_{t-k}(z_{t-k}) + \sum_{k=0}^{t-1} \Psi_k \epsilon_{t-k}(z_{t-k}),$$

where the Ψ_k 's are defined recursively by $\Psi_0 = 1$ and $\Psi_j = \sum_{k=1}^{\min\{p,j\}} \phi_k \Psi_{j-k}$. The coefficient Ψ_k is the k -step impulse-response coefficient from the autoregressive model of order p (Hamilton, 1994), which measures the effect of a one-time shock today on future outcomes. We can further write the potential outcome as

$$Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \Psi_k (\mu_{t-k}(1) - \mu_{t-k}(0)) z_{t-k} + \sum_{k=0}^{t-1} \Psi_k \mu_{t-k}(0) + \sum_{k=0}^{t-1} \Psi_k \epsilon_{t-k}(z_{t-k}).$$

Define $\beta_{t,k} = \Psi_k (\mu_{t-k}(1) - \mu_{t-k}(0))$ to see that it is a special case of Example 2.2.1. If the error term satisfies $\epsilon_t(1) = \epsilon_t(0)$ for all t , then by Example 2.2.1, we have

$$\tau_k = \Psi_k \left[\frac{1}{T-K} \sum_{t=K+1}^T (\mu_{t-k}(1) - \mu_{t-k}(0)) \right].$$

Therefore, τ_k reflects the average treatment effect on the mean function μ_t at lag k over time, multiplied by the k -step impulse-response coefficient from the autoregressive model of order p .

Example 2.2.3 (Moving-average model). Consider the moving-average model of order q for the potential outcome:

$$Y_t(z_{t:1}) = \mu_t(z_{t:1}) + \sum_{k=1}^q \theta_k \epsilon_{t-k}(z_{t-k:1}) + \epsilon_t(z_{t:1}),$$

for $t = 1, \dots, T$ and all treatment realizations $z_{t:1} \in \{0, 1\}^t$. Further assume $\mu_t(z_{t:1}) = \mu_t(z_t)$ and $\epsilon_t(z_{t:1}) = \epsilon_t(z_t)$ for all $t = 1, \dots, T$. Then we can write the potential outcome as

$$Y_t(z_{t:1}) = \mu_t(z_t) + \sum_{k=1}^q \theta_k \epsilon_{t-k}(z_{t-k}) + \epsilon_t(z_t).$$

We can further write the potential outcome as

$$\begin{aligned} Y_t(z_{t:1}) &= (\mu_t(1) - \mu_t(0) + \epsilon_t(1) - \epsilon_t(0)) z_t + \sum_{k=1}^q \theta_k (\epsilon_{t-k}(1) - \epsilon_{t-k}(0)) z_{t-k} \\ &\quad + \mu_t(0) + \epsilon_t(0) + \sum_{k=1}^q \theta_k \epsilon_{t-k}(0). \end{aligned}$$

Define $\beta_{t,0} = \mu_t(1) - \mu_t(0) + \epsilon_t(1) - \epsilon_t(0)$, $\beta_{t,k} = \theta_k(\epsilon_{t-k}(1) - \epsilon_{t-k}(0))$ for $0 < k \leq q$, and $\beta_{t,k} = 0$ for $k > q$ to see that it is a special case of Example 2.2.1. Then by Example 2.2.1, we have

$$\begin{aligned}\tau_0 &= \frac{1}{T-K} \sum_{t=K+1}^T (\mu_t(1) - \mu_t(0) + \epsilon_t(1) - \epsilon_t(0)), \\ \tau_k &= \theta_k \left[\frac{1}{T-K} \sum_{t=K+1}^T (\epsilon_{t-k}(1) - \epsilon_{t-k}(0)) \right], \quad \text{for } 0 < k \leq q, \\ \tau_k &= 0, \quad \text{for } k > q.\end{aligned}$$

Therefore, τ_0 is the average treatment effect on the mean function μ_t and the error term ϵ_t over time. For $0 < k \leq q$, τ_k is the average treatment effect on the error term ϵ_t at lag k over time, multiplied by θ_k , which is the k -step coefficient of the moving-average model. Moreover, $\tau_k = 0$ for k larger than the moving average order q .

Remark 2.2.1. In general, the estimand τ_k in (2.2.8) depends on the treatment assignment probabilities $(p_1, \dots, p_T)$ because for each $\tau_{t,k}$, the expectation is taken with respect to the joint distribution of $(Z_{t:t-k+1}, Z_{t-k-1:1})$. Hence, changing the design probabilities may change the estimand. However, under the linear model considered in Examples 2.2.1–2.2.3, the estimand τ_k is independent of the treatment assignment probabilities because the treatment effects are linear and the potential outcome model is additive in the treatment indicators Z_t 's. In this case, τ_k reflects the structural effect of lag- k treatment on the outcome, independent of the randomization probabilities.

2.3 Theory

2.3.1 Consistency and asymptotic normality

In this section, we analyze the asymptotic properties of $\hat{\tau}$ in (2.2.6) as $T \rightarrow \infty$ (Bojinov and Shephard, 2019; Liang and Recht, 2025), analogous to the asymptotic analysis in traditional randomized experiments by letting the sample size grow to infinity (Li and Ding, 2017). To build intuition for why $\hat{\tau}$ consistently estimates τ , consider first a simple marginal regression. Fix a lag k and consider the marginal OLS regression of Y_t on the normalized treatment indicator $\tilde{Z}_{t-k}$. Because the regressors in the full OLS regression (2.2.3) are mean-zero and orthogonal in expectation across lags, the coefficient from the marginal regression is close to the lag- k coefficient from the full regression. Let $\hat{\beta}_k$ denote the OLS coefficient from the marginal regression of Y_t on $\tilde{Z}_{t-k}$ without an intercept, using observations $t = K+1, \dots, T$. Then the marginal OLS estimator approximately recovers the lag- k treatment effect τ_k :

$$w_k^{-1} \hat{\beta}_k = w_k^{-1} \left(\sum_{t=K+1}^T \tilde{Z}_{t-k} Y_t \right) \left(\sum_{t=K+1}^T \tilde{Z}_{t-k}^2 \right)^{-1} \approx \frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} Y_t \approx \tau_k.$$

Therefore, using the representation of $\hat{\tau}$ in (2.2.6), we decompose $\hat{\tau}$ as:

$$\hat{\tau} = \tau + \mathbf{W}_K^{-1}[(T - K)^{-1} \mathbf{Z}_K^\top \mathbf{Z}_K]^{-1}[(T - K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)]. \quad (2.3.1)$$

Then the key is to analyze the asymptotic behaviors of $(T - K)^{-1} \mathbf{Z}_K^\top \mathbf{Z}_K$ and $(T - K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)$, since $\mathbf{W}_K$ is a fixed diagonal matrix. We impose the following assumption on the treatment probabilities of the time-series experiment.

Assumption 2.3.1 (Treatment probabilities condition). *There exist some constant $\epsilon \in (0, 1/2]$ such that $\epsilon \leq \inf_{t \geq 1} p_t \leq \sup_{t \geq 1} p_t \leq 1 - \epsilon$.*

Assumption 2.3.1 assumes that the treatment probabilities are uniformly bounded away from 0 and 1 for all time points, which rules out extreme treatment probabilities. This corresponds to the classic overlap assumption for causal inference with observational studies (D'Amour et al., 2021), which assumes the propensity scores are bounded away from 0 and 1 for almost all covariates.

Then we impose weak dependence conditions on the potential outcomes sequence $[Y_t = Y_t(Z_{t:1})]_{t=1}^\infty$, which rule out strong dependence of the observed outcomes on the past treatments. We consider two types of weak dependence: decaying carryover effects and m -dependence. Our results can be extended to more general weak dependence structures using the same proof techniques.

We first introduce the decaying carryover effects assumption, which assumes the carryover effects of the treatment decay over time.

Assumption 2.3.2 (decaying carryover effects condition). *Let*

$$\theta_{T,k} = \max_{k < t \leq T} \mathbb{E} \left[|Y_t(Z_{t:t-k+1}, Z_{t-k:1}) - Y_t(Z_{t:t-k+1}, Z'_{t-k:1})| \right],$$

where $Z'_{t-k:1}$ is an identically distributed copy of $Z_{t-k:1}$ and is independent of $Z_{t:1}$, and the expectation is taken over the joint distribution of $(Z_{t:t-k+1}, Z_{t-k:1})$ and the distribution of $Z'_{t-k:1}$. Assume that

$$(i) \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0 \text{ as } T \rightarrow \infty.$$

$$(ii) \theta_{T,k} = O(k^{-1-\delta}) \text{ for some } \delta > 0.$$

We then introduce the m -dependence assumption (Hoeffding and Robbins, 1948), which restricts the influence of past treatments to some given number of time points.

Assumption 2.3.3 (m -dependence condition). *There exist some $m = m(T)$ such that $Y_t = Y_t(Z_{t:1}) = Y_t(Z_{t:t-m})$ for all $m < t \leq T$ with*

$$(i) m/T \rightarrow 0.$$

$$(ii) m^4/T \rightarrow 0.$$

In both Assumption 2.3.2 and Assumption 2.3.3, the first condition is weaker than the second: the first ensures consistency of $\hat{\tau}$, whereas the second ensures the asymptotic normality of $\hat{\tau}$. Assumption 2.3.2 requires that the carryover effects of the treatment decay sufficiently fast. Similar conditions appear in the previous time-series experiments literature (Ni et al., 2023; Liang and Recht, 2025). The decaying carryover effects assumption is analogous to the approximate neighborhood interference assumption in network experiments (Leung, 2022), and can be viewed as a causal analogue of mixing conditions in traditional time series analysis (Rosenblatt, 1956; Bradley, 2005). Assumption 2.3.3 requires that the observed outcome at time t depends only on the most recent m treatments. Assumptions 2.3.2 and 2.3.3 are related but do not necessarily imply one another. Assumption 2.3.2 states that the carryover effects of the treatment decay over time, while still allowing dependence on all prior treatments. In contrast, Assumption 2.3.3 permits arbitrary dependence on past treatments but restricts this dependence to a given number of previous time points.

Remark 2.3.1. Recall the linear potential outcome model in Example 2.2.1, and suppose the error term ϵ_t does not depend on the treatment path $z_{t:1}$, that is,

$$Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \beta_{t,k} z_{t-k} + \epsilon_t. \quad (2.3.2)$$

Under the model (2.3.2), the quantity $\theta_{T,k}$ in the decaying carryover effects assumption (Assumption 2.3.2) can be expressed as

$$\begin{aligned} \theta_{T,k} &= \max_{k < t \leq T} \mathbb{E} \left[\left| \sum_{j=k}^{t-1} \beta_{t,j} Z_{t-j} - \sum_{j=k}^{t-1} \beta_{t,j} Z'_{t-j} \right| \right] \\ &\leq 2 \max_{k < t \leq T} \sum_{j=k}^{t-1} p_{t-j} (1 - p_{t-j}) |\beta_{t,j}| \leq \frac{1}{2} \max_{k < t \leq T} \sum_{j=k}^{t-1} |\beta_{t,j}|, \end{aligned}$$

where $(Z'_1, \dots, Z'_{t-k})$ is an independent and identically distributed copy of $(Z_1, \dots, Z_{t-k})$ for all $k < t \leq T$. This upper bound shows that the decay rate of $\theta_{T,k}$ in Assumption 2.3.2 is determined by how quickly the coefficients $\beta_{t,j}$ decrease as the lag j increases. In contrast, the m -dependence assumption (Assumption 2.3.3) is equivalent to for all $k < t \leq T$: $\beta_{t,k} = 0$ for all $m < k \leq t-1$. Thus, Assumption 2.3.3 enforces a strict truncation of the carryover effects after m lags.

Remark 2.3.2. The decaying carryover effects assumption (Assumption 2.3.2) involves the quantity $\theta_{T,k}$, which depends on the distribution of treatment assignments. One might therefore be concerned with this dependence. However, $\theta_{T,k}$ is upper bounded by

$$\theta_{T,k} \leq \max_{k < t \leq T} \max_{\substack{z_{t:t-k+1} \in \{0,1\}^k \\ z_{t-k:1}, z'_{t-k:1} \in \{0,1\}^{t-k}}} \left| Y_t(z_{t:t-k+1}, z_{t-k:1}) - Y_t(z_{t:t-k+1}, z'_{t-k:1}) \right|,$$

where the maximum is taken over all possible realizations of treatment assignments $z_{t:t-k+1}, z_{t-k:1}, z'_{t-k:1}$. Consequently, we may impose a stronger assumption that depends only on the potential outcomes but not on the distribution of treatment assignments.

Remark 2.3.3. What distinguishes time-series experiments from network experiments is that the primary causal estimands of interest are lagged treatment effects, i.e., the impact of current and past treatments on the current outcome, and the associated structure of potential outcomes. The temporal structure enables us to regress outcomes directly on normalized treatment indicators from multiple past periods. This necessitates asymptotic analysis of regression-based estimators and HAC variance estimation in this section. The causal estimands, identification strategy, and asymptotic theory in time-series experiments are fundamentally different from and not covered by the existing network experiments literature.

In the decomposition (2.3.1), we can write

$$\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) = \frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau). \quad (2.3.3)$$

Consider the sequence $[\tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau)]_{t=K+1}^\infty$. Under Assumption 2.3.2, for a fixed K , the carryover effects of the sequence decay with the same asymptotic order as $\theta_{T,k}$. Under Assumption 2.3.3, the sequence is $\max\{K, m\}$ -dependent. Therefore, we can establish the consistency and apply the central limit theorem (CLT) for the weak dependence sequence (Chen and Shao, 2004; Chandrasekhar et al., 2023).

To ensure well-behaved asymptotic properties of $\hat{\tau}$, we impose the following standard moment conditions on the potential outcomes.

Assumption 2.3.4 (Moment condition). *The potential outcomes $Y_t(z_{t:1})$ are uniformly bounded for all t and $z_{t:1} \in \{0, 1\}^t$, i.e., $\sup_{t \geq 1} \sup_{z_{t:1} \in \{0, 1\}^t} |Y_t(z_{t:1})| < \infty$. Additionally, $\|\tau\|_1$ is uniformly bounded for all T , i.e., $\sup_{T \geq 1} \|\tau\|_1 < \infty$.*

By definition in (2.2.8), τ depends on T implicitly. In Assumption 2.3.4, if we further assume K is fixed, the uniform boundedness of $\|\tau\|_1$ follows directly from the uniform boundedness of the potential outcomes Y_t . However, when K is diverging, the uniform boundedness of $\|\tau\|_1$ implies that τ_k decays to zero as $k \rightarrow \infty$, which indicates that the treatment effect decays over time. It is possible to relax the uniform boundedness assumption on τ , but it will complicate the asymptotic analysis. We impose Assumption 2.3.4 for the simplicity of presentation.

We first show the consistency of $\hat{\tau}$ in the following theorem.

Theorem 2.3.1 (Consistency). *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.1, Assumption 2.3.2(i) with $K \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0$ or Assumption 2.3.3(i) with $Km/T \rightarrow 0$, Assumption 2.3.4, and $K^2/T \rightarrow 0$. Then we have $\|\hat{\tau} - \tau\|_2 \xrightarrow{p} 0$.*

Since the dimension of τ can potentially grow with T , we establish the CLT for a linear projection of $\hat{\tau}$ rather than $\hat{\tau}$ itself. To characterize the asymptotic variance, we introduce a covariance matrix $\mathbf{V} \in \mathbb{R}^{(K+1) \times (K+1)}$ below. Since $\hat{\tau}$ is the estimate of τ and $\tilde{\tau} = \mathbf{W}_K \hat{\tau}$ by linear transformation (2.2.4), $\tilde{\tau}$ is the estimate of $\mathbf{W}_K \tau$. We define the oracle residual of OLS regression (2.2.3) at time t as

$$U_t = Y_t - \tilde{\mathbf{Z}}_{t:t-K}^\top \mathbf{W}_K \tau = Y_t - \sum_{k=0}^K \tilde{Z}_{t-k} w_k \tau_k. \quad (2.3.4)$$

Then by (2.3.1), for $k, k' = 1, \dots, K+1$, define the k, k' -th element of $\mathbf{V}$ as:

$$[\mathbf{V}]_{k,k'} = \frac{1}{T-K} \text{Cov} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k+1} U_t, \sum_{t=K+1}^T \tilde{Z}_{t-k'+1} U_t \right]. \quad (2.3.5)$$

By definition in (2.3.5), $\mathbf{V}$ depends on T but not on the observed treatment path and potential outcomes path. This covariance matrix captures the dependence structure of the oracle residuals across different lags.

We can then establish the CLT for the linear projection of $\hat{\tau}$ in the following theorem.

Theorem 2.3.2 (Asymptotic normality). *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.1, Assumption 2.3.2(ii) with K fixed or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, and Assumption 2.3.4. Then for any non-zero $\lambda_K \in \mathbb{R}^{K+1}$ satisfying $\liminf_{T \rightarrow \infty} \lambda_K^\top \mathbf{V} \lambda_K > 0$ and $\limsup_{T \rightarrow \infty} \|\lambda_K\|_1 < \infty$, we have*

$$\sqrt{T-K} (\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top (\hat{\tau} - \tau) \xrightarrow{d} N(0, 1).$$

The condition on λ_K in Theorem 2.3.2 prevents the linear contrast λ_K from spreading its mass over an increasing number of lags as K grows. When K is fixed, this condition holds automatically for any nonzero λ_K as long as $\liminf_{T \rightarrow \infty} \lambda_{\min}(\mathbf{V}) > 0$, where $\lambda_{\min}(\mathbf{V})$ is the smallest eigenvalue of $\mathbf{V}$. By the Cramér–Wold theorem, Theorem 2.3.2 directly implies a multivariate CLT for $\hat{\tau}$ when K is fixed.

Corollary 2.3.1. *Under the conditions of Theorem 2.3.2, if further K is fixed and $\liminf_{T \rightarrow \infty} \lambda_{\min}(\mathbf{V}) > 0$, we have*

$$\sqrt{T-K} \mathbf{V}^{-1/2} (\hat{\tau} - \tau) \xrightarrow{d} N(0, \mathbf{I}_{K+1}).$$

In the supplementary material, we establish three lemmas for proving Theorem 2.3.1 (Consistency) and Theorem 2.3.2 (Asymptotic normality). Recall the decomposition of $\hat{\tau}$ in (2.3.1). The first lemma shows that $(T-K)^{-1} \mathbf{Z}_K^\top \mathbf{Z}_K$ asymptotically converges to $\mathbf{W}_K^{-1}$ under Assumption 2.3.1. The second and third lemmas show the asymptotic behavior of $(T-K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)$ under all assumptions.

2.3.2 HAC variance estimation

Recall the oracle residual U_t defined in (2.3.4). Let

$$\widehat{U}_t = Y_t - \widetilde{Z}_{t:t-K}^\top \widetilde{\tau} \quad (2.3.6)$$

be the sample residual of OLS regression (2.2.3) at time t . Let $\mathbf{U}_K = \text{diag}(U_t : t = K + 1, \dots, T) \in \mathbb{R}^{(T-K) \times (T-K)}$ and $\widehat{\mathbf{U}}_K = \text{diag}(\widehat{U}_t : t = K + 1, \dots, T) \in \mathbb{R}^{(T-K) \times (T-K)}$ be the diagonal matrices of the oracle and sample residuals, respectively.

To account for autocorrelation, we introduce a kernel weighting matrix $\mathbf{Q}_K \in \mathbb{R}^{(T-K) \times (T-K)}$. Specifically, we use the Bartlett kernel (Bartlett, 1950), defined entrywise as

$$[\mathbf{Q}_K]_{ij} = \begin{cases} 1 - \frac{|j-i|}{L+1}, & |j-i| \leq L, \\ 0, & |j-i| > L, \end{cases}$$

for $1 \leq i, j \leq T-K$, where L is the bandwidth parameter controlling the effective dependence range. A common choice is $L = \lfloor T^{1/4} \rfloor$ (Greene, 2003). By construction, $\mathbf{Q}_K$ is positive semidefinite. The HAC estimator with the Bartlett kernel corresponds to Newey–West estimator (Newey and West, 1987). We can generalize the results to HAC estimators with other kernels.

To construct the HAC variance estimator for $\widehat{\tau}$, we first define the HAC variance estimator for $\widetilde{\tau}$ by OLS regression (2.2.3) as:

$$\widetilde{\mathbf{V}} = (T-K)(\mathbf{Z}_K^\top \mathbf{Z}_K)^{-1}(\mathbf{Z}_K^\top \widehat{\mathbf{U}}_K \mathbf{Q}_K \widehat{\mathbf{U}}_K \mathbf{Z}_K)(\mathbf{Z}_K^\top \mathbf{Z}_K)^{-1},$$

and then define the HAC variance estimator for $\widehat{\tau}$ by linear transformation (2.2.4) as:

$$\widehat{\mathbf{V}} = \mathbf{W}_K^{-1} \widetilde{\mathbf{V}} \mathbf{W}_K^{-1} = (T-K) \mathbf{W}_K^{-1} (\mathbf{Z}_K^\top \mathbf{Z}_K)^{-1} (\mathbf{Z}_K^\top \widehat{\mathbf{U}}_K \mathbf{Q}_K \widehat{\mathbf{U}}_K \mathbf{Z}_K) (\mathbf{Z}_K^\top \mathbf{Z}_K)^{-1} \mathbf{W}_K^{-1}. \quad (2.3.7)$$

Intuitively, the HAC variance estimator corrects the usual OLS variance formula by allowing residuals at nearby time points to be correlated. The kernel matrix $\mathbf{Q}_K$ assigns larger weights to correlations between residuals that are closer in time and gradually downweights correlations as the lag length increases, with the bandwidth L determining how far into the past these correlations are taken into account. This adjustment ensures that the estimated variance remains consistent under general forms of heteroskedasticity and autocorrelation in the model-based inference with time-series data. We borrow their intuition and analyze $\widehat{\mathbf{V}}$ from the design-based framework.

Recall the lag- k treatment effects defined in (2.2.7) and (2.2.8). Define $\mathbf{b}_K \in \mathbb{R}^{(T-K) \times (K+1)}$ with

$$[\mathbf{b}_K]_{t-K,k+1} = \tau_{t,k} - \frac{w_k}{p_{t-k}(1-p_{t-k})} \tau_k,$$

for $t = K+1, \dots, T$ and $k = 0, \dots, K$, which measures the heterogeneity of the treatment effect and treatment probability across time points. We then define the bias term

as $\mathbf{B}_K = \mathbf{b}_K^\top \mathbf{Q}_K \mathbf{b}_K$. By construction, $\mathbf{B}_K$ is positive semidefinite because $\mathbf{Q}_K$ is positive semidefinite under the Bartlett kernel. Intuitively, $\mathbf{B}_K$ captures the bias component in the HAC variance estimator that arises when treatment probabilities vary over time or when treatment effects are heterogeneous across time points. The following theorem establishes that $\hat{\mathbf{V}}$ is a conservative estimator of $\mathbf{V}$.

Theorem 2.3.3. *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.1, Assumption 2.3.2(ii) with K fixed, $L \rightarrow \infty$ or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0, Km^2/L \rightarrow 0$, Assumption 2.3.4 and $K^4L^2/T \rightarrow 0$. Then $\|\hat{\mathbf{V}} - \mathbf{V} - \mathbf{B}_K\|_F \xrightarrow{p} 0$.*

Further assume the treatment probabilities are constant across time ($p_t = p$ for all t) and the treatment effects are homogeneous over time ($\tau_{t,k} = \tau_k$ for all t and k). Then $\mathbf{b}_K = 0$, which implies $\mathbf{B}_K = 0$. In this case, the HAC variance estimator $\hat{\mathbf{V}}$ is consistent for $\mathbf{V}$, i.e., $\|\hat{\mathbf{V}} - \mathbf{V}\|_F \xrightarrow{p} 0$.

Theorem 2.3.3 generalizes the classical variance results of [Neyman \(1923\)](#) for completely randomized experiments. In the special case with constant treatment probabilities and homogeneous treatment effects, the additional term $\mathbf{B}_K$ vanishes, and the HAC variance estimator consistently estimates the true variance. This mirrors [Neyman \(1923\)](#)'s result that the design-based variance estimator is consistent under constant effects. When treatment effects or assignment probabilities vary over time, however, the extra positive semidefinite term $\mathbf{B}_K$ inflates the variance, making the HAC estimator conservative, just as the design-based variance estimator is conservative without the constant effects assumption. In this sense, our framework extends Neyman's design-based analysis from randomized experiments to time-series experiments.

2.3.3 Comparisons with the existing estimands

[Bojinov and Shephard \(2019\)](#) defined the general treatment effect of lag k at time t as:

$$\sum_{\substack{z_{t:t-k+1} \in \{0,1\}^k \\ z_{t-k-1:1} \in \{0,1\}^{t-k-1}}} a_{z_{t:t-k+1}, z_{t-k-1:1}} [Y_t(z_{t:t-k+1}, 1, z_{t-k-1:1}) - Y_t(z_{t:t-k+1}, 0, z_{t-k-1:1})], \quad (2.3.8)$$

for some non-stochastic weights satisfying $\sum_{z_{t:t-k+1}, z_{t-k-1:1}} a_{z_{t:t-k+1}, z_{t-k-1:1}} = 1$ and $a_{z_{t:t-k+1}, z_{t-k-1:1}} \geq 0$. Then the general ATE is defined as the corresponding average over all time points, analogous to the definition of τ_k based on $\tau_{t,k}$.

By the design under Assumption 2.2.2, we can write $\tau_{t,k}$ in (2.2.7) explicitly as

$$\tau_{t,k} = \sum_{\substack{z_{t:t-k+1} \in \{0,1\}^k \\ z_{t-k-1:1} \in \{0,1\}^{t-k-1}}} a_{z_{t:t-k+1}, z_{t-k-1:1}} [Y_t(z_{t:t-k+1}, 1, z_{t-k-1:1}) - Y_t(z_{t:t-k+1}, 0, z_{t-k-1:1})],$$

where

$$a_{z_{t:t-k+1}, z_{t-k-1:1}} = \prod_{s=1, s \neq t-k}^t p_s^{z_s} (1 - p_s)^{1-z_s}.$$

Therefore, $\tau_{t,k}$ is a special case of the general treatment effect (2.3.8) in [Bojinov and Shephard \(2019\)](#). However, [Bojinov and Shephard \(2019\)](#) argued that the general ATE (2.3.8) cannot be estimated without strong assumptions, so they changed the estimand to the causal effect as a function of the observed treatment path, which is random (see Section 3.2 of [Bojinov and Shephard \(2019\)](#)). [Rambachan and Shephard \(2019\)](#) also discussed the nonparametric causal interpretation of common time series estimands in terms of dynamic causal effects depending on the observed treatment path. In contrast, we focus on a special class of the general treatment effects, which does not depend on the observed treatment path, and can be estimated by OLS under time-series experiments with independent treatments.

In general, we can adjust the weights in $\tilde{Z}_t$ to estimate a broad class of ATEs. Consider a new normalization: $\tilde{Z}_{h,t} = h_t(p_1, \dots, p_t)(Z_t - \mathbb{E}[Z_t]) / \text{Var}[Z_t]$ for some function $h_t : \{0, 1\}^t \rightarrow \mathbb{R}$ at each time t . Define the weight as $w_{h,k} = [(T - K)^{-1} \sum_{t=K+1}^T h_t(p_1, \dots, p_t)^2 (p_{t-k}(1 - p_{t-k}))^{-1}]^{-1}$. We then consider the OLS regression without an intercept: $\text{lm}(Y_t \sim \tilde{Z}_{h,t} + \tilde{Z}_{h,t-1} + \dots + \tilde{Z}_{h,t-K})$ using observations from $t = K+1, \dots, T$, and let $(\tilde{\tau}_{h,0}, \dots, \tilde{\tau}_{h,K})$ denote the OLS coefficients. For each $k = 0, \dots, K$, we define $\hat{\tau}_{h,k} = w_{h,k}^{-1} \tilde{\tau}_{h,k}$. The corresponding estimands of this procedure are $(\tau_{h,0}, \dots, \tau_{h,K})$, where

$$\begin{aligned} \tau_{h,k} &= \frac{1}{T - K} \sum_{t=K+1}^T h_t(p_1, \dots, p_t) \tau_{t,k} \\ &= \frac{1}{T - K} \sum_{t=K+1}^T h_t(p_1, \dots, p_t) \mathbb{E}[Y_t(Z_{t:t-k+1}, 1, Z_{t-k-1:1}) - Y_t(Z_{t:t-k+1}, 0, Z_{t-k-1:1})]. \end{aligned}$$

[Liang and Recht \(2025\)](#) also studied time-series experiments and the corresponding design-based analysis. Our paper differs from theirs in three key ways. First, their analysis assumes a linear convolution model (their Section 2) and a circular convolution model (their Section 3), along with a Bernoulli randomization scheme where $p_t = 1/2$ for all t . In contrast, our framework does not impose any structural assumptions on the potential outcomes, allowing for arbitrary and unknown heterogeneous effects over time. Second, we consider more general experimental designs with time-varying treatment probabilities. Third, [Liang and Recht \(2025\)](#) proposed a method-of-moments estimator, established a central limit theorem, and proposed a variance estimator. In our work, we focus on the simpler OLS estimation and analyze its asymptotic properties, along with the performance of the HAC variance estimator derived from the same OLS procedure. Both OLS and HAC variance estimation are widely used in practice, and are implemented in standard statistical software routines. We provide theoretical support for their application in time-series experiments.

2.4 Extension to continuous treatment

Many modern experimental settings feature treatments that take discrete or continuous values rather than binary values. For instance, in online platforms, treatment intensity may

correspond to the proportion of user exposure to a new algorithm or the duration for which an experimental interface is displayed. In clinical trials, treatments often involve dosage levels or exposure durations, which are naturally measured on a continuous scale.

In settings where the experiment involves multiple discrete treatment levels in addition to a control group, one natural approach is to run separate OLS regressions for each treatment level relative to control. Our results in Section 2.2 and 2.3 can then be applied. Therefore, we only focus on the case where the treatment variable $Z_t \in \mathbb{R}$ is continuous. We introduce the following assumption, which is the continuous treatment analogue of Assumption 2.2.2.

Assumption 2.4.1 (Time-series experiment for continuous treatment). *Assume the treatments $Z_1, \dots, Z_T$ are independent. Assume the values of $(E[Z_1], \dots, E[Z_T])$ and $(\text{Var}[Z_1], \dots, \text{Var}[Z_T])$ exist and are known by the design.*

As in the binary treatment case, we normalize the treatment variable by considering $\tilde{Z}_t = (Z_t - E[Z_t]) / \text{Var}[Z_t]$. We then fit the following OLS regression model:

$$\text{lm}(Y_t \sim \tilde{Z}_t + \dots + \tilde{Z}_{t-K}),$$

where the coefficient vector is denoted by $\tilde{\tau}$ and its transformed version by $\hat{\tau} = \mathbf{W}_K^{-1} \tilde{\tau}$ as (2.2.6) before with the same $\mathbf{W}_K$. The causal estimand is the vector of lagged treatment effects $\tau = (\tau_0, \dots, \tau_K)^\top \in \mathbb{R}^{K+1}$, where the treatment effect of lag- k treatment on the outcome at time t is defined as

$$\tau_{t,k} = E \left[\frac{(Z_{t-k} - E[Z_{t-k}])}{\text{Var}[Z_{t-k}]} Y_t \right],$$

and the ATE of lag- k treatment on the outcome over time is defined as

$$\tau_k = \frac{1}{T-K} \sum_{t=K+1}^T \tau_{t,k},$$

which extend the notion of the treatment effect of lag- k in (2.2.7) and (2.2.8) to the continuous treatment setting. This estimand τ_k represents the average slope of the potential outcome with respect to the treatment assigned k periods earlier, while averaging over all other past and concurrent treatment assignments.

In the literature on continuous treatments, the potential outcome under treatment intensity z is referred to as the (causal) dose-response curve, denoted $Y(z)$, while its derivative $(d/dz)Y(z)$ is known as the causal derivative effect; see Zhang and Chen (2025) and references therein. Building on this perspective, we adapt the notion of derivative effects to the setting of time-series experiments. Specifically, we define the effect of treatment at lag- k on the outcome at time t as

$$\tau'_{t,k}(z) = \frac{d}{dz} E[Y_t(Z_{t:t-k+1}, z, Z_{t-k-1:1})],$$

where the expectation is taken over the joint distribution of $(Z_{t:t-k+1}, Z_{t-k-1:1})$. Intuitively, $\tau'_{t,k}(z)$ measures the local sensitivity of the expected outcome at time t to a marginal change in treatment intensity at lag- k , holding the rest of the treatment path fixed. Aggregating across time, we can equivalently represent the estimand τ_k as a weighted average of these local effects:

$$\tau_k = \frac{1}{T-K} \sum_{t=K+1}^T \left(\int w_{t,k}(z) dz \right)^{-1} \int \tau'_{t,k}(z) w_{t,k}(z) dz,$$

where the weighting function $w_{t,k}(z)$ is given by

$$w_{t,k}(z) = \{E[Y_t(Z_{t:1}) \mid Z_{t-k} \geq z] - E[Y_t(Z_{t:1}) \mid Z_{t-k} < z]\}P(Z_{t-k} \geq z)P(Z_{t-k} < z).$$

Thus, the estimand τ_k can be interpreted as the time-averaged, weighted average derivative effect at lag k . This places our estimand within the broader class of weighted average derivatives (Newey and Stoker, 1993), and connects naturally to discussions of how OLS and instrumental variables identify such objects (Yitzhaki, 1996; Borusyak and Hull, 2024). Importantly, both the dose-response curve and the derivative effect are nonregular parameters and cannot generally be estimated at parametric rates. In contrast, we consider the weighted average of the derivative effect, which admits parametric convergence rates.

We next impose regularity conditions on the treatment distribution and outcome to extend our asymptotic results to the continuous treatment setting. We consider the following assumption on the distribution of Z_t , which serves as a continuous treatment analogue of Assumption 2.3.1.

Assumption 2.4.2 (Treatment distributions condition). *There exists some constant $\epsilon > 0$ such that $\epsilon \leq \inf_{t \geq 1} \text{Var}[Z_t]$.*

We also consider the following assumption, which serves as a continuous treatment analogue of Assumption 2.3.4.

Assumption 2.4.3 (Moment condition). *The potential outcomes $Y_t(z_{t:1})$ are uniformly bounded for all t and all $z_{t:1} \in \mathbb{R}^t$, i.e., $\sup_{t \geq 1, z_{t:1} \in \mathbb{R}^t} |Y_t(z_{t:1})| < \infty$. Additionally, $\|\tau\|_1$ is uniformly bounded for all T , i.e., $\sup_T \|\tau\|_1 < \infty$.*

With these conditions in place, we obtain the consistency and asymptotic normality of the OLS estimator. These results are the continuous treatment counterparts to Theorem 2.3.1 and Theorem 2.3.2.

Theorem 2.4.1 (Consistency). *Assume Assumption 2.2.1, Assumption 2.4.1, Assumption 2.4.2, Assumption 2.3.2(i) with $K \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0$ or Assumption 2.3.3(i) with $Km/T \rightarrow 0$, Assumption 2.4.3, and $K^2/T \rightarrow 0$. Then we have $\|\hat{\tau} - \tau\|_2 \xrightarrow{p} 0$.*

Theorem 2.4.2 (Asymptotic normality). *Assume Assumption 2.2.1, Assumption 2.4.1, Assumption 2.4.2, Assumption 2.3.2(ii) with K fixed or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, and Assumption 2.4.3. Then for any non-zero $\lambda_K \in \mathbb{R}^{K+1}$ satisfying $\liminf_{T \rightarrow \infty} \lambda_K^\top \mathbf{V} \lambda_K > 0$ and $\limsup_{T \rightarrow \infty} \|\lambda_K\|_1 < \infty$, we have*

$$\sqrt{T-K}(\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top (\hat{\tau} - \tau) \xrightarrow{d} N(0, 1),$$

where $\mathbf{V}$ is defined in (2.3.5).

To estimate the asymptotic variance $\mathbf{V}$ in Theorem 2.4.2, we can use exactly the same HAC variance estimator $\hat{\mathbf{V}}$ in (2.3.7). Define $\mathbf{b}_K \in \mathbb{R}^{(T-K) \times (K+1)}$ with

$$[\mathbf{b}_K]_{t-K, k+1} = \tau_{t,k} - \frac{w_k}{\text{Var}[Z_{t-k}]} \tau_k,$$

for $t = K+1, \dots, T$ and $k = 0, \dots, K$, and define the bias term as $\mathbf{B}_K = \mathbf{b}_K^\top \mathbf{Q}_K \mathbf{b}_K$. The following theorem establishes that $\hat{\mathbf{V}}$ is a conservative estimator of $\mathbf{V}$, which is parallel to Theorem 2.3.3.

Theorem 2.4.3. *Assume Assumption 2.2.1, Assumption 2.4.1, Assumption 2.4.2, Assumption 2.3.2(ii) with K fixed, $L \rightarrow \infty$ or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, $Km^2/L \rightarrow 0$, Assumption 2.4.3 and $K^4 L^2/T \rightarrow 0$. Then $\|\hat{\mathbf{V}} - \mathbf{V} - \mathbf{B}_K\|_F \xrightarrow{p} 0$.*

If the treatment variances are constant across time ($\text{Var}[Z_t] = \sigma^2$ for all t) and the treatment effects are homogeneous over time ($\tau_{t,k} = \tau_k$ for all t and k), then $\mathbf{b}_K = 0$, which implies $\mathbf{B}_K = 0$. In this case, the HAC variance estimator $\hat{\mathbf{V}}$ is consistent for $\mathbf{V}$, i.e., $\|\hat{\mathbf{V}} - \mathbf{V}\|_F \xrightarrow{p} 0$.

2.5 Empirical studies

In this section, we conduct simulations to examine the finite-sample performance of $\hat{\tau}$ and the HAC variance estimator $\hat{\mathbf{V}}$ for estimating $\mathbf{V}$. We then illustrate the method on data from a trading experiment.

2.5.1 Simulations

We consider the autoregressive model in Example 2.2.2. Specifically, we take a first-order autoregressive model, $Y_t(z_{t:1}) = \mu_t(z_{t:1}) + \phi Y_{t-1}(z_{t-1:1}) + \epsilon_t(z_{t:1})$, for any treatment history $z_{t:1} \in \{0, 1\}^t$. We set $\mu_t(z_{t:1}) = \mu_t(z_t) = \mu(z_t)$ with $\mu(1) = 0.5$ and $\mu(0) = 0$, take $\phi = 0.5$, and let $\epsilon_t(z_{t:1}) = \epsilon_t \sim N(0, 1)$ i.i.d. across t . Under this specification, $Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \phi^k \mu(z_{t-k}) + \sum_{k=0}^{t-1} \phi^k \epsilon_{t-k}$. Then the true lag- k treatment effect is $\tau_k = \phi^k (\mu(1) - \mu(0))$.

We vary the time horizon $T \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$, fix the number of lags at $K = 5$, and use a constant treatment probability $p_t = p = 0.5$. Adopting a design-based perspective, randomness arises only from the assignment path $Z_1, \dots, Z_T$. For each

T , we therefore fix a realization $[\epsilon_1, \dots, \epsilon_T]$ and repeatedly resample $Z_1, \dots, Z_T$. For each draw we compute the OLS estimator $\hat{\tau}$ and the HAC variance estimator $\hat{\mathbf{V}}$ with the Bartlett kernel. A common choice is $L = \lfloor T^{1/4} \rfloor$ (Greene, 2003). Since the treatment probabilities are constant across time and the treatment effects are homogeneous over time, the HAC variance estimator $\hat{\mathbf{V}}$ is consistent for the asymptotic variance $\mathbf{V}$ by Theorem 2.3.3.

Figure 2.1 presents histograms of $\hat{\tau}_k - \tau_k$ for $k = 0, \dots, 5$ across replications under $T = 1000$. The empirical distributions are close to the normal distribution, supporting our central limit theorem (Theorem 2.3.2). This provides empirical evidence that the regression-based estimator achieves approximate normality even at moderate sample sizes.

Figure 2.2 displays violin plots of $\|\hat{\tau} - \tau\|_2$ as a function of T . The L_2 error decreases with larger sample sizes, aligning with the $T^{-1/2}$ convergence rate established in Theorem 2.3.2.

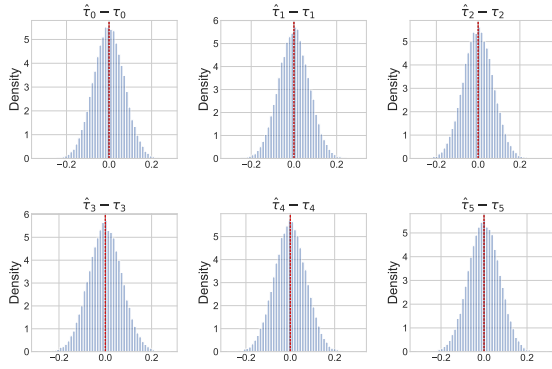


Figure 2.1: Histograms of $\hat{\tau}_k - \tau_k$ for $k = 0, \dots, 5$ under $T = 1000$.

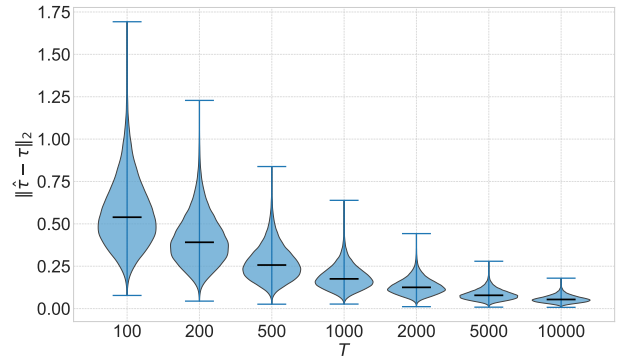


Figure 2.2: $\|\hat{\tau} - \tau\|_2$ versus T . The center bar is the median.

Figure 2.3 investigates the performance of the HAC variance estimator $\hat{\mathbf{V}}$. The plot shows the normalized Frobenius error $\|\hat{\mathbf{V}} - \mathbf{V}\|_F / \|\mathbf{V}\|_F$ for different kernel bandwidth choices $L \in \{0, 1, 5, \lfloor T^{1/4} \rfloor\}$, averaged across replications. When no autocorrelation adjustment is made ($L = 0$), the estimator remains biased and does not improve with T . By contrast, increasing the bandwidth significantly reduces error, with $L = \lfloor T^{1/4} \rfloor$ yielding the best finite-sample performance and approaching zero error as T grows. This confirms our theoretical variance results (Theorem 2.3.3) and demonstrates the importance of selecting an appropriate increasing bandwidth.

Table 2.1 reports the empirical coverage probabilities of nominal 95% confidence intervals for $\hat{\tau}_k$, constructed using the HAC variance estimator $\hat{\mathbf{V}}$ with bandwidth $L = T^{1/4}$. Across all lags $k = 0, \dots, 5$ and all sample sizes, the coverage probabilities remain very close to the nominal level, even for relatively small $T = 100$. As T increases, the coverage stabilizes tightly around 0.95, confirming that the asymptotic normal approximation in Theorem 2.3.2 provides accurate inference in finite samples.

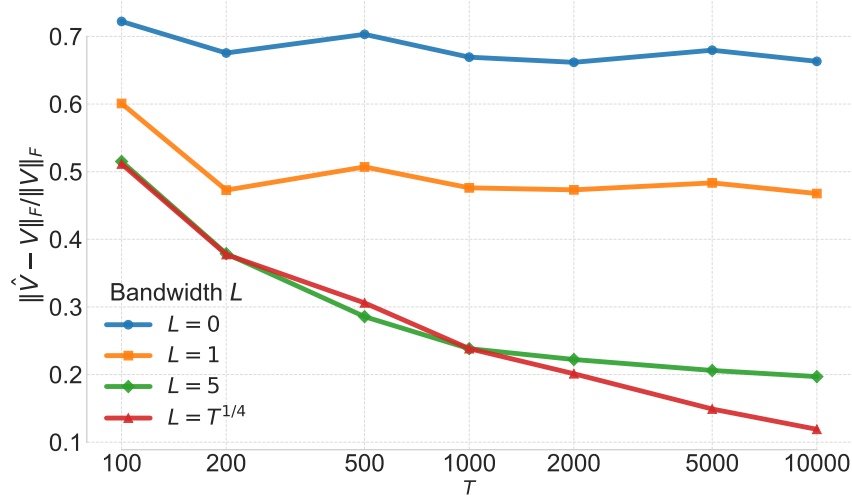


Figure 2.3: Normalized Frobenius error $\|\hat{\mathbf{V}} - \mathbf{V}\|_F / \|\mathbf{V}\|_F$ versus T for different bandwidths L .

Table 2.1: Empirical 95% coverage for $\hat{\tau}_k$ using $\hat{\mathbf{V}}$ with $L = \lfloor T^{1/4} \rfloor$.

T	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
100	0.934	0.935	0.932	0.933	0.934	0.934
200	0.944	0.943	0.940	0.943	0.942	0.943
500	0.943	0.946	0.947	0.945	0.948	0.945
1000	0.949	0.947	0.946	0.947	0.946	0.950
2000	0.949	0.949	0.948	0.949	0.950	0.950
5000	0.949	0.949	0.950	0.949	0.951	0.951
10000	0.950	0.950	0.950	0.949	0.950	0.950

2.5.2 Real data analysis

We analyze the trading experiment described in [Bojinov and Shephard \(2019\)](#), which was conducted by AHL Partners, a quantitative trading firm specializing in financial futures. The firm executes orders in one of two ways: using a human trader or using an algorithmic trader. For the experiment, AHL restricted attention to orders whose sizes lay between fixed lower and upper bounds (time-invariant within each market, but varying across markets). These eligible orders were randomly assigned to either a human trader or an algorithmic trader. Therefore, treatments were assigned independently across time, following an i.i.d. Bernoulli design, ensuring that the setup satisfied our no-anticipation assumption. AHL provided 10 independent datasets from 2016, each corresponding to a different equity index futures market spanning the US, Europe, and Asia. We analyze these markets separately

and regard each as a distinct unit. To protect confidentiality, AHL did not reveal which of methods “A” and “B” corresponds to human versus algorithmic traders, nor did they disclose the identities of the individual markets. For consistency, we label A as the control group and B as the treatment group. The outcome of interest Y is the slippage, which is a simple function of the trade prices and volumes at which an order was executed. Lower slippage indicates better execution quality and thus serves as a proxy for higher trading performance.

With $K = 4$, Figure 2.4 displays, for each market, the joint p-value for $H_0 : \tau = \mathbf{0}$, and the per-lag estimates $\hat{\tau}_k$ with their 95% confidence intervals. We find clear heterogeneity across markets: the null is rejected in Markets 4, 6, 7, and 9 under the nominal level 0.05, indicating the presence of dynamic treatment effects, while other markets show no joint evidence of impact. In both Market 6 and Market 9, “A” (control) performed significantly better than “B” (treatment), and there is no market where “B” performed significantly better than “A”, consistent with the findings of [Bojinov and Shephard \(2019\)](#). Market 4 shows lagged effects at longer lags, whereas Market 7 shows a negative lag-2 effect. Some isolated single-lag signals (e.g., Market 5’s positive lag-1 estimate) do not lead to a joint rejection.

2.6 Extensions

Beyond estimating the main effects of treatments at different lags, it is often of interest to investigate interaction effects between treatments administered at distinct time periods. Including interaction terms is particularly useful in settings where the effect of a treatment depends on its recent history or on combinations of past interventions. In Section B.1.2 of the supplementary material, we extend our framework to regression with interaction terms.

So far, we have proposed working models for the treatment effects through individual lags or interactions of consecutive lags. While this approach captures direct and pairwise temporal dependencies, it may be restrictive in settings where outcomes depend on more general features of the treatment path. To address this, we consider exposure mappings that summarize treatment assignments over multiple periods into lower-dimensional summary statistics. In Section B.1.3 of the supplementary material, we extend our framework to general exposure mapping.

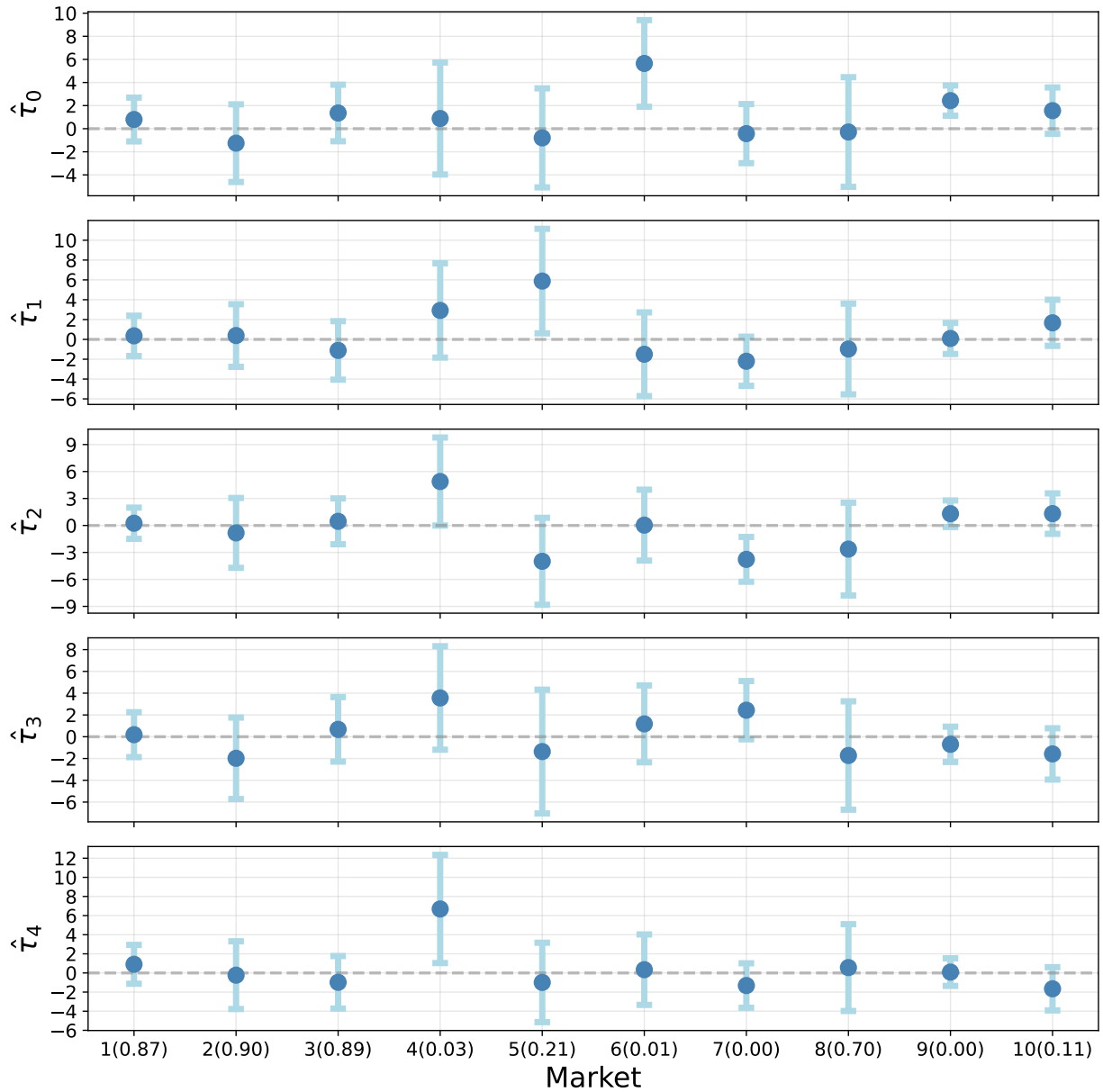


Figure 2.4: Lag-specific estimates $\hat{\tau}_k$ with 95% confidence intervals for $k = 0, \dots, 4$ under different markets. The number in parentheses following each market label reports the joint p-value for $H_0 : \tau = \mathbf{0}$.

Chapter 3

Estimation based on nearest neighbor matching: from density ratio to average treatment effect

3.1 Introduction

Matching methods ([Greenwood, 1945](#); [Chapin, 1947](#); [Cochran and Rubin, 1973](#); [Rubin, 2006](#); [Rosenbaum, 2010](#)) aim to balance observations from different groups through minimizing group differences in observed covariates. Such methods have proven their usefulness for causal inference in various disciplines, including economics ([Imbens, 2004](#)), epidemiology ([Brookhart et al., 2006](#)), political science ([Ho et al., 2007](#); [Sekhon, 2008](#)) and sociology ([Morgan and Harding, 2006](#)).

Among all the matching methods, nearest neighbor (NN) matching ([Rubin, 1973](#)) is likely the most frequently used and easiest to implement approach. In the simplest treatment-control study, NN matching assigns each treatment (control) individual to M control (treatment) individuals with the smallest distance to it. In this regard, two questions arise. First, how do we select the number of matches, M ? This is referred to in the literature as ratio matching, and is both important and delicate, well-known to be related to the bias-variance trade-off in nonparametric statistics ([Smith, 1997](#); [Rubin and Thomas, 2000](#); [Imbens and Rubin, 2015](#)). Second, how do we perform large-sample statistical inference for NN matching estimators? Such an analysis is usually nonstandard and technically challenging. Indeed, it was long-lacking in the literature until [Abadie and Imbens \(2006\)](#).

To answer the above two questions, a series of papers ([Abadie and Imbens, 2006, 2008, 2011, 2012](#)) established large-sample properties of M -NN matching for estimating the average treatment effect (ATE). These results are, however, only valid when, in ratio matching, M is fixed. The according message is then mixed. As a matter of fact, the ATE estimator based on M -NN matching with a fixed M is asymptotically biased and inefficient. While bias correction is now feasible to alleviate the first issue ([Abadie and Imbens, 2011](#)), the lack

of efficiency seems fundamental.

This manuscript revisits the study of Abadie and Imbens (2006) from a new perspective, bridging M -NN matching to density ratio estimation (Nguyen et al., 2010; Sugiyama et al., 2012) as well as double robustness (Scharfstein et al., 1999; Bang and Robins, 2005). To this end, our analysis stresses, in ratio matching, the benefits of forcing M to diverge with the sample size n in order to achieve statistical efficiency. Our claim is thus aligned with observations in the random graph-based inference literature (Wald and Wolfowitz, 1940; Friedman and Rafsky, 1979; Henze, 1988; Liu and Singh, 1993; Henze and Penrose, 1999; Berrett et al., 2019; Bhattacharya, 2019; Shi et al., 2024, 2022; Lin and Han, 2023).

The contributions of this manuscript are two-fold. First, we show that a statistic that plays a pivotal role in the analysis of Abadie and Imbens (2006), $K_M(x)$ (Abadie and Imbens (2006, Page 240); to be defined in (3.2.2) of Section 3.2), which measures the number of matched times of the covariate value x , actually gives rise to a consistent density ratio estimator in the two-sample setting. Furthermore, from the angle of density ratio estimation, this NN matching-based estimator is to our knowledge the first one that simultaneously satisfies being conceptually one-step, computationally efficient, and statistically rate-optimal. This estimator itself is thus an appealing alternative to existing density ratio estimators.

Getting back to the original ATE estimation problem, our second contribution is to use the above insights to bridge the bias-corrected matching estimator (Abadie and Imbens, 2011), doubly robust estimators (Scharfstein et al., 1999; Bang and Robins, 2005; Farrell, 2015), and double machine learning estimators (Chernozhukov et al., 2018). In fact, Abadie and Imbens (2011)’s bias-corrected estimator can be formulated as

$$\hat{\tau}_M^{\text{bc}} = \hat{\tau}^{\text{reg}} + \frac{1}{n} \left[\sum_{i=1, D_i=1}^n \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i - \sum_{i=1, D_i=0}^n \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i \right]$$

(see Lemma 3.3.1, with notation introduced in Section 3.3 and $K_M(i)$ representing the number of times the unit i is matched), and then $1 + K_M(i)/M$ converges to the inverses of the propensity scores $1 - e(X_i)$ and $e(X_i)$ for units with $D_i = 0$ and 1, respectively. One could then leverage the general double robustness and double machine learning theory to validate the following two properties of $\hat{\tau}_M^{\text{bc}}$.

- (1) *Consistency*: $\hat{\tau}_M^{\text{bc}}$ converges in probability to the population ATE, if either the density (propensity score) functions satisfy certain conditions or the outcome (regression) model is consistently estimated, with $M \log n/n \rightarrow 0$ and $M \rightarrow \infty$ as $n \rightarrow \infty$.
- (2) *Semiparametric efficiency*: $\hat{\tau}_M^{\text{bc}}$ is an asymptotically Normal estimator of τ with the asymptotic variance attaining the semiparametric efficiency lower bound (Hahn, 1998), if the density functions are sufficiently smooth, the outcome model is consistently estimated, and M scales with n at a proper rate. Furthermore, a simple consistent estimator of the asymptotic variance is available.

Although [Abadie and Imbens \(2006, Theorem 5\)](#) hints at the necessity of allowing M to diverge for gaining efficiency, we provide rigorous theory for their conjecture. Our results thus complement those made in [Abadie and Imbens \(2006, 2011\)](#) and provide additional theoretical justifications for practitioners to use NN matching for inferring the ATE.

Technically, our analysis hinges on a diverging M that grows with n . In contrast, existing results on NN matching for causal effects ([Abadie and Imbens, 2006, 2008, 2011, 2012](#)) all focused on a fixed M . Instead, we take a different route to establish nonasymptotic moment bounds on $K_M(x)$ with more flexibility in specifying the rate of M with respect to n (see [Lin and Han \(2023\)](#) for a similar idea in analyzing rank-based statistics).

Paper organization. The rest of this manuscript proceeds as follows. Section 3.2 gives a brief overview of the NN matching-based density ratio estimator. Section 3.3 revisits [Abadie and Imbens \(2011\)](#)’s bias-corrected NN matching-based estimator of the ATE, $\hat{\tau}_M^{\text{bc}}$. Section 3.4 elaborates on the double robustness and semiparametric efficiency of $\hat{\tau}_M^{\text{bc}}$ as well as its double machine learning version. Section 3.5 presents simulation studies to complement the theory. Section 3.6 includes some final remarks. We relegate technical details to the appendix as well as an online appendix. Appendices 3.7 and 3.8 introduce the algorithms and theory for the NN matching-based density ratio estimator. Appendix 3.9 and the online appendix present the proofs of results in the paper and in the appendix, respectively.

Notation. For any integers $n, d \geq 1$, let $\llbracket n \rrbracket = \{1, 2, \dots, n\}$, $n!$ be the factorial of n , and $\mathbb{R}^d$ be the d -dimensional real space. A set consisting of distinct elements $x_1, \dots, x_n$ is written as either $\{x_1, \dots, x_n\}$ or $\{x_i\}_{i=1}^n$, and its cardinality is written by $|\{x_i\}_{i=1}^n|$. The corresponding sequence is denoted by $[x_1, \dots, x_n]$ or $[x_i]_{i=1}^n$. Let $\mathbb{1}(\cdot)$ denote the indicator function. For any $a, b \in \mathbb{R}$, write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use $\xrightarrow{d}$ and $\xrightarrow{p}$ to denote convergence in distribution and in probability, respectively. For any sequence of random variables $\{X_n\}$, write $X_n = o_P(1)$ if $X_n \xrightarrow{p} 0$ and $X_n = O_P(1)$ if X_n is bounded in probability. Let P_Z represent the law of a random variable Z .

3.2 Density ratio estimation via NN-matching

Consider two general random vectors X, Z in $\mathbb{R}^d$ that are defined on the same probability space, with d to be a fixed positive integer. Let ν_0 and ν_1 represent the probability measures of X and Z , respectively. Assume ν_0 and ν_1 are absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$ equipped with the Euclidean norm $\|\cdot\|$; denote the corresponding densities (Radon–Nikodym derivatives) by f_0 and f_1 . Assume further that ν_1 is absolutely continuous with respect to ν_0 and write the corresponding density ratio, f_1/f_0 , as r ; set $0/0 = 0$ by default.

Assume $X_1, \dots, X_{N_0}$ are N_0 independent copies of X , $Z_1, \dots, Z_{N_1}$ are N_1 independent copies of Z , and $[X_i]_{i=1}^{N_0}$ and $[Z_j]_{j=1}^{N_1}$ are mutually independent. The problem of estimating

the density ratio r based on $\{X_1, \dots, X_{N_0}, Z_1, \dots, Z_{N_1}\}$ is fundamental in economics (Cunningham, 2021), information theory (Cover and Thomas, 2006), machine learning (Sugiyama et al., 2012), statistics (Imbens and Rubin, 2015), and other fields.

In density ratio estimation, NN-based estimators are advocated before due to its computational efficiency; cf. Lima et al. (2008), Póczos and Schneider (2011), Kremer et al. (2015), Noshad et al. (2017), Berrett et al. (2019), Zhao and Lai (2020), among many others. Based on Abadie and Imbens (2006, 2008, 2011, 2012)'s NN matching framework, we propose a new density ratio estimator based on NN matching. To this end, some necessary notation is introduced first.

Definition 3.2.1 (NN matching). For any $x, z \in \mathbb{R}^d$ and $M \in \llbracket N_0 \rrbracket$,

- (i) let $\mathcal{X}_{(M)}(\cdot) : \mathbb{R}^d \rightarrow \{X_i\}_{i=1}^{N_0}$ be the mapping that returns the value of the input z 's M -th NN in $\{X_i\}_{i=1}^{N_0}$, i.e., the value of $x \in \{X_i\}_{i=1}^{N_0}$ such that

$$\sum_{i=1}^{N_0} \mathbb{1}(\|X_i - z\| \leq \|x - z\|) = M; \quad (3.2.1)$$

- (ii) let $K_M(\cdot) : \mathbb{R}^d \rightarrow \{0\} \cup \llbracket N_1 \rrbracket$ be the mapping that returns the number of matched times of x , i.e.,

$$K_M(x) = K_M(x; \{X_i\}_{i=1}^{N_0}, \{Z_j\}_{j=1}^{N_1}) = \sum_{j=1}^{N_1} \mathbb{1}(\|x - Z_j\| \leq \|\mathcal{X}_{(M)}(Z_j) - Z_j\|); \quad (3.2.2)$$

- (iii) let $A_M(\cdot) : \mathbb{R}^d \rightarrow \mathcal{B}(\mathbb{R}^d)$ be the corresponding mapping from $\mathbb{R}^d$ to the class of all Borel sets in $\mathbb{R}^d$ so that

$$A_M(x) = A_M(x; \{X_i\}_{i=1}^{N_0}) = \left\{ z \in \mathbb{R}^d : \|x - z\| \leq \|\mathcal{X}_{(M)}(z) - z\| \right\} \quad (3.2.3)$$

returns the catchment area of x in the setting of (ii);

Because ν_0 is absolutely continuous with respect to the Lebesgue measure, (3.2.1) has a unique solution. Abadie and Imbens (2006, Pages 240 and 260) introduced the terms $K_M(\cdot)$ and $A_M(\cdot)$ to analyze the asymptotic behavior of their NN matching-based ATE estimator. We also adopt their terminology “catchment area” in Definition 3.2.1(iii). Proposition 3.2.1 below formally links $K_M(\cdot)$ to $A_M(\cdot)$. It was established in the proof of Abadie and Imbens (2006, Lemma 3), and is stated here to aid understanding.

Proposition 3.2.1. For any $x \in \mathbb{R}^d$, we have $K_M(x) = \sum_{j=1}^{N_1} \mathbb{1}(Z_j \in A_M(x))$.

Remark 3.2.1 (Relation between $A_M(X_i)$'s and Voronoi tessellation when $M = 1$). We can verify that, due to the absolute continuity of ν_0 , $[A_1(X_i)]_{i=1}^{N_0}$ are almost surely disjoint except for a Lebesgue measure zero area, and partition $\mathbb{R}^d$ into N_0 polygons. Furthermore,

we can also verify that $\{A_1(X_i)\}_{i=1}^{N_0}$ are exactly the Voronoi tessellation defined in [Voronoi \(1908\)](#), which plays a vital role in stochastic and computational geometry. In this case, each element $A_1(X_i)$ is a Voronoi cell from the definition of [\(3.2.3\)](#).

With these notation and concepts, we are now ready to introduce the following density ratio estimator based on NN matching.

Definition 3.2.2 (NN matching-based density ratio estimator). For any $M \in \llbracket N_0 \rrbracket$ and $x \in \mathbb{R}^d$, we define the following estimator for $r(x)$:

$$\hat{r}_M(x) = \hat{r}_M\left(x; \{X_i\}_{i=1}^{N_0}, \{Z_j\}_{j=1}^{N_1}\right) = \frac{N_0}{N_1} \frac{K_M(x)}{M}. \quad (3.2.4)$$

The estimator $\hat{r}_M(\cdot)$ is by construction a one-step estimator, and satisfies the following two properties simultaneously.

- (P1) *Computationally of low complexity*: it is of a sub-quadratic (and nearly linear when M is small) time complexity via a careful algorithmic formulation based on k -d trees (see Algorithms 1–2 and Theorem 3.7.1 in Appendix 3.7), and thus in many scientific applications is computationally more attractive than its optimization-based alternatives ([Lima et al., 2008](#); [Kremer et al., 2015](#); [Borgeaud et al., 2022](#)).
- (P2) *Statistically rate-optimal*: it is information-theoretically efficient in terms of achieving an upper bound of estimation accuracy that matches the corresponding minimax lower bound over a class of Lipschitz density functions (see Appendix 3.8).

3.3 Revisiting the bias-corrected matching estimator of the ATE

This section studies the bias-corrected NN matching-based estimator of the ATE, proposed in [Abadie and Imbens \(2011\)](#) to correct the asymptotic bias of the original matching-based estimator derived by [Abadie and Imbens \(2006\)](#). To this end, we leverage the new insights in Section 3.2 as well as the technical results in Appendices 3.7–3.8, and bridge the study to both the classic double robustness and the modern double machine learning frameworks.

We first review the setup for the NN matching-based estimator and its bias-corrected version. Following [Abadie and Imbens \(2006\)](#), let $[(X_i, D_i, Y_i)]_{i=1}^n$ be n independent copies of (X, D, Y) , where $D \in \{0, 1\}$ is a binary variable, let $X \in \mathbb{R}^d$ represent the individual covariates, assumed to be absolute continuous admitting a density f_X , and let $Y \in \mathbb{R}$ stand for the outcome variable.

For each unit $i \in \llbracket n \rrbracket$, we observe $D_i = 1$ if in the treated group and $D_i = 0$ if in the control group. Let $n_0 = \sum_{i=1}^n (1 - D_i)$ and $n_1 = \sum_{i=1}^n D_i$ be the numbers of control and treated units, respectively. Under the potential outcomes framework ([Rubin, 1974](#)), the unit i has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, but we observe only one of them:

$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. The goal is to estimate the population ATE, $\tau = E[Y_i(1) - Y_i(0)]$, based on the observations $\{(X_i, D_i, Y_i)\}_{i=1}^n$. To estimate ATE, we consider its empirical counterpart $\hat{\tau}_M = n^{-1} \sum_{i=1}^n [\hat{Y}_i(1) - \hat{Y}_i(0)]$, where $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ are the imputed outcomes of $Y_i(0)$ and $Y_i(1)$. Following [Abadie and Imbens \(2006\)](#), we focus on the matching-based estimator by imputing missing potential outcomes as

$$\hat{Y}_i(0) = \begin{cases} Y_i, & \text{if } D_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j, & \text{if } D_i = 1 \end{cases} \quad \text{and} \quad \hat{Y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j, & \text{if } D_i = 0, \\ Y_i, & \text{if } D_i = 1. \end{cases}$$

Here $\mathcal{J}_M(i)$ represents the index set of M -NNs of X_i in $\{X_j : D_j = 1 - D_i\}_{j=1}^n$, i.e., the set of all indices $j \in \llbracket n \rrbracket$ such that $D_j = 1 - D_i$ and $\sum_{\ell=1, D_\ell=1-D_i}^n \mathbf{1}(\|X_\ell - X_i\| \leq \|X_j - X_i\|) \leq M$. With a slight abuse of notation, let $K_M(i)$ represent the number of matched times for unit i , i.e., $K_M(i) = \sum_{j=1, D_j=1-D_i}^n \mathbf{1}(i \in \mathcal{J}_M(j))$. We can then rewrite the above matching-based estimator as

$$\hat{\tau}_M = \frac{1}{n} \left[\sum_{i=1, D_i=1}^n \left(1 + \frac{K_M(i)}{M}\right) Y_i - \sum_{i=1, D_i=0}^n \left(1 + \frac{K_M(i)}{M}\right) Y_i \right]. \quad (3.3.1)$$

However, when $d > 1$, the bias of $\hat{\tau}_M$ is asymptotically non-negligible ([Abadie and Imbens, 2006](#)). To fix this, [Abadie and Imbens \(2011\)](#) proposed the following bias-corrected version for $\hat{\tau}_M$. In detail, let $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ be mappings from $\mathbb{R}^d$ to $\mathbb{R}$ that estimate the conditional means of the outcomes $\mu_0(x) = E[Y | X = x, D = 0]$ and $\mu_1(x) = E[Y | X = x, D = 1]$, respectively, with the corresponding residuals $\hat{R}_i = Y_i - \hat{\mu}_{D_i}(X_i)$, $i \in \llbracket n \rrbracket$. Define the estimator based on outcome regressions as $\hat{\tau}^{\text{reg}} = n^{-1} \sum_{i=1}^n [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$. Consider the bias-corrected matching-based estimator in [Abadie and Imbens \(2011\)](#):

$$\hat{\tau}_M^{\text{bc}} = \frac{1}{n} \sum_{i=1}^n [\hat{Y}_i^{\text{bc}}(1) - \hat{Y}_i^{\text{bc}}(0)], \quad (3.3.2)$$

with

$$\hat{Y}_i^{\text{bc}}(0) = \begin{cases} Y_i, & \text{if } D_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)), & \text{if } D_i = 1, \end{cases}$$

and

$$\hat{Y}_i^{\text{bc}}(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_j)), & \text{if } D_i = 0, \\ Y_i, & \text{if } D_i = 1. \end{cases}$$

Lemma 3.3.1 below shows an equivalent form of $\hat{\tau}_M^{\text{bc}}$.

Lemma 3.3.1. *The bias-corrected matching-based estimator in (3.3.2) can be rewritten in terms of $\hat{\tau}^{\text{reg}}$ and the residuals $\hat{R}_i$'s as:*

$$\hat{\tau}_M^{\text{bc}} = \hat{\tau}^{\text{reg}} + \frac{1}{n} \left[\sum_{i=1, D_i=1}^n \left(1 + \frac{K_M(i)}{M}\right) \hat{R}_i - \sum_{i=1, D_i=0}^n \left(1 + \frac{K_M(i)}{M}\right) \hat{R}_i \right]. \quad (3.3.3)$$

Otsu and Rai (2017) derived another linear form of $\hat{\tau}_M^{\text{bc}}$ to motivate a bootstrap procedure for variance estimation. The form in (3.3.3) is related to doubly robust estimators reviewed shortly. In detail, we first have some outcome models and residuals defined in the same way as above, and then let $\hat{e}(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a generic estimator of the propensity score (Rosenbaum and Rubin, 1983b), $e(x) = P(D = 1 | X = x)$. The doubly robust estimator in Scharfstein et al. (1999) and Bang and Robins (2005) could then be formulated as

$$\hat{\tau}^{\text{dr}} = \hat{\tau}^{\text{reg}} + \frac{1}{n} \left[\sum_{i=1, D_i=1}^n \frac{1}{\hat{e}(X_i)} \hat{R}_i - \sum_{i=1, D_i=0}^n \frac{1}{1 - \hat{e}(X_i)} \hat{R}_i \right]. \quad (3.3.4)$$

Conditional on $(D_1, \dots, D_n)$, $[X_i : D_i = \omega]_{i=1}^n$ are n_ω i.i.d. random variables sampled from the distribution of $X | D = \omega$, and the two groups of sample points, $[X_i : D_i = 0]_{i=1}^n$ and $[X_i : D_i = 1]_{i=1}^n$, are mutually independent. Let $f_{X|D=\omega}$ denote the density of $X | D = \omega$. From the construction of $K_M(i)$ and results in Appendix 3.8, once allowing M to diverge to infinity, conditional on $(D_1, \dots, D_n)$, $n_0/n_1 \cdot K_M(i)/M$ and $n_1/n_0 \cdot K_M(i)/M$ are consistent estimators of $f_{X|D=1}(X_i)/f_{X|D=0}(X_i)$ and $f_{X|D=0}(X_i)/f_{X|D=1}(X_i)$ for units with $D_i = 0$ and $D_i = 1$, respectively. Because n_1/n_0 converges almost surely to $P(D = 1)/P(D = 0)$ by the law of large numbers, the statistic $1 + K_M(i)/M$ is then a consistent estimator of $1/(1 - e(X_i))$ and $1/e(X_i)$ for units with $D_i = 0$ and $D_i = 1$, respectively. Thus, in view of (3.3.4), the bias-corrected matching-based estimator $\hat{\tau}_M^{\text{bc}}$ in (3.3.3) is actually a doubly robust estimator of τ , and accordingly, should also enjoy all the corresponding desirable properties. This novel insight into $\hat{\tau}_M^{\text{bc}}$ allows us to establish its asymptotic properties with a diverging M .

3.4 Asymptotic analysis with diverging M

The theory for matching with a diverging M has been an important gap in the literature. With a univariate covariate, Abadie and Imbens (2006) provided a heuristic argument about the additional efficiency gain for $\hat{\tau}_M$ with larger M . With a general covariate, Abadie and Imbens (2011) used simulation to evaluate the finite-sample properties of $\hat{\tau}_M^{\text{bc}}$ and highlighted the importance of bias correction with large M . Nevertheless, existing theoretical results for NN matching estimators all focused on fixed M (Abadie and Imbens, 2006, 2008, 2011, 2016; Kallus, 2020; Armstrong and Kolesár, 2021a; Ferman, 2021). In this section, we will present the corresponding theory with a diverging M and also make connections between $\hat{\tau}_M^{\text{bc}}$ and double robustness/double machine learning estimators.

3.4.1 The original matching-based estimator

We first analyze the original bias-corrected matching-based estimator $\hat{\tau}_M^{\text{bc}}$. Let $U_\omega = Y(\omega) - \mu_\omega(X)$ for $\omega \in \{0, 1\}$ and $\mathbb{X}$ be the support of X . Let $\|\cdot\|_\infty$ denote the L_∞ norm of a function.

We need following assumptions to prove the consistency of $\hat{\tau}_M^{\text{bc}}$.

Assumption 3.4.1. (i) For almost all $x \in \mathbb{X}$, D is independent of $(Y(0), Y(1))$ conditional on $X = x$, and there exists some constant $\eta > 0$ such that $\eta < P(D = 1 | X = x) < 1 - \eta$.
(ii) $[(X_i, D_i, Y_i)]_{i=1}^n$ are i.i.d. following the joint distribution of (X, D, Y) .
(iii) $E[U_\omega^2 | X = x]$ is uniformly bounded for almost all $x \in \mathbb{X}$ and $\omega \in \{0, 1\}$.
(iv) $E[\mu_\omega^2(X)]$ is bounded for $\omega \in \{0, 1\}$.

Assumption 3.4.2. For $\omega \in \{0, 1\}$, there exists a deterministic function $\bar{\mu}_\omega(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $E[\bar{\mu}_\omega^2(X)]$ is bounded and the estimator $\hat{\mu}_\omega(x)$ satisfies $\|\hat{\mu}_\omega - \bar{\mu}_\omega\|_\infty = o_P(1)$.

Assumption 3.4.3. For $\omega \in \{0, 1\}$, the estimator $\hat{\mu}_\omega(x)$ satisfies $\|\hat{\mu}_\omega - \mu_\omega\|_\infty = o_P(1)$.

Assumption 3.4.1(i) is the unconfoundedness and overlap assumptions, and is often referred to as the strong ignorability condition (Rosenbaum and Rubin, 1983b). Assumption 3.4.2 allows for outcome model misspecification; for example, if $\hat{\mu}_\omega = \bar{\mu}_\omega = 0$, $\hat{\tau}_M^{\text{bc}}$ then reduces to $\hat{\tau}_M$. Assumption 3.4.3 assumes that the outcome models are consistently estimated.

We need the following assumptions to prove the efficiency of $\hat{\tau}_M^{\text{bc}}$.

Assumption 3.4.4. (i) $E[U_\omega^2 | X = x]$ is uniformly bounded away from zero for almost all $x \in \mathbb{X}$ and $\omega \in \{0, 1\}$.
(ii) There exists some constant $\kappa > 0$ such that $E[|U_\omega|^{2+\kappa} | X = x]$ is uniformly bounded for almost all $x \in \mathbb{X}$ and $\omega \in \{0, 1\}$.
(iii) $\max_{t \in \Lambda_{\lfloor d/2 \rfloor + 1}} \|\partial^t \mu_\omega\|_\infty$ is bounded, where for any positive integer k , Λ_k is the set of all d -dimensional vectors of nonnegative integers $t = (t_1, \dots, t_d)$ such that $\sum_{i=1}^d t_i = k$ and $\lfloor \cdot \rfloor$ stands for the floor function.

Assumption 3.4.5. For $\omega \in \{0, 1\}$, the estimator $\hat{\mu}_\omega(x)$ satisfies

$$\max_{t \in \Lambda_{\lfloor d/2 \rfloor + 1}} \|\partial^t \hat{\mu}_\omega\|_\infty = O_P(1) \quad \text{and} \quad \max_{t \in \Lambda_\ell} \|\partial^t \hat{\mu}_\omega - \partial^t \mu_\omega\|_\infty = O_P(n^{-\gamma_\ell}) \quad \text{for all } \ell \in \llbracket \lfloor d/2 \rfloor \rrbracket,$$

with some constants γ_ℓ 's satisfying $\gamma_\ell > \frac{1}{2} - \frac{\ell}{d}$ for $\ell = 1, 2, \dots, \lfloor d/2 \rfloor$.

Assumption 3.4.4 is comparable to Assumption A.4 and the assumptions in Abadie and Imbens (2011, Theorem 2). Compared with the assumptions in Abadie and Imbens (2011, Theorem 2), Assumption 3.4.4(iii) is weaker in the sense that it only requires a finite order of smoothness. Assumption 3.4.5 again assumes the approximation accuracy of the outcome models, with lower convergence rates required for higher order derivatives of the outcome models. Under some smoothness conditions on the outcome model as made in Abadie and Imbens (2011), Assumption 3.4.5 holds using power series approximation (Abadie and Imbens, 2011, Lemma A.1). Lastly, compared with Chernozhukov et al. (2018), we need approximation accuracy concerning derivatives of the outcome model estimator, which is not required in Chernozhukov et al. (2018); see Section 3.4.2 for more discussions.

Theorem 3.4.1 below presents the double robustness and semiparametric efficiency properties of $\hat{\tau}_M^{\text{bc}}$. Recall the semiparametric efficiency lower bound for estimating ATE (see Hahn (1998)):

$$\sigma^2 = \mathbb{E} \left[\mu_1(X) - \mu_0(X) + \frac{D(Y - \mu_1(X))}{e(X)} - \frac{(1 - D)(Y - \mu_0(X))}{1 - e(X)} - \tau \right]^2,$$

and introduce an estimator for σ^2 based on NN matching:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + (2D_i - 1) \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i - \hat{\tau}_M^{\text{bc}} \right]^2.$$

Theorem 3.4.1. (i) (Double robustness of $\hat{\tau}_M^{\text{bc}}$) On one hand, if the distribution of (X, D, Y) satisfies Assumptions 3.4.1, 3.4.2, either $(P_{X|D=0}, P_{X|D=1})$ or $(P_{X|D=1}, P_{X|D=0})$ satisfies Assumption 3.8.1 in the appendix, and $M \log n/n \rightarrow 0$ and $M \rightarrow \infty$ as $n \rightarrow \infty$, then $\hat{\tau}_M^{\text{bc}} - \tau \xrightarrow{\text{P}} 0$.

On the other hand, if the distribution of (X, D, Y) satisfies Assumptions 3.4.1 and 3.4.3, then $\hat{\tau}_M^{\text{bc}} - \tau \xrightarrow{\text{P}} 0$.

(ii) (Semiparametric efficiency of $\hat{\tau}_M^{\text{bc}}$) Assume the distribution of (X, D, Y) satisfies Assumptions 3.4.1, 3.4.4, 3.4.5 and either $(P_{X|D=0}, P_{X|D=1})$ or $(P_{X|D=1}, P_{X|D=0})$ satisfies Assumption 3.8.1 in the appendix. Define

$$\gamma = \left\{ \min_{\ell \in \llbracket [d/2] \rrbracket} \left[1 - \left(\frac{1}{2} - \gamma_\ell \right) \frac{d}{\ell} \right] \right\} \wedge \left[1 - \frac{1}{2} \frac{d}{[d/2] + 1} \right],$$

recalling that γ_ℓ 's were introduced in Assumption 3.4.5. Then, if $M \rightarrow \infty$ and $M/n^\gamma \rightarrow 0$ as $n \rightarrow \infty$, we have $\sqrt{n}(\hat{\tau}_M^{\text{bc}} - \tau) \xrightarrow{\text{d}} N(0, \sigma^2)$.

If in addition Assumption 3.4.3 holds, then $\hat{\sigma}^2 \xrightarrow{\text{P}} \sigma^2$.

Remark 3.4.1. To be in line with the double robustness terminology, we can call Assumptions 3.8.1 used in Theorem 3.4.1 the “density (or propensity) model assumptions” and Assumptions 3.4.3–3.4.5 the “outcome (or regression) model assumptions”.

Remark 3.4.2. The first part of Theorem 3.4.1(i) requires $M \rightarrow \infty$ for achieving the consistency of the propensity score model. When M is fixed and the outcome model is misspecified, $\hat{\tau}_M^{\text{bc}}$ is no longer doubly robust in the sense of Theorem 3.4.1. However, it does not imply that $\hat{\tau}_M^{\text{bc}}$ is inconsistent for estimating τ . In fact, Abadie and Imbens (2006, Theorem 3) showed that $\hat{\tau}_M^{\text{bc}}$ with a fixed M is still consistent even if we choose $\hat{\mu}_w = 0$ for $w = 0, 1$. One can further show that $\hat{\tau}_M^{\text{bc}}$ with a fixed M is consistent as long as the outcome models are smooth but misspecified since the matching discrepancy then converges to zero.

Remark 3.4.3. Theorem 3.4.1 has implications for practical data analysis. We discuss two. First, it highlights the importance of allowing M to diverge in asymptotic analysis.

Nevertheless, it is a challenging problem to choose M in finite samples. We use simulation to illustrate the choice of M . Second, it gives an alternative variance estimator $\hat{\sigma}^2$ for the bias-corrected matching estimator when M diverges. [Abadie and Imbens \(2006\)](#) gave another variance estimator for fixed M . While it is challenging to compare the two variance estimators in theory, we use simulation to compare them in finite samples. See Section 3.5 for the details of simulation.

If $d = 1$ and we pick $\hat{\mu}_\omega = 0$ for $\omega \in \{0, 1\}$, then Assumption 3.4.5 automatically holds and the bias-corrected estimator $\hat{\tau}_M^{\text{bc}}$ reduces to the original estimator $\hat{\tau}_M$ studied in [Abadie and Imbens \(2006\)](#). Theorem 3.4.1(ii) then directly implies the following corollary that corresponds to [Abadie and Imbens \(2006, Corollary 1\)](#) with one key difference that M goes to infinity here.

Corollary 3.4.1 (Semiparametric efficiency of $\hat{\tau}_M$ when $d = 1$). *Assume $d = 1$, the distribution of (X, D, Y) satisfies Assumptions 3.4.1, 3.4.4, and either $(P_{X|D=0}, P_{X|D=1})$ or $(P_{X|D=1}, P_{X|D=0})$ satisfies Assumption 3.8.1 in the appendix. If $M \rightarrow \infty$ and $M/n^{\frac{1}{2}} \rightarrow 0$ as $n \rightarrow \infty$, then $\sqrt{n}(\hat{\tau}_M - \tau) \xrightarrow{d} N(0, \sigma^2)$.*

Remark 3.4.4. By picking $\hat{\mu}_\omega = 0$ for $\hat{\tau}_M$, Assumption 3.4.3 is in general no longer satisfied. Accordingly, in Corollary 3.4.1, $\hat{\sigma}^2$ may not be a consistent estimator of σ^2 without additional assumptions. However, by decomposing σ^2 into the form of Theorem 1 in [Hahn \(1998\)](#), one could still estimate σ^2 via a similar and direct way as what is outlined in Section 4 in [Abadie and Imbens \(2006\)](#). We do not pursue this track in detail here as the case of $d = 1$ without Assumption 3.4.3 is beyond the main scope of this manuscript.

3.4.2 A double machine learning version of the matching

Assumptions 3.4.4 and 3.4.5 enforce arguably strong requirements on the smoothness of the outcome model. To weaken such assumptions, [Chernozhukov et al. \(2018\)](#) introduced the idea of double machine learning. In this section, we consider the option to combine NN matching with double machine learning.

Assume n is divisible by K for simplicity. Let $[I_k]_{k=1}^K$ be a K -fold random partition of $\llbracket n \rrbracket$, with each of size equal to $n' = n/K$. For each $k \in \llbracket K \rrbracket$ and $\omega \in \{0, 1\}$, construct $\hat{\mu}_{\omega,k}(\cdot)$ using data $[(X_j, D_j, Y_j)]_{j=1, j \notin I_k}^n$, and let $K_{M,k}(i)$ be the number of matched times for unit i by adding (X_i, D_i, Y_i) into $[(X_j, D_j, Y_j)]_{j=1, j \notin I_k}^n$. Define

$$\begin{aligned} \hat{\tau}_{M,k}^{\text{bc}} = & \frac{1}{n'} \sum_{i=1, i \in I_k}^n \left[\hat{\mu}_{1,k}(X_i) - \hat{\mu}_{0,k}(X_i) \right] + \frac{1}{n'} \left[\sum_{i=1, i \in I_k, D_i=1}^n \left(1 + \frac{K_{M,k}(i)}{M} \right) \right. \\ & \left. \left(Y_i - \hat{\mu}_{1,k}(X_i) \right) - \sum_{i=1, i \in I_k, D_i=0}^n \left(1 + \frac{K_{M,k}(i)}{M} \right) \left(Y_i - \hat{\mu}_{0,k}(X_i) \right) \right] \end{aligned}$$

for $k = 1, \dots, K$, and then define $\tilde{\tau}_{M,K}^{\text{bc}} = K^{-1} \sum_{k=1}^K \tilde{\tau}_{M,k}^{\text{bc}}$. We can use the same variance estimator $\hat{\sigma}^2$ for $\tilde{\tau}_{M,K}^{\text{bc}}$.

To analyze $\tilde{\tau}_{M,K}^{\text{bc}}$ instead of $\hat{\tau}_M^{\text{bc}}$, we replace Assumptions 3.4.4 and 3.4.5 with the following two assumptions.

Assumption 3.4.6. (i) $E[U_\omega^2]$ is bounded away from zero for $\omega \in \{0, 1\}$.

(ii) There exist some constants $\kappa > 0$ such that $E[|Y|^{2+\kappa}]$ is bounded.

Assumption 3.4.7. For $\omega \in \{0, 1\}$, the estimator $\hat{\mu}_\omega(x)$ satisfies $\|\hat{\mu}_\omega - \mu_\omega\|_\infty = o_P(n^{-d/(4+2d)})$.

Remark 3.4.5. Assumption 3.4.6 corresponds to Assumption 5.1 in Chernozhukov et al. (2018), and is similar to Assumption 4 in Abadie and Imbens (2006). Assumption 3.4.7 assumes approximation accuracy of the outcome model under the L_∞ norm. Abadie and Imbens (2011) used the power series approximation (Newey, 1997) to estimate the outcome model, which under some classic nonparametric statistics assumptions automatically satisfies Assumption 3.4.7 (cf. Lemma A.1 in Abadie and Imbens (2011)). The same conclusion also applies to spline and wavelet regression estimators; cf. Chen and Christensen (2015).

Remark 3.4.6. Assumption 3.4.7 assumes an approximation rate under L_∞ norm. This is different from the L_2 norm put in Chernozhukov et al. (2018, Assumption 5.1), but can be handled with some trivial modifications to the proof of Chernozhukov et al. (2018, Theorem 5.1) since one can replace the Cauchy–Schwarz inequality by the L_1 – L_∞ Hölder’s inequality. An L_1 -norm bound on $K_M(i)/M$, to be established in Theorem 3.8.4 in the appendix, can then be applied directly.

Theorem 3.4.2. (i) (Double robustness of $\tilde{\tau}_{M,K}^{\text{bc}}$) On one hand, if the distribution of (X, D, Y) satisfies Assumptions 3.4.1, 3.4.2, either $(P_{X|D=0}, P_{X|D=1})$ or $(P_{X|D=1}, P_{X|D=0})$ satisfies Assumption 3.8.1 in the appendix, and $M \log n/n \rightarrow 0$ and $M \rightarrow \infty$ as $n \rightarrow \infty$, then $\tilde{\tau}_{M,K}^{\text{bc}} - \tau \xrightarrow{P} 0$.

On the other hand, if the distribution of (X, D, Y) satisfies Assumptions 3.4.1 and 3.4.3, then $\tilde{\tau}_{M,K}^{\text{bc}} - \tau \xrightarrow{P} 0$.

(ii) (Semiparametric efficiency of $\tilde{\tau}_{M,K}^{\text{bc}}$) Assume the distribution of (X, D, Y) satisfies Assumptions 3.4.1, 3.4.6, 3.4.7 and either $(P_{X|D=0}, P_{X|D=1})$ or $(P_{X|D=1}, P_{X|D=0})$ satisfies Assumption 3.8.3 in the appendix. Then if we pick $M = \alpha n^{\frac{2}{2+d}}$ for some constant $\alpha > 0$, then $\sqrt{n}(\tilde{\tau}_{M,K}^{\text{bc}} - \tau) \xrightarrow{d} N(0, \sigma^2)$.

In addition, we have $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$.

Remark 3.4.7. There are two parts where Theorem 3.4.2(ii) requires stronger conditions than Theorem 3.4.1(ii). First, Theorem 3.4.2(ii) requires M to grow polynomially fast with n , whereas Theorem 3.4.1(ii) only requires M to (i) diverge not so fast for controlling the difference of matching units; and (ii) diverge to infinity (no matter how slowly it is) for achieving semiparametric efficiency. The assumptions in Theorems 3.4.1(ii) and 3.4.2(ii)

both ensure semiparametric efficiency for bias-corrected matching-based estimators. Second, Theorem 3.4.1(ii) only requires Assumption 3.8.1 for the density model. This is again weaker than the Lipschitz-type conditions (Assumption 3.8.3) assumed in Theorem 3.4.2(ii) but is in line with the observations made in Abadie and Imbens (2006) and Abadie and Imbens (2011). Of note, these relaxations are possible due to adding more smoothness assumptions on the outcome model (Assumptions 3.4.4–3.4.5 versus Assumptions 3.4.6–3.4.7).

Remark 3.4.8. Technically, to use Chernozhukov et al. (2018)’s Theorem 5.1 to establish Theorem 3.4.2, we need some modifications. This is because Chernozhukov et al. (2018) considered estimating $1/e(X)$ and $1/(1 - e(X))$ via plugging in an estimate of $e(X)$, whereas $\hat{\tau}_{M,K}^{\text{bc}}$ directly uses $1 + K_M(X)/M$ to estimate $1/e(X)$ and $1/(1 - e(X))$ for units with $D = 1$ and $D = 0$, respectively. We elaborate the modifications in the proof of Theorem 3.4.2(ii).

3.5 Simulation

This section uses simulation to complement the theory. We consider bias-corrected matching estimators with either a fixed or diverging M , with the asymptotic variance estimated by either $\hat{\sigma}^2$ or the estimator introduced in Abadie and Imbens (2006, Section 4).

The first data are from the National Supported Work (LaLonde, 1986). We use the specific sample studied in Dehejia and Wahba (1999). The data contain 185 treated and 260 control units. To simulate data from this study, we follow the Monte Carlo simulation design of Athey et al. (2024), and use the same pre-treatment variables that include “age”, “education”, “black”, “hispanic”, “married”, “nodegree”, “re74”, and “re75”. By using the conditional Wasserstein Generative Adversarial Networks (WGAN), one could then create a large population of observations similar to the real data, and have access to both potential outcomes for evaluating the treatment effect. Specifically, we directly use the conditional WGAN generated data available on the repository of Athey et al. (2024). There the population size is 1,000,000. For a given sample size n , we set $n_1 = n * 185 / (185 + 260)$ and $n_0 = n * 260 / (185 + 260)$, and draw samples from the generated data separately for treated and control groups. We consider $n \in \{600, 1200, 4800, 9600\}$.

The second data are from Shadish et al. (2008), which evaluated the effects of mathematical training on mathematics test performance. We use the data from the nonrandomized arm. The data contain 79 treated and 131 control units. We use nine pre-treatment covariates including “vocabulary pretest”, “mathematics pretest”, “number of prior mathematics courses”, “caucasian”, “age”, “male”, “mother education”, “father education”, and “high school GPA”. We follow the Monte Carlo simulation design of Athey et al. (2024) to generate new data with population size 1,000,000. For a given sample size n , we set $n_1 = n * 79 / (79 + 131)$ and $n_0 = n * 131 / (79 + 131)$. Other settings are the same as those of the first data.

We consider the estimator $\hat{\tau}_M^{\text{bc}}$ with both fixed $M \in \{1, 4, 16\}$ and diverging $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ of $d = 4$ for the first data and $d = 7$ for the second data; here the diverging rate is suggested by Theorem 3.8.4. In this study, we pick $\alpha \in \{0.5, 1, 2, 5, 10\}$. Notice

that here we choose $d = 4$ for the first data since in the eight pre-treatment variables there are only four continuous variables; it is straightforward to check that the rest four binary variables won't affect the asymptotic properties established in this manuscript as well as those in [Abadie and Imbens \(2006\)](#). We choose $d = 7$ for the second data based on the same reason. For the outcome models, we consider the second order power series. The estimator's asymptotic variance is estimated using either $\hat{\sigma}$ in Theorem 3.4.1(ii) (SE) or [Abadie and Imbens \(2006\)](#)'s (AISE). We implement 2,000 repetitions and Tables 3.1 and 3.2 report the calculated root-mean-squared-error (RMSE), bias, standard deviation (SD), mean-absolute-error (MAE), and the empirical coverage rate for nominal 95% and 90% confidence intervals. Tables 3.1 and 3.2 also provide, inside the parentheses, the root- n scaled RMSE, bias, SD, and MAE divided by σ^* , with σ^* computed as the sample size-scaled standard deviation of $\hat{\tau}_M^{\text{bc}}$ with the sample size chosen to be 100,000, $\alpha = 1$, and 2,000 Monte Carlo repetitions. Here we use the value $(\sigma^*)^2$ to approximate the semiparametric efficiency lower bound if the assumptions in Theorem 3.4.1 hold.

For the first data, two observations are in-line.

1. Regardless of which n is chosen, picking $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ with α set to be 1 nearly always achieves the smallest SD, RMSE, and MAE. The simulation results thus support our recommendation to increase M for achieving better statistical performance.
2. Although consistency is established under different requirements for M , the two considered asymptotic variance estimators (SE and AISE) both yield good empirical coverage rates. The coverage rates are both very close to the nominal ones when n is large and there is not much difference between the two.

Some similar observations can be found for the second data. Notably, although picking $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ with $\alpha = 1$ is not achieving the smallest RMSE this time, its RMSE is very close to the smallest. However, for the second data, AISE yields generally better coverage rates than SE, although SE's coverage rates are improving as n increases.

To conclude, the simulation results generally support (a) increasing M with the sample size n for minimizing the RMSE; and (b) exploiting [Abadie and Imbens \(2006\)](#)'s approach to estimating the asymptotic variance of $\hat{\tau}_M^{\text{bc}}$. For choosing the M , the simulation results favor $M = \lfloor \alpha n^{2/(2+d)} \rfloor$ with α selected to be 1, while calculating the theoretically optimal α is believed to be difficult and also is beyond the scope of this manuscript.

The codes to reproduce all the simulation results presented in this section are available on <https://github.com/zhexiaolin/Estimation-based-on-nearest-neighbor-matching>.

3.6 Some final remarks

Some alternative matching estimators can also achieve double robustness or semiparametric efficiency. [Yang and Zhang \(2023\)](#) proposed to use the NN matching based on the propensity score ([Rosenbaum and Rubin, 1983b](#); [Abadie and Imbens, 2016](#)) and the prognostic

Table 3.1: Simulation results, [LaLonde \(1986\)](#), $\sigma^* = 9.55$.

n	M	RMSE	Bias	SD	MAE	95%Coverage		90%Coverage	
						SE	AISE	SE	AISE
600	$M = 1$	1.055(2.71)	-0.039(-0.10)	1.054(2.70)	0.469(1.20)	0.930	0.913	0.868	0.858
	$M = 4$	1.043(2.68)	-0.038(-0.10)	1.042(2.67)	0.442(1.13)	0.926	0.911	0.862	0.847
	$M = 16$	1.037(2.66)	-0.027(-0.07)	1.036(2.66)	0.435(1.12)	0.931	0.913	0.873	0.858
	$\alpha = 0.5$	1.043(2.68)	-0.038(-0.10)	1.042(2.67)	0.442(1.13)	0.926	0.911	0.862	0.847
	$\alpha = 1$	1.039(2.67)	-0.034(-0.09)	1.039(2.67)	0.437(1.12)	0.928	0.911	0.864	0.845
	$\alpha = 2$	1.037(2.66)	-0.027(-0.07)	1.036(2.66)	0.435(1.12)	0.931	0.913	0.873	0.858
	$\alpha = 5$	1.037(2.66)	-0.022(-0.06)	1.037(2.66)	0.434(1.11)	0.948	0.927	0.891	0.869
	$\alpha = 10$	1.037(2.66)	-0.058(-0.15)	1.036(2.66)	0.433(1.11)	0.947	0.926	0.901	0.882
1200	$M = 1$	0.341(1.24)	-0.001(-0.00)	0.341(1.24)	0.272(0.99)	0.939	0.941	0.886	0.890
	$M = 4$	0.310(1.12)	0.002(0.01)	0.310(1.12)	0.248(0.90)	0.934	0.936	0.884	0.887
	$M = 16$	0.305(1.11)	0.014(0.05)	0.305(1.11)	0.244(0.88)	0.939	0.940	0.887	0.889
	$\alpha = 0.5$	0.309(1.12)	0.003(0.01)	0.309(1.12)	0.247(0.90)	0.940	0.942	0.882	0.883
	$\alpha = 1$	0.305(1.11)	0.008(0.03)	0.305(1.11)	0.244(0.89)	0.941	0.943	0.882	0.884
	$\alpha = 2$	0.306(1.11)	0.018(0.06)	0.305(1.11)	0.244(0.89)	0.939	0.939	0.886	0.887
	$\alpha = 5$	0.307(1.11)	0.029(0.10)	0.306(1.11)	0.246(0.89)	0.950	0.950	0.898	0.895
	$\alpha = 10$	0.307(1.11)	0.020(0.07)	0.306(1.11)	0.245(0.89)	0.955	0.956	0.908	0.907
4800	$M = 1$	0.163(1.18)	-0.001(-0.01)	0.163(1.18)	0.129(0.94)	0.949	0.948	0.900	0.901
	$M = 4$	0.149(1.08)	-0.000(-0.00)	0.149(1.08)	0.118(0.86)	0.951	0.951	0.897	0.897
	$M = 16$	0.145(1.05)	0.001(0.01)	0.145(1.05)	0.116(0.84)	0.952	0.950	0.903	0.903
	$\alpha = 0.5$	0.146(1.06)	-0.000(-0.00)	0.146(1.06)	0.117(0.85)	0.949	0.948	0.899	0.899
	$\alpha = 1$	0.145(1.05)	0.001(0.01)	0.145(1.05)	0.116(0.84)	0.952	0.950	0.903	0.903
	$\alpha = 2$	0.145(1.05)	0.006(0.04)	0.144(1.05)	0.116(0.84)	0.953	0.953	0.906	0.907
	$\alpha = 5$	0.145(1.05)	0.017(0.12)	0.144(1.05)	0.116(0.84)	0.957	0.957	0.906	0.903
	$\alpha = 10$	0.147(1.06)	0.027(0.19)	0.144(1.05)	0.117(0.85)	0.958	0.958	0.909	0.910
9600	$M = 1$	0.115(1.18)	-0.003(-0.03)	0.115(1.18)	0.092(0.94)	0.951	0.952	0.897	0.896
	$M = 4$	0.106(1.08)	-0.002(-0.02)	0.105(1.08)	0.084(0.86)	0.950	0.950	0.901	0.901
	$M = 16$	0.103(1.06)	-0.001(-0.01)	0.103(1.06)	0.082(0.84)	0.948	0.948	0.902	0.902
	$\alpha = 0.5$	0.104(1.07)	-0.001(-0.01)	0.104(1.07)	0.082(0.85)	0.953	0.951	0.904	0.905
	$\alpha = 1$	0.103(1.06)	-0.001(-0.01)	0.103(1.06)	0.082(0.84)	0.950	0.950	0.904	0.903
	$\alpha = 2$	0.103(1.05)	0.001(0.02)	0.103(1.05)	0.082(0.84)	0.954	0.953	0.906	0.906
	$\alpha = 5$	0.104(1.07)	0.010(0.11)	0.103(1.06)	0.083(0.85)	0.951	0.950	0.899	0.898
	$\alpha = 10$	0.106(1.09)	0.020(0.20)	0.104(1.07)	0.084(0.87)	0.951	0.951	0.899	0.900

Table 3.2: Simulation results, [Shadish et al. \(2008\)](#), $\sigma^* = 3.85$.

n	M	RMSE	Bias	SD	MAE	95%Coverage		90%Coverage	
						SE	AISE	SE	AISE
600	$M = 1$	0.190(1.21)	0.005(0.03)	0.190(1.21)	0.152(0.97)	0.879	0.942	0.806	0.888
	$M = 4$	0.182(1.16)	0.007(0.04)	0.182(1.16)	0.144(0.92)	0.879	0.926	0.799	0.875
	$M = 16$	0.180(1.14)	0.011(0.07)	0.179(1.14)	0.143(0.91)	0.865	0.921	0.785	0.866
	$\alpha = 0.5$	0.185(1.18)	0.006(0.04)	0.185(1.18)	0.147(0.94)	0.875	0.932	0.799	0.880
	$\alpha = 1$	0.182(1.16)	0.007(0.04)	0.182(1.16)	0.144(0.92)	0.879	0.926	0.799	0.875
	$\alpha = 2$	0.180(1.15)	0.009(0.06)	0.180(1.14)	0.143(0.91)	0.870	0.922	0.798	0.866
	$\alpha = 5$	0.179(1.14)	0.012(0.07)	0.179(1.14)	0.143(0.91)	0.863	0.922	0.780	0.863
	$\alpha = 10$	0.179(1.14)	0.012(0.08)	0.179(1.14)	0.142(0.91)	0.856	0.924	0.780	0.864
1200	$M = 1$	0.130(1.17)	0.006(0.05)	0.129(1.16)	0.103(0.93)	0.905	0.948	0.840	0.893
	$M = 4$	0.123(1.11)	0.008(0.07)	0.123(1.11)	0.098(0.88)	0.898	0.934	0.822	0.879
	$M = 16$	0.122(1.09)	0.013(0.12)	0.121(0.19)	0.097(0.87)	0.892	0.931	0.818	0.878
	$\alpha = 0.5$	0.125(1.13)	0.008(0.07)	0.125(1.13)	0.100(0.90)	0.898	0.943	0.836	0.885
	$\alpha = 1$	0.123(1.11)	0.008(0.07)	0.123(1.11)	0.098(0.88)	0.898	0.934	0.822	0.879
	$\alpha = 2$	0.122(1.10)	0.011(0.10)	0.121(1.09)	0.097(0.87)	0.896	0.934	0.819	0.877
	$\alpha = 5$	0.122(1.10)	0.015(0.13)	0.121(1.09)	0.097(0.87)	0.889	0.932	0.816	0.875
	$\alpha = 10$	0.121(1.09)	0.017(0.16)	0.120(1.08)	0.097(0.87)	0.880	0.930	0.799	0.876
4800	$M = 1$	0.064(1.15)	0.006(0.11)	0.063(1.14)	0.051(0.91)	0.918	0.943	0.858	0.890
	$M = 4$	0.060(1.09)	0.007(0.12)	0.060(1.08)	0.048(0.87)	0.912	0.939	0.839	0.877
	$M = 16$	0.060(1.08)	0.009(0.17)	0.059(1.07)	0.048(0.86)	0.902	0.926	0.825	0.865
	$\alpha = 0.5$	0.061(1.09)	0.006(0.12)	0.060(1.09)	0.048(0.87)	0.918	0.941	0.844	0.876
	$\alpha = 1$	0.060(1.08)	0.007(0.13)	0.060(1.07)	0.048(0.86)	0.908	0.933	0.838	0.870
	$\alpha = 2$	0.060(1.08)	0.009(0.16)	0.059(1.07)	0.048(0.86)	0.902	0.930	0.829	0.865
	$\alpha = 5$	0.060(1.08)	0.012(0.21)	0.059(1.06)	0.048(0.86)	0.895	0.920	0.824	0.861
	$\alpha = 10$	0.060(1.09)	0.015(0.26)	0.059(1.05)	0.048(0.87)	0.891	0.916	0.819	0.858
9600	$M = 1$	0.045(1.14)	0.005(0.14)	0.044(1.13)	0.036(0.91)	0.923	0.940	0.864	0.886
	$M = 4$	0.042(1.07)	0.006(0.15)	0.042(1.06)	0.034(0.86)	0.920	0.933	0.853	0.881
	$M = 16$	0.042(1.06)	0.008(0.20)	0.041(1.04)	0.033(0.85)	0.910	0.928	0.847	0.869
	$\alpha = 0.5$	0.042(1.08)	0.006(0.15)	0.042(1.06)	0.034(0.86)	0.922	0.938	0.856	0.882
	$\alpha = 1$	0.042(1.06)	0.006(0.16)	0.041(1.05)	0.033(0.85)	0.916	0.934	0.851	0.872
	$\alpha = 2$	0.042(1.06)	0.008(0.20)	0.041(1.04)	0.033(0.85)	0.912	0.929	0.847	0.869
	$\alpha = 5$	0.042(1.07)	0.010(0.25)	0.041(1.04)	0.034(0.86)	0.902	0.922	0.842	0.862
	$\alpha = 10$	0.042(1.08)	0.012(0.31)	0.041(1.03)	0.034(0.86)	0.897	0.916	0.829	0.860

score (Hansen, 2008) simultaneously, and established the double robustness of the resulting matching estimator. They focused on fixed M , and consequently, their estimator did not achieve semiparametric efficiency. Wang and Zubizarreta (2023) proposed a matching method based on integer programming to ensure global balance of the covariates, and established the efficiency of the resulting difference-in-means estimator. They focused on fixed M , and even with fixed M , their integer programming problem was computationally challenging compared with NN matching.

There are three additional questions addressed in Abadie and Imbens (2006, 2012). First, estimation of the average treatment effect on the treated (ATT) can be incorporated in the double robustness and double machine learning framework (Theorem 3.4.2) and matching framework (Theorem 3.4.1(ii)) in a similar way. Second, asymptotic normality (with an additional asymptotic bias term) of $\hat{\tau}_M$ in general d can be established as Theorem 3.4.1(ii). Third, unbalanced designs with n_0 much larger than n_1 cannot be incorporated in the double robustness and double machine learning framework, but can be studied in the same way as Theorem 3.4.1(ii).

3.7 Density ratio estimation I: computation

Additional notation. For any two real sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, write $a_n \lesssim b_n$ (or equivalently, $b_n \gtrsim a_n$) if there exists a universal constant $C > 0$ such that $a_n/b_n \leq C$ for all sufficiently large n , and write $a_n \prec b_n$ (or equivalently, $b_n \succ a_n$) if $a_n/b_n \rightarrow 0$ as n goes to infinity. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. We write $a_n = O(b_n)$ if $|a_n| \lesssim b_n$ and $a_n = o(b_n)$ if $|a_n| \prec b_n$. Denote the closed ball in $\mathbb{R}^d$ centered at x with radius δ by $B_{x,\delta}$. In the sequel, let $c, C, C', C'', C''', \dots$ be generic positive constants whose actual values may change at different locations.

This section discusses implementation and establishes Property (P1) for the proposed estimator $\hat{r}_M(\cdot)$. To this end, we separately discuss two cases:

Case I: estimating only the values of $\hat{r}_M(\cdot)$ at the observed data points $X_1, \dots, X_{N_0}$;

Case II: estimating the values of $\hat{r}_M(\cdot)$ at both the observed data points $X_1, \dots, X_{N_0}$ and n new points $x_1, \dots, x_n \in \mathbb{R}^d$.

Case I. In many applications, we are only interested in a functional of density ratios at observed sample points, i.e., the values of $\Phi(\{r(X_i)\}_{i=1}^{N_0})$ for some given functions Φ defined on $\mathbb{R}^{N_0}$. Check, e.g., in a slightly different but symmetric form – (3.3.3) for such an example on ATE estimation. To this end, it is natural to consider the plug-in estimator $\Phi(\{\hat{r}_M(X_i)\}_{i=1}^{N_0})$, for which it suffices to compute the values of $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$.

Built on the k -d tree structure (Bentley, 1975) for tracking NNs, Algorithm 1 below outlines an easy to implement algorithm to simultaneously compute all the values of $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$. This algorithm could be regarded as a direct extension of the celebrated Friedman-Bentley-Finkel algorithm (Friedman et al., 1977) to the NN matching setting.

Algorithm 1: Density ratio estimators at sample points.

Input: $\{X_i\}_{i=1}^{N_0}$, $\{Z_j\}_{j=1}^{N_1}$, and M .

Output: $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$.

Build a k -d tree using $\{X_i\}_{i=1}^{N_0}$;

for $j = 1 : N_1$ **do**

 Search the M -NNs of Z_j in $\{X_i\}_{i=1}^{N_0}$ using the k -d tree;
 Store the indices of the M -NNs of Z_j as S_j ;

Count and store the number of occurrence in $\bigcup_{j=1}^{N_1} S_j$ for each element in $\llbracket N_0 \rrbracket$,

 which is then $\{K_M(X_i)\}_{i=1}^{N_0}$;

Obtain $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$ based on (3.2.4).

Case II. Suppose we are interested in estimating density ratios at both the observed and n new points in $\mathbb{R}^d$. A naive algorithm is then to insert each new point into observed points and perform Algorithm 1 in order. However, this algorithm is not ideal as the corresponding time complexity would be n times the complexity of Algorithm 1, which could be computationally heavy with a large number of new points.

Instead, we develop a more sophisticated implementation. Let the new points be $\{x_i\}_{i=1}^n$. Algorithm 2 computes all the values of $\{\hat{r}_M(x_i)\}_{i=1}^n$ as well as $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$. The key message delivered here is that, compared with the aforementioned naive implementation, in Algorithm 2 we only need to construct one single k -d tree; the matching elements are then categorized to two different sets, corresponding to those with regard to X_i 's and x_i 's, separately. Such an implementation is thus intuitively much more efficient.

Theorem 3.7.1 below elaborates on the computational advantage of the proposed estimator.

Theorem 3.7.1. (1) *The average time complexity of Algorithm 1 to compute all the values of $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$ is $O\left((d + N_1 M / N_0) N_0 \log N_0\right)$.*

(2) *Assume $[x_i]_{i=1}^n$ are independent and identically distributed (i.i.d.) following ν_0 and are independent of $[X_i]_{i=1}^{N_0}$. Then the average time complexity of Algorithm 2 to compute all the values of $\{\hat{r}_M(x_i)\}_{i=1}^n$ and $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$ is $O\left((d + N_1 M / N_0)(N_0 + n) \log(N_0 + n)\right)$.*

Remark 3.7.1 (Comparison to non-NN-based estimators). Assuming $N_0 \asymp N_1 \asymp N$, it is worth noting that optimization-based methods are commonly of a time complexity $O(N^2)$ if not worse (Noshad et al., 2017). They are thus less appealing in terms of handling gigantic data as was argued in, e.g., astronomy (Lima et al., 2008; Kremer et al., 2015) and big text analysis (Borgeaud et al., 2022) applications.

Remark 3.7.2 (Comparison to the two-step NN-based density ratio estimator). Regarding Case I, a direct calculation yields that the time complexity of the simple two-step NN-based method, which separately estimates f_1 and f_0 based on individual M -NN density

Algorithm 2: Density ratio estimators at both sample and new points.

Input: $\{X_i\}_{i=1}^{N_0}$, $\{Z_j\}_{j=1}^{N_1}$, M , and new points $\{x_i\}_{i=1}^n$.

Output: $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$ and $\{\hat{r}_M(x_i)\}_{i=1}^n$.

Build a k -d tree using $\{X_i\}_{i=1}^{N_0} \cup \{x_i\}_{i=1}^n$;

for $j = 1 : N_1$ **do**

Set S_j and S'_j be two empty sets;

$m \leftarrow 1$;

while $|S_j| < M$ **do**

Search the m -th NN of Z_j in $\{X_i\}_{i=1}^{N_0} \cup \{x_i\}_{i=1}^n$;

if the m -th NN of Z_j is in $\{X_i\}_{i=1}^{N_0}$ **then**

add the index into S_j ;

else

add the index into S'_j ;

$m \leftarrow m + 1$;

Store the indices sets S_j and S'_j ;

Count and store the number of occurrence in $\bigcup_{j=1}^{N_1} S_j$ for each element in $\llbracket N_0 \rrbracket$,

which is then $\{K_M(X_i)\}_{i=1}^{N_0}$. Count and store the number of occurrence in $\bigcup_{j=1}^{N_1} S'_j$

for each element in $\llbracket n \rrbracket$, which is then $\{K_M(x_i)\}_{i=1}^n$;

Obtain $\{\hat{r}_M(X_i)\}_{i=1}^{N_0}$ and $\{\hat{r}_M(x_i)\}_{i=1}^n$ based on (3.2.4).

estimators, is $O(dN_0 \log N_0 + dN_1 \log N_1 + N_0 M \log N_0 + N_0 M \log N_1)$. It is thus of the same order as Algorithm 1 when $N_1 \asymp N_0$, while computationally heavier when $N_1 \prec N_0$. Regarding Case II, the time complexity of the simple two-step NN-based method is $O(dN_0 \log N_0 + dN_1 \log N_1 + (N_0 + n)M \log N_0 + (N_0 + n)M \log N_1)$. Thus, if n is of less or equal order of N_0 , it is of the same order when $N_1 \asymp N_0$, while computationally heavier than Algorithm 2 when $N_1 \prec N_0$.

Remark 3.7.3 (Comparison to the one-step NN-based density ratio estimator in Noshad et al. (2017)). To estimate f -divergence measures, Noshad et al. (2017) constructed another one-step NN-based estimator admitting the simple form: $\hat{r}_M(x) = (N_0/N_1)(\mathcal{M}_i/(\mathcal{N}_i + 1))$, where $\mathcal{N}_i$ and $\mathcal{M}_i$ are the numbers of points in $\{X_i\}_{i=1}^{N_0}$ and $\{Z_i\}_{i=1}^{N_1}$ among the M NNs of x ; cf. Noshad et al. (2017, Equ. (20)). For Case I, its time complexity is $O(d(N_0 + N_1) \log(N_0 + N_1) + N_0 M \log(N_0 + N_1))$; while for Case II it is $O(d(N_0 + N_1) \log(N_0 + N_1) + (N_0 + n)M \log(N_0 + N_1))$. Both are at the same order as the naive NN-based one, but unlike the naive approach, this estimator is indeed one-step. However, it is still theoretically unclear if this estimator is statistically efficient; see Remark 3.8.4 ahead for more details.

3.8 Density ratio estimation II: theory

This section introduces the theory for density ratio estimation based on NN matching. To this end, before establishing detailed theoretical properties (e.g., consistency and the rate of convergence) for $\hat{r}_M(\cdot)$, we first exhibit a lemma elaborating on the (asymptotic) L^p moments of $\nu_1(A_M(x))$, the ν_1 -measure of the catchment area. This novel result did not appear in Abadie and Imbens's analysis. It is also of independent interest in stochastic and computational geometry in light of Remark 3.2.1.

Lemma 3.8.1 (Asymptotic L^p moments of catchment areas's ν_1 -measure). *Assuming $M \log N_0/N_0 \rightarrow 0$ as $N_0 \rightarrow \infty$, we have $\lim_{N_0 \rightarrow \infty} (N_0/M) \mathbb{E}[\nu_1(A_M(x))] = r(x)$ holds for ν_0 -almost all x . If we further assume $M \rightarrow \infty$, then for any positive integer p , we have $\lim_{N_0 \rightarrow \infty} (N_0/M)^p \mathbb{E}[\nu_1^p(A_M(x))] = [r(x)]^p$ holds for ν_0 -almost all x .*

Remark 3.8.1 (Relation to the measure of Voronoi cells). When $M = 1$ and $\nu_0 = \nu_1$, the measure of catchment areas reduces to the measure of Voronoi cells as pointed out in Remark 3.2.1. Interestingly, in the stochastic geometry literature, Devroye et al. (2017) studied a related problem of bounding the moments of the measure of Voronoi cells (cf. Theorem 2.1 therein). Setting $M = 1$ and $\nu_0 = \nu_1$ in the first part of Lemma 3.8.1 and recalling Remark 3.2.1, we can derive their Theorem 2.1(i). On the other hand, Devroye et al. (2017, Theorem 2.1(ii)) showed that when $\nu_0 = \nu_1$, $p = 2$, and $d \leq 3$, $(M^{-1}N_0)^2 \mathbb{E}[\nu_1^2(A_M(x))]$ converges to 1 whereas $N_0^2 \mathbb{E}[\nu_1^2(A_1(x))]$ does not; cf. Devroye et al. (2017, Section 4.2). This supports the necessity of forcing $M \rightarrow \infty$ for stabilizing the moments of $\hat{r}_M(\cdot)$.

3.8.1 Consistency

We first establish the pointwise consistency of the estimator $\hat{r}_M(x)$ for $r(x)$. This requires nearly no assumption on ν_0, ν_1 except for those made at the beginning of Section 3.2, in line with similar observations made in NN-based density estimation (Biau and Devroye, 2015, Theorem 3.1).

Theorem 3.8.1 (Pointwise consistency). *Assume $M \log N_0/N_0 \rightarrow 0$ as $N_0 \rightarrow \infty$.*

- (i) (Asymptotic unbiasedness) *For ν_0 -almost all x , we have $\lim_{N_0 \rightarrow \infty} \mathbb{E}[\hat{r}_M(x)] = r(x)$.*
- (ii) (Pointwise L_p consistency) *Let p be any positive integer and assume further that $MN_1/N_0 \rightarrow \infty$ and $M \rightarrow \infty$ as $N_0 \rightarrow \infty$. Then for ν_0 -almost all x , we have $\lim_{N_0 \rightarrow \infty} \mathbb{E}[|\hat{r}_M(x) - r(x)|^p] = 0$.*

For evaluating the global consistency of the estimator, on the other hand, it is necessary to introduce the following (global) L_p risk:

$$L_p \text{ risk} = \mathbb{E} \left[\left| \hat{r}_M(X) - r(X) \right|^p \mid X_1, \dots, X_{N_0}, Z_1, \dots, Z_{N_1} \right] = \int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right|^p f_0(x) dx,$$

where X is a copy drawn from ν_0 that is independent of the data. For the L_p risk consistency of the estimator, we impose conditions on ν_0 and ν_1 further as follows.

Define the supports of ν_0 and ν_1 as S_0 and S_1 , respectively. For any set $S \subset \mathbb{R}^d$, define the diameter of S as $\text{diam}(S) = \sup_{x,z \in S} \|x - z\|$.

- Assumption 3.8.1.** (i) ν_0, ν_1 are two probability measures on $\mathbb{R}^d$, both are absolutely continuous with respect to λ , and ν_1 is absolutely continuous with respect to ν_0 .
(ii) There exists a constant $R > 0$ such that $\text{diam}(S_0) \leq R$.
(iii) There exist two constants $f_L, f_U > 0$ such that for any $x \in S_0$ and $z \in S_1$, $f_L \leq f_0(x) \leq f_U$ and $f_1(z) \leq f_U$.
(iv) There exists a constant $a \in (0, 1)$ such that for any $\delta \in (0, \text{diam}(S_0)]$ and $z \in S_1$, $\lambda(B_{z,\delta} \cap S_0) \geq a\lambda(B_{z,\delta})$, recalling that $B_{z,\delta}$ represents the closed ball in $\mathbb{R}^d$ with center at z and radius δ .

Remark 3.8.2. Assumption 3.8.1 is standard in the literature for establishing the global consistency of density ratio estimators. The regularity conditions on the support ensure that the angle of the support is not too sharp, which trivially hold for any d -dimensional cube. These conditions were also enforced in Nguyen et al. (2010, Theorem 1), Sugiyama et al. (2008, Assumption 1), Kpotufe (2017, Definition 1), among many others.

We then establish the L_p risk consistency of the estimator via the Hardy–Littlewood maximal inequality (Stein, 2016); cf. Lemma C.3.2 in the online appendix. Of note, this inequality was used in Han et al. (2020) in a relative manner in order to study the information-theoretic limit of entropy estimation.

Theorem 3.8.2 (L_p risk consistency). *Assume the pair of ν_0, ν_1 satisfies Assumption 3.8.1. Let p be any positive integer. Assume further that $M \log N_0/N_0 \rightarrow 0$, $MN_1/N_0 \rightarrow \infty$, and $M \rightarrow \infty$ as $N_0 \rightarrow \infty$. We then have*

$$\lim_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right|^p f_0(x) dx \right] = 0.$$

As a direct corollary of Theorem 3.8.2, one can obtain the limit of any finite moment of $\nu_1(A_M(\cdot))$ with a random center. This can be regarded as a global extension to Lemma 3.8.1.

Corollary 3.8.1. *Assume the same conditions as in Theorem 3.8.2. We then have $\lim_{N_0 \rightarrow \infty} (N_0/M)^p \mathbb{E}[\nu_1^p(A_M(W))] = \mathbb{E}([r(W)]^p)$, where W follows an arbitrary distribution that is absolutely continuous with respect to ν_0 and has density bounded above and below by two positive constants. In particular, it holds when W is drawn from ν_0 .*

3.8.2 Rates of convergence

In this section, we establish the rates of convergence for $\hat{r}(x)$ under both pointwise and global measures. We first consider the pointwise mean square error (MSE) convergence rate and show that $\hat{r}_M(\cdot)$ is minimax optimal in that regard. In the sequel, we fix an $x \in \mathbb{R}^d$ and consider the following local assumption on (ν_0, ν_1) .

Assumption 3.8.2 (Local assumption). *(i) ν_0, ν_1 are two probability measures on $\mathbb{R}^d$, both are absolutely continuous with respect to λ , and ν_1 is absolutely continuous with respect to ν_0 .*
(ii) There exist two constants $f_L, f_U > 0$ such that $f_0(x) \geq f_L$ and $f_1(x) \leq f_U$.
(iii) There exists a constant $\delta > 0$ such that for any $z \in B_{x,\delta}$, $|f_0(x) - f_0(z)| \vee |f_1(x) - f_1(z)| \leq L\|x - z\|$ for some constants $L > 0$.

Define the following probability class

$$\mathcal{P}_{x,p}(f_L, f_U, L, d, \delta) = \left\{ (\nu_0, \nu_1) : \text{Assumption 3.8.2 holds} \right\}.$$

The following theorem establishes the uniform pointwise convergence rate of $\hat{r}_M(\cdot)$.

Theorem 3.8.3 (Pointwise rates of convergence). *Assume $M \log N_0/N_0 \rightarrow 0$ and $M/\log N_0 \rightarrow \infty$ as $N_0 \rightarrow \infty$. Consider a sufficiently large N_0 .*

(i) Asymptotic bias:

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)} \left| \mathbb{E}[\hat{r}_M(x)] - r(x) \right| \leq C \left(\frac{M}{N_0} \right)^{1/d},$$

where $C > 0$ is a constant only depending on f_L, f_U, L, d .

Further assume $MN_1/N_0 \rightarrow \infty$ as $N_0 \rightarrow \infty$.

(ii) Asymptotic variance:

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)} \text{Var} [\hat{r}_M(x)] \leq C' \left[\left(\frac{1}{M} \right) + \left(\frac{N_0}{MN_1} \right) \right],$$

where $C' > 0$ is a constant only depending on f_L, f_U .

(iii) Asymptotic MSE:

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)} \mathbb{E}[\hat{r}_M(x) - r(x)]^2 \leq C'' \left[\left(\frac{M}{N_0} \right)^{2/d} + \left(\frac{1}{M} \right) + \left(\frac{N_0}{MN_1} \right) \right],$$

where $C'' > 0$ is a constant only depending on f_L, f_U, L, d .

Further assume $N_1^{-\frac{d}{2+d}} \log N_0 \rightarrow 0$ as $N_0 \rightarrow \infty$.

(iv) Fix $\alpha > 0$ and take $M = \alpha \cdot \{N_0^{\frac{2}{2+d}} \vee (N_0 N_1^{-\frac{d}{2+d}})\}$. We have

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)} \mathbb{E}[\widehat{r}_M(x) - r(x)]^2 \leq C'''(N_0 \wedge N_1)^{-\frac{2}{2+d}}, \quad (3.8.1)$$

where $C''' > 0$ is a constant only depending on f_L, f_U, L, d, α .

The final rate of convergence in (3.8.1) matches the established minimax lower bound in Lipschitz density function estimation (Tsybakov, 2009, Section 2). By some simple manipulation, the argument in Tsybakov (2009, Exercise 2.8) directly extends to density ratio as the latter is a harder statistical problem (Kpotufe, 2017, Remark 3). This is formally stated in the following proposition.

Proposition 3.8.1 (Pointwise MSE minimax lower bound). *For sufficiently large N_0 and N_1 ,*

$$\inf_{\widetilde{r}} \sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)} \mathbb{E}[\widetilde{r}(x) - r(x)]^2 \geq c(N_0 \wedge N_1)^{-\frac{2}{2+d}},$$

where $c > 0$ is a constant only depending on f_L, f_U, L, d and the infimum is taken over all measurable functions.

We then move on to the global risk and study the rates of convergence in this regard. To this end, a global assumption on (ν_0, ν_1) is given below.

Assumption 3.8.3 (Global assumption). (i) ν_0, ν_1 are two probability measures on $\mathbb{R}^d$, both are absolutely continuous with respect to λ , and ν_1 is absolutely continuous with respect to ν_0 .

(ii) There exists a constant $R > 0$ such that $\text{diam}(S_0) \leq R$.

(iii) There exist two constants $f_L, f_U > 0$ such that for any $x \in S_0$ and $z \in S_1$, $f_L \leq f_0(x) \leq f_U$ and $f_1(z) \leq f_U$.

(iv) There exists a constant $a \in (0, 1)$ such that for any $\delta \in (0, \text{diam}(S_0)]$ and any $z \in S_1$, $\lambda(B_{z,\delta} \cap S_0) \geq a\lambda(B_{z,\delta})$.

(v) There exists a constant $H > 0$ such that the surface area (Hausdorff measure, Evans and Garzepy (2018, Section 3.3)) of S_1 is bounded by H .

(vi) There exists a constant $L > 0$ such that for any $x, z \in S_1$, $|f_0(x) - f_0(z)| \vee |f_1(x) - f_1(z)| \leq L\|x - z\|$.

Remark 3.8.3. Assumption 3.8.3 is standard in the literature for establishing the global risk of density ratio estimators; similar assumptions were made in Zhao and Lai (2022, Assumption 1) and Zhao and Lai (2020, Assumption 1). Note that the regularity conditions on the support automatically hold for d -dimensional cubes, and the restriction on the surface area is added to control the boundary effect on NN-based methods.

Define the following probability class

$$\mathcal{P}_g(f_L, f_U, L, d, a, H, R) = \left\{ (\nu_0, \nu_1) : \text{Assumption 3.8.3 holds} \right\}. \quad (3.8.2)$$

The next theorem establishes the uniform rate of convergence of $\hat{r}(\cdot)$ within the above probability class under the L_1 risk. This rate is further matched by a minimax lower bound derived in Theorem 1 of [Zhao and Lai \(2022\)](#) using similar arguments as in the pointwise case.

Theorem 3.8.4 (Global rates of convergence under the L_1 risk). *Assume $M \log N_0/N_0 \rightarrow 0$, $M/\log N_0 \rightarrow \infty$, $MN_1/N_0 \rightarrow \infty$ as $N_0 \rightarrow \infty$. Consider a sufficiently large N_0 .*

(i) *We have the following uniform upper bound,*

$$\begin{aligned} & \sup_{(\nu_0, \nu_1) \in \mathcal{P}_g(f_L, f_U, L, d, a, H, R)} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right| f_0(x) dx \right] \\ & \leq C \left[\left(\frac{M}{N_0} \right)^{1/d} + \left(\frac{1}{M} \right)^{1/2} + \left(\frac{N_0}{MN_1} \right)^{1/2} \right], \end{aligned}$$

where $C > 0$ is a constant only depending on f_L, f_U, a, H, L, d .

(ii) *Further assume $N_1^{-\frac{d}{2+d}} \log N_0 \rightarrow 0$ as $N_0 \rightarrow \infty$, fix $\alpha > 0$, and take $M = \alpha \cdot \{N_0^{\frac{2}{2+d}} \vee (N_0 N_1^{-\frac{d}{2+d}})\}$. We then have*

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_g(f_L, f_U, L, d, a, H, R)} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right| f_0(x) dx \right] \leq C' (N_0 \wedge N_1)^{-\frac{1}{2+d}},$$

where $C' > 0$ is a constant only depending on $f_L, f_U, a, H, L, d, \alpha$.

Proposition 3.8.2 (Global minimax lower bound under the L_1 risk). *If a is sufficiently small and H, R are sufficiently large, then for sufficiently large N_0 and N_1 ,*

$$\inf_{\hat{r}} \sup_{(\nu_0, \nu_1) \in \mathcal{P}_g(f_L, f_U, L, d, a, H, R)} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}(x) - r(x) \right| f_0(x) dx \right] \geq c (N_0 \wedge N_1)^{-\frac{1}{2+d}},$$

where $c > 0$ is a constant only depending on f_L, f_U, L, d and the infimum is taken over all measurable functions.

Remark 3.8.4 (Comparison to the one-step estimator in [Noshad et al. \(2017\)](#)). The estimator introduced in Remark 3.7.3 by [Noshad et al. \(2017\)](#) is, to our knowledge, the only alternative density ratio estimator in the literature that is able to attain both the property (P1) and being one-step. However, the arguments in [Noshad et al. \(2017, Section III\)](#) can only yield the bound $\mathbb{E}[\hat{r}'_M(x) - r(x)]^2 \lesssim (M/N_0)^{1/d} + M^{-1}$ for $(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)$. This is via Equ. (21) therein, de-poissonizing the estimator, and further assuming N_1/N_0 converges to a positive constant. The above bound is strictly looser than the bound $(M/N_0)^{2/d} + M^{-1}$ for $\hat{r}_M(\cdot)$ shown in Theorem 3.8.3. However, it seems mathematically challenging to improve their analysis and accordingly, unlike $\hat{r}_M(\cdot)$, it is still theoretically unclear if the estimator $\hat{r}'_M(x)$ is a statistically efficient density ratio estimator.

3.9 Proofs of the results in Sections 3.3 and 3.4

3.9.1 Proof of Lemma 3.3.1

Proof of Lemma 3.3.1. By simple algebra, we have

$$\begin{aligned}
 \hat{\tau}_M^{\text{bc}} &= \frac{1}{n} \sum_{i=1}^n \left[\hat{Y}_i^{\text{bc}}(1) - \hat{Y}_i^{\text{bc}}(0) \right] \\
 &= \frac{1}{n} \sum_{i=1, D_i=1}^n \left[Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)) \right] \\
 &\quad + \frac{1}{n} \sum_{i=1, D_i=0}^n \left[\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_j)) - Y_i \right] \\
 &= \frac{1}{n} \sum_{i=1, D_i=1}^n \left[\hat{R}_i + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \hat{R}_j \right] \\
 &\quad + \frac{1}{n} \sum_{i=1, D_i=0}^n \left[\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \hat{R}_j - \hat{R}_i + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right] + \frac{1}{n} \left[\sum_{i=1, D_i=1}^n \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i - \sum_{i=1, D_i=0}^n \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i \right].
 \end{aligned}$$

This completes the proof. □

3.9.2 Proof of Theorem 3.4.1

Proof of Theorem 3.4.1(i). Part I. Suppose the density function is sufficiently smooth. For any $i \in \llbracket n \rrbracket$, let $\bar{R}_i = Y_i - \bar{\mu}_{D_i}(X_i)$. From (3.3.3),

$$\begin{aligned}
 \hat{\tau}_M^{\text{bc}} &= \hat{\tau}^{\text{reg}} + \frac{1}{n} \left[\sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i - \sum_{i=1}^n (1 - D_i) \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_1(X_i) - \bar{\mu}_1(X_i) \right] - \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_0(X_i) - \bar{\mu}_0(X_i) \right] \\
 &\quad + \frac{1}{n} \left[\sum_{i=1}^n (2D_i - 1) \left(1 + \frac{K_M(i)}{M} \right) \left(\bar{\mu}_{D_i}(X_i) - \hat{\mu}_{D_i}(X_i) \right) \right] \\
 &\quad + \frac{1}{n} \left[\sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M} - \frac{1}{e(X_i)} \right) \bar{R}_i - \sum_{i=1}^n (1 - D_i) \left(1 + \frac{K_M(i)}{M} - \frac{1}{1 - e(X_i)} \right) \bar{R}_i \right] \\
 &\quad + \frac{1}{n} \left[\sum_{i=1}^n \left(1 - \frac{D_i}{e(X_i)} \right) \bar{\mu}_1(X_i) - \sum_{i=1}^n \left(1 - \frac{1 - D_i}{1 - e(X_i)} \right) \bar{\mu}_0(X_i) \right]
 \end{aligned}$$

$$+ \frac{1}{n} \left[\sum_{i=1}^n \frac{D_i}{e(X_i)} Y_i - \sum_{i=1}^n \frac{1 - D_i}{1 - e(X_i)} Y_i \right]. \quad (3.9.1)$$

For each pair of terms, we only establish the first half part under treatment, and the second half under control can be established in the same way.

For the first term in (3.9.1),

$$\left| \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_1(X_i) - \bar{\mu}_1(X_i) \right] \right| \leq \|\hat{\mu}_1 - \bar{\mu}_1\|_\infty = o_P(1). \quad (3.9.2)$$

For the second term in (3.9.1),

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M} \right) \left(\bar{\mu}_1(X_i) - \hat{\mu}_1(X_i) \right) \right| \\ & \leq \|\hat{\mu}_1 - \bar{\mu}_1\|_\infty \frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M} \right) = \|\hat{\mu}_1 - \bar{\mu}_1\|_\infty = o_P(1), \end{aligned} \quad (3.9.3)$$

where the last step is due to $\sum_{i=1}^n D_i K_M(i) = n_0 M$.

Notice that from Assumption 3.4.1 (i), $P_{X|D=0}$ and $P_{X|D=1}$ share the same support, and their densities are both bounded and bounded away from zero as long as one is. Then $(P_{X|D=0}, P_{X|D=1})$ and $(P_{X|D=1}, P_{X|D=0})$ both satisfy Assumption 3.8.1 as long as one satisfies. For the third term in (3.9.1), by Theorem 3.8.2,

$$\begin{aligned} & \left\{ E \left[\left| \frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M} - \frac{1}{e(X_i)} \right) \bar{R}_i \right| \right] \right\}^2 \leq \left\{ E \left[\left| D_i \left(1 + \frac{K_M(i)}{M} - \frac{1}{e(X_i)} \right) \bar{R}_i \right| \right] \right\}^2 \\ & \leq E \left[1 + \frac{K_M(i)}{M} - \frac{1}{e(X_i)} \right]^2 E[D_i \bar{R}_i]^2 = E \left[1 + \frac{K_M(i)}{M} - \frac{1}{e(X_i)} \right]^2 E[D_i (Y_i(1) - \bar{\mu}_1(X_i))^2] \\ & \leq E \left[1 + \frac{K_M(i)}{M} - \frac{1}{e(X_i)} \right]^2 E \left[\sigma_1^2(X_i) + (\mu_1(X_i) - \bar{\mu}_1(X_i))^2 \right] = o(1), \end{aligned} \quad (3.9.4)$$

where $\sigma_1^2(x) = E[U_1^2 | X = x]$ for $x \in \mathbb{X}$.

For the fourth term in (3.9.1), notice that

$$E \left[\frac{1}{n} \sum_{i=1}^n \left(1 - \frac{D_i}{e(X_i)} \right) \bar{\mu}_1(X_i) \mid X_1, \dots, X_n \right] = 0,$$

and

$$\begin{aligned} \text{Var} \left[\frac{1}{n} \sum_{i=1}^n \left(1 - \frac{D_i}{e(X_i)} \right) \bar{\mu}_1(X_i) \right] &= E \left[\text{Var} \left[\frac{1}{n} \sum_{i=1}^n \left(1 - \frac{D_i}{e(X_i)} \right) \bar{\mu}_1(X_i) \mid X_1, \dots, X_n \right] \right] \\ &= \frac{1}{n} E \left[\bar{\mu}_1^2(X_i) \left(\frac{1}{e(X_i)} - 1 \right) \right] = O(n^{-1}). \end{aligned}$$

Then

$$\frac{1}{n} \sum_{i=1}^n \left(1 - \frac{D_i}{e(X_i)}\right) \bar{\mu}_1(X_i) = o_P(1). \quad (3.9.5)$$

For the fifth term in (3.9.1), notice that $E[Y^2]$ are bounded and $[(X_i, D_i, Y_i)]_{i=1}^n$ are i.i.d.. Using the weak law of large numbers yields

$$\frac{1}{n} \left[\sum_{i=1}^n \frac{D_i}{e(X_i)} Y_i - \sum_{i=1}^n \frac{1 - D_i}{1 - e(X_i)} Y_i \right] \xrightarrow{P} E[Y_i(1) - Y_i(0)] = \tau. \quad (3.9.6)$$

Plugging (3.9.2), (3.9.3), (3.9.4), (3.9.5), (3.9.6) into (3.9.1) completes the proof.

Part II. Suppose the outcome model is correct. By (3.3.3),

$$\begin{aligned} \hat{\tau}_M^{\text{bc}} &= \hat{\tau}^{\text{reg}} + \frac{1}{n} \left[\sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M}\right) \hat{R}_i - \sum_{i=1}^n (1 - D_i) \left(1 + \frac{K_M(i)}{M}\right) \hat{R}_i \right] \\ &= \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(X_i) - \mu_1(X_i)] - \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_0(X_i) - \mu_0(X_i)] \\ &\quad + \frac{1}{n} \left[\sum_{i=1}^n (2D_i - 1) \left(1 + \frac{K_M(i)}{M}\right) (\mu_{D_i}(X_i) - \hat{\mu}_{D_i}(X_i)) \right] \\ &\quad + \frac{1}{n} \left[\sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M}\right) (Y_i - \mu_1(X_i)) - \sum_{i=1}^n (1 - D_i) \left(1 + \frac{K_M(i)}{M}\right) (Y_i - \mu_0(X_i)) \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n [\mu_1(X_i) - \mu_0(X_i)]. \end{aligned} \quad (3.9.7)$$

For the first term in (3.9.7),

$$\left| \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(X_i) - \mu_1(X_i)] \right| \leq \|\hat{\mu}_1 - \mu_1\|_\infty = o_P(1). \quad (3.9.8)$$

For the second term in (3.9.7),

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M}\right) (\mu_1(X_i) - \hat{\mu}_1(X_i)) \right| \\ &\leq \|\hat{\mu}_1 - \mu_1\|_\infty \frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M}\right) = \|\hat{\mu}_1 - \mu_1\|_\infty = o_P(1). \end{aligned} \quad (3.9.9)$$

For the third term in (3.9.7), noticing that $K_M(\cdot)$ is a function of $(X_1, \dots, X_n)$ and $(D_1, \dots, D_n)$, we can obtain

$$E \left[\frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M}\right) (Y_i - \mu_1(X_i)) \mid X_1, \dots, X_n, D_1, \dots, D_n \right] = 0.$$

By a martingale representation (Abadie and Imbens, 2012) and then the martingale convergence theorem (which holds for both fixed and diverging M), we obtain

$$\frac{1}{n} \sum_{i=1}^n D_i \left(1 + \frac{K_M(i)}{M}\right) (Y_i - \mu_1(X_i)) = o_P(1). \quad (3.9.10)$$

For the fourth term in (3.9.7), notice that $E[\mu_\omega^2(X)]$ is bounded for $\omega \in \{0, 1\}$. Using the weak law of large number, we obtain

$$\frac{1}{n} \sum_{i=1}^n [\mu_1(X_i) - \mu_0(X_i)] \xrightarrow{P} E[\mu_1(X_i) - \mu_0(X_i)] = \tau. \quad (3.9.11)$$

Plugging (3.9.8), (3.9.9), (3.9.10), (3.9.11) into (3.9.7) completes the proof. $\square$

Proof of Theorem 3.4.1 (ii). For $\omega \in \{0, 1\}$ and $m \in \llbracket M \rrbracket$, let $j_m(i)$ represent the index of m th-NN of X_i in $\{X_j : D_j = 1 - D_i\}_{j=1}^n$, i.e., the index $j \in \llbracket n \rrbracket$ such that $D_j = 1 - D_i$ and $\sum_{\ell=1, D_\ell=1-D_i}^n \mathbb{1}(\|X_\ell - X_i\| \leq \|X_j - X_i\|) = m$. With a little abuse of notation, let $\epsilon_i = Y_i - \mu_{D_i}(X_i)$ for any $i \in \llbracket n \rrbracket$. By the definition of $\hat{\tau}_M^{\text{bc}}$ in (3.3.3), we can verify the decomposition $\hat{\tau}_M^{\text{bc}} = \bar{\tau}(\mathbf{X}) + E_M + B_M - \hat{B}_M$, where

$$\begin{aligned} \bar{\tau}(\mathbf{X}) &= \frac{1}{n} \sum_{i=1}^n [\mu_1(X_i) - \mu_0(X_i)], \\ E_M &= \frac{1}{n} \sum_{i=1}^n (2D_i - 1) \left(1 + \frac{K_M(i)}{M}\right) \epsilon_i, \\ B_M &= \frac{1}{n} \sum_{i=1}^n (2D_i - 1) \left[\frac{1}{M} \sum_{m=1}^M \left(\mu_{1-D_i}(X_i) - \mu_{1-D_i}(X_{j_m(i)}) \right) \right], \\ \hat{B}_M &= \frac{1}{n} \sum_{i=1}^n (2D_i - 1) \left[\frac{1}{M} \sum_{m=1}^M \left(\hat{\mu}_{1-D_i}(X_i) - \hat{\mu}_{1-D_i}(X_{j_m(i)}) \right) \right]. \end{aligned}$$

We have the following central limit theorem on $\bar{\tau}(\mathbf{X}) + E_M$.

Lemma 3.9.1. $\sqrt{n}\sigma^{-1/2}(\bar{\tau}(\mathbf{X}) + E_M - \tau) \xrightarrow{d} N(0, 1)$.

For the bias term $B_M - \hat{B}_M$, define $U_{m,i} = X_{j_m(i)} - X_i$ for any $i \in \llbracket n \rrbracket$ and $m \in \llbracket M \rrbracket$. We then have the following lemma bounding the moments of $U_{M,i}$.

Lemma 3.9.2. *Let p be any positive integer. Then there exists a constant $C_p > 0$ only depending on p such that for any $i \in \llbracket n \rrbracket$ and $M \in \llbracket n_{1-D_i} \rrbracket$, $E[\|U_{M,i}\|^p | D_1, \dots, D_n] \leq C_p (M/n_{1-D_i})^{p/d}$.*

In light of the smoothness conditions on μ_ω and approximation conditions on $\hat{\mu}_\omega$ for $\omega \in \{0, 1\}$, we can establish the following lemma using Lemma 3.9.2.

Lemma 3.9.3. $\sqrt{n}(B_M - \hat{B}_M) \xrightarrow{P} 0$.

Combining Lemma 3.9.1 and Lemma 3.9.3 completes the proof. $\square$

Proof of Theorem 3.4.1(ii), consistency of $\hat{\sigma}^2$. By definition, we can verify the decomposition $\hat{\sigma}^2 - \sigma^2 = T_1 + T_2 + T_3 + T_4$, where

$$\begin{aligned} T_1 &= \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + D_i \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i - (1 - D_i) \left(1 + \frac{K_M(i)}{M} \right) \hat{R}_i - \hat{\tau}_M^{\text{bc}} \right]^2 \\ &\quad - \frac{1}{n} \sum_{i=1}^n \left[\mu_1(X_i) - \mu_0(X_i) + D_i \left(1 + \frac{K_M(i)}{M} \right) (Y_i - \mu_1(X_i)) \right. \\ &\quad \left. - (1 - D_i) \left(1 + \frac{K_M(i)}{M} \right) (Y_i - \mu_0(X_i)) - \hat{\tau}_M^{\text{bc}} \right]^2, \\ T_2 &= \frac{1}{n} \sum_{i=1}^n \left[\mu_1(X_i) - \mu_0(X_i) + D_i \left(1 + \frac{K_M(i)}{M} \right) (Y_i - \mu_1(X_i)) \right. \\ &\quad \left. - (1 - D_i) \left(1 + \frac{K_M(i)}{M} \right) (Y_i - \mu_0(X_i)) - \hat{\tau}_M^{\text{bc}} \right]^2 \\ &\quad - \frac{1}{n} \sum_{i=1}^n \left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i}{e(X_i)} (Y_i - \mu_1(X_i)) - \frac{1 - D_i}{1 - e(X_i)} (Y_i - \mu_0(X_i)) - \hat{\tau}_M^{\text{bc}} \right]^2, \\ T_3 &= \frac{1}{n} \sum_{i=1}^n \left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i}{e(X_i)} (Y_i - \mu_1(X_i)) - \frac{1 - D_i}{1 - e(X_i)} (Y_i - \mu_0(X_i)) - \hat{\tau}_M^{\text{bc}} \right]^2 \\ &\quad - \frac{1}{n} \sum_{i=1}^n \left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i}{e(X_i)} (Y_i - \mu_1(X_i)) - \frac{1 - D_i}{1 - e(X_i)} (Y_i - \mu_0(X_i)) - \tau \right]^2, \\ T_4 &= \frac{1}{n} \sum_{i=1}^n \left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i}{e(X_i)} (Y_i - \mu_1(X_i)) - \frac{1 - D_i}{1 - e(X_i)} (Y_i - \mu_0(X_i)) - \tau \right]^2 - \sigma^2. \end{aligned}$$

By Assumption 3.4.3, Assumption 3.4.1, Theorem 3.8.2, and the fact that $\hat{\tau}_M^{\text{bc}} = O_P(1)$, we have $T_1 = o_P(1)$. By Assumption 3.4.1, Theorem 3.8.2, and $\hat{\tau}_M^{\text{bc}} = O_P(1)$, we have $T_2 = o_P(1)$. By Assumption 3.4.1 and $\hat{\tau}_M^{\text{bc}} - \tau = o_P(1)$, we have $T_3 = o_P(1)$. By the fact that $[(X_i, D_i, Y_i)]_{i=1}^n$ are i.i.d., Assumption 3.4.1 and the weak law of large numbers, we have $T_4 = o_P(1)$. Combining the above four facts together then completes the proof. $\square$

3.9.3 Proof of Theorem 3.4.2

Proof of Theorem 3.4.2. For Theorem 3.4.2 (i), analysis analogous to the proof of Theorem 3.4.1(i) can be performed on $\tilde{\tau}_{M,k}^{\text{bc}}$ for any $k \in \llbracket K \rrbracket$. Then the results apply to $\tilde{\tau}_{M,K}^{\text{bc}}$ automatically since K is fixed.

For Theorem 3.4.2 (ii), from Definition 3.1 in Chernozhukov et al. (2018), $\tilde{\tau}_{M,K}^{\text{bc}}$ is the double machine learning estimator. We then follow the proof of Theorem 5.1 (recalling

Remark 3.4.8) and use the notations in Chernozhukov et al. (2018), essentially checking Assumption 3.1 and 3.2 therein. In the following the notation in Chernozhukov et al. (2018) is adopted.

For estimating the ATE, from Equ. (5.3) in Chernozhukov et al. (2018), the score (or the efficient influence function (Tsiatis, 2006, Section 3.4)) is

$$\psi(X, D, Y; \tilde{\tau}, \tilde{\zeta}) = \tilde{\mu}_1(X) - \tilde{\mu}_0(X) + \frac{D(Y - \tilde{\mu}_1(X))}{\tilde{e}(X)} - \frac{(1 - D)(Y - \tilde{\mu}_0(X))}{1 - \tilde{e}(X)} - \tilde{\tau},$$

where $\tilde{\zeta}(x) = (\tilde{\mu}_0(x), \tilde{\mu}_1(x), \tilde{\rho}_0(x), \tilde{\rho}_1(x))$ are the nuisance parameters by letting $\tilde{\rho}_0(x) = 1/(1 - \tilde{e}(x))$ and $\tilde{\rho}_1(x) = 1/\tilde{e}(x)$. Let $\rho_0(x) = 1/(1 - e(x))$ and $\rho_1(x) = 1/e(x)$. Then the true value is $\zeta(x) = (\mu_0(x), \mu_1(x), \rho_0(x), \rho_1(x))$.

We can then write the score as

$$\psi(X, D, Y; \tilde{\tau}, \tilde{\zeta}) = \tilde{\mu}_1(X) - \tilde{\mu}_0(X) + D(Y - \tilde{\mu}_1(X))\tilde{\rho}_1(X) - (1 - D)(Y - \tilde{\mu}_0(X))\tilde{\rho}_0(X) - \tilde{\tau}.$$

For any $p > 0$, let $\|f\|_p = \|f(X, D, Y)\|_p = (\int |f(\omega)|^p dP_{(X, D, Y)}(\omega))^{1/p}$. For the κ in Assumption 3.4.1, let $q = 2 + \kappa/2$, $q_1 = 2 + \kappa$ and q_2 such that $q^{-1} = q_1^{-1} + q_2^{-1}$. Let $\mathcal{T}_n$ be the set consisting of all $\tilde{\zeta}$ such that for $\omega \in \{0, 1\}$,

$$\|\tilde{\mu}_\omega - \mu_\omega\|_\infty = o(n^{-d/(4+2d)}), \quad \|\tilde{\rho}_\omega - \rho_\omega\|_1 = O(n^{-1/(d+2)}), \quad \|\tilde{\rho}_\omega - \rho_\omega\|_{q_2} = o(1).$$

Then the selection of $\mathcal{T}_n$ satisfies Assumption 3.2(a) in Chernozhukov et al. (2018) from Assumption 3.4.7, Theorem 3.8.4, and Theorem 3.8.2, respectively.

For step 1 in the proof of Theorem 5.1 in Chernozhukov et al. (2018), we verify the Neyman orthogonality. We can show that $E\psi(X, D, Y; \tau, \zeta) = 0$. For any $\tilde{\zeta} \in \mathcal{T}_n$, the Gateaux derivative in the direction $\tilde{\zeta} - \zeta$ is

$$\begin{aligned} \partial_{\tilde{\zeta}} E\psi(X, D, Y; \tau, \zeta)[\tilde{\zeta} - \zeta] &= E[\tilde{\mu}_1(X) - \mu_1(X)] - E[\tilde{\mu}_0(X) - \mu_0(X)] \\ &\quad - E[D(\tilde{\mu}_1(X) - \mu_1(X))\rho_1(X)] + E[(1 - D)(\tilde{\mu}_0(X) - \mu_0(X))\rho_0(X)] \\ &\quad + E[D(Y - \mu_1(X))(\tilde{\rho}_1(X) - \rho_1(X))] - E[(1 - D)(Y - \mu_0(X))(\tilde{\rho}_0(X) - \rho_0(X))]. \end{aligned}$$

We can check that the above quantity is zero, which completes this step.

Step 2 and step 3 therein can be directly applied.

For step 4 therein, we can establish in the same way that for $\omega \in \{0, 1\}$, $\|\mu_\omega\|_{q_1} = O(1)$ from $\|Y\|_{q_1} = O(1)$, and $\tau = O(1)$. Then from Hölder's inequality and $\|\rho_\omega\|_\infty$ is bounded for $\omega \in \{0, 1\}$, for any $\tilde{\zeta} \in \mathcal{T}_n$,

$$\begin{aligned} \|\psi(X, D, Y; \tau, \tilde{\zeta})\|_q &= \|\tilde{\mu}_1(X) - \tilde{\mu}_0(X) + (2D - 1)(Y - \tilde{\mu}_D(X))\tilde{\rho}_D(X) - \tau\|_q \\ &\leq \|\tilde{\mu}_1(X)\|_q + \|\tilde{\mu}_0(X)\|_q + \|(Y - \tilde{\mu}_1(X))\tilde{\rho}_1(X)\|_q + \|(Y - \tilde{\mu}_0(X))\tilde{\rho}_0(X)\|_q + \tau \\ &\leq \|\mu_1\|_q + \|\tilde{\mu}_1 - \mu_1\|_\infty + \|\mu_0\|_q + \|\tilde{\mu}_0 - \mu_0\|_\infty + (\|Y\|_{q_1} + \|\mu_1\|_{q_1} + \|\tilde{\mu}_1 - \mu_1\|_\infty)\|\tilde{\rho}_1\|_{q_2} \\ &\quad + (\|Y\|_{q_1} + \|\mu_0\|_{q_1} + \|\tilde{\mu}_0 - \mu_0\|_\infty)\|\tilde{\rho}_0\|_{q_2} + \tau = O(1). \end{aligned}$$

The last step is from the definition of $\mathcal{T}_n$. Then we complete this step.

For step 5 therein, by Hölder's inequality, for any $\tilde{\zeta} \in \mathcal{T}_n$,

$$\begin{aligned} & \|\psi(X, D, Y; \tau, \tilde{\zeta}) - \psi(X, D, Y; \tau, \zeta)\|_2 \\ & \leq \|\tilde{\mu}_1 - \mu_1\|_2 + \|\tilde{\mu}_0 - \mu_0\|_2 + \|D(Y - \tilde{\mu}_1(X))\tilde{\rho}_1(X) - D(Y - \mu_1(X))\rho_1(X)\|_2 \\ & \quad + \|(1 - D)(Y - \tilde{\mu}_0(X))\tilde{\rho}_0(X) - (1 - D)(Y - \mu_0(X))\rho_0(X)\|_2 \\ & \leq \|\tilde{\mu}_1 - \mu_1\|_\infty + \|\tilde{\mu}_0 - \mu_0\|_\infty + (\|Y\|_{q_1} + \|\mu_1(X)\|_{q_1})\|\tilde{\rho}_1 - \rho_1\|_{q_2} + \|\tilde{\mu}_1 - \mu_1\|_\infty\|\tilde{\rho}_1\|_2 \\ & \quad + (\|Y\|_{q_1} + \|\mu_0(X)\|_{q_1})\|\tilde{\rho}_0 - \rho_0\|_{q_2} + \|\tilde{\mu}_0 - \mu_0\|_\infty\|\tilde{\rho}_0\|_2 = o(1). \end{aligned}$$

The last step is due to the definition of $\mathcal{T}_n$.

Notice that for any $t \in (0, 1)$,

$$\partial_t^2 \mathbb{E}\psi(X, D, Y; \tau, \zeta + t(\tilde{\zeta} - \zeta)) = -2\mathbb{E}[(2D - 1)(\tilde{\mu}_D(X) - \mu_D(X))(\tilde{\rho}_D(X) - \rho_D(X))].$$

Then by the definition of $\mathcal{T}_n$, for any $\tilde{\zeta} \in \mathcal{T}_n$,

$$|\partial_t^2 \mathbb{E}\psi(X, D, Y; \tau, \zeta + t(\tilde{\zeta} - \zeta))| \leq 2 \sum_{w \in \{0,1\}} \|\tilde{\mu}_w - \mu_w\|_\infty \|\tilde{\rho}_w - \rho_w\|_1 = o(n^{-1/2}).$$

We then complete this step and thus complete the proof.

The consistency of the variance estimator can be established in the same way as Theorem 3.4.1(ii). □

Bibliography

- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.
- Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.
- Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.
- Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843.
- Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.
- Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.
- Armstrong, T., Weidner, M., and Zeleneev, A. (2025a). Robust estimation and inference in panels with interactive fixed effects. *Journal of Political Economy*, to appear.
- Armstrong, T. B., Kline, P., and Sun, L. (2025b). Adapting to misspecification. *Econometrica*, 93(6):1981–2005.
- Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.
- Armstrong, T. B. and Kolesár, M. (2020). Simple and honest confidence intervals in non-parametric regression. *Quantitative Economics*, 11(1):1–39.
- Armstrong, T. B. and Kolesár, M. (2021a). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177.
- Armstrong, T. B. and Kolesár, M. (2021b). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108.

- Armstrong, T. B., Kolesár, M., and Kwon, S. (2020). Bias-aware inference in regularized regression models. *arXiv preprint arXiv:2012.14823*.
- Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.
- Athey, S., Chetty, R., and Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.
- Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2024). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 240(2):105076.
- Bancroft, T. and Han, C.-P. (1977). Inference based on conditional specification: a note and a bibliography. *International Statistical Review/Revue Internationale de Statistique*, pages 117–127.
- Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics*, 15(2):190–204.
- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.
- Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37(1/2):1–16.
- Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.
- Berger, J. (1981). *Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results*. Purdue University. Department of Statistics.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.
- Berrett, T. B., Samworth, R. J., and Yuan, M. (2019). Efficient multivariate entropy estimation via k -nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318.
- Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):575–602.
- Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer.

- Bickel, P. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics*, pages 511–528. Elsevier.
- Bickel, P. (1984). Parametric robustness: small biases can be worthwhile. *The Annals of Statistics*, 12(3):864–879.
- Billingsley, P. (2008). *Probability and Measure*. John Wiley and Sons.
- Bogachev, V. I. and Ruas, M. A. S. (2007). *Measure Theory*, volume 1. Springer.
- Bojinov, I. and Shephard, N. (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682.
- Bojinov, I., Simchi-Levi, D., and Zhao, J. (2023). Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lepiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR.
- Borusyak, K. and Hull, P. (2024). Negative weights are no concern in design-based specifications. In *AEA Papers and Proceedings*, volume 114, pages 597–600. American Economic Association.
- Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.
- Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144.
- Brantner, C. L., Chang, T.-H., Nguyen, T. Q., Hong, H., Di Stefano, L., and Stuart, E. A. (2023). Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical Science*, 38(4):640–654.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.
- Brown, R. A. (2015). Building k -d tree in $O(kn \log n)$ time. *Journal of Computer Graphics Techniques*, 4(1):50–68.
- Cattaneo, M. D., Han, F., and Lin, Z. (2025). On rosenbaum’s rank-based matching estimator. *Biometrika*, 112(1):asae062.

- Chandrasekhar, A. G., Jackson, M. O., McCormick, T. H., and Thiyageswaran, V. (2023). General covariance-based conditions for central limit theorems with dependent triangular arrays. *arXiv preprint arXiv:2308.12506*.
- Chapin, F. S. (1947). *Experimental Designs in Sociological Research*. Harper and Brothers.
- Chen, A., Owen, A. B., and Shi, M. (2015). Data enriched linear regression. *Electronic Journal of Statistics*, 9(1):1078–1112.
- Chen, L. H. and Shao, Q.-M. (2004). Normal approximation under local dependence. *The Annals of Probability*, 32(3):1985–2028.
- Chen, S., Li, S., Zhang, B., and Ye, T. (2025). Minimax rates and adaptivity in combining experimental and observational data. *Journal of Causal Inference*, 13(1):20240024.
- Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67.
- Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā, Series A*, 35(4):417–446.
- Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 39(1):165–191.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. MIT Press.
- Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203.
- Cover, T. M. and Thomas, J. (2006). *Elements of Information Theory (2nd Edition)*. John Wiley and Sons.
- Cunningham, S. (2021). *Causal Inference: the Mixtape*. Yale University Press.
- D’Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

- Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):727–753.
- Davidson, K. W., Cheung, Y. K., Friel, C. P., and Suls, J. (2022). Introducing data sciences to n-of-1 designs, statistics, use-cases, the future, and the moniker ‘n-of-1’trial. *Harvard Data Science Review*, 4(SI3):10–1162.
- de Chaisemartin, C. and D’Haultfoeulle, X. (2020). Empirical mse minimization to estimate a scalar parameter. *arXiv preprint arXiv:2006.14667*.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.
- Devroye, L., Györfi, L., Lugosi, G., and Walk, H. (2017). On the measure of Voronoi cells. *Journal of Applied Probability*, 54(2):394–408.
- Ding, P. (2024). *A First Course in Causal Inference*. Chapman and Hall/CRC.
- Ding, P. (2025). What randomization can and cannot guarantee. *Observational studies*, 11(1):27–40.
- Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27(3):368–377.
- Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.
- Duan, Y. and Wang, K. (2023). Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039.
- Evans, L. C. and Garzepy, R. F. (2018). *Measure Theory and Fine Properties of Functions*. Routledge.
- Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.
- Ferman, B. (2021). Matching estimators with few treated and many control observations. *Journal of Econometrics*, 225(2):295–307.
- Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4):994–1000.
- Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226.

- Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717.
- Gabler, N., Duan, N., Vohra, S., and Kravitz, R. (2011). N-of-1 trials in the medical literature: a systematic review. *Medical Care*, 49(8):761–768.
- Gao, C. and Yang, S. (2023). Pretest estimation in combining probability and non-probability samples. *Electronic Journal of Statistics*, 17(1):1492–1546.
- Gao, M. and Ding, P. (2025). Causal inference in network experiments: regression-based analysis and design-based properties. *Journal of Econometrics*, 252:106119.
- Giles, J. A. and Giles, D. E. (1993). Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys*, 7(2):145–197.
- Green, E. J. and Strawderman, W. E. (1991). A james-stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006.
- Green, E. J., Strawderman, W. E., Amateis, R. L., and Reams, G. A. (2005). Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1):1–6.
- Greene, W. H. (2003). *Econometric analysis*. Pretence Hall.
- Greenwood, E. (1945). *Experimental Sociology*. Columbia University Press.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.
- Han, Y., Jiao, J., Weissman, T., and Wu, Y. (2020). Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228–3250.
- Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.
- Hawthornth, O., Chatters, R., Julious, S., Cook, A., Biggs, K., Solaiman, K., Quah, M. C., and Cheong, S. C. (2024). A methodological review of randomised n-of-1 trials. *Trials*, 25(1):263.
- Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783.
- Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298.

- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.
- Hoeffding, W. and Robbins, H. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(1):773–780.
- Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302.
- Hotelling, H. (1931). The generalization of student’s ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.
- Hwang, J. T. and Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. *The Annals of Statistics*, 10(3):868–881.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.
- Kaplan, D. M. and Liu, X. (2024). Confidence intervals for intentionally biased estimators. *Econometric Reviews*, 43(2-4):197–214.
- Kline, P. and Santos, A. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the us wage structure. *Quantitative Economics*, 4(2):231–267.
- Kolesár, M. and Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304.
- Koo, T. and Pashley, N. E. (2026). Design-based causal inference for incomplete block designs. *Biometrika*, page asag013.
- Kpotufe, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *International Conference on Artificial Intelligence and Statistics*, pages 1320–1328. PMLR.
- Kremer, J., Gieseke, F., Pedersen, K. S., and Igel, C. (2015). Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing*, 12:67–72.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620.

- Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 129–157. University of California Press.
- Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293.
- Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.
- Liang, T. and Recht, B. (2025). Randomization inference when n equals one. *Biometrika*, 112(2):asaf013.
- Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., and Schork, N. J. (2011). The n -of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine*, 8(2):161–173.
- Lima, M., Cunha, C. E., Oyaizu, H., Frieman, J., Lin, H., and Sheldon, E. S. (2008). Estimating the redshift distribution of photometric galaxy samples. *Monthly Notices of the Royal Astronomical Society*, 390(1):118–130.
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman’s critique. *The Annals of Applied Statistics*, 7(1):295–318.
- Lin, Z., Bickel, P. J., and Ding, P. (2026). Introducing the b -value: combining unbiased and biased estimators from a sensitivity analysis perspective. *arXiv preprint arXiv:2602.16310*.
- Lin, Z. and Ding, P. (2025). Unifying regression-based and design-based causal inference in time-series experiments. *arXiv preprint arXiv:2510.22864*.
- Lin, Z., Ding, P., and Han, F. (2023). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217.
- Lin, Z. and Han, F. (2023). On boosting the power of Chatterjee’s rank correlation. *Biometrika*, 110(2):283–299.
- Lin, Z. and Han, F. (2025). On regression-adjusted imputation estimators of average treatment effects. *Journal of Econometrics*, 251:106080.
- Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260.
- Malenica, I., Guo, Y., Gan, K., and Konigorski, S. (2023). Anytime-valid inference in n -of-1 trials. In *Machine Learning for Health (ML4H)*, pages 307–322. PMLR.

- Masten, M. A. and Poirier, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111.
- Menchetti, F., Cipollini, F., and Mealli, F. (2023). Combining counterfactual outcomes and arima models for policy evaluation. *The Econometrics Journal*, 26(1):1–24.
- Middleton, J. A. and Aronow, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6(1-2):39–75.
- Mirza, R. D., Punja, S., Vohra, S., and Guyatt, G. (2017). The history and development of n-of-1 trials. *Journal of the Royal Society of Medicine*, 110(8):330–340.
- Morgan, S. L. and Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods and Research*, 35(1):3–60.
- Mukerjee, R. and Dasgupta, T. (2022). Causal inference from possibly unbalanced split-plot designs. *Statistica Sinica*, 32:591–612.
- Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.
- Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472.
- Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.
- Ni, T., Bojinov, I., and Zhao, J. (2023). Design of panel experiments with spatial and temporal interference. *Available at SSRN 4466598*.
- Nikles, J. and Mitchell, G. (2015). *The Essential Guide to N-of-1 Trials in Health*. Springer.
- Noack, C. and Rothe, C. (2024). Bias-aware inference in fuzzy regression discontinuity designs. *Econometrica*, 92(3):687–711.
- Noshad, M., Moon, K. R., Sekeh, S. Y., and Hero, A. O. (2017). Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 903–907.

- Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732.
- Póczos, B. and Schneider, J. (2011). On the estimation of alpha-divergences. In *International Conference on Artificial Intelligence and Statistics*, PMLR, pages 609–617.
- Rambachan, A. and Shephard, N. (2019). When do common time series estimands have nonparametric causal meaning? *arXiv preprint arXiv:1903.01637*.
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.
- Romanovsky, V. (1923). Note on the moments of a binomial $(p + q)^n$ about its mean. *Biometrika*, 15(3/4):410–412.
- Rosenbaum, P. and Rubin, D. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 45(2):212–218.
- Rosenbaum, P. R. (2002). *Observational Studies*. Springer.
- Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika*, 91(1):153–164.
- Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.
- Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, 42(1):43–47.
- Rosenman, E. T., Basse, G., Owen, A. B., and Baiocchi, M. (2023a). Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973.
- Rosenman, E. T., Dominici, F., and Miratrix, L. (2023b). Empirical bayes double shrinkage for combining biased and unbiased causal estimates. *arXiv preprint arXiv:2309.06727*.
- Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.
- Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

- Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.
- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110.
- Schmid, C. and Yang, J. (2022). Bayesian models for n-of-1 trials. *Harvard Data Science Review*, 2022(SI3):10–1162.
- Schwartz, D., Saha, R., Ventz, S., and Trippa, L. (2026). Harmonized estimation of subgroup-specific treatment effects in randomized trials: The use of external control data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1):143–170.
- Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, 42(7):1–52.
- Senn, S. (2017). Sample size considerations for n-of-1 trials. *Statistical Methods in Medical Research*, 28(2):372–383.
- Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344.
- Shi, H., Drton, M., and Han, F. (2022). On the power of Chatterjee’s rank correlation. *Biometrika*, 109(2):317–333.
- Shi, H., Drton, M., and Han, F. (2024). On azadkia–chatterjee’s conditional dependence coefficient. *Bernoulli*, 30(2):851–877.
- Shi, L. and Li, X. (2024). Some theoretical foundations for the design and analysis of randomized experiments. *Journal of Causal Inference*, 12(1):20230067.
- Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.
- Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27(1):325–353.
- Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 24(2):265–285.

- Stein, E. M. (2016). *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press.
- Su, F. and Ding, P. (2021). Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):994–1015.
- Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.
- Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Büna, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746.
- Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer.
- Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274.
- Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les paralléloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1908(134):198–287.
- Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205.
- Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.
- Wallace, T. D. (1977). Pretest estimation in regression: A survey. *American Journal of Agricultural Economics*, 59(3):431–443.
- Wang, Y. and Zubizarreta, J. R. (2023). Large sample properties of matching for balance. *Statistica Sinica*, 33(3):1789–1808.
- Yang, S. and Zhang, Y. (2023). Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*, 50(1):235–265.
- Yang, X., Lin, L., Athey, S., Jordan, M. I., and Imbens, G. W. (2025). Cross-validated causal inference: a modern method to combine experimental and observational data. *arXiv preprint arXiv:2511.00727*.

- Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statistics*, 14(4):478–486.
- Zhang, Y. and Chen, Y.-C. (2025). Doubly robust inference on causal derivative effects for continuous treatments. *arXiv preprint arXiv:2501.06969*.
- Zhao, A. and Ding, P. (2022). Reconciling design-based and model-based causal inferences for split-plot experiments. *The Annals of Statistics*, 50(2):1170–1192.
- Zhao, A. and Ding, P. (2023). Covariate adjustment in multiarmed, possibly factorial experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):1–23.
- Zhao, P. and Lai, L. (2020). Minimax optimal estimation of KL divergence for continuous distributions. *IEEE Transactions on Information Theory*, 66(12):7787–7811.
- Zhao, P. and Lai, L. (2022). Analysis of KNN density estimation. *IEEE Transactions on Information Theory*, 68(12):7971–7995.
- Zucker, D. R., Ruthazer, R., and Schmid, C. H. (2010). Individual (n-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of Clinical Epidemiology*, 63(12):1312–1323.

Appendix A

Appendix for Chapter 1

Appendix [A.1](#) contains several additional discussions that complement the main text. Appendix [A.1.1](#) discusses how to compute the b-value efficiently. Appendix [A.1.2](#) extends our framework to one-sided confidence bounds for τ . Appendix [A.1.3](#) discusses the advantages of using prespecified point estimators instead of bias-dependent point estimators. Appendix [A.1.4](#) provides further details on the generalization to the dependent case, complementing the discussion in Section [1.3.5](#). Appendix [A.1.5](#) discusses further details on the relationship between the generic soft-thresholding estimator and the estimator in [Berger \(1981\)](#) and [Bickel \(1984\)](#), complementing the discussion in Section [1.4.2](#).

Appendix [A.2](#) contains the proofs of the results in the main paper.

Appendix [A.3](#) contains the proofs of the results in the appendix.

A.1 More discussions

A.1.1 Computing the b-value efficiently

In this section, we discuss how to compute the b-value efficiently.

Consider a general combined estimator $\hat{\tau}$ with symmetric fixed-length centered confidence interval $[\hat{\tau} - c(b, \zeta, \sigma_0^2, \sigma_1^2), \hat{\tau} + c(b, \zeta, \sigma_0^2, \sigma_1^2)]$. By Definition [1.2.2](#), the b-value $b^*(\zeta) = b^*(\zeta, \hat{\tau}, \sigma_0^2, \sigma_1^2) \geq 0$ is the smallest bias level at which the null value 0 is contained in the confidence interval. Equivalently, it is the solution to the equation of b :

$$\begin{aligned} b^*(\zeta) &= \inf \{b \geq 0 : 0 \in [\hat{\tau} - c(b, \zeta, \sigma_0^2, \sigma_1^2), \hat{\tau} + c(b, \zeta, \sigma_0^2, \sigma_1^2)]\} \\ &= \inf \{b \geq 0 : c(b, \zeta, \sigma_0^2, \sigma_1^2) \geq |\hat{\tau}|\}. \end{aligned}$$

Therefore, the b-value $b^*(\zeta)$ is 0 if $c(0, \zeta, \sigma_0^2, \sigma_1^2) > |\hat{\tau}|$, or is ∞ if $c(\infty, \zeta, \sigma_0^2, \sigma_1^2) < |\hat{\tau}|$. Otherwise, if we further assume $c(b, \zeta, \sigma_0^2, \sigma_1^2)$ is strictly increasing in b for given $\zeta, \sigma_0^2, \sigma_1^2$, then the b-value $b^*(\zeta)$ is the unique solution to the equation of b : $c(b, \zeta, \sigma_0^2, \sigma_1^2) = |\hat{\tau}|$. In this case, we can compute $b^*(\zeta)$ efficiently using standard one-dimensional root-finding methods such as the bisection algorithm. In many settings, the function $c(b, \zeta, \sigma_0^2, \sigma_1^2)$ does not admit

a closed-form expression. Instead, we solve an equation of L : $g(L; b, \zeta, \sigma_0^2, \sigma_1^2) = 0$ to obtain $c(b, \zeta, \sigma_0^2, \sigma_1^2)$ for some function g depending on $b, \zeta, \sigma_0^2, \sigma_1^2$. In such cases, we can compute the b-value by solving $g(|\hat{\tau}|; b, \zeta, \sigma_0^2, \sigma_1^2) = 0$ as an equation of b . This reduces the computation of the b-value to a single one-dimensional root-finding problem.

The above discussion provides a general method for computing the b-value for any combined estimator with a symmetric fixed-length centered confidence interval. Now we apply this method to compute the b-value for the precision-weighted estimator $\hat{\tau}_{PW}$, the pretest estimator $\hat{\tau}_{PT}$, and the soft-thresholding estimator $\hat{\tau}_{ST}$.

First, the following theorem provides the b-value $b_{PW}^*(\zeta)$ for the precision-weighted estimator $\hat{\tau}_{PW}$ with confidence interval $[\hat{\tau}_{PW} - \hat{L}_{PW}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{PW} + \hat{L}_{PW}(1 + \gamma)^{-1/2}\sigma_0]$.

Theorem A.1.1. *The b-value $b_{PW}^*(\zeta)$ is 0 if*

$$\Phi\left(\frac{|\hat{\tau}_{PW}|}{(1 + \gamma)^{-1/2}\sigma_0}\right) - \Phi\left(-\frac{|\hat{\tau}_{PW}|}{(1 + \gamma)^{-1/2}\sigma_0}\right) < 1 - \zeta.$$

Otherwise, the b-value $b_{PW}^(\zeta)$ can be equivalently written as the solution to the equation of b :*

$$\Phi\left(\frac{|\hat{\tau}_{PW}|}{(1 + \gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1 + \gamma}}b\right) - \Phi\left(-\frac{|\hat{\tau}_{PW}|}{(1 + \gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1 + \gamma}}b\right) = 1 - \zeta.$$

Second, the following theorem provides the b-value $b_{PT}^*(\zeta)$ for the pretest estimator $\hat{\tau}_{PT}$ with confidence interval $[\hat{\tau}_{PT} - \hat{L}_{PT}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1 + \gamma)^{-1/2}\sigma_0]$.

Theorem A.1.2. *The b-value $b_{PT}^*(\zeta)$ is 0 if*

$$P_{\Delta/\sigma_0=0}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| | \hat{\tau}_{PT}) < 1 - \zeta,$$

and is ∞ if

$$\min_{t \geq 0} P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| | \hat{\tau}_{PT}) > 1 - \zeta.$$

Otherwise, the b-value $b_{PT}^(\zeta)$ can be equivalently written as the solution to the equation of b :*

$$\min_{0 \leq t \leq b} P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| | \hat{\tau}_{PT}) = 1 - \zeta,$$

where $\tilde{\tau}_{PT}$ is independent and identically distributed as $\hat{\tau}_{PT}$, with

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| | \hat{\tau}_{PT}) \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1 + \gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1 + \gamma}}t\right) \right] \\ & \quad \left[\Phi\left(\frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1 + \gamma}}t\right) - \Phi\left(-\frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1 + \gamma}}t\right) \right] \\ & \quad + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1 + \gamma}}t} \left[\Phi\left(\frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}u\right) - \Phi\left(-\frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}u\right) \right] \phi(u) du \\ & \quad + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1 + \gamma}}t}^{\infty} \left[\Phi\left(\frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}u\right) - \Phi\left(-\frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}u\right) \right] \phi(u) du. \end{aligned}$$

Third, the following theorem provides the b-value $b_{\text{ST}}^*(\zeta)$ for the soft-thresholding estimator $\widehat{\tau}_{\text{ST}}$ with confidence interval $[\widehat{\tau}_{\text{ST}} - \widehat{L}_{\text{ST}}(1 + \gamma)^{-1/2}\sigma_0, \widehat{\tau}_{\text{ST}} + \widehat{L}_{\text{ST}}(1 + \gamma)^{-1/2}\sigma_0]$.

Theorem A.1.3. *The b-value $b_{\text{ST}}^*(\zeta)$ is 0 if*

$$P_{\Delta/\sigma_0=0}(|\widetilde{\tau}_{\text{ST}} - \tau| \leq |\widehat{\tau}_{\text{ST}}| \mid \widehat{\tau}_{\text{ST}}) < 1 - \zeta,$$

and is ∞ if

$$P_{\Delta/\sigma_0=\infty}(|\widetilde{\tau}_{\text{ST}} - \tau| \leq |\widehat{\tau}_{\text{ST}}| \mid \widehat{\tau}_{\text{ST}}) > 1 - \zeta.$$

Otherwise, the b-value $b_{\text{ST}}^(\zeta)$ can be equivalently written as the solution to the equation of b:*

$$P_{\Delta/\sigma_0=b}(|\widetilde{\tau}_{\text{ST}} - \tau| \leq |\widehat{\tau}_{\text{ST}}| \mid \widehat{\tau}_{\text{ST}}) = 1 - \zeta,$$

where $\widetilde{\tau}_{\text{ST}}$ is independent and identically distributed as $\widehat{\tau}_{\text{ST}}$, with

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(|\widetilde{\tau}_{\text{ST}} - \tau| \leq |\widehat{\tau}_{\text{ST}}| \mid \widehat{\tau}_{\text{ST}}) \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) \right] \\ & \quad \left[\Phi\left(\frac{|\widehat{\tau}_{\text{ST}}|}{(1+\gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1+\gamma}}t\right) - \Phi\left(-\frac{|\widehat{\tau}_{\text{ST}}|}{(1+\gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \right] \\ & + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \left[\Phi\left(\frac{|\widehat{\tau}_{\text{ST}}|}{(1+\gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}(u + c_{\alpha/2})\right) - \Phi\left(-\frac{|\widehat{\tau}_{\text{ST}}|}{(1+\gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}(u + c_{\alpha/2})\right) \right] \phi(u) du \\ & + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \left[\Phi\left(\frac{|\widehat{\tau}_{\text{ST}}|}{(1+\gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}(u - c_{\alpha/2})\right) - \Phi\left(-\frac{|\widehat{\tau}_{\text{ST}}|}{(1+\gamma)^{-1/2}\sigma_0} + \sqrt{\gamma}(u - c_{\alpha/2})\right) \right] \phi(u) du. \end{aligned}$$

A.1.2 One-sided confidence bounds

In this section, we discuss how to construct one-sided confidence bounds within our framework. We only focus on the lower confidence bound since the upper confidence bound can be constructed in an analogous manner. By using the lower confidence bound, we can conduct one-sided hypothesis testing, such as testing $\tau = 0$ versus $\tau > 0$ or testing $\tau \leq 0$ versus $\tau > 0$. The computation of the b-value in the one-sided setting follows the same logic as in the two-sided case, so here we focus on the construction of the one-sided confidence bounds.

First, the following theorem provides the sequence of lower confidence bounds based on $\widehat{\tau}_{\text{PW}}$, analogous to Theorem 1.3.1.

Theorem A.1.4. *Let $\widehat{L}'_{\text{PW}} = \widehat{L}'_{\text{PW}}(b, \zeta, \gamma) = c_{\zeta} + \frac{\gamma}{\sqrt{1+\gamma}}b$. The shortest length lower confidence bound based on $\widehat{\tau}_{\text{PW}}$ for τ satisfying (1.2.1) is given by $[\widehat{\tau}_{\text{PW}} - \widehat{L}'_{\text{PW}}(1 + \gamma)^{-1/2}\sigma_0, \infty)$.*

Second, the following theorem provides the sequence of lower confidence bounds based on $\hat{\tau}_{PT}$, analogous to Theorem 1.3.2.

Theorem A.1.5. $\hat{L}'_{PT} = \hat{L}'_{PT}(b, \zeta, \gamma, \alpha)$ is the solution to the following equation of L :

$$\min_{0 \leq t \leq b} P_{\Delta/\sigma_0=t}(\hat{\tau}_{PT} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta,$$

where

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(\hat{\tau}_{PT} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0) \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) \right] \Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \\ &+ \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \Phi\left(L + \sqrt{\gamma}u\right) \phi(u) du + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \Phi\left(L + \sqrt{\gamma}u\right) \phi(u) du. \end{aligned}$$

The shortest length lower confidence bound based on $\hat{\tau}_{PT}$ for τ satisfying (1.2.1) is given by $[\hat{\tau}_{PT} - \hat{L}'_{PT}(1 + \gamma)^{-1/2}\sigma_0, \infty)$.

Third, the following theorem provides the sequence of lower confidence bounds based on $\hat{\tau}_{ST}$, analogous to Theorem 1.3.3.

Theorem A.1.6. For any $L > 0$, $P_{\Delta}(\hat{\tau}_{ST} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0)$ is monotonically decreasing in Δ . Then $\hat{L}'_{ST} = \hat{L}'_{ST}(b, \zeta, \gamma, \alpha)$ is the solution to the equation of L :

$$P_{\Delta/\sigma_0=b}(\hat{\tau}_{ST} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta,$$

where

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(\hat{\tau}_{ST} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0) \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) \right] \Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \\ &+ \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \Phi\left(L + \sqrt{\gamma}(u + c_{\alpha/2})\right) \phi(u) du + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \Phi\left(L + \sqrt{\gamma}(u - c_{\alpha/2})\right) \phi(u) du. \end{aligned}$$

The shortest length lower confidence bound based on $\hat{\tau}_{ST}$ for τ satisfying (1.2.1) is given by $[\hat{\tau}_{ST} - \hat{L}'_{ST}(1 + \gamma)^{-1/2}\sigma_0, \infty)$.

A.1.3 Bias-dependent point estimator

A natural question arises: since the combined estimator $\hat{\tau}$ itself does not depend on the unknown true bias Δ , could we instead construct confidence intervals using a bias-dependent point estimator, i.e., one that explicitly depends on Δ ? More generally, suppose we have

an estimator whose expectation is $\tau + g(\Delta)$, where $g(\cdot)$ is a known function of the bias. There are two natural approaches to subtract the bias from the point estimator. The first approach is to estimate the bias and subtract it from the point estimator. Specifically, we can construct an adaptive estimator of the form $\hat{\tau} - \widehat{g(\Delta)}$, where $\widehat{g(\Delta)}$ is an estimate of $g(\Delta)$. The second approach is to construct a point estimator $\hat{\tau} - g(\Delta)$ and its corresponding confidence interval for each possible value of Δ within the bias bound, and then take the union of all such intervals. However, the union of these confidence intervals is conservative since the biased estimator cannot simultaneously exhibit multiple bias values within the bound. See also [Armstrong and Kolesár \(2020\)](#) for a related argument: they show that using critical values to account for potential bias is more efficient than subtracting an estimate of the bias from the point estimator.

We now illustrate why constructing confidence intervals using a prespecified point estimator is more efficient than using a bias-dependent one.

For the first approach, a natural estimate of the bias is $\hat{\Delta} = \hat{\tau}_1 - \hat{\tau}_0$. We take the precision-weighted estimator as an example. Subtracting this estimated bias from the precision-weighted estimator gives $\hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \frac{\gamma}{1+\gamma}\hat{\Delta} = \hat{\tau}_0$. Thus, after bias correction, the estimator collapses to the unbiased estimator $\hat{\tau}_0$. In other words, this approach discards the more precise estimator. The issue is that, when constructing the confidence interval, we need to account for the randomness in $\hat{\Delta}$, which makes the confidence interval wider. As a result, this approach provides no efficiency gain over the unbiased estimator.

For the second approach, suppose we know the true bias Δ . Then the bias-corrected combined estimator is $\hat{\tau}_\Delta = \hat{\tau} - g(\Delta)$. Suppose the shortest length symmetric centered confidence interval based on $\hat{\tau}_\Delta$ for τ is given by $[\hat{\tau}_\Delta - \hat{L}_\zeta, \hat{\tau}_\Delta + \hat{L}_\zeta]$, where $\hat{L}_\zeta$ does not depend on Δ . When we only have $|\Delta| \leq b$, we take the union of these confidence intervals over all possible Δ with $|\Delta| \leq b$. The resulting confidence interval is

$$\bigcup_{\Delta: |\Delta| \leq b} [\hat{\tau}_\Delta - \hat{L}_\zeta, \hat{\tau}_\Delta + \hat{L}_\zeta] = \left[\hat{\tau} - \left(\max_{\Delta: |\Delta| \leq b} g(\Delta) + \hat{L}_\zeta \right), \hat{\tau} + \left(\max_{\Delta: |\Delta| \leq b} g(\Delta) + \hat{L}_\zeta \right) \right]. \quad (\text{A.1.1})$$

By Definition 1.2.1, the shortest length symmetric centered confidence interval for τ based on $\hat{\tau}$ is given by $[\hat{\tau} - \hat{L}_\zeta(b), \hat{\tau} + \hat{L}_\zeta(b)]$, where

$$\hat{L}_\zeta(b) = \operatorname{argmin}_{L \geq 0} \left\{ \inf_{\Delta: |\Delta| \leq b} P_\Delta(|\hat{\tau} - \tau| \leq L) \geq 1 - \zeta \right\}. \quad (\text{A.1.2})$$

Comparing the confidence interval based on bias-dependent point estimator (A.1.1) with the confidence interval based on prespecified point estimator (A.1.2), we note that

$$\left(\max_{\Delta: |\Delta| \leq b} g(\Delta) + \hat{L}_\zeta \right) \in \left\{ \inf_{\Delta: |\Delta| \leq b} P_\Delta(|\hat{\tau} - \tau| \leq L) \geq 1 - \zeta \right\}.$$

Indeed, for any Δ' with $|\Delta'| \leq b$,

$$P_{\Delta'} \left(|\hat{\tau} - \tau| \leq \max_{\Delta: |\Delta| \leq b} g(\Delta) + \hat{L}_\zeta \right) \geq P_{\Delta'} \left(|\hat{\tau} - \tau| \leq g(\Delta') + \hat{L}_\zeta \right) \geq 1 - \zeta.$$

Therefore

$$\widehat{L}_\zeta(b) \leq \max_{\Delta: |\Delta| \leq b} g(\Delta) + \widehat{L}_\zeta,$$

with equality only when $b = 0$. Hence, the confidence interval based on prespecified point estimator is strictly narrower when the bias bound is nonzero. This comparison highlights a key principle of our framework: instead of accounting for the bias in the point estimation through subtracting the bias, it is more efficient to account for the bias in the confidence interval construction through critical values.

A.1.4 More details for the dependence case

In this section, we provide more details on the generalization to the dependence case in Section 1.3.5 of the main paper. In the reparametrization in (1.3.5), one may be concerned about how scaling the bias works and how to interpret the relative bias. We compute the confidence intervals and the b-value by the following three steps.

- (a) Reparametrization: We apply the transformation in (1.3.5) to obtain the transformed biased estimator $\widehat{\tau}'_1$. Under this transformation, $\widehat{\tau}_0$ remains an unbiased estimator of τ , and $\widehat{\tau}'_1$ becomes a biased estimator that is independent of $\widehat{\tau}_0$. We then construct the combined estimator $\widehat{\tau}$ using these two independent components.
- (b) Compute the confidence intervals and the b-value in the transformed problem: We next work in the transformed problem. Let the relative bias in the transformed parametrization be Δ'/σ_0 with bound b' . Using Definition 1.2.1, we construct the confidence interval $\widehat{\tau} - \mathcal{I}(b', \zeta)$ such that

$$\inf_{\Delta': |\Delta'/\sigma_0| \leq b'} P_{\Delta'}(\tau \in \widehat{\tau} - \mathcal{I}(b', \zeta)) \geq 1 - \zeta,$$

and compute the corresponding b-value $b^{*'} in the transformed problem.$

- (c) Transform the confidence intervals and the b-value back to the original problem: Finally, we transform the results back to the original problem. Under the transformation in (1.3.5), the relative biases in the two parametrizations are related by

$$|\Delta'/\sigma_0| \leq b' \iff |\Delta/\sigma_0| \leq |1 - \rho\sigma_1/\sigma_0|b'.$$

Thus, a bias bound b' in the transformed parametrization corresponds to a bias bound $b = |1 - \rho\sigma_1/\sigma_0|b'$ in the original parametrization. Using this relationship, the confidence interval $\widehat{\tau} - \mathcal{I}(b, \zeta)$ satisfies Definition 1.2.1 in the original parametrization:

$$\inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(\tau \in \widehat{\tau} - \mathcal{I}(b, \zeta)) \geq 1 - \zeta,$$

and the b-value in the original parametrization is $b^* = |1 - \rho\sigma_1/\sigma_0|b^{*'}$.

A.1.5 Relationship between the generic soft-thresholding estimator and the estimator in Berger (1981); Bickel (1984)

In this section, we discuss the relationship between the generic soft-thresholding estimator and the estimator in Berger (1981); Bickel (1984) in Section 1.4.2 of the main paper. We consider the special setting studied in Berger (1981) and Bickel (1984), where the covariance matrices are proportional and isotropic. Specifically, assume $\Sigma_0 = \sigma_0^2 \mathbf{I}_d$ and $\Sigma_1 = \gamma \Sigma_0$ for some $\gamma > 0$. The γ here generalizes the variance ratio σ_0^2/σ_1^2 in the univariate setting. Let $C \geq 0$ be a user-chosen constant and let $\rho_C(r)$ denote the ratio of Bessel functions, whose explicit form is given in Lemma 3 of Berger (1981). The constant C here plays the same role as the critical value $c_{\alpha/2}$ in the univariate setting. The estimator in Berger (1981) and Bickel (1984) is then given by choosing the shrinkage function $h_q^*(r) = \rho_{C/\sigma_0^2}(r)$ where the threshold $q = q(C)$ is defined as the solution to

$$h_q^*(q) = 1 \iff \rho_{C/\sigma_0^2}(q) = 1.$$

Suppose we choose the function $h_q^*(\cdot)$ as the estimator in Berger (1981) and Bickel (1984), i.e., $h_q^*(r) = \rho_{C/\sigma_0^2}(r)$ with q given above. In the univariate case $d = 1$, the estimator $\hat{\tau}_{\text{ST}}$ reduces to the univariate soft-thresholding estimator $\hat{\tau}_{\text{ST}}$. In this case, the user-chosen constant C in $\hat{\tau}_{\text{ST}}$ corresponds one-to-one to the significance level α , or equivalently to the critical value $c_{\alpha/2}$ in $\hat{\tau}_{\text{ST}}$. Consider the special case $C = 0$. When $d \leq 2$, the estimator $\hat{\tau}_{\text{ST}}$ reduces to the unbiased estimator $\hat{\tau}_0$. However, when $d \geq 3$, the behavior of $\hat{\tau}_{\text{ST}}$ is different. When $C = 0$ and $d \geq 3$, we have $\rho_{C/\sigma_0^2}(r) = \rho_0(r) = 2(d-2)/r$, and therefore the threshold q is given by $q = 2(d-2)$. If $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q$, or equivalently, $\|\hat{\tau}_1 - \hat{\tau}_0\|_2^2 \leq 2(d-2)(\sigma_0^2 + \sigma_1^2)$, the estimator $\hat{\tau}_{\text{ST}}$ reduces to the precision-weighted estimator $\hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0)$. If instead $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q$, or equivalently, $\|\hat{\tau}_1 - \hat{\tau}_0\|_2^2 > 2(d-2)(\sigma_0^2 + \sigma_1^2)$, the estimator $\hat{\tau}_{\text{ST}}$ takes the James–Stein shrinkage form (Green and Strawderman, 1991):

$$\hat{\tau}_0 + h_q^*(\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2)(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}\Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = \hat{\tau}_0 + \frac{2(d-2)\sigma_0^2}{\|\hat{\tau}_1 - \hat{\tau}_0\|_2^2}(\hat{\tau}_1 - \hat{\tau}_0).$$

A.2 Proofs of the main results

A.2.1 Proof of Theorem 1.3.1

Proof of Theorem 1.3.1. By the definition of $\hat{\tau}_{\text{PW}}$, under Assumption 1.2.1, we have

$$\hat{\tau}_{\text{PW}} \sim N\left(\tau + \frac{\gamma}{1+\gamma}\Delta, \frac{1}{1+\gamma}\sigma_0^2\right).$$

Equivalently, after standardization,

$$(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{\text{PW}} - \tau) \sim N\left(\frac{\gamma}{\sqrt{1+\gamma}}\frac{\Delta}{\sigma_0}, 1\right).$$

We now evaluate the worst-case coverage probability of the confidence interval

$$[\hat{\tau}_{\text{PW}} - L(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{PW}} + L(1 + \gamma)^{-1/2}\sigma_0].$$

We have

$$\begin{aligned} & \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta(\tau \in [\hat{\tau}_{\text{PW}} - L(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{PW}} + L(1 + \gamma)^{-1/2}\sigma_0]) \\ &= \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta(|\hat{\tau}_{\text{PW}} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta(|(1 + \gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{\text{PW}} - \tau)| \leq L) \\ &= \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta\left(\left|N\left(\frac{\gamma}{\sqrt{1 + \gamma}}\frac{\Delta}{\sigma_0}, 1\right)\right| \leq L\right) = \mathbb{P}\left(\left|N\left(\sup_{\Delta: |\Delta/\sigma_0| \leq b} \frac{\gamma}{\sqrt{1 + \gamma}}\frac{\Delta}{\sigma_0}, 1\right)\right| \leq L\right) \\ &= \mathbb{P}\left(\left|N\left(\frac{\gamma}{\sqrt{1 + \gamma}}b, 1\right)\right| \leq L\right) = \Phi\left(L - \frac{\gamma}{\sqrt{1 + \gamma}}b\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1 + \gamma}}b\right). \end{aligned}$$

Therefore, the proof is complete since $\hat{L}_{\text{PW}}$ is the solution to

$$\Phi\left(L - \frac{\gamma}{\sqrt{1 + \gamma}}b\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1 + \gamma}}b\right) = 1 - \zeta.$$

□

A.2.2 Proof of Theorem 1.3.2

Proof of Theorem 1.3.2. Recall that the pretest estimator $\hat{\tau}_{\text{PT}}$ is defined as

$$\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + \frac{\gamma}{1 + \gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}).$$

To characterize the distribution of $\hat{\tau}_{\text{PT}} - \tau$, note that

$$\begin{pmatrix} \hat{\tau}_0 + \frac{\gamma}{1 + \gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \tau \\ \hat{\tau}_1 - \hat{\tau}_0 \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\gamma}{1 + \gamma}\Delta \\ \Delta \end{pmatrix}, \begin{pmatrix} \frac{1}{1 + \gamma}\sigma_0^2 & 0 \\ 0 & \sigma_0^2 + \sigma_1^2 \end{pmatrix}\right).$$

We first consider the event that the pretest fails to reject, $|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}$. Conditional on this event, $\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + \frac{\gamma}{1 + \gamma}(\hat{\tau}_1 - \hat{\tau}_0)$, and therefore

$$(\hat{\tau}_{\text{PT}} - \tau) | \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\} \sim N\left(\frac{\gamma}{1 + \gamma}\Delta, \frac{1}{1 + \gamma}\sigma_0^2\right).$$

Equivalently, after standardization,

$$(1 + \gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{\text{PT}} - \tau) | \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\} \sim N\left(\frac{\gamma}{\sqrt{1 + \gamma}}\frac{\Delta}{\sigma_0}, 1\right).$$

We next consider the event that the pretest rejects, $|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}$. Conditional on $\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u$, we have

$$\frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) = \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0 u.$$

Since $\hat{\tau}_{PT} = \hat{\tau}_0$ conditional on this event, we have

$$\begin{aligned} & (\hat{\tau}_{PT} - \tau) \mid \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\} \\ &= (\hat{\tau}_0 - \tau) \mid \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\} \\ &= (\hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \tau) \mid \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\} - \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0 u \\ &\sim N\left(\frac{\gamma}{1+\gamma}\Delta - \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0 u, \frac{1}{1+\gamma}\sigma_0^2\right), \end{aligned}$$

which implies

$$(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau) \mid \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\} \sim N\left(\frac{\gamma}{\sqrt{1+\gamma}}\frac{\Delta}{\sigma_0} - \sqrt{\gamma}u, 1\right).$$

We now evaluate the worst-case coverage probability of the confidence interval

$$[\hat{\tau}_{PT} - L(1+\gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + L(1+\gamma)^{-1/2}\sigma_0].$$

We have

$$\begin{aligned} & \inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(\tau \in [\hat{\tau}_{PT} - L(1+\gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + L(1+\gamma)^{-1/2}\sigma_0]) \\ &= \inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0). \end{aligned}$$

Fixing $\Delta/\sigma_0 = t$, we decompose the probability according to whether the pretest accepts or rejects:

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) \\ &= P_{\Delta/\sigma_0=t}(|(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau)| \leq L) \\ &= P_{\Delta/\sigma_0=t}(|(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau)| \leq L, |\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}) \\ & \quad + P_{\Delta/\sigma_0=t}(|(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau)| \leq L, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}) \\ &= P\left(\left|N\left(\frac{\gamma}{\sqrt{1+\gamma}}t, 1\right)\right| \leq L\right) P\left(\left|N\left(\sqrt{\frac{\gamma}{1+\gamma}}t, 1\right)\right| \leq c_{\alpha/2}\right) \\ & \quad + E_{U \sim N(\sqrt{\frac{\gamma}{1+\gamma}}t, 1)}\left[P\left(\left|N\left(\frac{\gamma}{\sqrt{1+\gamma}}t - \sqrt{\gamma}U, 1\right)\right| \leq L\right) \mathbf{1}(|U| > c_{\alpha/2})\right] \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right)\right] \left[\Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1+\gamma}}t\right)\right] \end{aligned}$$

$$\begin{aligned}
& + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \left[\Phi(L + \sqrt{\gamma}u) - \Phi(-L + \sqrt{\gamma}u) \right] \phi(u) du \\
& + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \left[\Phi(L + \sqrt{\gamma}u) - \Phi(-L + \sqrt{\gamma}u) \right] \phi(u) du,
\end{aligned}$$

which yields exactly the expression in (1.3.2).

Finally, since the coverage probability is symmetric in Δ , the worst case over $|\Delta/\sigma_0| \leq b$ is attained for some $t \in [0, b]$. Choosing $\hat{L}_{PT}$ to satisfy

$$\min_{0 \leq t \leq b} P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) = 1 - \zeta$$

completes the proof. □

A.2.3 Proof of Theorem 1.3.3

Proof of Theorem 1.3.3. Recall that the soft-thresholding estimator $\hat{\tau}_{ST}$ is defined as

$$\hat{\tau}_{ST} = \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}) + \frac{\gamma}{1+\gamma} \sigma c_{\alpha/2} \operatorname{sign}(\hat{\tau}_1 - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}).$$

As before, note that

$$\begin{pmatrix} \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \tau \\ \hat{\tau}_1 - \hat{\tau}_0 \end{pmatrix} \sim N \left(\begin{pmatrix} \frac{\gamma}{1+\gamma} \Delta \\ \Delta \end{pmatrix}, \begin{pmatrix} \frac{1}{1+\gamma} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 + \sigma_1^2 \end{pmatrix} \right).$$

We first consider the event that the pretest accepts, $|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}$. Conditional on this event, $\hat{\tau}_{ST} = \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0)$, and therefore

$$(\hat{\tau}_{ST} - \tau) | \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\} \sim N \left(\frac{\gamma}{1+\gamma} \Delta, \frac{1}{1+\gamma} \sigma_0^2 \right).$$

Equivalently, after standardization,

$$(1+\gamma)^{1/2} \sigma_0^{-1} (\hat{\tau}_{ST} - \tau) | \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\} \sim N \left(\frac{\gamma}{\sqrt{1+\gamma}} \frac{\Delta}{\sigma_0}, 1 \right).$$

We next consider the event that the pretest rejects, $|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}$. Conditional on $\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u$, we have

$$\frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) = \sqrt{\frac{\gamma}{1+\gamma}} \sigma_0 u.$$

Since $\hat{\tau}_{ST} = \hat{\tau}_0 + \frac{\gamma}{1+\gamma} \sigma c_{\alpha/2} \operatorname{sign}(\hat{\tau}_1 - \hat{\tau}_0)$ conditional on this event, we have

$$(\hat{\tau}_{ST} - \tau) | \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\}$$

$$\begin{aligned}
&= \widehat{\tau}_0 - \tau + \frac{\gamma}{1+\gamma} \sigma c_{\alpha/2} \operatorname{sign}(\widehat{\tau}_1 - \widehat{\tau}_0) \mid \{\sigma^{-1}(\widehat{\tau}_1 - \widehat{\tau}_0) = u, |\widehat{\tau}_1 - \widehat{\tau}_0| > \sigma c_{\alpha/2}\} \\
&= \widehat{\tau}_0 + \frac{\gamma}{1+\gamma} (\widehat{\tau}_1 - \widehat{\tau}_0) - \tau \mid \{\sigma^{-1}(\widehat{\tau}_1 - \widehat{\tau}_0) = u, |\widehat{\tau}_1 - \widehat{\tau}_0| > \sigma c_{\alpha/2}\} \\
&\quad - \sqrt{\frac{\gamma}{1+\gamma}} \sigma_0 [u - c_{\alpha/2} \operatorname{sign}(u)] \\
&\sim N \left(\frac{\gamma}{1+\gamma} \Delta - \sqrt{\frac{\gamma}{1+\gamma}} \sigma_0 [u - c_{\alpha/2} \operatorname{sign}(u)], \frac{1}{1+\gamma} \sigma_0^2 \right),
\end{aligned}$$

which implies

$$\begin{aligned}
&(1+\gamma)^{1/2} \sigma_0^{-1} (\widehat{\tau}_{\text{ST}} - \tau) \mid \{\sigma^{-1}(\widehat{\tau}_1 - \widehat{\tau}_0) = u, |\widehat{\tau}_1 - \widehat{\tau}_0| > \sigma c_{\alpha/2}\} \\
&\sim N \left(\frac{\gamma}{\sqrt{1+\gamma}} \frac{\Delta}{\sigma_0} - \sqrt{\gamma} [u - c_{\alpha/2} \operatorname{sign}(u)], 1 \right).
\end{aligned}$$

We now evaluate the worst-case coverage probability of the confidence interval

$$[\widehat{\tau}_{\text{ST}} - L(1+\gamma)^{-1/2} \sigma_0, \widehat{\tau}_{\text{ST}} + L(1+\gamma)^{-1/2} \sigma_0].$$

We have

$$\begin{aligned}
&\inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_{\Delta}(\tau \in [\widehat{\tau}_{\text{ST}} - L(1+\gamma)^{-1/2} \sigma_0, \widehat{\tau}_{\text{ST}} + L(1+\gamma)^{-1/2} \sigma_0]) \\
&= \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_{\Delta}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1+\gamma)^{-1/2} \sigma_0).
\end{aligned}$$

Fixing $\Delta/\sigma_0 = t$, we decompose the probability according to whether the pretest accepts or rejects:

$$\begin{aligned}
&\mathbb{P}_{\Delta/\sigma_0=t}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1+\gamma)^{-1/2} \sigma_0) \\
&= \mathbb{P}_{\Delta/\sigma_0=t}(|(1+\gamma)^{1/2} \sigma_0^{-1} (\widehat{\tau}_{\text{ST}} - \tau)| \leq L) \\
&= \mathbb{P}_{\Delta/\sigma_0=t}(|(1+\gamma)^{1/2} \sigma_0^{-1} (\widehat{\tau}_{\text{ST}} - \tau)| \leq L, |\widehat{\tau}_1 - \widehat{\tau}_0| \leq \sigma c_{\alpha/2}) \\
&\quad + \mathbb{P}_{\Delta/\sigma_0=t}(|(1+\gamma)^{1/2} \sigma_0^{-1} (\widehat{\tau}_{\text{ST}} - \tau)| \leq L, |\widehat{\tau}_1 - \widehat{\tau}_0| > \sigma c_{\alpha/2}) \\
&= \mathbb{P} \left(\left| N \left(\frac{\gamma}{\sqrt{1+\gamma}} t, 1 \right) \right| \leq L \right) \mathbb{P} \left(\left| N \left(\sqrt{\frac{\gamma}{1+\gamma}} t, 1 \right) \right| \leq c_{\alpha/2} \right) \\
&\quad + \mathbb{E}_{U \sim N(\sqrt{\frac{\gamma}{1+\gamma}} t, 1)} \left[\mathbb{P} \left(\left| N \left(\frac{\gamma}{\sqrt{1+\gamma}} t - \sqrt{\gamma} [U - c_{\alpha/2} \operatorname{sign}(U)], 1 \right) \right| \leq L \right) \mathbf{1}(|U| > c_{\alpha/2}) \right] \\
&= \left[\Phi \left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}} t \right) - \Phi \left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}} t \right) \right] \left[\Phi \left(L - \frac{\gamma}{\sqrt{1+\gamma}} t \right) - \Phi \left(-L - \frac{\gamma}{\sqrt{1+\gamma}} t \right) \right] \\
&\quad + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}} t} \left[\Phi \left(L + \sqrt{\gamma} (u + c_{\alpha/2}) \right) - \Phi \left(-L + \sqrt{\gamma} (u + c_{\alpha/2}) \right) \right] \phi(u) du
\end{aligned}$$

$$+ \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \left[\Phi\left(L + \sqrt{\gamma}(u - c_{\alpha/2})\right) - \Phi\left(-L + \sqrt{\gamma}(u - c_{\alpha/2})\right) \right] \phi(u) du.$$

which yields exactly the expression in (1.3.4).

To establish the monotonicity of the coverage probability in $|\Delta|$, fix $L > 0$ and write $t = \Delta/\sigma_0$. Since $P_{\Delta}(|\hat{\tau}_{\text{ST}} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0)$ is symmetric about $\Delta = 0$, it suffices to consider $t \geq 0$. By Lemma 1.3.2, we have

$$(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{\text{ST}} - \tau) \stackrel{d}{=} Z + \frac{\gamma}{\sqrt{1+\gamma}}t - \sqrt{\gamma}S(U),$$

where $Z \sim N(0, 1)$ is independent of $U \sim N(\sqrt{\frac{\gamma}{1+\gamma}}t, 1)$ and $S(u) = \text{sign}(u)(|u| - c_{\alpha/2})_+$. Consequently, we can write the coverage probability as

$$P_{\Delta/\sigma_0=t}(|\hat{\tau}_{\text{ST}} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) = E[\Phi(L - \mu_t(U)) - \Phi(-L - \mu_t(U))],$$

with $\mu_t(U) = \frac{\gamma}{\sqrt{1+\gamma}}t - \sqrt{\gamma}S(U)$. The function $\mu \mapsto \Phi(L - \mu) - \Phi(-L - \mu)$ is even and strictly decreasing in $|\mu|$. Moreover, increasing t shifts the distribution of U away from zero, which increases $|\mu_t(U)|$ in the sense of stochastic order. Therefore, the coverage probability is nonincreasing in $t \geq 0$, or equivalently, monotonically decreasing in $|\Delta|$.

Finally, since the coverage probability is symmetric in Δ and monotonically decreasing in $|\Delta|$, the worst case over $|\Delta/\sigma_0| \leq b$ is attained at $\Delta/\sigma_0 = b$. Choosing $\hat{L}_{\text{ST}}$ to satisfy

$$P_{\Delta/\sigma_0=b}(|\hat{\tau}_{\text{ST}} - \tau| \leq L(1+\gamma)^{-1/2}\sigma_0) = 1 - \zeta$$

completes the proof. □

A.2.4 Proof of Theorem 1.4.1

Proof of Theorem 1.4.1. By the definition of $\hat{\tau}_{\text{PW}}$, under Assumption 1.4.1, we have

$$\hat{\tau}_{\text{PW}} \sim N(\boldsymbol{\tau} + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Delta}, (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1}).$$

Equivalently, after standardization,

$$(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{1/2}(\hat{\tau}_{\text{PW}} - \boldsymbol{\tau}) \sim N((\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Delta}, \mathbf{I}_d),$$

which implies

$$(\hat{\tau}_{\text{PW}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\hat{\tau}_{\text{PW}} - \boldsymbol{\tau}) \sim \chi_d^2 \left(\left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Delta} \right\|_2^2 \right),$$

where $\chi_d^2(\cdot)$ denotes the noncentral chi-squared distribution with d degrees of freedom.

We now evaluate the worst-case coverage probability of the quadratic confidence region

$$\{\boldsymbol{\tau} \in \mathbb{R}^d : (\widehat{\boldsymbol{\tau}}_{\text{PW}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PW}} - \boldsymbol{\tau}) \leq M\}.$$

We have

$$\begin{aligned} & \inf_{\boldsymbol{\Delta}: |\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}|_j \leq b_j, \forall j=1,2,\dots,d} \mathbb{P}_{\boldsymbol{\Delta}} \left((\widehat{\boldsymbol{\tau}}_{\text{PW}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PW}} - \boldsymbol{\tau}) \leq M \right) \\ &= \inf_{\boldsymbol{\Delta}: |\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}|_j \leq b_j, \forall j=1,2,\dots,d} \mathbb{P}_{\boldsymbol{\Delta}} \left(\chi_d^2 \left(\left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta} \right\|_2^2 \right) \leq M \right) \\ &= \mathbb{P} \left(\chi_d^2 \left(\sup_{\boldsymbol{\Delta}: |\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}|_j \leq b_j, \forall j=1,2,\dots,d} \left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta} \right\|_2^2 \right) \leq M \right). \end{aligned}$$

To evaluate the supremum in the noncentrality parameter, note that

$$\left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta} \right\|_2^2 = \left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_1^{1/2} (\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}) \right\|_2^2,$$

which is a convex quadratic function of $\mathbf{u} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}$. Since the constraint $|u_j| \leq b_j$ defines a hyperrectangle, the maximum of this convex quadratic over the constraint set is attained at a vertex. Therefore,

$$\begin{aligned} & \sup_{\boldsymbol{\Delta}: |\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}|_j \leq b_j, \forall j=1,2,\dots,d} \left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta} \right\|_2^2 \\ &= \sup_{\mathbf{s} \in \{\pm 1\}^d} \left\| (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_1^{1/2} \mathbf{b} \odot \mathbf{s} \right\|_2^2. \end{aligned}$$

Finally, choosing $\widehat{M}_{\text{PW}}$ as the $(1 - \zeta)$ upper quantile of the noncentral chi-squared distribution with d degrees of freedom and the above noncentrality parameter ensures that the confidence region attains the desired coverage level. This completes the proof. $\square$

A.2.5 Proof of Theorem 1.4.2

Proof of Theorem 1.4.2. Recall that the pretest estimator $\widehat{\boldsymbol{\tau}}_{\text{PT}}$ is defined as

$$\widehat{\boldsymbol{\tau}}_{\text{PT}} = \widehat{\boldsymbol{\tau}}_0 + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1} (\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0) \mathbb{1} \left(\left\| (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} (\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0) \right\|_2^2 \leq q \right).$$

A direct calculation shows that

$$\widehat{\boldsymbol{\tau}}_0 + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1} (\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0) \perp\!\!\!\perp \widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0.$$

Moreover,

$$\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0 \sim N(\boldsymbol{\Delta}, \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1),$$

and

$$\hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} (\hat{\tau}_1 - \hat{\tau}_0) \sim N(\tau + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} \Delta, (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}).$$

We first consider the event that the pretest accepts, $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q$. Conditional on this event, $\hat{\tau}_{PT} = \hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} (\hat{\tau}_1 - \hat{\tau}_0)$, and therefore

$$(\hat{\tau}_{PT} - \tau) | \{ \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q \} \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} \Delta, (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}).$$

After standardization,

$$\begin{aligned} & (\Sigma_0^{-1} + \Sigma_1^{-1})^{1/2} (\hat{\tau}_{PT} - \tau) | \{ \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q \} \\ & \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} \Delta, I_d), \end{aligned}$$

which implies

$$\begin{aligned} & (\hat{\tau}_{PT} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{PT} - \tau) | \{ \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q \} \\ & \sim \chi_d^2 \left(\|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} \Delta\|_2^2 \right). \end{aligned}$$

We next consider the event that the pretest rejects, $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q$. Conditional on $(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}$, we have

$$(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} (\hat{\tau}_1 - \hat{\tau}_0) = (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} (\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}.$$

Since $\hat{\tau}_{PT} = \hat{\tau}_0$ conditional on this event, we have

$$\begin{aligned} & (\hat{\tau}_{PT} - \tau) | \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & = \hat{\tau}_0 | \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & = \hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} (\hat{\tau}_1 - \hat{\tau}_0) | \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \\ & \quad \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} - (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} (\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u} \\ & \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1} [\Delta - (\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}], (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}). \end{aligned}$$

After standardization,

$$\begin{aligned} & (\Sigma_0^{-1} + \Sigma_1^{-1})^{1/2} (\hat{\tau}_{PT} - \tau) | \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} [\Delta - (\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}], I_d), \end{aligned}$$

which implies

$$\begin{aligned} & (\hat{\tau}_{PT} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{PT} - \tau) | \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \\ & \quad \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & \sim \chi_d^2 \left(\|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} [\Delta - (\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}]\|_2^2 \right). \end{aligned}$$

We now evaluate the worst-case coverage probability of the confidence region

$$\{\boldsymbol{\tau} \in \mathbb{R}^d : (\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M\}.$$

We decompose the coverage probability according to whether the pretest accepts or rejects:

$$\begin{aligned} & P_{\boldsymbol{\Delta}} \left((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M \right) \\ &= P_{\boldsymbol{\Delta}} \left((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M, \|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0)\|_2^2 \leq q \right) \\ &\quad + P_{\boldsymbol{\Delta}} \left((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M, \|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0)\|_2^2 > q \right) \\ &= \Psi \left(M; \|(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta}\|_2^2 \right) \Psi \left(q; \|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} \boldsymbol{\Delta}\|_2^2 \right) \\ &\quad + \int_{\|\mathbf{u}\|_2^2 > q} \Psi \left(M; \|(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1/2} \boldsymbol{\Sigma}_1^{-1} [\boldsymbol{\Delta} - (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{1/2} \mathbf{u}]\|_2^2 \right) \phi(\mathbf{u}; (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2} \boldsymbol{\Delta}) d\mathbf{u}, \end{aligned}$$

which yields exactly the expression in (1.4.4) by taking $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta} = \mathbf{t}$.

Finally, choosing $\widehat{M}_{\text{PT}}$ to satisfy

$$\inf_{\boldsymbol{\Delta}: \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}\| \leq b} P_{\boldsymbol{\Delta}}((\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau})^\top (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})(\widehat{\boldsymbol{\tau}}_{\text{PT}} - \boldsymbol{\tau}) \leq M) = 1 - \zeta$$

completes the proof. □

A.2.6 Proof of Theorem 1.4.3

Proof of Theorem 1.4.3. Recall that the soft-thresholding estimator $\widehat{\boldsymbol{\tau}}_{\text{ST}}$ is defined as

$$\widehat{\boldsymbol{\tau}}_{\text{ST}} = \widehat{\boldsymbol{\tau}}_0 + h_q(\|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0)\|_2^2)(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0).$$

A direct calculation shows that

$$\widehat{\boldsymbol{\tau}}_0 + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0) \perp\!\!\!\perp \widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0.$$

Moreover,

$$\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0 \sim N(\boldsymbol{\Delta}, \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1),$$

and

$$\widehat{\boldsymbol{\tau}}_0 + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0) \sim N(\boldsymbol{\tau} + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta}, (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1}).$$

We first consider the event that the pretest accepts, $\|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0)\|_2^2 \leq q$. Conditional on this event, $\widehat{\boldsymbol{\tau}}_{\text{ST}} = \widehat{\boldsymbol{\tau}}_0 + (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0)$, and therefore

$$(\widehat{\boldsymbol{\tau}}_{\text{ST}} - \boldsymbol{\tau}) \mid \{\|(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1/2}(\widehat{\boldsymbol{\tau}}_1 - \widehat{\boldsymbol{\tau}}_0)\|_2^2 \leq q\} \sim N((\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta}, (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1}).$$

After standardization,

$$\begin{aligned} & (\Sigma_0^{-1} + \Sigma_1^{-1})^{1/2}(\hat{\tau}_{\text{ST}} - \tau) \mid \{ \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q \} \\ & \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} \Delta, I_d), \end{aligned}$$

which implies

$$\begin{aligned} & (\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1})(\hat{\tau}_{\text{ST}} - \tau) \mid \{ \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q \} \\ & \sim \chi_d^2 \left(\|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} \Delta\|_2^2 \right). \end{aligned}$$

We next consider the event that the pretest rejects, $\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q$. Conditional on $(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}$, we have

$$(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}(\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}.$$

Since $\hat{\tau}_{\text{ST}} = \hat{\tau}_0 + h_q^*(\|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2)(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0)$ conditional on this event, we have

$$\begin{aligned} & (\hat{\tau}_{\text{ST}} - \tau) \mid \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & = \hat{\tau}_0 + h_q^*(\|\mathbf{u}\|_2^2)(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0) \mid \\ & \quad \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & = \hat{\tau}_0 + (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}(\hat{\tau}_1 - \hat{\tau}_0) \mid \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \\ & \quad \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} - (1 - h_q^*(\|\mathbf{u}\|_2^2))(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}(\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u} \\ & \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1} \Sigma_1^{-1}[\Delta - (1 - h_q^*(\|\mathbf{u}\|_2^2))(\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}], (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}). \end{aligned}$$

After standardization,

$$\begin{aligned} & (\Sigma_0^{-1} + \Sigma_1^{-1})^{1/2}(\hat{\tau}_{\text{ST}} - \tau) \mid \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & \sim N((\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1}[\Delta - (1 - h_q^*(\|\mathbf{u}\|_2^2))(\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}], I_d), \end{aligned}$$

which implies

$$\begin{aligned} & (\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1})(\hat{\tau}_{\text{ST}} - \tau) \mid \{ (\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0) = \mathbf{u}, \\ & \quad \|(\Sigma_0 + \Sigma_1)^{-1/2}(\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \} \\ & \sim \chi_d^2 \left(\|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1}[\Delta - (1 - h_q^*(\|\mathbf{u}\|_2^2))(\Sigma_0 + \Sigma_1)^{1/2} \mathbf{u}]\|_2^2 \right). \end{aligned}$$

We now evaluate the worst-case coverage probability of the confidence region

$$\{ \tau \in \mathbb{R}^d : (\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1})(\hat{\tau}_{\text{ST}} - \tau) \leq M \}.$$

We decompose the coverage probability according to whether the pretest accepts or rejects:

$$P_\Delta \left((\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1})(\hat{\tau}_{\text{ST}} - \tau) \leq M \right)$$

$$\begin{aligned}
&= P_{\Delta} \left((\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{\text{ST}} - \tau) \leq M, \|(\Sigma_0 + \Sigma_1)^{-1/2} (\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 \leq q \right) \\
&\quad + P_{\Delta} \left((\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{\text{ST}} - \tau) \leq M, \|(\Sigma_0 + \Sigma_1)^{-1/2} (\hat{\tau}_1 - \hat{\tau}_0)\|_2^2 > q \right) \\
&= \Psi \left(M; \|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} \Delta\|_2^2 \right) \Psi \left(q; \|(\Sigma_0 + \Sigma_1)^{-1/2} \Delta\|_2^2 \right) \\
&\quad + \int_{\|u\|_2^2 > q} \Psi \left(M; \|(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1/2} \Sigma_1^{-1} [\Delta - (1 - h_q^*(\|u\|_2^2))(\Sigma_0 + \Sigma_1)^{1/2} u]\|_2^2 \right) \\
&\quad \phi(u; (\Sigma_0 + \Sigma_1)^{-1/2} \Delta) du,
\end{aligned}$$

which yields exactly the expression in (1.4.5) by taking $\Sigma^{-1/2} \Delta = \mathbf{t}$.

Let $\mathbf{t} = \Sigma^{-1/2} \Delta$ and denote by

$$p(\mathbf{t}) := P_{\Sigma^{-1/2} \Delta = \mathbf{t}} \left((\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{\text{ST}} - \tau) \leq M \right).$$

By construction of the soft-thresholding rule, $p(\mathbf{t})$ is invariant under coordinate-wise sign flips, i.e., $p(\mathbf{t}) = p(\mathbf{t} \odot \mathbf{s})$ for all $\mathbf{s} \in \{\pm 1\}^d$. Moreover, for each $j \in \{1, \dots, d\}$ and any fixed values of the remaining coordinates, the map $u \mapsto p(t_1, \dots, t_{j-1}, u, t_{j+1}, \dots, t_d)$ is non-increasing in $|u|$. Therefore,

$$\inf_{\mathbf{t}: |t_j| \leq b_j} p(\mathbf{t}) = \inf_{\mathbf{s} \in \{\pm 1\}^d} p(\mathbf{b} \odot \mathbf{s}),$$

i.e., the worst-case coverage probability over the hyperrectangle is attained at a vertex.

Finally, choosing $\widehat{M}_{\text{ST}}$ to satisfy

$$\inf_{\Delta: \Sigma^{-1/2} \Delta = \mathbf{b} \odot \mathbf{s}, \mathbf{s} \in \{-1, 1\}^d} P_{\Delta} \left((\hat{\tau}_{\text{ST}} - \tau)^\top (\Sigma_0^{-1} + \Sigma_1^{-1}) (\hat{\tau}_{\text{ST}} - \tau) \leq M \right) = 1 - \zeta$$

completes the proof. □

A.2.7 Proof of Theorem 1.5.1

Proof of Theorem 1.5.1. By the definition of $\hat{\tau}_{\text{PW}}$, under Assumption 1.5.1, we have

$$\hat{\tau}_{\text{PW}} \sim N \left(\tau + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\gamma\|_1} \Delta_j, \frac{1}{1 + \|\gamma\|_1} \sigma_0^2 \right).$$

Equivalently, after standardization,

$$(1 + \|\gamma\|_1)^{1/2} \sigma_0^{-1} (\hat{\tau}_{\text{PW}} - \tau) \sim N \left(\sum_{j=1}^K \frac{\gamma_j}{\sqrt{1 + \|\gamma\|_1}} \frac{\Delta_j}{\sigma_0}, 1 \right).$$

We now evaluate the worst-case coverage probability of the confidence interval

$$[\hat{\tau}_{\text{PW}} - L(1 + \|\gamma\|_1)^{-1/2} \sigma_0, \hat{\tau}_{\text{PW}} + L(1 + \|\gamma\|_1)^{-1/2} \sigma_0].$$

We have

$$\begin{aligned}
& \inf_{\mathbf{\Delta}: |\Delta_j/\sigma_0| \leq b_j, \forall j=1, \dots, K} P_{\mathbf{\Delta}}(|\hat{\tau}_{\text{PW}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2}\sigma_0) \\
&= \inf_{\mathbf{\Delta}: |\Delta_j/\sigma_0| \leq b_j, \forall j=1, \dots, K} P_{\mathbf{\Delta}} \left(\left| N \left(\sum_{j=1}^K \frac{\gamma_j}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \frac{\Delta_j}{\sigma_0}, 1 \right) \right| \leq L \right) \\
&= P \left(\left| N \left(\sup_{\mathbf{\Delta}: |\Delta_j/\sigma_0| \leq b_j, \forall j=1, \dots, K} \sum_{j=1}^K \frac{\gamma_j}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \frac{\Delta_j}{\sigma_0}, 1 \right) \right| \leq L \right) \\
&= P \left(\left| N \left(\frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}}, 1 \right) \right| \leq L \right) \\
&= \Phi \left(L - \frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) - \Phi \left(-L - \frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right).
\end{aligned}$$

Therefore, the proof is complete since $\hat{L}_{\text{PW}}$ is the solution to

$$\Phi \left(L - \frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) - \Phi \left(-L - \frac{\langle \boldsymbol{\gamma}, \mathbf{b} \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) = 1 - \zeta.$$

□

A.2.8 Proof of Theorem 1.5.2

Proof of Theorem 1.5.2. Recall that the pretest estimator $\hat{\tau}_{\text{PT}}$ is defined as

$$\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| \leq (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}).$$

Then

$$(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 \sim N(\mathbf{\Delta} / \sigma_0, \mathbf{V}), \quad \mathbf{V}_{ij} = 1 + \gamma_i^{-1} \mathbf{1}(i = j),$$

and moreover

$$\hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) \perp\!\!\!\perp (\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top.$$

Rewrite $\hat{\tau}_{\text{PT}}$ as

$$\hat{\tau}_{\text{PT}} = \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| \leq (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2})$$

$$=\hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) - \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| > (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}).$$

Consequently, conditioning on $(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}$ yields

$$\begin{aligned} & \hat{\tau}_{\text{PT}} \mid \{(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}\} \\ &= \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) - \sum_{j: |u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}} \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \sigma_0 u_j, \\ & \sim N \left(\tau + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \Delta_j - \sum_{j: |u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}} \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \sigma_0 u_j, \frac{1}{1 + \|\boldsymbol{\gamma}\|_1} \sigma_0^2 \right). \end{aligned}$$

Equivalently,

$$\begin{aligned} & (1 + \|\boldsymbol{\gamma}\|_1)^{1/2} \sigma_0^{-1} (\hat{\tau}_{\text{PT}} - \tau) \mid \{(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}\} \\ & \sim N \left(\sum_{j=1}^K \frac{\gamma_j}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \frac{\Delta_j}{\sigma_0} - \sum_{j: |u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}} \frac{\gamma_j}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} u_j, 1 \right). \end{aligned}$$

We now evaluate the worst-case coverage probability of the confidence interval

$$[\hat{\tau}_{\text{PT}} - L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0, \hat{\tau}_{\text{PT}} + L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0].$$

Define $\mathbf{u}' \in \mathbb{R}^K$ coordinate-wise by $u'_j = u_j \mathbf{1}(|u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2})$ for $j = 1, \dots, K$. We have

$$\begin{aligned} & P_{\boldsymbol{\Delta}/\sigma_0=\mathbf{t}}(|\hat{\tau}_{\text{PT}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0) \\ &= \int_{\mathbb{R}^K} P_{\boldsymbol{\Delta}/\sigma_0=\mathbf{t}}(|\hat{\tau}_{\text{PT}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0 \mid (\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}) \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u} \\ &= \int_{\mathbb{R}^K} P \left(\left| N \left(\frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}}, 1 \right) \right| \leq L \right) \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u} \\ &= \int_{\mathbb{R}^K} \left[\Phi \left(L - \frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) - \Phi \left(-L - \frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) \right] \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u}, \end{aligned}$$

which yields exactly the expression in the theorem.

Finally, choosing $\hat{L}_{\text{PT}}$ to satisfy

$$\inf_{\boldsymbol{\Delta}: |\boldsymbol{\Delta}/\sigma_0| \leq \mathbf{b}} P_{\boldsymbol{\Delta}}(|\hat{\tau}_{\text{PT}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0) = 1 - \zeta$$

completes the proof. □

A.2.9 Proof of Theorem 1.5.3

Proof of Theorem 1.5.3. Recall that the soft-thresholding estimator $\hat{\tau}_{\text{ST}}$ is defined as

$$\begin{aligned} \hat{\tau}_{\text{ST}} = \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} & \left[(\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| \leq (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}) \right. \\ & \left. + (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2} \text{sign}(\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| > (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}) \right]. \end{aligned}$$

Then

$$(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 \sim N(\boldsymbol{\Delta} / \sigma_0, \mathbf{V}), \quad \mathbf{V}_{ij} = 1 + \gamma_i^{-1} \mathbf{1}(i = j),$$

and moreover

$$\hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) \perp\!\!\!\perp (\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top.$$

Rewrite $\hat{\tau}_{\text{ST}}$ as

$$\begin{aligned} \hat{\tau}_{\text{ST}} &= \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \left[(\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| \leq (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}) \right. \\ & \quad \left. + (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2} \text{sign}(\hat{\tau}_j - \hat{\tau}_0) \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| > (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}) \right] \\ &= \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) - \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \\ & \quad [(\hat{\tau}_j - \hat{\tau}_0) - (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2} \text{sign}(\hat{\tau}_j - \hat{\tau}_0)] \mathbf{1}(|\hat{\tau}_j - \hat{\tau}_0| > (1 + \gamma_j^{-1})^{1/2} \sigma_0 c_{\alpha/2}). \end{aligned}$$

Consequently, conditioning on $(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}$ yields

$$\begin{aligned} & \hat{\tau}_{\text{ST}} \mid \{(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}\} \\ &= \hat{\tau}_0 + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} (\hat{\tau}_j - \hat{\tau}_0) - \sum_{j: |u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}} \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \sigma_0 [u_j - (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2} \text{sign}(u_j)] \\ &\sim N\left(\tau + \sum_{j=1}^K \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \Delta_j - \sum_{j: |u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}} \frac{\gamma_j}{1 + \|\boldsymbol{\gamma}\|_1} \sigma_0 [u_j - (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2} \text{sign}(u_j)], \right. \\ & \quad \left. \frac{1}{1 + \|\boldsymbol{\gamma}\|_1} \sigma_0^2\right). \end{aligned}$$

Equivalently,

$$(\hat{\tau}_{\text{ST}} - \tau) / [(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0] \mid \{(\hat{\tau}_1 - \hat{\tau}_0, \dots, \hat{\tau}_K - \hat{\tau}_0)^\top / \sigma_0 = \mathbf{u}\}$$

$$\sim N \left(\sum_{j=1}^K \frac{\gamma_j}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \frac{\Delta_j}{\sigma_0} - \sum_{j: |u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2}} \frac{\gamma_j}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} [u_j - (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2} \text{sign}(u_j)], 1 \right).$$

We now evaluate the worst-case coverage probability of the confidence interval

$$[\widehat{\tau}_{\text{ST}} - L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0, \widehat{\tau}_{\text{ST}} + L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0].$$

Define $\mathbf{u}' \in \mathbb{R}^K$ coordinate-wise by $u'_j = [u_j - (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2} \text{sign}(u_j)] \mathbf{1}(|u_j| > (1 + \gamma_j^{-1})^{1/2} c_{\alpha/2})$ for $j = 1, \dots, K$. We have

$$\begin{aligned} & P_{\boldsymbol{\Delta}/\sigma_0=\mathbf{t}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0) \\ &= \int_{\mathbb{R}^K} P_{\boldsymbol{\Delta}/\sigma_0=\mathbf{t}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0 \mid (\widehat{\tau}_1 - \widehat{\tau}_0, \dots, \widehat{\tau}_K - \widehat{\tau}_0)^\top / \sigma_0 = \mathbf{u}) \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u} \\ &= \int_{\mathbb{R}^K} P \left(\left| N \left(\frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}}, 1 \right) \right| \leq L \right) \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u} \\ &= \int_{\mathbb{R}^K} \left[\Phi \left(L - \frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) - \Phi \left(-L - \frac{\langle \boldsymbol{\gamma}, \mathbf{t} - \mathbf{u}' \rangle}{\sqrt{1 + \|\boldsymbol{\gamma}\|_1}} \right) \right] \phi_{\mathbf{t}, \mathbf{V}}(\mathbf{u}) d\mathbf{u}, \end{aligned}$$

which yields exactly the expression in the theorem.

Let $\mathbf{t} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Delta}$ and denote by

$$p(\mathbf{t}) := P_{\boldsymbol{\Delta}/\sigma_0=\mathbf{t}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0).$$

Analogous to the proof of Theorem 1.3.3, we can show that $p(\mathbf{t})$ is symmetric about $\mathbf{t} = \mathbf{0}$ and monotonically decreasing in $|t_j|$ for all $j = 1, \dots, K$. Therefore, the worst-case coverage probability is attained at $\mathbf{t} = \mathbf{b}$.

Finally, choosing $\widehat{L}_{\text{ST}}$ to satisfy

$$P_{\boldsymbol{\Delta}/\sigma_0=\mathbf{b}}(|\widehat{\tau}_{\text{ST}} - \tau| \leq L(1 + \|\boldsymbol{\gamma}\|_1)^{-1/2} \sigma_0) = 1 - \zeta$$

completes the proof. □

A.3 Proofs of additional results

A.3.1 Proof of Lemma 1.3.1 and Lemma 1.3.2

Proof of Lemma 1.3.1 and Lemma 1.3.2. By Assumption 1.2.1,

$$\begin{pmatrix} \widehat{\tau}_0 - \tau \\ \widehat{\tau}_1 - \tau \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ \Delta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} \right).$$

Hence, by linearity of multivariate normals,

$$\begin{pmatrix} \hat{\tau}_0 - \tau \\ \hat{\tau}_1 - \hat{\tau}_0 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ \Delta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & -\sigma_0^2 \\ -\sigma_0^2 & \sigma_0^2 + \sigma_1^2 \end{pmatrix} \right).$$

Another linear transformation yields

$$\begin{pmatrix} \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \tau \\ \hat{\tau}_1 - \hat{\tau}_0 \end{pmatrix} \sim N \left(\begin{pmatrix} \frac{\gamma}{1+\gamma}\Delta \\ \Delta \end{pmatrix}, \begin{pmatrix} \frac{1}{1+\gamma}\sigma_0^2 & 0 \\ 0 & \sigma_0^2 + \sigma_1^2 \end{pmatrix} \right).$$

Therefore, there exist independent standard normals $\mathcal{Z}_1, \mathcal{Z}_2$ such that

$$\hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \tau \stackrel{d}{=} \frac{\gamma}{1+\gamma}\Delta + \frac{1}{\sqrt{1+\gamma}}\sigma_0\mathcal{Z}_1, \quad \hat{\tau}_1 - \hat{\tau}_0 \stackrel{d}{=} \Delta + \sigma\mathcal{Z}_2.$$

Next, by the definitions of $\hat{\tau}_{PT}$ and $\hat{\tau}_{ST}$,

$$\begin{aligned} \hat{\tau}_{PT} &= \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}) \\ &= \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}), \end{aligned}$$

and

$$\begin{aligned} \hat{\tau}_{ST} &= \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}) + \frac{\gamma}{1+\gamma}\sigma c_{\alpha/2} \text{sign}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}) \\ &= \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \frac{\gamma}{1+\gamma}[(\hat{\tau}_1 - \hat{\tau}_0) - \sigma c_{\alpha/2} \text{sign}(\hat{\tau}_1 - \hat{\tau}_0)]\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}). \end{aligned}$$

For the pretest estimator,

$$\begin{aligned} \hat{\tau}_{PT} - \tau &= \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \tau - \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0)\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}) \\ &\stackrel{d}{=} \frac{\gamma}{1+\gamma}\Delta + \frac{1}{\sqrt{1+\gamma}}\sigma_0\mathcal{Z}_1 - \frac{\gamma}{1+\gamma}(\Delta + \sigma\mathcal{Z}_2)\mathbf{1}\left(\left|\mathcal{Z}_2 + \frac{\Delta}{\sigma}\right| > c_{\alpha/2}\right) \\ &= \frac{1}{\sqrt{1+\gamma}}\sigma_0\mathcal{Z}_1 + \frac{\gamma}{1+\gamma}\Delta\mathbf{1}\left(\left|\mathcal{Z}_2 + \sqrt{\frac{\gamma}{1+\gamma}}\frac{\Delta}{\sigma_0}\right| \leq c_{\alpha/2}\right) \\ &\quad - \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0\mathcal{Z}_2\mathbf{1}\left(\left|\mathcal{Z}_2 + \sqrt{\frac{\gamma}{1+\gamma}}\frac{\Delta}{\sigma_0}\right| > c_{\alpha/2}\right), \end{aligned}$$

which proves Lemma 1.3.1.

Similarly, for the soft-thresholding estimator,

$$\begin{aligned} \hat{\tau}_{ST} - \tau &= \hat{\tau}_0 + \frac{\gamma}{1+\gamma}(\hat{\tau}_1 - \hat{\tau}_0) - \frac{\gamma}{1+\gamma}[(\hat{\tau}_1 - \hat{\tau}_0) - \sigma c_{\alpha/2} \text{sign}(\hat{\tau}_1 - \hat{\tau}_0)]\mathbf{1}(|\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}) \\ &\stackrel{d}{=} \frac{\gamma}{1+\gamma}\Delta + \frac{1}{\sqrt{1+\gamma}}\sigma_0\mathcal{Z}_1 - \frac{\gamma}{1+\gamma}[(\Delta + \sigma\mathcal{Z}_2) - \sigma c_{\alpha/2} \text{sign}(\Delta + \sigma\mathcal{Z}_2)]\mathbf{1}\left(\left|\mathcal{Z}_2 + \frac{\Delta}{\sigma}\right| > c_{\alpha/2}\right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{1+\gamma}}\sigma_0\mathcal{Z}_1 + \frac{\gamma}{1+\gamma}\Delta\mathbb{1}\left(\left|\mathcal{Z}_2 + \sqrt{\frac{\gamma}{1+\gamma}}\frac{\Delta}{\sigma_0}\right| \leq c_{\alpha/2}\right) \\
&\quad - \sqrt{\frac{\gamma}{1+\gamma}}\sigma_0\left(\mathcal{Z}_2 - c_{\alpha/2}\operatorname{sign}\left(\mathcal{Z}_2 + \sqrt{\frac{\gamma}{1+\gamma}}\frac{\Delta}{\sigma_0}\right)\right)\mathbb{1}\left(\left|\mathcal{Z}_2 + \sqrt{\frac{\gamma}{1+\gamma}}\frac{\Delta}{\sigma_0}\right| > c_{\alpha/2}\right),
\end{aligned}$$

which proves Lemma 1.3.2. $\square$

A.3.2 Proof of Theorem A.1.1

Proof of Theorem A.1.1. By Theorem 1.3.1, the shortest symmetric centered confidence interval based on $\hat{\tau}_{\text{PW}}$ is given by $[\hat{\tau}_{\text{PW}} - \hat{L}_{\text{PW}}(1+\gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{PW}} + \hat{L}_{\text{PW}}(1+\gamma)^{-1/2}\sigma_0]$, where $\hat{L}_{\text{PW}}$ is the unique solution to

$$\Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}b\right) - \Phi\left(-L - \frac{\gamma}{\sqrt{1+\gamma}}b\right) = 1 - \zeta.$$

By the definition of the b-value,

$$\begin{aligned}
b_{\text{PW}}^*(\zeta) &= \inf\left\{b \geq 0 : 0 \in [\hat{\tau}_{\text{PW}} - \hat{L}_{\text{PW}}(1+\gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{PW}} + \hat{L}_{\text{PW}}(1+\gamma)^{-1/2}\sigma_0]\right\} \\
&= \inf\left\{b \geq 0 : \hat{L}_{\text{PW}} \geq \frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}\right\}
\end{aligned}$$

Now evaluate the confidence interval at $b = 0$. In this case $\hat{L}_{\text{PW}}(0)$ solves $\Phi(L) - \Phi(-L) = 1 - \zeta$. If

$$\Phi\left(\frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}\right) - \Phi\left(-\frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}\right) < 1 - \zeta,$$

then $\hat{L}_{\text{PW}}(0) > \frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}$ and therefore $b_{\text{PW}}^*(\zeta) = 0$.

Otherwise,

$$\Phi\left(\frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}\right) - \Phi\left(-\frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}\right) \geq 1 - \zeta,$$

so $b_{\text{PW}}^*(\zeta) \in (0, \infty)$. By definition of the infimum, at $b = b_{\text{PW}}^*(\zeta)$ we must have the boundary condition

$$\hat{L}_{\text{PW}}(b_{\text{PW}}^*(\zeta)) = \frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}.$$

Plugging $L = \frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0}$ into the defining equation of $\hat{L}_{\text{PW}}$ yields that $b = b_{\text{PW}}^*(\zeta)$ solves

$$\Phi\left(\frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1+\gamma}}b\right) - \Phi\left(-\frac{|\hat{\tau}_{\text{PW}}|}{(1+\gamma)^{-1/2}\sigma_0} - \frac{\gamma}{\sqrt{1+\gamma}}b\right) = 1 - \zeta,$$

which is the desired characterization. $\square$

A.3.3 Proof of Theorem A.1.2

Proof of Theorem A.1.2. By Theorem 1.3.2, the shortest symmetric centered confidence interval based on $\hat{\tau}_{PT}$ is given by $[\hat{\tau}_{PT} - \hat{L}_{PT}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1 + \gamma)^{-1/2}\sigma_0]$, where $\hat{L}_{PT}$ is the solution to

$$\min_{0 \leq t \leq b} P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta.$$

By the definition of the b-value,

$$\begin{aligned} b_{PT}^*(\zeta) &= \inf \left\{ b \geq 0 : 0 \in [\hat{\tau}_{PT} - \hat{L}_{PT}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{PT} + \hat{L}_{PT}(1 + \gamma)^{-1/2}\sigma_0] \right\} \\ &= \inf \left\{ b \geq 0 : \hat{L}_{PT} \geq \frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0} \right\} \end{aligned}$$

At $b = 0$, the defining equation for $\hat{L}_{PT}$ becomes $P_{\Delta/\sigma_0=0}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta$. If

$$P_{\Delta/\sigma_0=0}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| |\hat{\tau}_{PT}|) < 1 - \zeta,$$

then $\hat{L}_{PT}(0) > \frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0}$ and therefore $b_{PT}^*(\zeta) = 0$.

At $b = \infty$, the defining equation for $\hat{L}_{PT}$ becomes $\min_{t \geq 0} P_{\Delta/\sigma_0=t}(|\hat{\tau}_{PT} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta$. If

$$\min_{t \geq 0} P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| |\hat{\tau}_{PT}|) > 1 - \zeta,$$

then $\hat{L}_{PT}(\infty) < \frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0}$ and therefore $b_{PT}^*(\zeta) = \infty$.

Otherwise,

$$\min_{t \geq 0} P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| |\hat{\tau}_{PT}|) \leq 1 - \zeta, \quad \text{and} \quad P_{\Delta/\sigma_0=0}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| |\hat{\tau}_{PT}|) \geq 1 - \zeta,$$

so $b_{PW}^*(\zeta) \in (0, \infty)$. By definition of the infimum, at $b = b_{PT}^*(\zeta)$ we must have the boundary condition

$$\hat{L}_{PT}(b_{PT}^*(\zeta)) = \frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0}.$$

Plugging $L = \frac{|\hat{\tau}_{PT}|}{(1 + \gamma)^{-1/2}\sigma_0}$ into the defining equation of $\hat{L}_{PT}$ yields that $b = b_{PT}^*(\zeta)$ solves

$$\min_{0 \leq t \leq b} P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| |\hat{\tau}_{PT}|) = 1 - \zeta,$$

which is the desired characterization.

The explicit form of $P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{PT} - \tau| \leq |\hat{\tau}_{PT}| |\hat{\tau}_{PT}|)$ is directly given by Theorem 1.3.2. $\square$

A.3.4 Proof of Theorem A.1.3

Proof of Theorem A.1.3. By Theorem 1.3.3, the shortest symmetric centered confidence interval based on $\hat{\tau}_{\text{ST}}$ is given by $[\hat{\tau}_{\text{ST}} - \hat{L}_{\text{ST}}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{ST}} + \hat{L}_{\text{ST}}(1 + \gamma)^{-1/2}\sigma_0]$, where $\hat{L}_{\text{ST}}$ is the solution to

$$P_{\Delta/\sigma_0=b}(|\hat{\tau}_{\text{ST}} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta.$$

By the definition of the b-value,

$$\begin{aligned} b_{\text{ST}}^*(\zeta) &= \inf \left\{ b \geq 0 : 0 \in [\hat{\tau}_{\text{ST}} - \hat{L}_{\text{ST}}(1 + \gamma)^{-1/2}\sigma_0, \hat{\tau}_{\text{ST}} + \hat{L}_{\text{ST}}(1 + \gamma)^{-1/2}\sigma_0] \right\} \\ &= \inf \left\{ b \geq 0 : \hat{L}_{\text{ST}} \geq \frac{|\hat{\tau}_{\text{ST}}|}{(1 + \gamma)^{-1/2}\sigma_0} \right\} \end{aligned}$$

At $b = 0$, the defining equation for $\hat{L}_{\text{ST}}$ becomes $P_{\Delta/\sigma_0=0}(|\hat{\tau}_{\text{ST}} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta$. If

$$P_{\Delta/\sigma_0=0}(|\tilde{\tau}_{\text{ST}} - \tau| \leq |\hat{\tau}_{\text{ST}}| \mid \hat{\tau}_{\text{ST}}) < 1 - \zeta,$$

then $\hat{L}_{\text{ST}}(0) > \frac{|\hat{\tau}_{\text{ST}}|}{(1 + \gamma)^{-1/2}\sigma_0}$ and therefore $b_{\text{ST}}^*(\zeta) = 0$.

At $b = \infty$, the defining equation for $\hat{L}_{\text{ST}}$ becomes $P_{\Delta/\sigma_0=\infty}(|\hat{\tau}_{\text{ST}} - \tau| \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta$. If

$$P_{\Delta/\sigma_0=\infty}(|\tilde{\tau}_{\text{ST}} - \tau| \leq |\hat{\tau}_{\text{ST}}| \mid \hat{\tau}_{\text{ST}}) > 1 - \zeta,$$

then $\hat{L}_{\text{ST}}(\infty) < \frac{|\hat{\tau}_{\text{ST}}|}{(1 + \gamma)^{-1/2}\sigma_0}$ and therefore $b_{\text{ST}}^*(\zeta) = \infty$.

Otherwise,

$$P_{\Delta/\sigma_0=\infty}(|\tilde{\tau}_{\text{ST}} - \tau| \leq |\hat{\tau}_{\text{ST}}| \mid \hat{\tau}_{\text{ST}}) \leq 1 - \zeta, \quad \text{and} \quad P_{\Delta/\sigma_0=0}(|\tilde{\tau}_{\text{ST}} - \tau| \leq |\hat{\tau}_{\text{ST}}| \mid \hat{\tau}_{\text{ST}}) \geq 1 - \zeta,$$

so $b_{\text{ST}}^*(\zeta) \in (0, \infty)$. By definition of the infimum, at $b = b_{\text{ST}}^*(\zeta)$ we must have the boundary condition

$$\hat{L}_{\text{ST}}(b_{\text{ST}}^*(\zeta)) = \frac{|\hat{\tau}_{\text{ST}}|}{(1 + \gamma)^{-1/2}\sigma_0}.$$

Plugging $L = \frac{|\hat{\tau}_{\text{ST}}|}{(1 + \gamma)^{-1/2}\sigma_0}$ into the defining equation of $\hat{L}_{\text{ST}}$ yields that $b = b_{\text{ST}}^*(\zeta)$ solves

$$P_{\Delta/\sigma_0=b}(|\tilde{\tau}_{\text{ST}} - \tau| \leq |\hat{\tau}_{\text{ST}}| \mid \hat{\tau}_{\text{ST}}) = 1 - \zeta,$$

which is the desired characterization.

The explicit form of $P_{\Delta/\sigma_0=t}(|\tilde{\tau}_{\text{ST}} - \tau| \leq |\hat{\tau}_{\text{ST}}| \mid \hat{\tau}_{\text{ST}})$ is directly given by Theorem 1.3.3. $\square$

A.3.5 Proof of Theorem A.1.4

Proof of Theorem A.1.4. We follow the argument in the proof of Theorem 1.3.1. From that proof, we know that

$$(1 + \gamma)^{1/2} \sigma_0^{-1} (\hat{\tau}_{\text{PW}} - \tau) \sim N \left(\frac{\gamma}{\sqrt{1 + \gamma}} \frac{\Delta}{\sigma_0}, 1 \right).$$

Consider the lower confidence bound

$$[\hat{\tau}_{\text{PW}} - L(1 + \gamma)^{-1/2} \sigma_0, \infty).$$

The worst-case coverage probability of the lower confidence bound is

$$\begin{aligned} & \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta(\tau \in [\hat{\tau}_{\text{PW}} - L(1 + \gamma)^{-1/2} \sigma_0, \infty)) \\ &= \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta(\hat{\tau}_{\text{PW}} - \tau \leq L(1 + \gamma)^{-1/2} \sigma_0) = \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta((1 + \gamma)^{1/2} \sigma_0^{-1} (\hat{\tau}_{\text{PW}} - \tau) \leq L) \\ &= \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_\Delta \left(N \left(\frac{\gamma}{\sqrt{1 + \gamma}} \frac{\Delta}{\sigma_0}, 1 \right) \leq L \right) = \mathbb{P} \left(N \left(\sup_{\Delta: |\Delta/\sigma_0| \leq b} \frac{\gamma}{\sqrt{1 + \gamma}} \frac{\Delta}{\sigma_0}, 1 \right) \leq L \right) \\ &= \mathbb{P} \left(N \left(\frac{\gamma}{\sqrt{1 + \gamma}} b, 1 \right) \leq L \right) = \Phi \left(L - \frac{\gamma}{\sqrt{1 + \gamma}} b \right). \end{aligned}$$

Therefore, the proof is complete since $\hat{L}'_{\text{PW}} = c_\zeta + \frac{\gamma}{\sqrt{1 + \gamma}} b$ is the solution to

$$\Phi \left(L - \frac{\gamma}{\sqrt{1 + \gamma}} b \right) = 1 - \zeta.$$

□

A.3.6 Proof of Theorem A.1.5

Proof of Theorem A.1.5. We follow the argument in the proof of Theorem 1.3.2. From that proof, we know that

$$(1 + \gamma)^{1/2} \sigma_0^{-1} (\hat{\tau}_{\text{PT}} - \tau) \mid \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\} \sim N \left(\frac{\gamma}{\sqrt{1 + \gamma}} \frac{\Delta}{\sigma_0}, 1 \right),$$

and

$$(1 + \gamma)^{1/2} \sigma_0^{-1} (\hat{\tau}_{\text{PT}} - \tau) \mid \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\} \sim N \left(\frac{\gamma}{\sqrt{1 + \gamma}} \frac{\Delta}{\sigma_0} - \sqrt{\gamma} u, 1 \right).$$

Consider the lower confidence bound

$$[\hat{\tau}_{\text{PT}} - L(1 + \gamma)^{-1/2} \sigma_0, \infty).$$

The worst-case coverage probability of the lower confidence bound is

$$\inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(\tau \in [\hat{\tau}_{PT} - L(1+\gamma)^{-1/2}\sigma_0, \infty)) = \inf_{\Delta: |\Delta/\sigma_0| \leq b} P_{\Delta}(\hat{\tau}_{PT} - \tau \leq L(1+\gamma)^{-1/2}\sigma_0).$$

For a fixed $t = \Delta/\sigma_0$, we decompose the probability according to whether the pretest accepts or rejects:

$$\begin{aligned} & P_{\Delta/\sigma_0=t}(\hat{\tau}_{PT} - \tau \leq L(1+\gamma)^{-1/2}\sigma_0) = P_{\Delta/\sigma_0=t}((1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau) \leq L) \\ & = P_{\Delta/\sigma_0=t}((1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau) \leq L, |\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}) \\ & \quad + P_{\Delta/\sigma_0=t}((1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{PT} - \tau) \leq L, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}) \\ & = P\left(N\left(\frac{\gamma}{\sqrt{1+\gamma}}t, 1\right) \leq L\right) P\left(\left|N\left(\sqrt{\frac{\gamma}{1+\gamma}}t, 1\right)\right| \leq c_{\alpha/2}\right) \\ & \quad + E_{U \sim N(\sqrt{\frac{\gamma}{1+\gamma}}t, 1)}\left[P\left(N\left(\frac{\gamma}{\sqrt{1+\gamma}}t - \sqrt{\gamma}U, 1\right) \leq L\right) \mathbb{1}(|U| > c_{\alpha/2})\right] \\ & = \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right)\right] \Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \\ & \quad + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \Phi\left(L + \sqrt{\gamma}u\right) \phi(u) du + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \Phi\left(L + \sqrt{\gamma}u\right) \phi(u) du. \end{aligned}$$

Finally, the worst-case coverage probability over $|\Delta/\sigma_0| \leq b$ is attained for some $t \in [0, b]$. Choosing $\hat{L}'_{PT}$ to satisfy

$$\min_{0 \leq t \leq b} P_{\Delta/\sigma_0=t}(\hat{\tau}_{PT} - \tau \leq L(1+\gamma)^{-1/2}\sigma_0) = 1 - \zeta$$

completes the proof. □

A.3.7 Proof of Theorem A.1.6

Proof of Theorem A.1.6. We follow the argument in the proof of Theorem 1.3.3. From that proof, we know that

$$(1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{ST} - \tau) | \{|\hat{\tau}_1 - \hat{\tau}_0| \leq \sigma c_{\alpha/2}\} \sim N\left(\frac{\gamma}{\sqrt{1+\gamma}}\frac{\Delta}{\sigma_0}, 1\right),$$

and

$$\begin{aligned} & (1+\gamma)^{1/2}\sigma_0^{-1}(\hat{\tau}_{ST} - \tau) | \{\sigma^{-1}(\hat{\tau}_1 - \hat{\tau}_0) = u, |\hat{\tau}_1 - \hat{\tau}_0| > \sigma c_{\alpha/2}\} \\ & \sim N\left(\frac{\gamma}{\sqrt{1+\gamma}}\frac{\Delta}{\sigma_0} - \sqrt{\gamma}[u - c_{\alpha/2} \text{sign}(u)], 1\right). \end{aligned}$$

Consider the lower confidence bound

$$[\widehat{\tau}_{\text{ST}} - L(1 + \gamma)^{-1/2}\sigma_0, \infty).$$

The worst-case coverage probability of the lower confidence bound is

$$\inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_{\Delta}(\tau \in [\widehat{\tau}_{\text{ST}} - L(1 + \gamma)^{-1/2}\sigma_0, \infty)) = \inf_{\Delta: |\Delta/\sigma_0| \leq b} \mathbb{P}_{\Delta}(\widehat{\tau}_{\text{ST}} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0).$$

For a fixed $t = \Delta/\sigma_0$, we decompose the probability according to whether the pretest accepts or rejects:

$$\begin{aligned} & \mathbb{P}_{\Delta/\sigma_0=t}(\widehat{\tau}_{\text{ST}} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0) = \mathbb{P}_{\Delta/\sigma_0=t}((1 + \gamma)^{1/2}\sigma_0^{-1}(\widehat{\tau}_{\text{ST}} - \tau) \leq L) \\ &= \mathbb{P}_{\Delta/\sigma_0=t}((1 + \gamma)^{1/2}\sigma_0^{-1}(\widehat{\tau}_{\text{ST}} - \tau) \leq L, |\widehat{\tau}_1 - \widehat{\tau}_0| \leq \sigma c_{\alpha/2}) \\ & \quad + \mathbb{P}_{\Delta/\sigma_0=t}((1 + \gamma)^{1/2}\sigma_0^{-1}(\widehat{\tau}_{\text{ST}} - \tau) \leq L, |\widehat{\tau}_1 - \widehat{\tau}_0| > \sigma c_{\alpha/2}) \\ &= \mathbb{P}\left(N\left(\frac{\gamma}{\sqrt{1+\gamma}}t, 1\right) \leq L\right) \mathbb{P}\left(\left|N\left(\sqrt{\frac{\gamma}{1+\gamma}}t, 1\right)\right| \leq c_{\alpha/2}\right) \\ & \quad + \mathbb{E}_{U \sim N(\sqrt{\frac{\gamma}{1+\gamma}}t, 1)}\left[\mathbb{P}\left(N\left(\frac{\gamma}{\sqrt{1+\gamma}}t - \sqrt{\gamma}[U - c_{\alpha/2} \text{sign}(U)], 1\right) \leq L\right) \mathbb{1}(|U| > c_{\alpha/2})\right] \\ &= \left[\Phi\left(c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right) - \Phi\left(-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t\right)\right] \Phi\left(L - \frac{\gamma}{\sqrt{1+\gamma}}t\right) \\ & \quad + \int_{-\infty}^{-c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t} \Phi\left(L + \sqrt{\gamma}(u + c_{\alpha/2})\right) \phi(u) du + \int_{c_{\alpha/2} - \sqrt{\frac{\gamma}{1+\gamma}}t}^{\infty} \Phi\left(L + \sqrt{\gamma}(u - c_{\alpha/2})\right) \phi(u) du. \end{aligned}$$

Finally, since the coverage probability is monotonically decreasing in Δ , the worst case over $|\Delta/\sigma_0| \leq b$ is attained at $\Delta/\sigma_0 = b$. Choosing $\widehat{L}'_{\text{ST}}$ to satisfy

$$\mathbb{P}_{\Delta/\sigma_0=b}(\widehat{\tau}_{\text{ST}} - \tau \leq L(1 + \gamma)^{-1/2}\sigma_0) = 1 - \zeta$$

completes the proof. □

Appendix B

Appendix for Chapter 2

In Section B.1.1, we discuss the relationship between full OLS, marginal OLS and weighted least squares. We also provide an alternative OLS regression formulation that is equivalent to the full OLS regression.

In Section B.1.2, we discuss the extension to the regression with interaction terms.

In Section B.1.3, we discuss the extension to the general exposure mapping.

In Section B.2, we provide the lemmas for the main results and their proofs.

In Section B.3, we provide the proofs of all the results.

Additional notation. For any $a, b \in \mathbb{R}$, write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any two real sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \lesssim b_n$ (or equivalently, $b_n \gtrsim a_n$) if $a_n = O(b_n)$.

B.1 Additional results

B.1.1 Relationship between full OLS, marginal OLS and weighted least squares

Although the OLS procedure in (2.2.3) involves regressing Y_t on the full set of transformed treatment indicators $\tilde{Z}_t, \dots, \tilde{Z}_{t-K}$, all of our results remain valid when regressing Y_t on a subset of these transformed treatment indicators because the transformed regressors are jointly independent. Suppose we are interested in estimating the treatment effect at lag k . There are two natural approaches. One approach is the full OLS in (2.2.3), where we regress Y_t on $\tilde{Z}_t, \dots, \tilde{Z}_{t-K}$ to obtain the estimate $\tilde{\tau}_k$, and then apply the linear transformation (2.2.4) to obtain the estimate $\hat{\tau}_k = w_k^{-1} \tilde{\tau}_k$. The other approach is the marginal OLS, where we regress Y_t on $\tilde{Z}_{t-k}$, i.e.,

$$\text{lm}(Y_t \sim \tilde{Z}_{t-k}), \quad (\text{B.1.1})$$

to obtain the estimate $\tilde{\tau}_{\text{marginal},k}$ and then apply the linear transformation (2.2.4) to obtain the estimate $\hat{\tau}_{\text{marginal},k} = w_k^{-1} \tilde{\tau}_{\text{marginal},k}$. Both approaches yield consistent and asymptotic

normal estimators ($\hat{\tau}_k$ and $\hat{\tau}_{\text{marginal},k}$) for τ_k with conservative variance estimators by HAC variance estimation. The key differences lie in their efficiency and variance estimation. First, the two approaches have different asymptotic variances. Second, the two approaches have different variance estimators, since the variance estimator in the marginal OLS is based on noisier residuals, which inflates the standard errors. As a direct result from Theorem 2.3.3, both approaches share the same bias term $\mathbf{B}_K$ in HAC variance estimation. Thus, comparing the HAC variance estimators $\hat{\mathbf{V}}$ reduces to comparing the asymptotic variances $\mathbf{V}$, since the two approaches have the same level of conservativeness. See Appendix for more technical details on this discussion. Therefore, we focus on comparing the asymptotic variances below.

By Theorem 2.3.2, the asymptotic variance from the full OLS (2.2.3) is $(T - K)^{-1} \text{Var}[\sum_{t=K+1}^T \tilde{Z}_{t-k}(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell)]$, while the asymptotic variance from the marginal OLS (B.1.1) is $(T - K)^{-1} \text{Var}[\sum_{t=K+1}^T \tilde{Z}_{t-k}(Y_t - \tilde{Z}_{t-k} w_k \tau_k)]$. Intuitively, the asymptotic variance depends on the residual $Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell$ for full OLS and $Y_t - \tilde{Z}_{t-k} w_k \tau_k$ for marginal OLS. By adding more regressors in full OLS, the residual should have smaller sum of squares, which implies the asymptotic variance will be smaller. To compare these asymptotic variances explicitly, we consider the linear model on the potential outcomes and establish the following result.

Proposition B.1.1. *Consider a linear model with homogeneous effects and nonrandom error: $Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \beta_k z_{t-k} + \epsilon_t$. Assume a constant treatment probability $p_t = p$ for all t . Then, the asymptotic variance of $\hat{\tau}_k$ from the full OLS is*

$$\frac{1}{T - K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \beta_\ell^2 + \frac{1}{T - K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right)^2,$$

while the asymptotic variance of $\hat{\tau}_{\text{marginal},k}$ from the marginal OLS is

$$\frac{1}{T - K} \sum_{t=K+1}^T \sum_{\ell=0, \ell \neq k}^{t-1} \beta_\ell^2 + \frac{1}{T - K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right)^2.$$

By Proposition B.1.1, under the linear model with homogeneous effects and nonrandom error, full OLS has a smaller asymptotic variance than marginal OLS. The variance difference is $(T - K)^{-1} \sum_{t=K+1}^T \sum_{\ell=0, \ell \neq k}^K \beta_\ell^2 = \sum_{\ell=0, \ell \neq k}^K \beta_\ell^2$ which is typically large, as smaller lags often have strong effects. Furthermore, full OLS achieves lower variance unless $\beta_\ell = 0$ for all $\ell > K$. However, including too many lags may lead to high model complexity. Therefore, it is crucial for practitioners to carefully choose the number of lags in the model. This selection can be guided by standard model selection techniques from the linear regression literature.

Remark B.1.1. Gao and Ding (2025) consider the network experiment setting and estimate the average of potential outcomes separately for each group. Compared with the weighted least squares (WLS) approach in Gao and Ding (2025), our method differs in two key ways. First, we estimate the causal effect directly rather than estimating the average potential

outcomes. Second, we normalize the treatment indicator directly, instead of applying weights to each observation with an unnormalized treatment indicator. Our OLS approach allows us to estimate treatment effects across multiple lags simultaneously. Since treatment indicators at different time points should receive different weights due to variations in propensity scores, and each observation contains treatment indicators from multiple time points, assigning a single weight per observation is not feasible. However, the marginal OLS estimator (B.1.1) is asymptotically equivalent to a version of WLS analogous to Gao and Ding (2025). If further the propensity scores are constant, the two estimators coincide.

B.1.1.1 Alternative OLS

Equivalently, from (2.2.3) and (2.2.4), $\hat{\tau}$ can be obtained directly from the following OLS regression without an intercept:

$$\mathbf{lm}(Y_t \sim w_0 \tilde{Z}_t + w_1 \tilde{Z}_{t-1} + \dots + w_K \tilde{Z}_{t-K}). \quad (\text{B.1.2})$$

In this formulation, the outcome vector remains $\mathbf{Y}_K$, but the regressor matrix is rescaled to $\mathbf{Z}_K \mathbf{W}_K$. The only difference lies in whether the scaling is applied before or after the OLS regression. Thus, the regression in (B.1.2) yields the same estimator as in (2.2.3) after applying the linear transformation (2.2.4), and both procedures lead to the same HAC variance estimator. When the p_t 's are constant across time $t = 1, \dots, T$, all treatments share the same normalization with the same centering and scaling. In this case, there is no need for the normalization and the OLS fit is equivalent to regressing the observed outcomes on unnormalized treatment indicators with intercepts: $\mathbf{lm}(Y_t \sim 1 + Z_t + Z_{t-1} + \dots + Z_{t-K})$ using samples with $t = K + 1, \dots, T$.

B.1.1.2 Full OLS and marginal OLS

By Theorem 2.3.2, the full OLS estimator satisfies the following CLT:

$$\sqrt{T - K} \mathbf{V}_{k+1, k+1}^{-1/2} (\hat{\tau}_k - \tau_k) \xrightarrow{d} N(0, 1),$$

where $\mathbf{V}$ denotes the asymptotic variance matrix of the full OLS estimator defined in (2.3.5). The corresponding asymptotic variance component for lag k can be written as

$$\mathbf{V}_{k+1, k+1} = \frac{1}{T - K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} U_t \right] = \frac{1}{T - K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right].$$

Define the asymptotic variance of the marginal OLS estimator as

$$V_{\text{marginal}, k} = \frac{1}{T - K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} (Y_t - \tilde{Z}_{t-k} w_k \tau_k) \right].$$

Analogous to Theorem 2.3.2, the marginal OLS estimator also satisfies the following CLT:

$$\sqrt{T-K} V_{\text{marginal},k}^{-1/2} (\hat{\tau}_k - \tau_k) \xrightarrow{d} N(0, 1).$$

Both $\mathbf{V}$ and $V_{\text{marginal},k}$ can be estimated using the HAC variance estimator. By Theorem 2.3.3, we have for the full OLS estimator:

$$\|\hat{\mathbf{V}} - \mathbf{V} - \mathbf{B}_K\|_F \xrightarrow{p} 0.$$

Let $b_k \in \mathbb{R}^{T-K}$ be defined by

$$[b_k]_{t-K} = \tau_{t,k} - \frac{w_k}{p_{t-k}(1-p_{t-k})} \tau_k, \quad t = K+1, \dots, T.$$

Then $[\mathbf{B}_K]_{k+1,k+1} = b_k^\top \mathbf{Q}_K b_k$, which implies for the full OLS estimator:

$$|\hat{\mathbf{V}}_{k+1,k+1} - \mathbf{V}_{k+1,k+1} - b_k^\top \mathbf{Q}_K b_k| \xrightarrow{p} 0.$$

Similarly, let $\hat{\mathbf{V}}_{\text{marginal},k}$ denote the HAC variance estimator for the marginal OLS. Then for the marginal OLS estimator, following the proof of Theorem 2.3.3, we have

$$|\hat{\mathbf{V}}_{\text{marginal},k} - V_{\text{marginal},k} - b_k^\top \mathbf{Q}_K b_k| \xrightarrow{p} 0,$$

since by the definition of b_k , for $t = K+1, \dots, T$,

$$\mathbb{E}[\tilde{Z}_{t-k}(Y_t - \tilde{Z}_{t-k} w_k \tau_k)] = \tau_{t,k} - \frac{w_k}{p_{t-k}(1-p_{t-k})} \tau_k = [b_k]_{t-K}.$$

Hence, both $\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}_{\text{marginal},k}$ share the same bias term $b_k^\top \mathbf{Q}_K b_k$ when estimating $\mathbf{V}_{k+1,k+1}$ and $V_{\text{marginal},k}$, respectively. The comparison between full and marginal OLS variance estimators therefore reduces to their asymptotic true variances.

B.1.1.3 Marginal OLS and WLS

Assume without loss of generality that we are interested in the treatment effect at lag 0. To directly estimate the causal effect, we center the treatment and assign a weight of $1/[p_t(1-p_t)]$ to each time point t . For comparison, Gao and Ding (2025) did not center the treatment and use weight $1/p_t$ and $1/(1-p_t)$ since they estimated the average potential outcomes for each group separately. The only distinction between marginal OLS and WLS is that marginal OLS applies weights to the treatment indicators, whereas WLS applies weights to the observations. Let $\tilde{\mathbf{Z}} = (Z_1 - p_1, \dots, Z_T - p_T)^\top$, $\tilde{\mathbf{W}} = \text{diag}(1/[p_t(1-p_t)], t = 1, \dots, T)$, and $\tilde{\mathbf{Y}} = (Y_1, \dots, Y_T)^\top$. Then the WLS estimator can be written as

$$(\tilde{\mathbf{Z}}^\top \tilde{\mathbf{W}} \tilde{\mathbf{Z}})^{-1} (\tilde{\mathbf{Z}}^\top \tilde{\mathbf{W}} \tilde{\mathbf{Y}}) = \left(\frac{1}{T} \sum_{t=1}^T \frac{(Z_t - p_t)^2}{p_t(1-p_t)} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{Z}_t Y_t \right).$$

In contrast, the marginal OLS estimator can be written as

$$\hat{\tau} = \left(\left(\frac{1}{T} \sum_{t=1}^T \frac{1}{p_t(1-p_t)} \right)^{-1} \frac{1}{T} \sum_{t=1}^T \frac{(Z_t - p_t)^2}{p_t^2(1-p_t)^2} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \tilde{Z}_t Y_t \right).$$

The numerators of these two estimators are identical, and under certain conditions, both denominators converge to 1. If the p_t values are constant, the two estimators coincide. However, when we infer treatment effects across multiple lags using full OLS, WLS cannot be generalized.

B.1.2 Regression with interaction terms

In practice, we may be interested in studying the interaction effects between treatments at different lags. In online platforms, repeated exposures (e.g., showing a user the same advertisement or recommendation on consecutive days) may produce reinforcement or fatigue, leading to nonlinear cumulative effects. In clinical and behavioral studies, the efficacy of an intervention (e.g., medication, exercise, or reminders) often depends on treatment consistency, with consecutive applications generating biological accumulation or tolerance effects. An additional advantage of the regression-based analysis is its flexibility to incorporate interaction terms. For instance, one can include regressors such as $\tilde{Z}_{t-k} \tilde{Z}_{t-k'}$ to study whether the joint presence of treatments at two different lags produces an effect beyond the sum of their individual contributions. To accommodate this possibility, we extend the OLS specification by including interaction terms between consecutive lags:

$$\text{lm}(Y_t \sim \tilde{Z}_t + \cdots + \tilde{Z}_{t-K} + \tilde{Z}_t \tilde{Z}_{t-1} + \cdots + \tilde{Z}_{t-K+1} \tilde{Z}_{t-K}).$$

We denote the corresponding estimator by $\tilde{\tau}_{\text{int}} = (\tilde{\tau}_{\text{int},0}, \dots, \tilde{\tau}_{\text{int},K}, \tilde{\tau}_{\text{int},0,1}, \dots, \tilde{\tau}_{\text{int},K-1,K})^\top \in \mathbb{R}^{2K+1}$. The extended regression specification includes $K+1$ main-effect terms corresponding to individual lags and K interaction terms between consecutive lags. In total, the model has $2K+1$ regressors. This framework can be further generalized to incorporate higher-order interactions by augmenting the regression with corresponding interaction terms, similar to regression with interactions in classic factorial experiments (Zhao and Ding, 2023). To illustrate the key idea in time-series experiments, we focus here on two-way consecutive interactions.

We recall the normalized treatment variable $\tilde{Z}_t$ defined in (2.2.2) and the main effect weights w_k defined in (2.2.5). We set $\hat{\tau}_{\text{int},k} = w_k^{-1} \tilde{\tau}_{\text{int},k}$ as (2.2.4). For the interaction terms, we introduce analogous weights that account for the joint variation in the two treatment lags. Specifically, we define

$$w_{k-1,k} = \left[\frac{1}{T-K} \sum_{t=K+1}^T (\text{Var}[Z_{t-k}] \text{Var}[Z_{t-k+1}])^{-1} \right]^{-1}$$

$$= \left[\frac{1}{T-K} \sum_{t=K+1}^T [p_{t-k} p_{t-k+1} (1-p_{t-k})(1-p_{t-k+1})]^{-1} \right]^{-1},$$

and define $\hat{\tau}_{\text{int},k-1,k} = w_{k-1,k}^{-1} \tilde{\tau}_{\text{int},k-1,k}$. Collecting the main and interaction effects, we obtain the transformed OLS estimator $\hat{\tau}_{\text{int}} = (\hat{\tau}_{\text{int},0}, \dots, \hat{\tau}_{\text{int},K}, \hat{\tau}_{\text{int},0,1}, \dots, \hat{\tau}_{\text{int},K-1,K})^\top \in \mathbb{R}^{2K+1}$. The causal estimand is $\tau_{\text{int}} = (\tau_0, \dots, \tau_K, \tau_{0,1}, \dots, \tau_{K-1,K})^\top$, which jointly captures both lagged main effects and interaction effects. The lag- k main effect τ_k is defined as (2.2.8). The interaction effect between lag $k-1$ and lag k is given by

$$\begin{aligned} \tau_{k-1,k} = \frac{1}{T-K} \sum_{t=K+1}^T & \mathbb{E}[Y_t(Z_{t:t-k+2}, 1, 1, Z_{t-k-1:1}) - Y_t(Z_{t:t-k+2}, 1, 0, Z_{t-k-1:1}) \\ & - Y_t(Z_{t:t-k+2}, 0, 1, Z_{t-k-1:1}) + Y_t(Z_{t:t-k+2}, 0, 0, Z_{t-k-1:1})], \end{aligned}$$

where the expectation is taken over the joint distribution of $(Z_{t:t-k+2}, Z_{t-k-1:1})$. Intuitively, $\tau_{k-1,k}$ measures the incremental effect of applying treatment at both lags $k-1$ and lag k simultaneously, beyond the sum of their individual effects.

To write $\hat{\tau}_{\text{int}}$ explicitly, recall the outcome vector $\mathbf{Y}_K = (Y_{K+1}, \dots, Y_T)^\top \in \mathbb{R}^{T-K}$ and define the regressor matrix $\mathbf{Z}_{\text{int},K} = (\tilde{Z}_{\text{int},K+1:1}^\top, \dots, \tilde{Z}_{\text{int},T:T-K}^\top)^\top \in \mathbb{R}^{(T-K) \times (2K+1)}$ where $\tilde{Z}_{\text{int},t:t'} = (\tilde{Z}_t, \tilde{Z}_{t-1}, \dots, \tilde{Z}_{t'}, \tilde{Z}_t \tilde{Z}_{t-1}, \dots, \tilde{Z}_{t'+1} \tilde{Z}_{t'})^\top$ for $t > t'$. Let $\mathbf{W}_{\text{int},K} \in \mathbb{R}^{(2K+1) \times (2K+1)}$ be the diagonal weight matrix $\mathbf{W}_{\text{int},K} = \text{diag}(w_k : k = 0, \dots, K, w_{k-1,k} : k = 1, \dots, K)$. Then, $\hat{\tau}_{\text{int}}$ can be expressed as

$$\hat{\tau}_{\text{int}} = \mathbf{W}_{\text{int},K}^{-1} \tilde{\tau}_{\text{int}} = \mathbf{W}_{\text{int},K}^{-1} (\mathbf{Z}_{\text{int},K}^\top \mathbf{Z}_{\text{int},K})^{-1} \mathbf{Z}_{\text{int},K}^\top \mathbf{Y}_K.$$

Analogous to the decomposition of $\hat{\tau}_k$ in (2.3.1), we can decompose $\hat{\tau}_{\text{int}}$ as

$$\hat{\tau}_{\text{int}} = \tau + \mathbf{W}_{\text{int},K}^{-1} [(T-K)^{-1} \mathbf{Z}_{\text{int},K}^\top \mathbf{Z}_{\text{int},K}]^{-1} [(T-K)^{-1} \mathbf{Z}_{\text{int},K}^\top (\mathbf{Y}_K - \mathbf{Z}_{\text{int},K} \mathbf{W}_{\text{int},K} \tau)].$$

To study the asymptotic properties of the OLS estimator $\hat{\tau}_{\text{int}}$, analogous to (2.3.4), we define the oracle residual of OLS regression at time t as

$$U_{\text{int},t} = Y_t - \sum_{k=0}^K \tilde{Z}_{t-k} w_k \tau_k - \sum_{k=1}^K \tilde{Z}_{t-k} \tilde{Z}_{t-k+1} w_{k-1,k} \tau_{k-1,k}.$$

We then define the asymptotic variance matrix $\mathbf{V}_{\text{int}} \in \mathbb{R}^{(2K+1) \times (2K+1)}$ analogous to (2.3.5) as

$$\begin{aligned} [\mathbf{V}_{\text{int}}]_{k,k'} &= \frac{1}{T-K} \text{Cov} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k+1} U_{\text{int},t}, \sum_{t=K+1}^T \tilde{Z}_{t-k'+1} U_{\text{int},t} \right], \\ [\mathbf{V}_{\text{int}}]_{K+1+k, K+1+k'} &= \frac{1}{T-K} \text{Cov} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \tilde{Z}_{t-k+1} U_{\text{int},t}, \sum_{t=K+1}^T \tilde{Z}_{t-k'} \tilde{Z}_{t-k'+1} U_{\text{int},t} \right], \end{aligned}$$

$$[\mathbf{V}_{\text{int}}]_{K+1+k,k'} = [\mathbf{V}_{\text{int}}]_{k',K+1+k} = \frac{1}{T-K} \text{Cov} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \tilde{Z}_{t-k+1} U_{\text{int},t}, \sum_{t=K+1}^T \tilde{Z}_{t-k'+1} U_{\text{int},t} \right].$$

The design assumptions and treatment probability conditions carry over unchanged from the main effects model. Under Assumption 2.3.4, the uniform boundedness of $\|\tau_{\text{int}}\|_1$ across T follows immediately, ensuring that both main and interaction effects remain well-behaved.

With these assumptions in place, we obtain the consistency and asymptotic normality of the extended estimator $\hat{\tau}_{\text{int}}$. These results parallel Theorem 2.3.1 and Theorem 2.3.2, now incorporating both main and interaction effects.

Theorem B.1.1 (Consistency). *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.1, Assumption 2.3.2(i) with $K \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0$ or Assumption 2.3.3(i) with $Km/T \rightarrow 0$, Assumption 2.3.4, and $K^2/T \rightarrow 0$. Then we have*

$$\|\hat{\tau}_{\text{int}} - \tau_{\text{int}}\|_2 \xrightarrow{\text{p}} 0.$$

Theorem B.1.2 (Asymptotic normality). *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.1, Assumption 2.3.2(ii) with K fixed or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, and Assumption 2.3.4. Then for any non-zero $\lambda_K \in \mathbb{R}^{K+1}$ satisfying $\liminf_{T \rightarrow \infty} \lambda_K^\top \mathbf{V}_{\text{int}} \lambda_K > 0$ and $\limsup_{T \rightarrow \infty} \|\lambda_K\|_1 < \infty$, we have*

$$\sqrt{T-K} (\lambda_K^\top \mathbf{V}_{\text{int}} \lambda_K)^{-1/2} \lambda_K^\top (\hat{\tau}_{\text{int}} - \tau_{\text{int}}) \xrightarrow{\text{d}} N(0, 1).$$

To estimate the asymptotic variance $\mathbf{V}_{\text{int}}$ in Theorem B.1.2, we apply the HAC variance estimator introduced in Section 2.3.2. Let

$$\hat{U}_{\text{int},t} = Y_t - \sum_{k=0}^K \tilde{Z}_{t-k} w_k \hat{\tau}_{\text{int},k} - \sum_{k=1}^K \tilde{Z}_{t-k} \tilde{Z}_{t-k+1} w_{k-1,k} \hat{\tau}_{\text{int},k-1,k}$$

be the sample residual at time t analogous to (2.3.6). Using these residuals, we can then construct the HAC variance estimator in the same way as in (2.3.7), and the consistency result is parallel to Theorem 2.3.3, since the inclusion of interaction terms simply augments the number of regressors from $K+1$ to $2K+1$.

We note that the k -th element of both τ and τ_{int} coincides with τ_k . This raises a natural question: do we actually need to include the interaction terms? The answer closely parallels the comparison between full OLS and marginal OLS in Section B.1.1: both procedures target the same estimand, but differ in their asymptotic variances and variance estimators.

By Theorems 2.3.2 and B.1.2, the asymptotic variance of $\hat{\tau}_k$ without interaction terms is $(T-K)^{-1} \text{Var}[\sum_{t=K+1}^T \tilde{Z}_{t-k} (Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell)]$, while the asymptotic variance with interaction terms is $(T-K)^{-1} \text{Var}[\sum_{t=K+1}^T \tilde{Z}_{t-k} (Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell - \sum_{\ell=1}^K \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} w_{\ell-1,\ell} \tau_{\ell-1,\ell})]$. To compare these variances explicitly, we consider a linear potential outcome model with interaction terms.

Proposition B.1.2. *Consider a linear model with homogeneous effects and nonrandom error: $Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \beta_k z_{t-k} + \sum_{k=1}^{t-1} \beta_{k-1,k} z_{t-k} z_{t-k+1} + \epsilon_t$. Assume a constant treatment probability $p_t = p$. Then, the asymptotic variance for OLS with interaction terms is*

$$\begin{aligned} & \frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} (Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} - \sum_{\ell=1}^K \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} w_{\ell-1,\ell} \tau_{\ell-1,\ell}) \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} (\beta_{\ell}^2 + p^2 \beta_{\ell-1,\ell}^2 + p^2 \beta_{\ell,\ell+1}^2) + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} p(1-p) \beta_{\ell-1,\ell}^2 \\ &+ \frac{1}{T-K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right)^2, \end{aligned}$$

while the asymptotic variance for OLS without interaction terms is

$$\begin{aligned} & \frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} (Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell}) \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} (\beta_{\ell}^2 + p^2 \beta_{\ell-1,\ell}^2 + p^2 \beta_{\ell,\ell+1}^2) + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=1}^{t-1} p(1-p) \beta_{\ell-1,\ell}^2 \\ &+ \frac{1}{T-K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right)^2. \end{aligned}$$

We face a natural trade-off when deciding whether to include interaction terms in the regression. Even if the primary goal is to estimate the main effects, Proposition B.1.2 shows that incorporating interaction terms can reduce asymptotic variance: the gain is exactly $(T-K)^{-1} \sum_{t=K+1}^T \sum_{\ell=1}^K p(1-p) \beta_{\ell-1,\ell}^2 = \sum_{\ell=1}^K p(1-p) \beta_{\ell-1,\ell}^2$, which is strictly positive whenever the interaction effects $\beta_{\ell-1,\ell}$ are non-negligible. Thus, including interactions improves asymptotic efficiency in the presence of interaction effects. However, this benefit comes at a cost. When the true interactions are small or absent, the variance reduction is negligible, while the inclusion of additional regressors increases model complexity and inflates finite-sample variance. Therefore, this creates a trade-off: including interactions can improve asymptotic efficiency but may sacrifice finite-sample performance.

B.1.3 General exposure mapping

In practice, we may be interested in using the exposure mappings to summarize the treatment history. Specifically, we divide the past treatment assignments into several time blocks, and the treatments within each block are aggregated into a single exposure mapping. For example, in behavioral or clinical studies, treatments (e.g., reminders or interventions) often have cumulative effects that depend on how consistently the treatment has been received in

the past rather than on any single recent treatment. This allows estimation of short-term, medium-term and long-term persistence in treatment effects. Define a collection of exposure mappings:

$$g_{t,0} = g_0(Z_{t:t-k_0}), \quad g_{t,1} = g_1(Z_{t-k_0-1:t-k_1}), \dots, \quad g_{t,S} = g_S(Z_{t-k_{S-1}-1:t-k_S}),$$

where $S \geq 0$ indexes the exposure mappings. The exposure lags satisfy $0 \leq k_0 < k_1 < \dots < k_S = K$, and each function g_s is a known mapping from a block of treatment assignments to $\{0, 1\}$: $g_0 : \mathbb{R}^{k_0+1} \rightarrow \{0, 1\}$, $g_1 : \mathbb{R}^{k_1-k_0} \rightarrow \{0, 1\}$, $\dots$, $g_S : \mathbb{R}^{k_S-k_{S-1}} \rightarrow \{0, 1\}$. Intuitively, we divide the treatment history into $S+1$ consecutive time blocks, and within each block, the sequence of past treatments is summarized by a binary value from the exposure mapping. We assume that the potential outcome can be expressed in terms of these exposure mappings: $Y_t(Z_{t:1}) = Y_t(g_{t,0}, g_{t,1}, \dots, g_{t,S})$. We treat these exposure mappings as treatment indicators in Section 2.2. For each exposure mapping $g_{t,s}$, we define a normalized version $\tilde{g}_{t,s} = (g_{t,s} - \mathbb{E}[g_{t,s}]) / \text{Var}[g_{t,s}]$ and corresponding weight

$$w_{g,s} = \left[\frac{1}{T-K} \sum_{t=K+1}^T (\text{Var}[g_{t,s}])^{-1} \right]^{-1},$$

for $s = 0, 1, \dots, S$. We consider the following OLS regression:

$$\mathbf{lm}(Y_t \sim \tilde{g}_{t,0} + \tilde{g}_{t,1} + \dots + \tilde{g}_{t,S}).$$

We denote the corresponding estimator by $\tilde{\tau}_g = (\tilde{\tau}_{g,0}, \dots, \tilde{\tau}_{g,S})$. We define $\hat{\tau}_{g,s} = w_{g,s}^{-1} \tilde{\tau}_{g,s}$ for $s = 0, 1, \dots, S$ analogous to (2.2.4), and obtain the transformed OLS estimator $\hat{\tau}_g = (\hat{\tau}_{g,0}, \dots, \hat{\tau}_{g,S})$. The causal estimand is $\tau_g = (\tau_{g,0}, \dots, \tau_{g,S})$, where each component $\tau_{g,s}$ captures the causal effect of the s -th exposure mapping. Formally,

$$\tau_{g,s} = \frac{1}{T-K} \sum_{t=K+1}^T \mathbb{E}[Y_t(g_{t,0}, \dots, g_{t,s-1}, 1, g_{t,s+1}, \dots, g_{t,S}) - Y_t(g_{t,0}, \dots, g_{t,s-1}, 0, g_{t,s+1}, \dots, g_{t,S})],$$

where the expectation averages over $(g_{t,0}, \dots, g_{t,s-1}, g_{t,s+1}, \dots, g_{t,S})$, holding the exposure mapping $g_{t,s}$ fixed at either 1 or 0.

This exposure mapping framework unifies and generalizes many commonly studied models. If we take $g_0(z) = \dots = g_S(z) = z$, and let $k_s = s$ with $S = K$, then each exposure mapping corresponds to a single treatment indicator at a lag. In this case, the model reduces to the OLS regression introduced in Section 2.2, where each treatment indicator enters separately as a regressor. If instead we set $S = 0$, $k_0 = K$, and define $g_0(z) = g(z)$ for some function $g : \{0, 1\}^K \rightarrow \{0, 1\}$, then the potential outcomes depend only on a single transformation of the treatment history over the most recent K periods. This corresponds to models in which treatment effects are summarized by a single exposure mapping.

To write $\hat{\tau}_g$ explicitly, recall the outcome vector $\mathbf{Y}_K = (Y_{K+1}, \dots, Y_T)^\top \in \mathbb{R}^{T-K}$ and define the regressor matrix $\mathbf{Z}_{g,K} = (\tilde{Z}_{g,K+1}^\top, \dots, \tilde{Z}_{g,T}^\top)^\top \in \mathbb{R}^{(T-K) \times (S+1)}$ where $\tilde{Z}_{g,t} = (g_{t,0}, \dots, g_{t,S})^\top$

for $t = K + 1, \dots, T$. Let $\mathbf{W}_{g,K} \in \mathbb{R}^{(S+1) \times (S+1)}$ be the diagonal weight matrix $\mathbf{W}_{g,K} = \text{diag}(w_{g,s} : s = 0, \dots, S)$. Then, $\hat{\tau}_g$ can be expressed as

$$\hat{\tau} = \mathbf{W}_{g,K}^{-1} \tilde{\tau}_g = \mathbf{W}_{g,K}^{-1} (\mathbf{Z}_{g,K}^\top \mathbf{Z}_{g,K})^{-1} \mathbf{Z}_{g,K}^\top \mathbf{Y}_K.$$

Analogous to the decomposition of $\hat{\tau}_k$ in (2.3.1), we can decompose $\hat{\tau}_g$ as

$$\hat{\tau}_g = \tau + \mathbf{W}_{g,K}^{-1} [(T - K)^{-1} \mathbf{Z}_{g,K}^\top \mathbf{Z}_{g,K}]^{-1} [(T - K)^{-1} \mathbf{Z}_{g,K}^\top (\mathbf{Y}_K - \mathbf{Z}_{g,K} \mathbf{W}_{g,K} \tau)].$$

To study the asymptotic properties of the OLS estimator $\hat{\tau}_g$, analogous to (2.3.4), we define the oracle residual of OLS regression at time t as

$$U_{g,t} = Y_t - \sum_{s=0}^S w_{g,s} \tau_{g,s} \tilde{g}_{t,s}.$$

We then define the asymptotic variance matrix $\mathbf{V}_g \in \mathbb{R}^{(S+1) \times (S+1)}$ analogous to (2.3.5) as

$$[\mathbf{V}_g]_{s,s'} = \frac{1}{T - K} \text{Cov} \left[\sum_{t=K+1}^T g_{t,s} U_{g,t}, \sum_{t=K+1}^T g_{t,s'} U_{g,t} \right].$$

To ensure well-defined and non-degenerate exposure mappings, we impose the following assumption, which is the analogue of Assumption 2.3.1 in the exposure setting.

Assumption B.1.1 (Treatment distributions condition). *There exists some constant $\epsilon > 0$ such that $\epsilon \leq \inf_{t \geq 1, s=0,1,\dots,S} \text{Var}[g_{t,s}]$.*

We also impose a boundedness condition on potential outcomes and exposure effects, analogous to Assumption 2.3.4.

Assumption B.1.2 (Moment condition). *The potential outcomes $Y_t(z_{t:1})$ are uniformly bounded for all t and all $z_{t:1} \in \mathbb{R}^t$, i.e., $\sup_{t \geq 1, z_{t:1} \in \mathbb{R}^t} |Y_t(z_{t:1})| < \infty$. Additionally, $\|\tau_g\|_1$ is uniformly bounded for all T , i.e., $\sup_T \|\tau_g\|_1 < \infty$.*

With these assumptions, we obtain the consistency and asymptotic normality of the exposure-effect estimator $\hat{\tau}_g$.

Theorem B.1.3 (Consistency). *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption B.1.1, Assumption 2.3.2(i) with $K \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0$ or Assumption 2.3.3(i) with $Km/T \rightarrow 0$, Assumption B.1.2, and $K^2/T \rightarrow 0$. Then we have*

$$\|\hat{\tau}_g - \tau_g\|_2 \xrightarrow{P} 0.$$

Theorem B.1.4 (Asymptotic normality). *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption B.1.1, Assumption 2.3.2(ii) with K fixed or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, and Assumption B.1.2. Then for any non-zero $\lambda_K \in \mathbb{R}^{K+1}$ satisfying $\liminf_{T \rightarrow \infty} \lambda_K^\top \mathbf{V}_g \lambda_K > 0$ and $\limsup_{T \rightarrow \infty} \|\lambda_K\|_1 < \infty$, we have*

$$\sqrt{T - K} (\lambda_K^\top \mathbf{V}_g \lambda_K)^{-1/2} \lambda_K^\top (\hat{\tau}_g - \tau_g) \xrightarrow{d} N(0, 1).$$

To estimate the asymptotic variance $\mathbf{V}_g$ in Theorem B.1.4, we apply the HAC variance estimator introduced in Section 2.3.2. Let

$$\widehat{U}_{g,t} = Y_t - \sum_{s=0}^S w_{g,s} \widehat{g}_{g,s} \widetilde{g}_{t,s}$$

be the sample residual at time t analogous to (2.3.6). Using these residuals, we can then construct the HAC variance estimator in the same way as in (2.3.7), and the consistency result is parallel to Theorem 2.3.3, since the inclusion of exposure mappings simply changes the number of regressors from $K + 1$ to $S + 1$.

B.2 Lemmas for the main results

We need the following lemmas to prove the main results.

Lemma B.2.1. *Assume Assumption 2.2.1, Assumption 2.2.2 and Assumption 2.3.1. If $K^2/T \rightarrow 0$, then*

$$\left\| \frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right\|_{\text{F}} \xrightarrow{\text{p}} 0.$$

Lemma B.2.2. *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.2(i) with $K \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0$ or Assumption 2.3.3(i) with $Km/T \rightarrow 0$, Assumption 2.3.4 and $K^2/T \rightarrow 0$. Then we have*

$$\left\| \frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right\|_2 \xrightarrow{\text{p}} 0.$$

Lemma B.2.3. *Assume Assumption 2.2.1, Assumption 2.2.2, Assumption 2.3.2(ii) with K fixed or Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, and Assumption 2.3.4. Then for any non-zero $\lambda_K \in \mathbb{R}^{K+1}$ satisfying $\liminf_{T \rightarrow \infty} \lambda_K^\top \mathbf{V} \lambda_K > 0$ and $\limsup_{T \rightarrow \infty} \|\lambda_K\|_1 < \infty$, we have*

$$\sqrt{T-K} (\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right] \xrightarrow{\text{d}} N(0, 1).$$

Now we prove these lemmas in the following subsections.

B.2.1 Proof of Lemma B.2.1

Proof of Lemma B.2.1. Recall

$$\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K = \frac{1}{T-K} \sum_{t=K+1}^T \widetilde{Z}_{t:t-K} \widetilde{Z}_{t:t-K}^\top,$$

with the (i, j) -th entry

$$\left[\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K \right]_{ij} = \frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1}.$$

Under Assumptions 2.2.2 and 2.3.1, the sequence Z_t is independent over t , and $E[\tilde{Z}_t] = 0$, $E[\tilde{Z}_t^2] = \text{Var}[Z_t]^{-1} = (p_t(1-p_t))^{-1}$, with $p_t \in (0, 1)$ uniformly bounded away from 0 and 1. If $i \neq j$, independence and centering imply $E[\tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1}] = 0$. If $i = j$, $E[\tilde{Z}_{t-i+1}^2] = (p_{t-i+1}(1-p_{t-i+1}))^{-1}$. Therefore

$$E \left[\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K \right] = \mathbf{W}_K^{-1}.$$

For $i \neq j$, we have $E[\tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1}] = 0$ and $\sup_t E[(\tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1})^2] < \infty$ since the treatment probability p_t is uniformly bounded away from 0 and 1. Moreover, because $\tilde{Z}_t$ is independent across t , we have for $t \neq t'$,

$$\text{Cov}(\tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1}, \tilde{Z}_{t'-i+1} \tilde{Z}_{t'-j+1}) = 0.$$

Consequently,

$$\begin{aligned} \text{Var} \left[\left[\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right]_{ij} \right] &= \text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1} \right] \\ &= \frac{1}{(T-K)^2} \sum_{t=K+1}^T \text{Var}[\tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1}] \leq \frac{1}{T-K} \sup_t E[(\tilde{Z}_{t-i+1} \tilde{Z}_{t-j+1})^2] \lesssim \frac{1}{T-K}. \end{aligned}$$

For $i = j$, we have $\sup_t E[\tilde{Z}_{t-i+1}^2] < \infty$ and $\tilde{Z}_t$ is independent across t . Thus

$$\begin{aligned} \text{Var} \left[\left[\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right]_{ij} \right] &= \text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-i+1}^2 \right] \\ &= \frac{1}{(T-K)^2} \sum_{t=K+1}^T \text{Var}[\tilde{Z}_{t-i+1}^2] \leq \frac{1}{T-K} \sup_t E[\tilde{Z}_{t-i+1}^2] \lesssim \frac{1}{T-K}. \end{aligned}$$

Combine the entrywise bounds:

$$E \left\| \frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right\|_F^2 = \sum_{i,j=1}^{K+1} \text{Var} \left[\left[\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right]_{ij} \right] \lesssim \frac{K^2}{T-K}.$$

By Markov's inequality, if $K^2/T \rightarrow 0$, then

$$\left\| \frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right\|_F \xrightarrow{\text{p}} 0.$$

This completes the proof. □

B.2.2 Proof of Lemma B.2.2

Proof of Lemma B.2.2. Expanding the term, we have:

$$\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) = \frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau). \quad (\text{B.2.1})$$

We first consider the mean of (B.2.1). By the independence of $\tilde{Z}_t$ and $\mathbb{E}[\tilde{Z}_t] = 0$, we can show that $\mathbb{E}[(T-K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)] = 0$ since for any $k = 0, \dots, K$,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau \right) \right] &= \mathbb{E} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \left(\mathbb{E}[\tilde{Z}_{t-k} Y_t] - \text{Var}[\tilde{Z}_{t-k}] w_k \tau_k \right) \\ &= \left(\frac{1}{T-K} \sum_{t=K+1}^T \mathbb{E}[\tilde{Z}_{t-k} Y_t] \right) - \left(\frac{1}{T-K} \sum_{t=K+1}^T \text{Var}[\tilde{Z}_{t-k}] \right) w_k \tau_k \\ &= \left(\frac{1}{T-K} \sum_{t=K+1}^T \tau_{t,k} \right) - \left(\frac{1}{T-K} \sum_{t=K+1}^T \text{Var}[\tilde{Z}_{t-k}]^{-1} \right) w_k \tau_k = \tau_k - \tau_k = 0. \end{aligned}$$

We then consider the variance of (B.2.1). For any $k = 0, \dots, K$, we have

$$\begin{aligned} &\text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau \right) \right] \\ &= \text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right] \\ &= \frac{1}{(T-K)^2} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right] \\ &\quad + \frac{1}{(T-K)^2} \sum_{t,t'=K+1, t \neq t'}^T \text{Cov} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right), \tilde{Z}_{t'-k} \left(Y_{t'} - \sum_{\ell=0}^K \tilde{Z}_{t'-\ell} w_\ell \tau_\ell \right) \right]. \end{aligned}$$

By Assumption 2.3.4,

$$\sup_t \text{Var} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right] \lesssim \sup_t |Y_t|^2 + \sup_t \|\tau\|_1^2 < \infty,$$

which implies

$$\frac{1}{(T-K)^2} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right] \lesssim \frac{1}{T-K}.$$

By the independence of $\tilde{Z}_t$, for any $t, t' \geq K + 1$ with $|t - t'| > K$, we have

$$\begin{aligned} & \text{Cov} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right), \tilde{Z}_{t'-k} \left(Y_{t'} - \sum_{\ell=0}^K \tilde{Z}_{t'-\ell} w_\ell \tau_\ell \right) \right] \\ &= \text{Cov} \left[\tilde{Z}_{t-k} Y_t, \tilde{Z}_{t'-k} Y_{t'} \right] \\ & \quad - \text{Cov} \left[\tilde{Z}_{t-k} Y_t, \tilde{Z}_{t'-k} \sum_{\ell=0}^K \tilde{Z}_{t'-\ell} w_\ell \tau_\ell \right] - \text{Cov} \left[\tilde{Z}_{t-k} \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell, \tilde{Z}_{t'-k} Y_{t'} \right]. \end{aligned}$$

We handle the dependence structure in two cases corresponding to Assumptions 2.3.2 and 2.3.3, respectively.

Decaying carryover effects condition (Assumption 2.3.2)

Assume $t > t'$ without loss of generality. Consider $Z'_{t':1} = (Z'_{t'}, \dots, Z'_1)^\top$ independent and identically distributed with $Z'_{t':1}$. Then

$$\begin{aligned} & \mathbb{E} \left[\tilde{Z}_{t-k} Y_t \tilde{Z}_{t'-k} Y_{t'} \right] = \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] \\ &= \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] - \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:t'+1}, Z'_{t':1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] \\ & \quad + \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:t'+1}, Z'_{t':1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right]. \end{aligned}$$

Since $(Z_{t:t'+1}, Z'_{t':1})$ is independent of $Z'_{t':1}$, we have

$$\begin{aligned} & \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:t'+1}, Z'_{t':1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] \\ &= \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:1}) \right] \mathbb{E} \left[\tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] = \mathbb{E} \left[\tilde{Z}_{t-k} Y_t \right] \mathbb{E} \left[\tilde{Z}_{t'-k} Y_{t'} \right]. \end{aligned}$$

By Assumption 2.3.4 and Assumption 2.3.2, we have

$$\begin{aligned} & \left| \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] - \mathbb{E} \left[\tilde{Z}_{t-k} Y_t(Z_{t:t'+1}, Z'_{t':1}) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] \right| \\ &= \left| \mathbb{E} \left[\tilde{Z}_{t-k} (Y_t(Z_{t:1}) - Y_t(Z_{t:t'+1}, Z'_{t':1})) \tilde{Z}_{t'-k} Y_{t'}(Z'_{t':1}) \right] \right| \\ &\leq \mathbb{E} [|Y_t(Z_{t:1}) - Y_t(Z_{t:t'+1}, Z'_{t':1})|] \lesssim \theta_{T,|t-t'|}. \end{aligned}$$

Therefore, we have

$$\left| \text{Cov} \left[\tilde{Z}_{t-k} Y_t, \tilde{Z}_{t'-k} Y_{t'} \right] \right| = \left| \mathbb{E} \left[\tilde{Z}_{t-k} Y_t \tilde{Z}_{t'-k} Y_{t'} \right] - \mathbb{E} \left[\tilde{Z}_{t-k} Y_t \right] \mathbb{E} \left[\tilde{Z}_{t'-k} Y_{t'} \right] \right| \lesssim \theta_{T,|t-t'|}.$$

By Assumption 2.3.4 and Assumption 2.3.2, we can show similarly that

$$\left| \text{Cov} \left[\tilde{Z}_{t-k} Y_t, \tilde{Z}_{t'-k} \sum_{\ell=0}^K \tilde{Z}_{t'-\ell} w_\ell \tau_\ell \right] \right| \lesssim \theta_{T,|t-t'|}.$$

By $t > t'$ and $|t - t'| > K$, we have

$$\text{Cov} \left[\tilde{Z}_{t-k} \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell}, \tilde{Z}_{t'-k} Y_{t'} \right] = 0.$$

Therefore, we have

$$\left| \text{Cov} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} \right), \tilde{Z}_{t'-k} \left(Y_{t'} - \sum_{\ell=0}^K \tilde{Z}_{t'-\ell} w_{\ell} \tau_{\ell} \right) \right] \right| \lesssim \theta_{T,|t-t'|}.$$

Summing over all time pairs gives

$$\text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \tilde{Z}_{t:t-K}^{\top} \mathbf{W}_K \tau \right) \right] \lesssim \frac{1}{T-K} \left(K + \sum_{k=K+1}^{T-K} \theta_{T,k} \right).$$

m-dependence condition (Assumption 2.3.3)

By Assumption 2.3.4 and Assumption 2.3.3, we have

$$\left| \text{Cov} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} \right), \tilde{Z}_{t'-k} \left(Y_{t'} - \sum_{\ell=0}^K \tilde{Z}_{t'-\ell} w_{\ell} \tau_{\ell} \right) \right] \right| \lesssim \mathbf{1}(|t - t'| \leq m).$$

Summing over all time pairs gives

$$\text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \tilde{Z}_{t:t-K}^{\top} \mathbf{W}_K \tau \right) \right] \lesssim \frac{1}{T-K} (K + m).$$

Final step

By Assumption 2.3.2(i) with $K \sum_{k=1}^T \theta_{T,k}/T \rightarrow 0$ or Assumption 2.3.3(i) with $Km/T \rightarrow 0$, and $K^2/T \rightarrow 0$, we have

$$\mathbb{E} \left\| \frac{1}{T-K} \mathbf{Z}_K^{\top} (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right\|_2^2 = \sum_{k=0}^K \text{Var} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \tilde{Z}_{t:t-K}^{\top} \mathbf{W}_K \tau \right) \right] \rightarrow 0.$$

This completes the proof by Markov's inequality. $\square$

B.2.3 Proof of Lemma B.2.3

Proof of Lemma B.2.3. In the same way as the proof of Lemma B.2.2, expanding the term, we have:

$$\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) = \frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau).$$

We can write

$$\begin{aligned} & \sqrt{T-K} (\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right] \\ &= \sqrt{T-K} (\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau) \right] \\ &= \sum_{t=K+1}^T (T-K)^{-1/2} (\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau). \end{aligned}$$

Define the triangular array $[U_{T,t}]_{t=K+1}^T$ by

$$U_{T,t} = (T-K)^{-1/2} (\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau).$$

As shown earlier in the proof of Lemma B.2.2,

$$\mathbb{E} \left[\tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau) \right] = 0,$$

hence $\mathbb{E}[U_{T,t}] = 0$. By the definition of $\mathbf{V}$ in (2.3.5), we have

$$\text{Var} \left[\sum_{t=K+1}^T U_{T,t} \right] = 1.$$

We then establish the moment bound for $U_{T,t}$. For $p \geq 2$, using $(\lambda_K^\top \mathbf{V} \lambda_K) \geq \lambda_{\min}(\mathbf{V}) \|\lambda_K\|_2^2$, the p -th moment of $U_{T,t}$ is bounded by

$$\begin{aligned} \mathbb{E}[|U_{T,t}|^p] &= (T-K)^{-p/2} (\lambda_K^\top \mathbf{V} \lambda_K)^{-p/2} \mathbb{E}[|\lambda_K^\top \tilde{Z}_{t:t-K} (Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau)|^p] \\ &= (T-K)^{-p/2} (\lambda_K^\top \mathbf{V} \lambda_K)^{-p/2} \mathbb{E}[|\lambda_K^\top \tilde{Z}_{t:t-K}|^p |Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau|^p] \\ &\lesssim (T-K)^{-p/2} (\lambda_K^\top \mathbf{V} \lambda_K)^{-p/2} \|\lambda_K\|_1^p \mathbb{E}[|Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau|^p] \\ &\lesssim (T-K)^{-p/2} \mathbb{E}[|Y_t|^p + \|\tau\|_1^p] \\ &\lesssim (T-K)^{-p/2}, \end{aligned}$$

where we used the uniform boundedness in Assumption 2.3.4, $\liminf_{T \rightarrow \infty} \lambda_K^\top \mathbf{V} \lambda_K > 0$ and $\limsup_{T \rightarrow \infty} \|\lambda_K\|_1 < \infty$. So the triangular array $[(T-K)^{1/2} U_{T,t}]_{t=K+1}^T$ has uniformly bounded p -th moments.

We now handle the dependence structure in two cases corresponding to Assumptions 2.3.2 and 2.3.3, respectively.

Decaying carryover effects condition (Assumption 2.3.2).

Before starting the proof, we first introduce the CLT for dependent triangular arrays with general covariance-based conditions in [Chandrasekhar et al. \(2023\)](#).

Lemma B.2.4 (Theorem 1 in [Chandrasekhar et al. \(2023\)](#) under dimension one). *Let $[Z_{n,i}]_{i=1}^n$ be a triangular array of random variables with mean zero. For each n , we define the affinity set $\mathcal{A}_i^n \subset \{1, \dots, n\}$ for each i . Let $\Omega_n := \sum_{i=1}^n \sum_{j \in \mathcal{A}_i^n} \text{Cov}[Z_{n,i}, Z_{n,j}]$. Let $Z_{n,-i} := \sum_{j \notin \mathcal{A}_i^n} Z_{n,j}$. Assume that*

1. $\sum_{i=1}^n \sum_{j,k \in \mathcal{A}_i^n} \mathbb{E}[|Z_{n,i}| Z_{n,j} Z_{n,k}] = o(|\Omega_n|^{3/2})$.
2. $\sum_{i,j=1}^n \sum_{k \in \mathcal{A}_i^n, \ell \in \mathcal{A}_j^n} \text{Cov}[Z_{n,i} Z_{n,k}, Z_{n,j} Z_{n,\ell}] = o(|\Omega_n|^2)$.
3. $\sum_{i=1}^n \mathbb{E}[|\mathbb{E}[Z_{n,i} Z_{n,-i} | Z_{n,-i}]|] = \sum_{i=1}^n \mathbb{E}[|Z_{n,-i} \mathbb{E}[Z_{n,i} | Z_{n,-i}]|] = o(|\Omega_n|)$.

Then the triangular array $[Z_{n,i}]_{i=1}^n$ satisfies the CLT:

$$\Omega_n^{-1/2} \sum_{i=1}^n Z_{n,i} \xrightarrow{d} N(0, 1).$$

To apply Lemma B.2.4, we follow the proof in Section 3 of [Chandrasekhar et al. \(2023\)](#). We define $\mathcal{A}_t^T = \{t' : |t - t'| \leq D\}$, where $D = T^\epsilon \rightarrow \infty$ as $T \rightarrow \infty$ for some small $\epsilon > 0$.

By the selection of D , we have

$$\left| \sum_{t=K+1}^T \sum_{t' \notin \mathcal{A}_t^T} \text{Cov}[U_{T,t}, U_{T,t'}] \right| \lesssim \sum_{t=K+1}^T \sum_{t' \notin \mathcal{A}_t^T} \frac{1}{T-K} \theta_{T,|t-t'|} \lesssim \sum_{k=D}^T \theta_{T,k} \lesssim D^{-\delta} = o(1),$$

which implies that

$$\begin{aligned} \Omega_T &= \sum_{t=K+1}^T \sum_{t' \in \mathcal{A}_t^T} \text{Cov}[U_{T,t}, U_{T,t'}] = \text{Var} \left[\sum_{t=K+1}^T U_{T,t} \right] - \sum_{t=K+1}^T \sum_{t' \notin \mathcal{A}_t^T} \text{Cov}[U_{T,t}, U_{T,t'}] \\ &= 1 - o(1) \rightarrow 1. \end{aligned}$$

For the first condition in Lemma B.2.4, we have

$$\begin{aligned} \left| \sum_{t=K+1}^T \sum_{t', t'' \in \mathcal{A}_t^T} \mathbb{E}[|U_{T,t}| U_{T,t'} U_{T,t''}] \right| &\lesssim D^2 \sum_{t=K+1}^T \mathbb{E}[|U_{T,t}|^3] \lesssim D^2 (T-K) (T-K)^{-3/2} \\ &= D^2 (T-K)^{-1/2} = o(1) = o(|\Omega_T|^{3/2}). \end{aligned}$$

For the second condition in Lemma B.2.4, we have

$$\left| \sum_{s,t=K+1}^T \sum_{s' \in \mathcal{A}_s^T, t' \in \mathcal{A}_t^T} \text{Cov}[U_{T,s}U_{T,s'}, U_{T,t}U_{T,t'}] \right| \lesssim \sum_{t=K+1}^T \sum_{|s-t| \leq D, |t'-t| \leq D, |s'-t| \leq 2D} \mathbb{E}[|U_{T,t}|^4] \\ \lesssim D^3(T-K)(T-K)^{-2} = D^3(T-K)^{-1} = o(1) = o(|\Omega_T|^2).$$

For the third condition in Lemma B.2.4, defining the random variable $\xi_{t,-t}$ such that $\xi_{t,-t} = 1$ if $\mathbb{E}[U_{T,t} | U_{T,-t}] > 0$ and $\xi_{t,-t} = -1$ otherwise, then we have

$$\left| \sum_{t=K+1}^T \mathbb{E}[| \mathbb{E}[U_{T,t}U_{T,-t} | U_{T,-t}] |] \right| = \left| \sum_{t=K+1}^T \sum_{t' \notin \mathcal{A}_t^T} \text{Cov}[\xi_{t,-t}U_{T,t'}, U_{T,t}] \right| \\ \lesssim \sum_{t=K+1}^T \sum_{t' \notin \mathcal{A}_t^T} \frac{1}{T-K} \theta_{T,|t-t'|} \lesssim \sum_{k=D}^T \theta_{T,k} \lesssim D^{-\delta} = o(1) = o(|\Omega_T|).$$

Then by Lemma B.2.4 and $\Omega_T \rightarrow 1$, we have

$$\sum_{t=K+1}^T U_{T,t} \xrightarrow{d} N(0, 1),$$

which is exactly

$$\sqrt{T-K}(\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right] \xrightarrow{d} N(0, 1).$$

m-dependence condition (Assumption 2.3.3).

Before starting the proof, we first introduce the CLT for m -dependent triangular arrays in Chen and Shao (2004).

Lemma B.2.5 (Theorem 2.6 in Chen and Shao (2004)). *Let $[Z_{n,i}]_{i=1}^n$ be a m -dependent triangular array of random variables with mean zero, $\text{Var}[\sum_{i=1}^n Z_{n,i}] = 1$, and finite $\mathbb{E}[|Z_{n,i}|^p] < \infty$ for $2 < p \leq 3$. Let F be the distribution function of $\sum_{i=1}^n Z_{n,i}$ and Φ be the distribution function of $N(0, 1)$. Then*

$$\sup_z |F(z) - \Phi(z)| \leq 75(10m+1)^{p-1} \sum_{i=1}^n \mathbb{E}[|Z_{n,i}|^p].$$

To apply Lemma B.2.5, note that by Assumption 2.3.3, the triangular array $[U_{T,t}]_{t=K+1}^T$ is $(K \vee m)$ -dependent since each $U_{T,t}$ is measurable with respect to $\sigma(Y_t, Z_t, \dots, Z_{t-K}) = \sigma(Y_t(Z_{t:t-m}), Z_t, \dots, Z_{t-K})$. From the moment bound already established for $U_{T,t}$ (take $p =$

3), $\sup_t \mathbb{E}[|U_{T,t}|^3] \lesssim (T-K)^{-3/2}$. Let F_T be the distribution function of $\sum_{t=K+1}^T U_{T,t}$, then by Lemma B.2.5 with $p = 3$, we have

$$\sup_z |F_T(z) - \Phi(z)| \lesssim (K \vee m)^2 (T-K)(T-K)^{-3/2} = (K \vee m)^2 (T-K)^{-1/2}.$$

Then by Assumption 2.3.3(ii) with $K^4/T \rightarrow 0$, we have

$$\sum_{t=K+1}^T U_{T,t} \xrightarrow{d} N(0, 1),$$

which is exactly

$$\sqrt{T-K}(\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right] \xrightarrow{d} N(0, 1).$$

□

B.3 Proofs

B.3.1 Proof of Theorem 2.3.1

Proof of Theorem 2.3.1. Recall (2.3.1) that

$$\hat{\tau} = \tau + \mathbf{W}_K^{-1}[(T-K)^{-1} \mathbf{Z}_K^\top \mathbf{Z}_K]^{-1}[(T-K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)].$$

By Lemma B.2.1, under the assumptions of the theorem, we have

$$\left\| \frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1} \right\|_F \xrightarrow{p} 0.$$

Since $\mathbf{W}_K^{-1}$ is positive definite with eigenvalues uniformly bounded away from 0 by Assumption 2.3.1, we have

$$\left\| \mathbf{W}_K^{-1} \left(\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K \right)^{-1} \right\|_2 \xrightarrow{p} \|\mathbf{I}_{K+1}\|_2 = 1.$$

By Lemma B.2.2, under the assumptions of the theorem, we have

$$\left\| \frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right\|_2 \xrightarrow{p} 0.$$

Then we obtain

$$\|\hat{\tau} - \tau\|_2 \leq \left\| \mathbf{W}_K^{-1} \left(\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K \right)^{-1} \right\|_2 \left\| \frac{1}{T-K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right\|_2 \xrightarrow{p} 0,$$

which completes the proof. □

B.3.2 Proof of Theorem 2.3.2

Proof of Theorem 2.3.2. Recall (2.3.1) that

$$\hat{\tau} = \tau + \mathbf{W}_K^{-1}[(T - K)^{-1} \mathbf{Z}_K^\top \mathbf{Z}_K]^{-1}[(T - K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)].$$

Then we have the following decomposition:

$$\begin{aligned} \lambda_K^\top (\hat{\tau} - \tau) &= \lambda_K^\top \mathbf{W}_K^{-1}[(T - K)^{-1} \mathbf{Z}_K^\top \mathbf{Z}_K]^{-1}[(T - K)^{-1} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau)] \\ &= \lambda_K^\top \left[\frac{1}{T - K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right] \\ &\quad + \lambda_K^\top \left[\mathbf{W}_K^{-1} \left(\frac{1}{T - K} \mathbf{Z}_K^\top \mathbf{Z}_K \right)^{-1} - \mathbf{I}_{K+1} \right] \left[\frac{1}{T - K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right]. \end{aligned}$$

By Lemma B.2.3, under the assumptions of the theorem, we have

$$\sqrt{T - K}(\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T - K} \mathbf{Z}_K^\top (\mathbf{Y}_K - \mathbf{Z}_K \mathbf{W}_K \tau) \right] \xrightarrow{d} N(0, 1).$$

By Lemma B.2.1, under the assumptions of the theorem, we have

$$\left\| \mathbf{W}_K^{-1} \left(\frac{1}{T - K} \mathbf{Z}_K^\top \mathbf{Z}_K \right)^{-1} - \mathbf{I}_{K+1} \right\|_2 \xrightarrow{p} 0.$$

Combining the above two results, we have

$$\sqrt{T - K}(\lambda_K^\top \mathbf{V} \lambda_K)^{-1/2} \lambda_K^\top (\hat{\tau} - \tau) \xrightarrow{d} N(0, 1),$$

which completes the proof. □

B.3.3 Proof of Theorem 2.3.3

Proof of Theorem 2.3.3. By Lemma B.2.1, we have $\|\frac{1}{T-K} \mathbf{Z}_K^\top \mathbf{Z}_K - \mathbf{W}_K^{-1}\|_F \xrightarrow{p} 0$, and then we have the asymptotic approximation:

$$\hat{\mathbf{V}} = \frac{1}{T - K} \mathbf{Z}_K^\top \hat{\mathbf{U}}_K \mathbf{Q}_K \hat{\mathbf{U}}_K \mathbf{Z}_K + o_P(1).$$

We decompose $\hat{\mathbf{V}}$ as follows:

$$\begin{aligned} &\hat{\mathbf{V}} - \mathbf{V} - \mathbf{B}_K \\ &= \frac{1}{T - K} \left(\mathbf{Z}_K^\top \hat{\mathbf{U}}_K \mathbf{Q}_K \hat{\mathbf{U}}_K \mathbf{Z}_K - \mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K \right) \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{T-K} (\mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K - \mathbf{B}_K - (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])) \\
& + \frac{1}{T-K} \left((\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right. \\
& \quad \left. - \mathbb{E}[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])] \right) \\
& + \frac{1}{T-K} (\mathbb{E}[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])] - \mathbf{V}) + o_P(1). \tag{B.3.1}
\end{aligned}$$

We bound the Frobenius norm of each term in (B.3.1).

First term.

For entries (k, k') , by the definition of $\widehat{\mathbf{U}}_K, \mathbf{U}_K$ and $\mathbf{Q}_K$, we have

$$\left[\mathbf{Z}_K^\top \widehat{\mathbf{U}}_K \mathbf{Q}_K \widehat{\mathbf{U}}_K \mathbf{Z}_K \right]_{k,k'} = \sum_{t=K+1}^T \sum_{t'=K+1}^T \tilde{Z}_{t-k+1} \widehat{U}_t \tilde{Z}_{t'-k'+1} \widehat{U}_{t'} \left(1 - \frac{|t-t'|}{L+1} \right) \mathbf{1}(|t-t'| \leq L),$$

and

$$\left[\mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K \right]_{k,k'} = \sum_{t=K+1}^T \sum_{t'=K+1}^T \tilde{Z}_{t-k+1} U_t \tilde{Z}_{t'-k'+1} U_{t'} \left(1 - \frac{|t-t'|}{L+1} \right) \mathbf{1}(|t-t'| \leq L).$$

Then we obtain

$$\begin{aligned}
& \left| \left[\frac{1}{T-K} \left(\mathbf{Z}_K^\top \widehat{\mathbf{U}}_K \mathbf{Q}_K \widehat{\mathbf{U}}_K \mathbf{Z}_K - \mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K \right) \right]_{k,k'} \right| \\
& = \left| \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L} \left(\tilde{Z}_{t-k+1} \widehat{U}_t \tilde{Z}_{t'-k'+1} \widehat{U}_{t'} - \tilde{Z}_{t-k+1} U_t \tilde{Z}_{t'-k'+1} U_{t'} \right) \left(1 - \frac{|t-t'|}{L+1} \right) \right| \\
& \leq \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L} |\tilde{Z}_{t-k+1} \tilde{Z}_{t'-k'+1}| (|\widehat{U}_t - U_t| |\widehat{U}_{t'}| + |\widehat{U}_{t'} - U_{t'}| |U_t|) \left(1 - \frac{|t-t'|}{L+1} \right).
\end{aligned}$$

By Theorem 2.3.2,

$$\left| \widehat{U}_t - U_t \right| = \left| \tilde{Z}_{t:t-K}^\top \mathbf{W}_K (\widehat{\tau} - \tau) \right| \lesssim \|\widehat{\tau} - \tau\|_1 = O_P \left(\frac{K}{\sqrt{T-K}} \right),$$

and by Assumption 2.3.4,

$$|U_t| + \left| \widehat{U}_t \right| = \left| Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \widehat{\tau} \right| + \left| Y_t - \tilde{Z}_{t:t-K}^\top \mathbf{W}_K \tau \right| \lesssim |Y_t| + \|\tau\|_1 = O_P(1).$$

Then we have

$$\left| \left[\frac{1}{T-K} \left(\mathbf{Z}_K^\top \hat{\mathbf{U}}_K \mathbf{Q}_K \hat{\mathbf{U}}_K \mathbf{Z}_K - \mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K \right) \right]_{k,k'} \right| = O_P \left(\frac{KL}{\sqrt{T-K}} \right),$$

and then

$$\left\| \frac{1}{T-K} \left(\mathbf{Z}_K^\top \hat{\mathbf{U}}_K \mathbf{Q}_K \hat{\mathbf{U}}_K \mathbf{Z}_K - \mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K \right) \right\|_F = O_P \left(\frac{K^2 L}{\sqrt{T-K}} \right) = o_P(1),$$

when $K^2 L / T^{1/2} \rightarrow 0$.

Second term.

Since $\mathbf{b}_K = \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]$ and $\mathbf{B}_K = \mathbf{b}_K^\top \mathbf{Q}_K \mathbf{b}_K$, we have

$$\begin{aligned} & \mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K - \mathbf{B}_K - (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \\ &= 2(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]. \end{aligned}$$

For entries (k, k') , by the definition of $\mathbf{U}_K$, $\mathbf{Z}_K$ and $\mathbf{Q}_K$, we have

$$\begin{aligned} & [(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]]_{k,k'} \\ &= \sum_{t=K+1}^T \sum_{t'=K+1}^T \left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right) \mathbb{E}[\tilde{Z}_{t-k'+1} U_{t'}] \left(1 - \frac{|t-t'|}{L+1} \right) \mathbb{1}(|t-t'| \leq L), \end{aligned}$$

which implies

$$\begin{aligned} & \left[\frac{1}{T-K} ((\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right]_{k,k'} \\ &= \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L}^T \left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right) \mathbb{E}[\tilde{Z}_{t-k'+1} U_{t'}] \left(1 - \frac{|t-t'|}{L+1} \right) \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \left[\sum_{t'=K+1, |t'-t| \leq L}^T \mathbb{E}[\tilde{Z}_{t-k'+1} U_{t'}] \left(1 - \frac{|t-t'|}{L+1} \right) \right] \left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right). \end{aligned}$$

Since each term in the sum is independent and mean zero, by Assumption 2.3.4, we have

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{T-K} ((\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right]_{k,k'}^2 \\ &= \frac{1}{(T-K)^2} \sum_{t=K+1}^T \left[\sum_{t'=K+1, |t'-t| \leq L}^T \mathbb{E}[\tilde{Z}_{t-k'+1} U_{t'}] \left(1 - \frac{|t-t'|}{L+1} \right) \right]^2 \text{Var}[\tilde{Z}_{t-k+1} U_t] \end{aligned}$$

$$\begin{aligned} &\lesssim \frac{1}{(T-K)^2} \sum_{t=K+1}^T \left[\sum_{t'=K+1, |t-t'|\leq L}^T \left(1 - \frac{|t-t'|}{L+1}\right) \right]^2 \\ &\lesssim \frac{L^2}{T-K}, \end{aligned}$$

and then by Markov's inequality,

$$\begin{aligned} &\left\| \frac{1}{T-K} (\mathbf{Z}_K^\top \mathbf{U}_K \mathbf{Q}_K \mathbf{U}_K \mathbf{Z}_K - \mathbf{B}_K - (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])) \right\|_{\text{F}} \\ &= O_{\text{P}} \left(\frac{KL}{\sqrt{T-K}} \right) = o_{\text{P}}(1), \end{aligned}$$

when $KL/T^{1/2} \rightarrow 0$.

Third term.

For entries (k, k') , by the definition of $\mathbf{U}_K$, $\mathbf{Z}_K$ and $\mathbf{Q}_K$, we have

$$\begin{aligned} &\left[\frac{1}{T-K} ((\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])) \right]_{k,k'} \\ &= \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'|\leq L}^T \left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right) \left(\tilde{Z}_{t'-k'+1} U_{t'} - \mathbb{E}[\tilde{Z}_{t'-k'+1} U_{t'}] \right) \left(1 - \frac{|t-t'|}{L+1} \right), \end{aligned}$$

Since $\tilde{Z}_t$ are independent, by Assumption 2.3.4, we have

$$\begin{aligned} &\text{Var} \left[\frac{1}{T-K} \sum_{t,t'=K+1, |t-t'|\leq L}^T \left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right) \left(\tilde{Z}_{t'-k'+1} U_{t'} - \mathbb{E}[\tilde{Z}_{t'-k'+1} U_{t'}] \right) \left(1 - \frac{|t-t'|}{L+1} \right) \right] \\ &= \frac{1}{(T-K)^2} \sum_{t,t'=K+1, |t-t'|\leq L}^T \text{Var} \left[\left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right) \left(\tilde{Z}_{t'-k'+1} U_{t'} - \mathbb{E}[\tilde{Z}_{t'-k'+1} U_{t'}] \right) \left(1 - \frac{|t-t'|}{L+1} \right) \right] \\ &\lesssim \frac{1}{(T-K)^2} (T-K)L = \frac{L}{T-K}. \end{aligned}$$

Then

$$\mathbb{E} \left[\frac{1}{T-K} ((\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])) \right]_{k,k'}^2 \lesssim \frac{L}{T-K},$$

and then by Markov's inequality,

$$\left\| \frac{1}{T-K} \left((\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) - \mathbb{E}[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])] \right) \right\|_F = O_P\left(\frac{K\sqrt{L}}{\sqrt{T-K}}\right) = o_P(1),$$

when $K\sqrt{L}/T^{1/2} \rightarrow 0$.

Fourth term.

For entries (k, k') , by the definition of $\mathbf{U}_K$, $\mathbf{Z}_K$ and $\mathbf{Q}_K$, we have

$$\begin{aligned} & \left[\frac{1}{T-K} \mathbb{E} [(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])] \right]_{k,k'} \\ &= \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L}^T \\ & \quad \mathbb{E} \left[\left(\tilde{Z}_{t-k+1} U_t - \mathbb{E}[\tilde{Z}_{t-k+1} U_t] \right) \left(\tilde{Z}_{t'-k'+1} U_{t'} - \mathbb{E}[\tilde{Z}_{t'-k'+1} U_{t'}] \right) \right] \left(1 - \frac{|t-t'|}{L+1} \right) \\ &= \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L}^T \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right] \left(1 - \frac{|t-t'|}{L+1} \right), \end{aligned}$$

and by the definition of $\mathbf{V}$ in (2.3.5), we have

$$\begin{aligned} [\mathbf{V}]_{k,k'} &= \frac{1}{T-K} \text{Cov} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k+1} U_t, \sum_{t'=K+1}^T \tilde{Z}_{t'-k'+1} U_{t'} \right] \\ &= \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L}^T \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right] \\ & \quad + \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| > L}^T \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right]. \end{aligned}$$

Then the entrywise difference is

$$\begin{aligned} & \left| \left[\frac{1}{T-K} \mathbb{E} [(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])] - \mathbf{V} \right]_{k,k'} \right| \\ & \leq \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'| \leq L}^T \frac{|t-t'|}{L+1} \left| \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right] \right| \end{aligned}$$

$$+ \frac{1}{T-K} \sum_{t,t'=K+1, |t-t'|>L}^T \left| \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right] \right|.$$

We now handle the dependence structure in two cases.

Assume decaying carryover effects condition (Assumption 2.3.2). Assume $L > K$. Consider t, t' such that $|t - t'| > K$, by Assumption 2.3.4 and the covariance inequality for the decaying carryover effects sequence in the proof of Theorem 2.3.1, we have

$$\left| \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right] \right| \lesssim \theta_{T, |t-t'|}.$$

Then by Assumption 2.3.2(ii), we have

$$\begin{aligned} & \left| \left[\frac{1}{T-K} \mathbb{E} \left[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right] - \mathbf{V} \right]_{k,k'} \right| \\ & \lesssim \frac{K^2}{L} + \frac{1}{L} \sum_{k=1}^L k \theta_{T,k} + \sum_{k=L+1}^T \theta_{T,k} \lesssim \frac{K^2}{L} + \frac{1}{L} + \frac{1}{L^\delta}, \end{aligned}$$

and then

$$\left\| \frac{1}{T-K} \mathbb{E} \left[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right] - \mathbf{V} \right\|_{\text{F}} \lesssim \frac{K^3}{L} + \frac{K}{L^\delta} \rightarrow 0,$$

when $K/L^{(1/3) \wedge \delta} \rightarrow 0$.

Assume m -dependence condition (Assumption 2.3.3). Assume $L > K$. Consider t, t' such that $|t - t'| > K$, by Assumption 2.3.4 and the definition of m -dependence, we have

$$\left| \text{Cov} \left[\tilde{Z}_{t-k+1} U_t, \tilde{Z}_{t'-k'+1} U_{t'} \right] \right| \lesssim \mathbb{1}(|t - t'| \leq m).$$

Assume further $L > m$. Then by Assumption 2.3.3, we have

$$\left| \left[\frac{1}{T-K} \mathbb{E} \left[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right] - \mathbf{V} \right]_{k,k'} \right| \lesssim \frac{m^2}{L},$$

and then

$$\left\| \frac{1}{T-K} \mathbb{E} \left[(\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K])^\top \mathbf{Q}_K (\mathbf{U}_K \mathbf{Z}_K - \mathbb{E}[\mathbf{U}_K \mathbf{Z}_K]) \right] - \mathbf{V} \right\|_{\text{F}} \lesssim \frac{K m^2}{L} \rightarrow 0,$$

when $K m^2 / L \rightarrow 0$.

Final step.

Combining the four terms using the decomposition of $\hat{\mathbf{V}}$ completes the proof. $\square$

B.3.4 Proofs of Theorem 2.4.1 and Theorem 2.4.2

Proofs of Theorem 2.4.1 and Theorem 2.4.2. The proofs of Theorem 2.4.1 and Theorem 2.4.2 are similar to the proofs of Theorem 2.3.1 and Theorem 2.3.2, we only outline the key steps here.

Recall that

$$\hat{\tau}_{\text{cont}} = \tau + \mathbf{W}_{\text{cont},K}^{-1}[(T-K)^{-1}\mathbf{Z}_{\text{cont},K}^{\top}\mathbf{Z}_{\text{cont},K}]^{-1}[(T-K)^{-1}\mathbf{Z}_{\text{cont},K}^{\top}(\mathbf{Y}_K - \mathbf{Z}_{\text{cont},K}\mathbf{W}_{\text{cont},K}\tau)].$$

We can establish Lemma B.2.1, Lemma B.2.2, Lemma B.2.3 in the same way, except replacing the mean $(p_1, \dots, p_T)$ and the variance $(p_1(1-p_1), \dots, p_T(1-p_T))$ with the mean $(\mathbb{E}[Z_1], \dots, \mathbb{E}[Z_T])$ and the variance $(\text{Var}[Z_1], \dots, \text{Var}[Z_T])$. To see why the estimand is

$$\tau_k = \frac{1}{T-K} \sum_{t=K+1}^T \mathbb{E} \left[Y_t(Z_{t:1}) \frac{(Z_{t-k} - \mathbb{E}[Z_{t-k}])}{\text{Var}[Z_{t-k}]} \right],$$

we show $\mathbb{E}[(T-K)^{-1}\mathbf{Z}_K^{\top}(\mathbf{Y}_K - \mathbf{Z}_K\mathbf{W}_K\tau)] = 0$ by following the proof of Lemma B.2.2. For any $k = 0, \dots, K$, we have

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \tilde{Z}_{t:t-K}^{\top} \mathbf{W}_K \tau \right) \right] = \mathbb{E} \left[\frac{1}{T-K} \sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} \right) \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \left(\mathbb{E}[\tilde{Z}_{t-k} Y_t] - \text{Var}[\tilde{Z}_{t-k}] w_k \tau_k \right) \\ &= \left(\frac{1}{T-K} \sum_{t=K+1}^T \mathbb{E}[\tilde{Z}_{t-k} Y_t] \right) - \left(\frac{1}{T-K} \sum_{t=K+1}^T \text{Var}[\tilde{Z}_{t-k}] \right) w_k \tau_k \\ &= \left(\frac{1}{T-K} \sum_{t=K+1}^T \mathbb{E} \left[Y_t(Z_{t:1}) \frac{(Z_{t-k} - \mathbb{E}[Z_{t-k}])}{\text{Var}[Z_{t-k}]} \right] \right) - \left(\frac{1}{T-K} \sum_{t=K+1}^T \text{Var}[Z_{t-k}]^{-1} \right) w_k \tau_k \\ &= \left(\frac{1}{T-K} \sum_{t=K+1}^T \mathbb{E} \left[Y_t(Z_{t:1}) \frac{(Z_{t-k} - \mathbb{E}[Z_{t-k}])}{\text{Var}[Z_{t-k}]} \right] \right) - \tau_k = 0. \end{aligned}$$

We then complete the proofs. □

B.3.5 Proof of Theorem 2.4.3

Proof of Theorem 2.4.3. The proof is similar to the proof of Theorem 2.3.3 by considering the decomposition of $\hat{\mathbf{V}}$ in (B.3.1). □

B.3.6 Proofs of Theorem B.1.1 and Theorem B.1.2

Proofs of Theorem B.1.1 and Theorem B.1.2. The proofs of Theorem B.1.1 and Theorem B.1.2 are similar to the proofs of Theorem 2.3.1 and Theorem 2.3.2, we only outline the key steps here.

Recall that

$$\hat{\tau}_{\text{int}} = \tau + \mathbf{W}_{\text{int},K}^{-1}[(T-K)^{-1}\mathbf{Z}_{\text{int},K}^{\top}\mathbf{Z}_{\text{int},K}]^{-1}[(T-K)^{-1}\mathbf{Z}_{\text{int},K}^{\top}(\mathbf{Y}_K - \mathbf{Z}_{\text{int},K}\mathbf{W}_{\text{int},K}\tau)].$$

Similar to Lemma B.2.1, we can show that under the same assumptions of Lemma B.2.1, we have

$$\left\| \frac{1}{T-K}\mathbf{Z}_{\text{int},K}^{\top}\mathbf{Z}_{\text{int},K} - \mathbf{W}_{\text{int},K}^{-1} \right\|_{\text{F}} \xrightarrow{\text{p}} 0.$$

Similar to Lemma B.2.2, we can show that under the same assumptions of Lemma B.2.2, we have

$$\left\| \frac{1}{T-K}\mathbf{Z}_{\text{int},K}^{\top}(\mathbf{Y}_K - \mathbf{Z}_{\text{int},K}\mathbf{W}_{\text{int},K}\tau) \right\|_2 \xrightarrow{\text{p}} 0.$$

As we have different asymptotic variance: $\mathbf{V}_{\text{int}}$ instead of $\mathbf{V}$, similar to Lemma B.2.3, we can show that under the same assumptions of Lemma B.2.3, we have

$$\sqrt{T-K}(\lambda_K^{\top}\mathbf{V}_{\text{int}}\lambda_K)^{-1/2}\lambda_K^{\top} \left[\frac{1}{T-K}\mathbf{Z}_{\text{int},K}^{\top}(\mathbf{Y}_K - \mathbf{Z}_{\text{int},K}\mathbf{W}_{\text{int},K}\tau) \right] \xrightarrow{\text{d}} N(0, 1).$$

We then complete the proofs. □

B.3.7 Proofs of Theorem B.1.3 and Theorem B.1.4

Proofs of Theorem B.1.3 and Theorem B.1.4. The proofs of Theorem B.1.3 and Theorem B.1.4 are similar to the proofs of Theorem 2.3.1 and Theorem 2.3.2, we only outline the key steps here.

Recall that

$$\hat{\tau}_{\text{g}} = \tau + \mathbf{W}_{\text{g},K}^{-1}[(T-K)^{-1}\mathbf{Z}_{\text{g},K}^{\top}\mathbf{Z}_{\text{g},K}]^{-1}[(T-K)^{-1}\mathbf{Z}_{\text{g},K}^{\top}(\mathbf{Y}_K - \mathbf{Z}_{\text{g},K}\mathbf{W}_{\text{g},K}\tau)].$$

Similar to Lemma B.2.1, we can show that under the same assumptions of Lemma B.2.1 except replacing Assumption 2.3.1 with Assumption B.1.1, we have

$$\left\| \frac{1}{T-K}\mathbf{Z}_{\text{g},K}^{\top}\mathbf{Z}_{\text{g},K} - \mathbf{W}_{\text{g},K}^{-1} \right\|_{\text{F}} \xrightarrow{\text{p}} 0.$$

Similar to Lemma B.2.2, we can show that under the same assumptions of Lemma B.2.2 except replacing Assumption 2.3.1, 2.3.4 with Assumption B.1.1, B.1.2, we have

$$\left\| \frac{1}{T-K} \mathbf{Z}_{g,K}^\top (\mathbf{Y}_K - \mathbf{Z}_{g,K} \mathbf{W}_{g,K} \tau) \right\|_2 \xrightarrow{p} 0.$$

As we have different asymptotic variance: $\mathbf{V}_g$ instead of $\mathbf{V}$, similar to Lemma B.2.3, we can show that under the same assumptions of Lemma B.2.3, we have

$$\sqrt{T-K} (\lambda_K^\top \mathbf{V}_g \lambda_K)^{-1/2} \lambda_K^\top \left[\frac{1}{T-K} \mathbf{Z}_{g,K}^\top (\mathbf{Y}_K - \mathbf{Z}_{g,K} \mathbf{W}_{g,K} \tau) \right] \xrightarrow{d} N(0, 1).$$

We then complete the proofs. □

B.3.8 Proof of Proposition B.1.1

Proof of Proposition B.1.1. By the model assumption, we can equivalently write the model as

$$Y_t(z_{t:1}) = \sum_{k=0}^{t-1} \beta_k z_{t-k} + \epsilon_t = p(1-p) \sum_{k=0}^{t-1} \beta_k \frac{z_{t-k} - p}{p(1-p)} + p \sum_{k=0}^{t-1} \beta_k + \epsilon_t.$$

By the model assumption, we have $w_k = p(1-p)$ and $\tau_k = \beta_k$. Then the asymptotic variance of full OLS is

$$\begin{aligned} & \frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_\ell \tau_\ell \right) \right] \\ &= \frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \left(p(1-p) \sum_{\ell=K+1}^{t-1} \beta_\ell \tilde{Z}_{t-\ell} + p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(p(1-p) \sum_{\ell=K+1}^{t-1} \beta_\ell \tilde{Z}_{t-\ell} + p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[p(1-p) \tilde{Z}_{t-k} \sum_{\ell=K+1}^{t-1} \beta_\ell \tilde{Z}_{t-\ell} \right] + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \text{Var} \left[p(1-p) \tilde{Z}_{t-k} \beta_\ell \tilde{Z}_{t-\ell} \right] + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\ &= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \beta_\ell^2 + \frac{1}{T-K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right)^2. \end{aligned}$$

The asymptotic variance of marginal OLS is

$$\begin{aligned}
& \frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} (Y_t - \tilde{Z}_{t-k} w_k \tau_k) \right] \\
&= \frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \left(p(1-p) \sum_{\ell=0, \ell \neq k}^{t-1} \beta_\ell \tilde{Z}_{t-\ell} + p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(p(1-p) \sum_{\ell=0, \ell \neq k}^{t-1} \beta_\ell \tilde{Z}_{t-\ell} + p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[p(1-p) \tilde{Z}_{t-k} \sum_{\ell=0, \ell \neq k}^{t-1} \beta_\ell \tilde{Z}_{t-\ell} \right] + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=0, \ell \neq k}^{t-1} \text{Var} \left[p(1-p) \tilde{Z}_{t-k} \beta_\ell \tilde{Z}_{t-\ell} \right] + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=0, \ell \neq k}^{t-1} \beta_\ell^2 + \frac{1}{T-K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + \epsilon_t \right)^2.
\end{aligned}$$

□

B.3.9 Proof of Proposition B.1.2

Proof of Proposition B.1.2. By the model assumption, we can equivalently write the model as

$$\begin{aligned}
Y_t(z_{t:1}) &= \sum_{k=0}^{t-1} \beta_k z_{t-k} + \sum_{k=1}^{t-1} \beta_{k-1,k} z_{t-k} z_{t-k+1} + \epsilon_t \\
&= p(1-p) \sum_{k=0}^{t-1} \beta_k \frac{z_{t-k} - p}{p(1-p)} + p^2(1-p)^2 \sum_{k=1}^{t-1} \beta_{k-1,k} \frac{z_{t-k} - p}{p(1-p)} \frac{z_{t-k+1} - p}{p(1-p)} \\
&\quad + p \sum_{k=0}^{t-1} \beta_k + p \sum_{k=1}^{t-1} \beta_{k-1,k} (z_{t-k} + z_{t-k+1} - p) + \epsilon_t \\
&= p(1-p) \sum_{k=0}^{t-1} \beta_k \frac{z_{t-k} - p}{p(1-p)} + p^2(1-p)^2 \sum_{k=1}^{t-1} \beta_{k-1,k} \frac{z_{t-k} - p}{p(1-p)} \frac{z_{t-k+1} - p}{p(1-p)} \\
&\quad + p^2(1-p) \sum_{k=1}^{t-1} \beta_{k-1,k} \frac{z_{t-k} - p}{p(1-p)} + p^2(1-p) \sum_{k=0}^{t-2} \beta_{k,k+1} \frac{z_{t-k} - p}{p(1-p)} + p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t.
\end{aligned}$$

By the definition of main effect weights and interaction effect weights, and the model assumption, we have $w_k = p(1 - p)$ and $w_{k-1,k} = p^2(1 - p)^2$. The main effect is

$$\begin{aligned}\tau_k &= \frac{1}{T - K} \sum_{t=K+1}^T \mathbb{E}[Y_t(Z_{t:t-k+1}, 1, Z_{t-k-1:1}) - Y_t(Z_{t:t-k+1}, 0, Z_{t-k-1:1})] \\ &= \frac{1}{T - K} \sum_{t=K+1}^T \mathbb{E}[\beta_k + \beta_{k-1,k} z_{t-k+1} \mathbf{1}(k \neq 0) + \beta_{k,k+1} z_{t-k-1} \mathbf{1}(k \neq T - 1)] \\ &= \beta_k + p\beta_{k-1,k} \mathbf{1}(k \neq 0) + p\beta_{k,k+1} \mathbf{1}(k \neq T - 1),\end{aligned}$$

and the interaction effect is

$$\tau_{k-1,k} = \beta_{k-1,k}.$$

Then the residual of OLS with interaction is

$$\begin{aligned}Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} - \sum_{\ell=1}^K \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} w_{\ell-1,\ell} \tau_{\ell-1,\ell} \\ = p(1 - p) \sum_{\ell=K+1}^{t-1} \beta_{\ell} \tilde{Z}_{t-\ell} + p^2(1 - p)^2 \sum_{\ell=K+1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} \\ + p^2(1 - p) \sum_{\ell=K+1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} + p^2(1 - p) \sum_{\ell=K+1}^{t-2} \beta_{\ell,\ell+1} \tilde{Z}_{t-\ell} + p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t.\end{aligned}$$

Then the asymptotic variance of OLS with interaction is

$$\begin{aligned}& \frac{1}{T - K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} - \sum_{\ell=1}^K \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} w_{\ell-1,\ell} \tau_{\ell-1,\ell} \right) \right] \\ &= \frac{1}{T - K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} - \sum_{\ell=1}^K \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} w_{\ell-1,\ell} \tau_{\ell-1,\ell} \right) \right] \\ &= \frac{1}{T - K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(p(1 - p) \sum_{\ell=K+1}^{t-1} \beta_{\ell} \tilde{Z}_{t-\ell} + p^2(1 - p) \left(\sum_{\ell=K+1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} + \sum_{\ell=K+1}^{t-2} \beta_{\ell,\ell+1} \tilde{Z}_{t-\ell} \right) \right) \right] \\ &+ \frac{1}{T - K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} p^2(1 - p)^2 \sum_{\ell=K+1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} \right] \\ &+ \frac{1}{T - K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right) \tilde{Z}_{t-k} \right]\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \text{Var} \left[p(1-p) \tilde{Z}_{t-k} \beta_{\ell} \tilde{Z}_{t-\ell} \right] \\
&\quad + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \text{Var} \left[p^2(1-p) \tilde{Z}_{t-k} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \right] \\
&\quad + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-2} \text{Var} \left[p^2(1-p) \tilde{Z}_{t-k} \beta_{\ell,\ell+1} \tilde{Z}_{t-\ell} \right] \\
&\quad + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \text{Var} \left[p^2(1-p)^2 \tilde{Z}_{t-k} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} \right] \\
&\quad + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} (\beta_{\ell}^2 + p^2 \beta_{\ell-1,\ell}^2 + p^2 \beta_{\ell,\ell+1}^2) + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} p(1-p) \beta_{\ell-1,\ell}^2 \\
&\quad + \frac{1}{T-K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right)^2.
\end{aligned}$$

On the other hand, the residual of OLS without interaction is

$$\begin{aligned}
&Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} \\
&= p(1-p) \sum_{\ell=K+1}^{t-1} \beta_{\ell} \tilde{Z}_{t-\ell} + p^2(1-p)^2 \sum_{\ell=1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} \\
&\quad + p^2(1-p) \sum_{\ell=K+1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} + p^2(1-p) \sum_{\ell=K+1}^{t-2} \beta_{\ell,\ell+1} \tilde{Z}_{t-\ell} + p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t.
\end{aligned}$$

Then the asymptotic variance of OLS without interaction is

$$\begin{aligned}
&\frac{1}{T-K} \text{Var} \left[\sum_{t=K+1}^T \tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} \right) \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} \left(Y_t - \sum_{\ell=0}^K \tilde{Z}_{t-\ell} w_{\ell} \tau_{\ell} \right) \right] \\
&= \frac{1}{T-K} \sum_{t=K+1}^T
\end{aligned}$$

$$\begin{aligned}
& \text{Var} \left[\tilde{Z}_{t-k} \left(p(1-p) \sum_{\ell=K+1}^{t-1} \beta_{\ell} \tilde{Z}_{t-\ell} + p^2(1-p) \left(\sum_{\ell=K+1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} + \sum_{\ell=K+1}^{t-2} \beta_{\ell,\ell+1} \tilde{Z}_{t-\ell} \right) \right) \right] \\
& + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\tilde{Z}_{t-k} p^2(1-p)^2 \sum_{\ell=1}^{t-1} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} \right] \\
& + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\
& = \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \text{Var} \left[p(1-p) \tilde{Z}_{t-k} \beta_{\ell} \tilde{Z}_{t-\ell} \right] \\
& + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} \text{Var} \left[p^2(1-p) \tilde{Z}_{t-k} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \right] \\
& + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-2} \text{Var} \left[p^2(1-p) \tilde{Z}_{t-k} \beta_{\ell,\ell+1} \tilde{Z}_{t-\ell} \right] \\
& + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=1}^{t-1} \text{Var} \left[p^2(1-p)^2 \tilde{Z}_{t-k} \beta_{\ell-1,\ell} \tilde{Z}_{t-\ell} \tilde{Z}_{t-\ell+1} \right] \\
& + \frac{1}{T-K} \sum_{t=K+1}^T \text{Var} \left[\left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right) \tilde{Z}_{t-k} \right] \\
& = \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=K+1}^{t-1} (\beta_{\ell}^2 + p^2 \beta_{\ell-1,\ell}^2 + p^2 \beta_{\ell,\ell+1}^2) + \frac{1}{T-K} \sum_{t=K+1}^T \sum_{\ell=1}^{t-1} p(1-p) \beta_{\ell-1,\ell}^2 \\
& + \frac{1}{T-K} \frac{1}{p(1-p)} \sum_{t=K+1}^T \left(p \sum_{k=0}^{t-1} \beta_k + p^2 \sum_{k=1}^{t-1} \beta_{k-1,k} + \epsilon_t \right)^2.
\end{aligned}$$

□

Appendix C

Appendix for Chapter 3

C.1 Proofs of the results in Sections 3.3 and 3.4

Additional notation. We use $\mathbf{X}$ and $\mathbf{Z}$ to represent $(X_1, X_2, \dots, X_{N_0})$ and $(Z_1, Z_2, \dots, Z_{N_1})$, respectively. Let $U(0, 1)$ denote the uniform distribution on $[0, 1]$. Let $U \sim U(0, 1)$ and $U_{(M)}$ be the M -th order statistic of N_0 independent random variables from $U(0, 1)$, assumed to be mutually independent and both independent of $(\mathbf{X}, \mathbf{Z})$. It is well known that $U_{(M)} \sim \text{Beta}(M, N_0 + 1 - M)$. Let $\text{Bin}(\cdot, \cdot)$ denote the binomial distribution. Let $L_1(\mathbb{R}^d)$ denote the space of all functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\int |f(x)| dx < \infty$. For any $x \in \mathbb{R}^d$ and function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we say x is a Lebesgue point (Bogachev and Ruas, 2007, Theorem 5.6.2) of f if

$$\lim_{\delta \rightarrow 0^+} \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f(x) - f(z)| dz = 0.$$

C.2 Proofs of the results in Section 3.7

C.2.1 Proof of Theorem 3.7.1

Proof of Theorem 3.7.1. We consider the complexities of two algorithms separately.

Algorithm 1.

The worst-case computation complexity of building a balanced k -d tree is $O(dN_0 \log N_0)$ (cf. Brown (2015)) since the size of the k -d tree is N_0 .

The average computation complexity of searching a NN is $O(\log N_0)$ from Friedman et al. (1977), and then the average computation complexity of search M -NNs in $\{X_i\}_{i=1}^{N_0}$ for all $\{Z_j\}_{j=1}^{N_1}$ is $O(MN_1 \log N_0)$.

Notice that $|S_j| = M$ for any $j \in \llbracket N_1 \rrbracket$ and then $|\bigcup_{j=1}^{N_1} S_j| \leq N_1 M$. Since the elements of each S_j are in $\llbracket N_0 \rrbracket$, the largest integer in $\bigcup_{j=1}^{N_1} S_j$ is N_0 . Then the computation complexity of counting step is $O(N_1 M + N_0)$ due to the counting sort algorithm (Cormen et al., 2009, Section 8.2).

Combining the above three steps completes the proof for Algorithm 1.

Algorithm 2.

The computation complexity of building a k -d tree is $O(d(N_0 + n) \log(N_0 + n))$ from Algorithm 1 since the size of the k -d tree is $N_0 + n$.

For the searching step, for each $j \in \llbracket N_1 \rrbracket$, the number of NNs to be searched is $M + \sum_{i=1}^n \mathbf{1}(\|x_i - Z_j\| \leq \|\mathcal{X}_{(M)}(Z_j) - Z_j\|)$. Then from (3.2.2), the total number of NNs searched for all $j \in \llbracket N_1 \rrbracket$ is $\sum_{j=1}^{N_1} (M + \sum_{i=1}^n \mathbf{1}(\|x_i - Z_j\| \leq \|\mathcal{X}_{(M)}(Z_j) - Z_j\|)) = N_1 M + \sum_{i=1}^n K_M(x_i)$. Let X, Z be two independent copies from ν_0, ν_1 , respectively, and are independent of the data. Since $[Z_j]_{j=1}^{N_1}$ are i.i.d. and $[X_i]_{i=1}^{N_0} \cup [x_i]_{i=1}^n$ are i.i.d, we have $E[\sum_{i=1}^n K_M(x_i)] = nE[K_M(X)] = N_1 n E[\nu_1(A_M(X))] = N_1 n \frac{M}{N_0+1}$ since $E[\nu_1(A_M(X))] = P(\|X - Z\| \leq \|\mathcal{X}_{(M)}(Z) - Z\|) = P(U \leq U_{(M)}) = \frac{M}{N_0+1}$ by using the probability integral transform. Then the average computation complexity for the searching step is $O(N_0^{-1} N_1 M (N_0 + n) \log(N_0 + n))$.

For the counting step, the computation complexity for counting $\bigcup_{j=1}^{N_1} S_j$ is $O(N_0 + N_1 M)$ since the cardinality of $\bigcup_{j=1}^{N_1} S_j$ is at most $N_1 M$ and the largest integer is N_0 . The average computation complexity for counting $\bigcup_{j=1}^{N_1} S'_j$ is $O(N_0^{-1} N_1 M n + n)$ since the average cardinality of $\bigcup_{j=1}^{N_1} S'_j$ is at most $N_0^{-1} N_1 M n$ and the largest integer is n .

Combining the above three steps completes the proof for Algorithm 2. $\square$

C.3 Proofs of the results in Section 3.8

C.3.1 Proof of Lemma 3.8.1

Proof of Lemma 3.8.1. From the Lebesgue differentiation theorem, for any $f \in L_1(\mathbb{R}^d)$, x is a Lebesgue point of f for λ -almost all x . Then for ν_0 -almost all x , we have $f_0(x) > 0$ and x is a Lebesgue point of f_0 and f_1 from the absolute continuity of ν_0 and ν_1 . We then only need to consider those $x \in \mathbb{R}^d$ such that $f_0(x) > 0$ and x is a Lebesgue point of f_0 and f_1 .

We first introduce a lemma about the Lebesgue point.

Lemma C.3.1. *Let ν be a probability measure on $\mathbb{R}^d$ admitting a density f with respect to the Lebesgue measure. Let $x \in \mathbb{R}^d$ be a Lebesgue point of f . Then for any $\epsilon \in (0, 1)$, there exists $\delta = \delta_x > 0$ such that for any $z \in \mathbb{R}^d$ satisfying $\|z - x\| \leq \delta$, we have*

$$\left| \frac{\nu(B_{x, \|z-x\|})}{\lambda(B_{x, \|z-x\|})} - f(x) \right| \leq \epsilon, \quad \left| \frac{\nu(B_{z, \|z-x\|})}{\lambda(B_{z, \|z-x\|})} - f(x) \right| \leq \epsilon.$$

Part I. This part proves the first claim. We separate the proof of Part I into two cases based on the value of $f_1(x)$.

Case I.1. $f_1(x) > 0$. Since x is a Lebesgue point of ν_0 and ν_1 , by Lemma C.3.1, for any $\epsilon \in (0, 1)$, there exists some $\delta = \delta_x > 0$ such that for any $z \in \mathbb{R}^d$ with $\|z - x\| \leq \delta$, we have for $w \in \{0, 1\}$,

$$\left| \frac{\nu_w(B_{x, \|z-x\|})}{\lambda(B_{x, \|z-x\|})} - f_w(x) \right| \leq \epsilon f_w(x), \quad \left| \frac{\nu_w(B_{z, \|z-x\|})}{\lambda(B_{z, \|z-x\|})} - f_w(x) \right| \leq \epsilon f_w(x).$$

Accordingly, if $\|z - x\| \leq \delta$, by $\lambda(B_{z,\|x-z\|}) = \lambda(B_{x,\|x-z\|})$, we have

$$\frac{1 - \epsilon f_0(x)}{1 + \epsilon f_1(x)} \leq \frac{\nu_0(B_{z,\|x-z\|})}{\lambda(B_{z,\|x-z\|})} \frac{\lambda(B_{x,\|x-z\|})}{\nu_1(B_{x,\|x-z\|})} = \frac{\nu_0(B_{z,\|x-z\|})}{\nu_1(B_{x,\|x-z\|})} \leq \frac{1 + \epsilon f_0(x)}{1 - \epsilon f_1(x)}. \quad (\text{C.3.1})$$

On the other hand, for any $z \in \mathbb{R}^d$ such that $\|z - x\| > \delta$, $\nu_0(B_{z,\|z-x\|}) \geq \nu_0(B_{z^*,\delta}) \geq (1 - \epsilon)f_0(x)\lambda(B_{z^*,\delta}) = (1 - \epsilon)f_0(x)\lambda(B_{0,\delta})$, where z^* is the intersection point of the surface of $B_{x,\delta}$ and the line connecting z and x .

Let $\eta_N = 4 \log(N_0/M)$. Since $M \log N_0/N_0 \rightarrow 0$, we can take N_0 large enough so that $\eta_N \frac{M}{N_0} = 4 \frac{M}{N_0} \log(\frac{N_0}{M}) < (1 - \epsilon)f_0(x)\lambda(B_{0,\delta})$. Then for any $z \in \mathbb{R}^d$ such that $\nu_0(B_{z,\|z-x\|}) \leq \eta_N M/N_0$, we have $\|z - x\| \leq \delta$ since otherwise it would contradict the selection of N_0 .

Let Z be a copy from ν_1 independent of the data. Then

$$\mathbb{E}[\nu_1(A_M(x))] = \mathbb{P}(Z \in A_M(x)) = \mathbb{P}(\nu_0(B_{Z,\|x-Z\|}) \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|})). \quad (\text{C.3.2})$$

For any given $z \in \mathbb{R}^d$, $[\nu_0(B_{z,\|X_i-z\|})]_{i=1}^{N_0}$ are i.i.d. from $U(0,1)$ since $[X_i]_{i=1}^{N_0}$ are i.i.d. from ν_0 and we use the probability integral transform. Then $\nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|})$ has the same distribution as $U_{(M)}$ and is independent of Z .

Upper bound. With a slight abuse of notation, we define $W = \nu_0(B_{Z,\|x-Z\|})$. We then have, from (C.3.1) and (C.3.2),

$$\begin{aligned} \mathbb{E}[\nu_1(A_M(x))] &= \mathbb{P}(W \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|})) \\ &\leq \mathbb{P}\left(W \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|}) \leq \eta_N \frac{M}{N_0}\right) + \mathbb{P}\left(\nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|}) > \eta_N \frac{M}{N_0}\right) \\ &= \mathbb{P}\left(W \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|}) \leq \eta_N \frac{M}{N_0}, \|Z - x\| \leq \delta\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) \\ &\leq \mathbb{P}\left(\nu_0(B_{Z,\|x-Z\|}) \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|}), \|Z - x\| \leq \delta\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) \\ &\leq \mathbb{P}\left(\frac{1 - \epsilon f_0(x)}{1 + \epsilon f_1(x)} \nu_1(B_{x,\|x-Z\|}) \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|}), \|Z - x\| \leq \delta\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) \\ &\leq \mathbb{P}\left(\frac{1 - \epsilon f_0(x)}{1 + \epsilon f_1(x)} \nu_1(B_{x,\|x-Z\|}) \leq \nu_0(B_{Z,\|\mathcal{X}_{(M)}(Z)-Z\|})\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) \\ &= \mathbb{P}\left(\frac{1 - \epsilon f_0(x)}{1 + \epsilon f_1(x)} U \leq U_{(M)}\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right). \end{aligned} \quad (\text{C.3.3})$$

For the second term in (C.3.3), notice that $\eta_N \rightarrow \infty$ as $N_0 \rightarrow \infty$. Then from the Chernoff bound and for N_0 sufficiently large, we have

$$\begin{aligned} \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) &= \frac{N_0}{M} \mathbb{P}\left(\text{Bin}\left(N_0, \eta_N \frac{M}{N_0}\right) < M\right) \\ &\leq \frac{N_0}{M} \exp\left((1 + \log \eta_N - \eta_N)M\right) \leq \frac{N_0}{M} \exp\left(-\frac{1}{2}\eta_N M\right) = \left(\frac{N_0}{M}\right)^{1-2M}. \end{aligned}$$

Since $M/N_0 \rightarrow 0$ and $M \geq 1$, we then obtain

$$\lim_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) = 0. \quad (\text{C.3.4})$$

For the first term in (C.3.3), we have

$$\begin{aligned} \frac{N_0}{M} \mathbb{P}\left(\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} U \leq U_{(M)}\right) &= \frac{N_0}{M} \int_0^1 \mathbb{P}\left(U_{(M)} \geq \frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} t\right) dt \\ &= \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \int_0^{\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} \frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt \leq \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \int_0^\infty \mathbb{P}\left(\frac{N_0}{M} U_{(M)} \geq t\right) dt \\ &= \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \frac{N_0}{M} \mathbb{E}[U_{(M)}] = \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \frac{N_0}{N_0+1}. \end{aligned} \quad (\text{C.3.5})$$

We then obtain

$$\limsup_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{P}\left(\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} U \leq U_{(M)}\right) \leq \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)}. \quad (\text{C.3.6})$$

Plugging (C.3.4) and (C.3.6) to (C.3.3) then yields

$$\limsup_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))] \leq \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)}. \quad (\text{C.3.7})$$

Lower bound. We have, from (C.3.1) and (C.3.2),

$$\begin{aligned} \mathbb{E}[\nu_1(A_M(x))] &= \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})\right) \geq \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq \eta_N \frac{M}{N_0}\right) \\ &= \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq \eta_N \frac{M}{N_0}, \|Z - x\| \leq \delta\right) \\ &\geq \mathbb{P}\left(\frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} \nu_1(B_{x, \|x - Z\|}) \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq \eta_N \frac{M}{N_0}, \|Z - x\| \leq \delta\right) \\ &= \mathbb{P}\left(\frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} \nu_1(B_{x, \|x - Z\|}) \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq \eta_N \frac{M}{N_0}\right) \\ &\geq \mathbb{P}\left(\frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} \nu_1(B_{x, \|x - Z\|}) \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})\right) - \mathbb{P}\left(\nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) > \eta_N \frac{M}{N_0}\right) \\ &= \mathbb{P}\left(\frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} U \leq U_{(M)}\right) - \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right). \end{aligned} \quad (\text{C.3.8})$$

The second last equality is from the fact that for $\|Z - x\| > \delta$,

$$\frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} \nu_1(B_{x, \|x - Z\|}) \geq \frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} \nu_1(B_{x, \delta}) \geq \frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} f_1(x) (1-\epsilon) \lambda(B_{0, \delta}) > \eta_N \frac{M}{N_0},$$

and then that $\frac{1+\epsilon}{1-\epsilon} \frac{f_0(x)}{f_1(x)} \nu_1(B_{x, \|x - Z\|}) \leq \eta_N \frac{M}{N_0}$ implies $\|Z - x\| \leq \delta$.

For the first term in (C.3.8), we have

$$\frac{N_0}{M} \mathbb{P}\left(\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} U \leq U_{(M)}\right) = \frac{1-\epsilon f_1(x)}{1+\epsilon f_0(x)} \int_0^{\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} \frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt.$$

If $\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} \geq 1$, then by $U_{(M)} \in [0, 1]$, we have

$$\frac{N_0}{M} \mathbb{P}\left(\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} U \leq U_{(M)}\right) = \frac{1-\epsilon f_1(x)}{1+\epsilon f_0(x)} \frac{N_0}{M} \mathbb{E}[U_{(M)}] = \frac{1-\epsilon f_1(x)}{1+\epsilon f_0(x)} \frac{N_0}{N_0+1}.$$

If $\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} < 1$, from the Chernoff bound,

$$\begin{aligned} & \int_0^{\frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt \leq \left[1 - \frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)}\right] \frac{N_0}{M} \mathbb{P}\left(U_{(M)} \geq \frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)}\right) \\ & \leq \left[1 - \frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)}\right] \frac{N_0}{M} \exp\left[M - \frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} N_0 - M \log M + M \log\left(\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} N_0\right)\right]. \end{aligned}$$

Since $f_0(x) > 0$ and $M \log N_0/N_0 \rightarrow 0$, we obtain

$$\lim_{N_0 \rightarrow \infty} \int_0^{\frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt = 0.$$

Then we always have

$$\lim_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{P}\left(\frac{1+\epsilon f_0(x)}{1-\epsilon f_1(x)} U \leq U_{(M)}\right) = \frac{1-\epsilon f_1(x)}{1+\epsilon f_0(x)}.$$

Using the above identity along with (C.3.4) to (C.3.8) yields

$$\liminf_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))] \geq \frac{1-\epsilon f_1(x)}{1+\epsilon f_0(x)}. \quad (\text{C.3.9})$$

Lastly, combining (C.3.7) with (C.3.9) and noticing that ϵ is arbitrary, we obtain

$$\lim_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))] = \frac{f_1(x)}{f_0(x)} = r(x). \quad (\text{C.3.10})$$

Case I.2. $f_1(x) = 0$. Again, for any $\epsilon \in (0, 1)$, by Lemma C.3.1, there exists some $\delta = \delta_x > 0$ such that for any $z \in \mathbb{R}^d$ with $\|z - x\| \leq \delta$, we have

$$\left| \frac{\nu_0(B_{z, \|z-x\|})}{\lambda(B_{z, \|z-x\|})} - f_0(x) \right| \leq \epsilon f_0(x), \quad \left| \frac{\nu_1(B_{x, \|z-x\|})}{\lambda(B_{x, \|z-x\|})} \right| \leq \epsilon.$$

Recall that $W = \nu_0(B_{Z, \|x-Z\|})$. Then if $\|Z - x\| \leq \delta$, we have

$$W \geq (1 - \epsilon)f_0(x)\lambda(B_{Z, \|x-Z\|}) = (1 - \epsilon)f_0(x)\lambda(B_{x, \|x-Z\|}) \geq \epsilon^{-1}(1 - \epsilon)f_0(x)\nu_1(B_{x, \|x-Z\|}).$$

Proceeding in the same way as (C.3.3), we obtain

$$\begin{aligned} \mathbb{E}[\nu_1(A_M(x))] &\leq \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})\eta_N \frac{M}{N_0}, \|Z - x\| \leq \delta\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) \\ &\leq \mathbb{P}\left(\frac{1 - \epsilon}{\epsilon}f_0(x)U \leq U_{(M)}\right) + \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right). \end{aligned}$$

For the first term above,

$$\begin{aligned} \frac{N_0}{M}\mathbb{P}\left(\frac{1 - \epsilon}{\epsilon}f_0(x)U \leq U_{(M)}\right) &= \frac{\epsilon}{1 - \epsilon} \frac{1}{f_0(x)} \int_0^{\frac{1 - \epsilon}{\epsilon}f_0(x) \frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0}t\right) dt \\ &\leq \frac{\epsilon}{1 - \epsilon} \frac{1}{f_0(x)} \int_0^\infty \mathbb{P}\left(\frac{N_0}{M}U_{(M)} \geq t\right) dt = \frac{\epsilon}{1 - \epsilon} \frac{1}{f_0(x)} \frac{N_0}{M} \mathbb{E}[U_{(M)}] = \frac{\epsilon}{1 - \epsilon} \frac{1}{f_0(x)} \frac{N_0}{N_0 + 1}. \end{aligned}$$

By (C.3.4) and noticing ϵ is arbitrary, we have

$$\lim_{N_0 \rightarrow \infty} \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))] = 0 = r(x). \quad (\text{C.3.11})$$

Combining (C.3.10) and (C.3.11) completes the proof of the first claim.

Part II. This part proves the second claim. We also separate the proof of Part II into two cases based on the value of $f_1(x)$.

Case II.1. $f_1(x) > 0$. Again, for any $\epsilon \in (0, 1)$, we take δ in the same way as in Case I.1. Let $\eta_N = \eta_{N,p} = 4p \log(N_0/M)$. We also take N_0 sufficiently large so that $\eta_N \frac{M}{N_0} = 4p \frac{M}{N_0} \log(\frac{N_0}{M}) < (1 - \epsilon)f_0(x)\lambda(B_{0,\delta})$.

Let $\tilde{Z}_1, \dots, \tilde{Z}_p$ be p independent copies that are drawn from ν_1 independent of the data. Then

$$\begin{aligned} \mathbb{E}[\nu_1^p(A_M(x))] &= \mathbb{P}\left(\tilde{Z}_1, \dots, \tilde{Z}_p \in A_M(x)\right) \\ &= \mathbb{P}\left(\nu_0(B_{\tilde{Z}_1, \|x - \tilde{Z}_1\|}) \leq \nu_0(B_{\tilde{Z}_1, \|\mathcal{X}_{(M)}(\tilde{Z}_1) - \tilde{Z}_1\|}), \dots, \nu_0(B_{\tilde{Z}_p, \|x - \tilde{Z}_p\|}) \leq \nu_0(B_{\tilde{Z}_p, \|\mathcal{X}_{(M)}(\tilde{Z}_p) - \tilde{Z}_p\|})\right). \end{aligned}$$

Let $W_k = \nu_0(B_{\tilde{Z}_k, \|x - \tilde{Z}_k\|})$ and $V_k = \nu_0(B_{\tilde{Z}_k, \|\mathcal{X}_{(M)}(\tilde{Z}_k) - \tilde{Z}_k\|})$ for any $k \in \llbracket p \rrbracket$. Then $[W_k]_{k=1}^p$ are i.i.d. since $[\tilde{Z}_k]_{k=1}^p$ are i.i.d.. For any $k \in \llbracket p \rrbracket$ and $\tilde{Z}_k \in \mathbb{R}^d$ given, $V_k \mid \tilde{Z}_k$ has the same distribution as $U_{(M)}$. Then for any $k \in \llbracket p \rrbracket$, V_k has the same distribution as $U_{(M)}$, and V_k is independent of $\tilde{Z}_k$.

Let $W_{\max} = \max_{k \in \llbracket p \rrbracket} W_k$ and $V_{\max} = \max_{k \in \llbracket p \rrbracket} V_k$. Then

$$\mathbb{E}[\nu_1^p(A_M(x))] \leq \mathbb{P}\left(W_{\max} \leq V_{\max}\right) \leq \mathbb{P}\left(W_{\max} \leq V_{\max} \leq \eta_N \frac{M}{N_0}\right) + \mathbb{P}\left(V_{\max} > \eta_N \frac{M}{N_0}\right). \quad (\text{C.3.12})$$

For the second term in (C.3.12),

$$\mathbb{P}\left(V_{\max} > \eta_N \frac{M}{N_0}\right) \leq \sum_{k=1}^p \mathbb{P}\left(V_k > \eta_N \frac{M}{N_0}\right) = p \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right).$$

Proceeding as (C.3.4),

$$\left(\frac{N_0}{M}\right)^p \mathbb{P}\left(U_{(M)} > \eta_N \frac{M}{N_0}\right) \leq \left(\frac{N_0}{M}\right)^p \exp\left(-\frac{1}{2}\eta_N M\right) = \left(\frac{N_0}{M}\right)^{p(1-2M)}.$$

We then obtain

$$\lim_{N_0 \rightarrow \infty} \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(V_{\max} > \eta_N \frac{M}{N_0}\right) = 0. \quad (\text{C.3.13})$$

For the first term in (C.3.12), notice that $[\nu_1(B_{x, \|\tilde{Z}_k - x\|})]_{k=1}^p$ are i.i.d. from $U(0, 1)$ since $[\tilde{Z}_k]_{k=1}^p$ are i.i.d. We then have

$$\begin{aligned} & \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(W_{\max} \leq V_{\max} \leq \eta_N \frac{M}{N_0}\right) = \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(W_{\max} \leq V_{\max} \leq \eta_N \frac{M}{N_0}, \max_{k \in \llbracket p \rrbracket} \|\tilde{Z}_k - x\| \leq \delta\right) \\ & \leq \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) \leq V_{\max} \leq \eta_N \frac{M}{N_0}, \max_{k \in \llbracket p \rrbracket} \|\tilde{Z}_k - x\| \leq \delta\right) \\ & \leq \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) \leq V_{\max}\right) \\ & = \left(\frac{N_0}{M}\right)^p \int_0^1 p t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = t\right) dt \\ & = p \left(\frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)}\right)^p \int_0^{\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} \frac{N_0}{M}} t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \frac{M}{N_0} t\right) dt \\ & = p \left(\frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)}\right)^p \left[\int_0^1 t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \frac{M}{N_0} t\right) dt \right. \\ & \quad \left. + \int_1^{\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} \frac{N_0}{M}} t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \frac{M}{N_0} t\right) dt \right]. \end{aligned}$$

For the first term,

$$\int_0^1 t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = \frac{1+\epsilon}{1-\epsilon} \frac{f_1(x)}{f_0(x)} \frac{M}{N_0} t\right) dt \leq \int_0^1 t^{p-1} dt = \frac{1}{p}.$$

For the second term, using the Chernoff bound, conditional on $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_p)$,

$$\int_1^{\frac{1-\epsilon}{1+\epsilon} \frac{f_0(x)}{f_1(x)} \frac{N_0}{M}} t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} t \mid \tilde{\mathbf{Z}}\right) dt \leq \int_0^\infty (1+t)^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} (1+t) \mid \tilde{\mathbf{Z}}\right) dt$$

$$\begin{aligned}
&\leq \int_0^\infty (1+t)^{p-1} \left[\sum_{k=1}^p \mathbb{P}\left(V_k \geq \frac{M}{N_0}(1+t) \mid \tilde{\mathbf{Z}}\right) \right] dt = p \int_0^\infty (1+t)^{p-1} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0}(1+t)\right) dt \\
&\leq p \int_0^\infty (1+t)^{p-1} (1+t)^M \exp(-tM) dt \leq \sqrt{2\pi} p M^{-1/2} \left(1 + \frac{1}{M}\right)^{p-1} (1+o(1)),
\end{aligned}$$

where the last step follows from Stirling's approximation with $M \rightarrow \infty$.

Then we obtain

$$\limsup_{N_0 \rightarrow \infty} \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(W_{\max} \leq V_{\max}, V_{\max} \leq \eta_N \frac{M}{N_0}\right) \leq \left(\frac{1+\epsilon f_1(x)}{1-\epsilon f_0(x)}\right)^p. \quad (\text{C.3.14})$$

Plugging (C.3.13) and (C.3.14) into (C.3.12) yields

$$\limsup_{N_0 \rightarrow \infty} \left(\frac{N_0}{M}\right)^p \mathbb{E}[\nu_1^p(A_M(x))] \leq \left(\frac{1+\epsilon f_1(x)}{1-\epsilon f_0(x)}\right)^p = \left(\frac{1+\epsilon}{1-\epsilon} r(x)\right)^p. \quad (\text{C.3.15})$$

Lastly, using Hölder's inequality,

$$\left(\frac{N_0}{M}\right)^p \mathbb{E}[\nu_1^p(A_M(x))] \geq \left[\frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))]\right]^p.$$

Employing the first claim, we have

$$\liminf_{N_0 \rightarrow \infty} \left(\frac{N_0}{M}\right)^p \mathbb{E}[\nu_1^p(A_M(x))] \geq [r(x)]^p. \quad (\text{C.3.16})$$

Combining (C.3.15) with (C.3.16) and noting that ϵ is arbitrary, we obtain

$$\lim_{N_0 \rightarrow \infty} \left(\frac{N_0}{M}\right)^p \mathbb{E}[\nu_1^p(A_M(x))] = [r(x)]^p. \quad (\text{C.3.17})$$

Case II.2. $f_1(x) = 0$. For any $\epsilon \in (0, 1)$, we take δ in the same way as in the proof of Case I.2 and take η_N as in the proof of Case II.1.

By (C.3.12),

$$\left(\frac{N_0}{M}\right)^p \mathbb{E}[\nu_1^p(A_M(x))] \leq \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(W_{\max} \leq V_{\max} \leq \eta_N \frac{M}{N_0}\right) + \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(V_{\max} > \eta_N \frac{M}{N_0}\right).$$

For the first term,

$$\begin{aligned}
&\left(\frac{N_0}{M}\right)^p \mathbb{P}\left(W_{\max} \leq V_{\max} \leq \eta_N \frac{M}{N_0}\right) \leq \left(\frac{N_0}{M}\right)^p \mathbb{P}\left(\frac{1-\epsilon}{\epsilon} f_0(x) \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) \leq V_{\max}\right) \\
&= \left(\frac{N_0}{M}\right)^p \int_0^1 p t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{1-\epsilon}{\epsilon} f_0(x) t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = t\right) dt \\
&= p \left(\frac{\epsilon}{1-\epsilon} \frac{1}{f_0(x)}\right)^p \int_0^{\frac{1-\epsilon}{\epsilon} f_0(x) \frac{N_0}{M}} t^{p-1} \mathbb{P}\left(V_{\max} \geq \frac{M}{N_0} t \mid \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x, \|\tilde{Z}_k - x\|}) = t\right) dt.
\end{aligned}$$

Then proceeding in the same way as (C.3.14), we have

$$\limsup_{N_0 \rightarrow \infty} \left(\frac{N_0}{M} \right)^p \mathbb{P} \left(W_{\max} \leq V_{\max} \leq \eta_N \frac{M}{N_0} \right) \leq \left(\frac{\epsilon}{1 - \epsilon} \frac{1}{f_0(x)} \right)^p.$$

Lastly, using (C.3.13) and noting again that ϵ is arbitrary, we obtain

$$\lim_{N_0 \rightarrow \infty} \left(\frac{N_0}{M} \right)^p \mathbb{E}[\nu_1^p(A_M(x))] = 0 = [r(x)]^p. \quad (\text{C.3.18})$$

Combining (C.3.17) and (C.3.18) then completes the proof of the second claim. $\square$

C.3.2 Proof of Theorem 3.8.1

Proof of Theorem 3.8.1(i). By (3.2.4) and that $[Z_j]_{j=1}^{N_1}$ are i.i.d,

$$\mathbb{E}[\widehat{r}_M(x)] = \mathbb{E}\left[\frac{N_0}{N_1} \frac{K_M(x)}{M}\right] = \frac{N_0}{N_1 M} \mathbb{E}\left[\sum_{j=1}^{N_1} \mathbf{1}(Z_j \in A_M(x))\right] = \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))].$$

Employing Lemma 3.8.1 then completes the proof. $\square$

Proof of Theorem 3.8.1(ii). By Hölder's inequality, it suffices to consider the case when p is even. Because x^p is convex for $p > 1$ and $x > 0$, we have

$$\mathbb{E}[|\widehat{r}_M(x) - r(x)|^p] \leq 2^{p-1} \left(\mathbb{E}\left[\left|\widehat{r}_M(x) - \mathbb{E}[\widehat{r}_M(x) | \mathbf{X}]\right|^p\right] + \mathbb{E}\left[\left|\mathbb{E}[\widehat{r}_M(x) | \mathbf{X}] - r(x)\right|^p\right] \right). \quad (\text{C.3.19})$$

For the second term in (C.3.19), Lemma 3.8.1 implies

$$\lim_{N_0 \rightarrow \infty} \mathbb{E}\left[\left|\mathbb{E}[\widehat{r}_M(x) | \mathbf{X}] - r(x)\right|^p\right] = \lim_{N_0 \rightarrow \infty} \mathbb{E}\left[\left|\frac{N_0}{M} \nu_1(A_M(x)) - r(x)\right|^p\right] = 0 \quad (\text{C.3.20})$$

by expanding the product term.

For the first term in (C.3.19), noticing that $[Z_j]_{j=1}^{N_1}$ are i.i.d, we have $K_M(x) | \mathbf{X} \sim \text{Bin}(N_1, \nu_1(A_M(x)))$. Using Lemma 3.8.1 and $MN_1/N_0 \rightarrow \infty$, for any positive integers p and q , we have

$$\begin{aligned} \lim_{N_0 \rightarrow \infty} \left(\frac{N_0}{N_1 M} \right)^p \mathbb{E}[N_1^p \nu_1^p(A_M(x))] &= [r(x)]^p, \\ \lim_{N_0 \rightarrow \infty} \left(\frac{N_0}{N_1 M} \right)^p \left(\frac{N_0}{M} \right)^q \mathbb{E}[N_1^p \nu_1^{p+q}(A_M(x))] &= [r(x)]^{p+q}, \end{aligned}$$

and then $\mathbb{E}[N_1^p \nu_1^p(A_M(x))]$ is the dominated term among $[\mathbb{E}[N_1^k \nu_1^{k+q}(A_M(x))]]_{k \leq p, q \geq 0}$.

To complete the proof, for any positive integer c and $Z \sim \text{Bin}(n, p')$, let $\mu_c = \mathbb{E}[(Z - \mathbb{E}[Z])^c]$ be the c -th central moment. By [Romanovsky \(1923\)](#), we have

$$\mu_{c+1} = p'(1 - p') \left(nc\mu_{c-1} + \frac{d\mu_c}{dp'} \right).$$

Then for even p , we obtain

$$\mathbb{E} \left[\left(K_M(x) - N_1 \nu_1(A_M(x)) \right)^p \right] \lesssim \mathbb{E} [N_1 \nu_1(A_M(x))]^{p/2} \lesssim \left(\frac{N_1 M}{N_0} \right)^{p/2}.$$

The first term in (C.3.19) then satisfies

$$\mathbb{E} \left[\left| \hat{r}_M(x) - \mathbb{E}[\hat{r}_M(x) | \mathbf{X}] \right|^p \right] = \left(\frac{N_0}{N_1 M} \right)^p \mathbb{E} \left[\left(K_M(x) - N_1 \nu_1(A_M(x)) \right)^p \right] \lesssim \left(\frac{N_0}{N_1 M} \right)^{p/2}.$$

Since $MN_1/N_0 \rightarrow \infty$, we obtain

$$\lim_{N_0 \rightarrow \infty} \mathbb{E} \left[\left| \hat{r}_M(x) - \mathbb{E}[\hat{r}_M(x) | \mathbf{X}] \right|^p \right] = 0. \quad (\text{C.3.21})$$

Plugging (C.3.20) and (C.3.21) into (C.3.19) then completes the proof. $\square$

C.3.3 Proof of Theorem 3.8.2

Proof of Theorem 3.8.2. We first cite the Hardy–Littlewood maximal inequality.

Lemma C.3.2 (Hardy–Littlewood maximal inequality ([Stein, 2016](#))). *For any locally integrable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, define*

$$\mathbf{M}f(x) = \sup_{\delta > 0} \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f(z)| dz.$$

Then for $d \geq 1$, there exists a constant $C_d > 0$ only depending on d such that for all $t > 0$ and $f \in L_1(\mathbb{R}^d)$, we have

$$\lambda(\{x : \mathbf{M}f(x) > t\}) < \frac{C_d}{t} \|f\|_{L_1},$$

where $\|\cdot\|_{L_1}$ stands for the function L_1 norm.

Let $\epsilon > 0$ be given. We assume $\epsilon \leq f_L$. From Assumption 3.8.1, S_0 and S_1 are bounded, then ν_0 and ν_1 are compactly supported. Since $f_0, f_1 \in L_1$ and the class of continuous functions are dense in the class of compactly supported L_1 functions from simple use of Lusin's theorem, we can find g_0, g_1 such that g_0, g_1 are continuous and $\|f_0 - g_0\|_{L_1} \leq \epsilon^3$ and $\|f_1 - g_1\|_{L_1} \leq \epsilon^3$.

Since g_0, g_1 are continuous with compact supports, they are uniformly continuous, that is, there exists $\delta > 0$ such that for any $x, z \in \mathbb{R}^d$ and $\|z - x\| \leq \delta$, we have $|g_0(x) - g_0(z)| \leq \frac{\epsilon^2}{3}$ and $|g_1(x) - g_1(z)| \leq \frac{\epsilon^2}{3}$.

For any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
& \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_0(x) - f_0(z)| dz \\
& \leq \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} [|f_0(x) - g_0(x)| + |g_0(x) - g_0(z)| + |f_0(z) - g_0(z)|] dz \\
& = |f_0(x) - g_0(x)| + \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |g_0(x) - g_0(z)| dz + \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_0(z) - g_0(z)| dz.
\end{aligned} \tag{C.3.22}$$

For the first term in (C.3.22), using Markov's inequality, we have

$$\lambda(\{x : |f_0(x) - g_0(x)| > \epsilon^2/3\}) \leq 3\epsilon^{-2} \|f_0 - g_0\|_{L_1} \leq 3\epsilon. \tag{C.3.23}$$

For the second term in (C.3.22), by the selection of δ ,

$$\frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |g_0(x) - g_0(z)| dz \leq \max_{z \in B_{x,\delta}} |g_0(x) - g_0(z)| \leq \frac{\epsilon^2}{3}. \tag{C.3.24}$$

For the third term,

$$\frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_0(z) - g_0(z)| dz \leq \sup_{\delta > 0} \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_0(z) - g_0(z)| dz = \mathbf{M}(f_0 - g_0)(x).$$

Lemma C.3.2 then yields

$$\lambda(\{x : \mathbf{M}(f_0 - g_0)(x) > \epsilon^2/3\}) < 3C_d \epsilon^{-2} \|f_0 - g_0\|_{L_1} \leq 3C_d \epsilon. \tag{C.3.25}$$

We can establish similar results for f_1, g_1 .

Let

$$\begin{aligned}
A_1 = & \{x : |f_0(x) - g_0(x)| > \epsilon^2/3\} \cup \{x : |f_1(x) - g_1(x)| > \epsilon^2/3\} \cup \\
& \{x : \mathbf{M}(f_0 - g_0)(x) > \epsilon^2/3\} \cup \{x : \mathbf{M}(f_1 - g_1)(x) > \epsilon^2/3\}.
\end{aligned}$$

Plugging (C.3.23), (C.3.24), (C.3.25) into (C.3.22), for any $x \notin A_1$ and $\|z - x\| \leq \delta$, we have

$$\frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_0(x) - f_0(z)| dz \leq \epsilon^2, \quad \frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_1(x) - f_1(z)| dz \leq \epsilon^2,$$

and $\lambda(A_1) \leq 6(C_d + 1)\epsilon$.

Let $A_2 = \{x : f_1(x) \leq \epsilon\}$. We then separate the proof into three cases. In the following, it suffices to consider $f_0(x) > 0$ due to the definition of L_p risk.

Case I. $x \notin A_1 \cup A_2$. By $\epsilon \leq f_L$ and the definition of A_2 , for any $x \notin A_1 \cup A_2$ and $\|z - x\| \leq \delta$,

$$\frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_0(x) - f_0(z)| dz \leq \epsilon^2 \leq \epsilon f_L \leq \epsilon f_0(x),$$

$$\frac{1}{\lambda(B_{x,\delta})} \int_{B_{x,\delta}} |f_1(x) - f_1(z)| dz \leq \epsilon^2 \leq \epsilon f_1(x).$$

We then obtain for $w \in \{0, 1\}$,

$$\left| \frac{\nu_w(B_{x,\|z-x\|})}{\lambda(B_{x,\|z-x\|})} - f_w(x) \right| \leq \epsilon f_w(x), \quad \left| \frac{\nu_w(B_{z,\|z-x\|})}{\lambda(B_{z,\|z-x\|})} - f_w(x) \right| \leq \epsilon f_w(x).$$

Let $\eta_N = \eta_{N,p} = 4p \log(N_0/M)$. We also take N_0 large enough so that $\eta_N \frac{M}{N_0} = 4p \frac{M}{N_0} \log(\frac{N_0}{M}) < (1 - \epsilon) f_L \lambda(B_{0,\delta})$. Then for any $x \in \mathbb{R}^d$ such that $f_0(x) > 0$, we have $\eta_N \frac{M}{N_0} < (1 - \epsilon) f_0(x) \lambda(B_{0,\delta})$.

Proceeding as in the proof of Case II.1 of Lemma 3.8.1 and also Theorem 3.8.1 by using Fubini's theorem, since ϵ is arbitrary, we obtain

$$\lim_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right|^p f_0(x) \mathbf{1}(x \notin A_1 \cup A_2) dx \right] = 0. \quad (\text{C.3.26})$$

Case II. $x \in A_2 \setminus A_1$. In this case, we have

$$\begin{aligned} \left| \frac{\nu_0(B_{x,\|z-x\|})}{\lambda(B_{x,\|z-x\|})} - f_0(x) \right| &\leq \epsilon f_0(x), \quad \left| \frac{\nu_0(B_{z,\|z-x\|})}{\lambda(B_{z,\|z-x\|})} - f_0(x) \right| \leq \epsilon f_0(x), \\ \left| \frac{\nu_1(B_{x,\|z-x\|})}{\lambda(B_{x,\|z-x\|})} - f_1(x) \right| &\leq \epsilon^2, \quad \left| \frac{\nu_1(B_{z,\|z-x\|})}{\lambda(B_{z,\|z-x\|})} - f_1(x) \right| \leq \epsilon^2. \end{aligned}$$

Take η_N and take N_0 sufficiently large as in Case I above. Proceeding as the proof of Case II.2 of Lemma 3.8.1 and also Theorem 3.8.1 by using Fubini's theorem, since ϵ is arbitrary, we obtain

$$\lim_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right|^p f_0(x) \mathbf{1}(x \in A_2 \setminus A_1) dx \right] = 0. \quad (\text{C.3.27})$$

Case III. $x \in A_1$. In this case, for any $x \in A_1$ and $z \in S_1$, $\nu_0(B_{z,\|z-x\|}) \geq f_L \lambda(B_{z,\|z-x\|} \cap S_0) \geq a f_L \lambda(B_{z,\|z-x\|}) \geq \frac{a f_L}{f_U} \nu_1(B_{x,\|z-x\|})$. Then for any $x \in A_1$, from (C.3.12) and in the same way as (C.3.14),

$$\begin{aligned} \left(\frac{N_0}{M} \right)^p \mathbb{E} [\nu_1^p(A_M(x))] &\leq \left(\frac{N_0}{M} \right)^p \mathbb{P}(W_{\max} \leq V_{\max}) \\ &\leq \left(\frac{N_0}{M} \right)^p \mathbb{P} \left(\frac{a f_L}{f_U} \max_{k \in \llbracket p \rrbracket} \nu_1(B_{x,\|\tilde{Z}_k - x\|}) \leq V_{\max} \right) \leq \left(\frac{f_U}{a f_L} \right)^p (1 + o(1)) = O(1). \end{aligned}$$

Proceeding as in the proof of Theorem 3.8.1, and due to the boundedness assumptions on f_0 and f_1 , for any $x \in A_1$ and p even,

$$\mathbb{E} \left[\left| \hat{r}_M(x) - r(x) \right|^p \right] \lesssim \mathbb{E} \left[\left| \hat{r}_M(x) - \mathbb{E}[\hat{r}_M(x) | \mathbf{X}] \right|^p \right] + \mathbb{E} \left[\left(\mathbb{E}[\hat{r}_M(x) | \mathbf{X}] \right)^p \right] + |r(x)|^p \lesssim 1.$$

Then

$$\mathbb{E} \left[\int_{\mathbb{R}^d} \left| \widehat{r}_M(x) - r(x) \right|^p f_0(x) \mathbf{1}(x \in A_1) dx \right] \lesssim f_U \lambda(A_1) \lesssim \epsilon.$$

Since ϵ is arbitrary, we have

$$\lim_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \widehat{r}_M(x) - r(x) \right|^p f_0(x) \mathbf{1}(x \in A_1) dx \right] = 0. \quad (\text{C.3.28})$$

Combining (C.3.26), (C.3.27), and (C.3.28) completes the proof. $\square$

C.3.4 Proof of Corollary 3.8.1

Proof of Corollary 3.8.1. Corollary 3.8.1 can be established following the same way as that of Theorem 3.8.2 but with less effort since we only have to show

$$\lim_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \mathbb{E}[\widehat{r}_M(x) \mid \mathbf{X}] - r(x) \right|^p f_0(x) dx \right] = 0.$$

In detail, denote the Radon-Nikodym derivative of the probability measure of W with respect to ν_0 by r_W . We then have

$$\begin{aligned} & \limsup_{N_0 \rightarrow \infty} \mathbb{E} \left[\left| \frac{N_0}{M} \nu_1(A_M(W)) - r(W) \right|^p \right] \\ &= \limsup_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \frac{N_0}{M} \nu_1(A_M(x)) - r(x) \right|^p r_W(x) f_0(x) dx \right] \\ &\lesssim \limsup_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \frac{N_0}{M} \nu_1(A_M(x)) - r(x) \right|^p f_0(x) dx \right] \\ &= \limsup_{N_0 \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \mathbb{E}[\widehat{r}_M(x) \mid \mathbf{X}] - r(x) \right|^p f_0(x) dx \right] = 0, \end{aligned}$$

where the last line has been established in the proof of Theorem 3.8.2. Noticing that $\mathbb{E}[r(W)]^p$ is bounded under Assumption 3.8.1, the proof is thus complete. $\square$

C.3.5 Proof of Theorem 3.8.3

We only have to prove the first two claims as the rest are trivial.

Proof of Theorem 3.8.3(i). For any $z \in \mathbb{R}^d$ such that $\|z - x\| \leq \delta/2$, since $B_{z, \|z-x\|} \subset B_{x, 2\|z-x\|} \subset B_{x, \delta}$, we have

$$\begin{aligned} \left| \frac{\nu_0(B_{z, \|z-x\|})}{\lambda(B_{z, \|z-x\|})} - f_0(x) \right| &\leq \frac{1}{\lambda(B_{z, \|z-x\|})} \int_{B_{z, \|z-x\|}} |f_0(y) - f_0(x)| dy \leq 2L \|z - x\|, \\ \left| \frac{\nu_1(B_{x, \|z-x\|})}{\lambda(B_{x, \|z-x\|})} - f_1(x) \right| &\leq \frac{1}{\lambda(B_{x, \|z-x\|})} \int_{B_{x, \|z-x\|}} |f_1(y) - f_1(x)| dy \leq L \|z - x\|. \end{aligned}$$

Consider any $\delta_N > 0$ such that $\delta_N \leq \delta/2$. If $\|z - x\| \leq \delta_N$ and $f_0(x) > 2L\delta_N$, then

$$\frac{f_0(x) - 2L\delta_N}{f_1(x) + L\delta_N} \leq \frac{\nu_0(B_{z,\|x-z\|})}{\lambda(B_{z,\|x-z\|})} \frac{\lambda(B_{x,\|x-z\|})}{\nu_1(B_{x,\|x-z\|})}.$$

If further $f_1(x) > L\delta_N$, then

$$\frac{\nu_0(B_{z,\|x-z\|})}{\lambda(B_{z,\|x-z\|})} \frac{\lambda(B_{x,\|x-z\|})}{\nu_1(B_{x,\|x-z\|})} \leq \frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N}.$$

On the other hand, if $\|z - x\| \geq \delta_N$ and $f_0(x) > 2L\delta_N$, $\nu_0(B_{z,\|z-x\|}) \geq (f_0(x) - 2L\delta_N)\lambda(B_{0,\delta_N}) = (f_0(x) - 2L\delta_N)V_d\delta_N^d$, where V_d is the Lebesgue measure of the unit ball on $\mathbb{R}^d$.

Let $\delta_N = (\frac{4}{f_L V_d})^{1/d} (\frac{M}{N_0})^{1/d}$. Since $M/N_0 \rightarrow 0$, we have $\delta_N \rightarrow 0$ as $N_0 \rightarrow \infty$. Taking N_0 large enough so that $\delta_N < f_L/(4L)$ and $\delta_N \leq \delta/2$, then $2LV_d\delta_N^{d+1} = \frac{M}{N_0} \frac{8L}{f_L} \delta_N < 2\frac{M}{N_0}$. Then for any $(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)$,

$$(f_0(x) - 2L\delta_N)V_d\delta_N^d > 4\frac{f_0(x)}{f_L} \frac{M}{N_0} - 2\frac{M}{N_0} \geq 2\frac{M}{N_0}.$$

With a slight abuse of notation, let $W = \nu_0(B_{z,\|x-z\|})$. Then $W \leq 2\frac{M}{N_0}$ implies that $\|Z - x\| \leq \delta_N$.

Depending on the value of $f_1(x)$, the proof is separated into two cases.

Case I. $f_1(x) > L\delta_N$.

Upper bound. Proceeding similar to (C.3.3), we have

$$\begin{aligned} \mathbb{E}[\hat{r}_M(x)] &= \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))] = \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z,\|x_{(M)}(Z)-Z\|})\right) \\ &\leq \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z,\|x_{(M)}(Z)-Z\|}) \leq 2\frac{M}{N_0}\right) + \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \\ &\leq \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z,\|x_{(M)}(Z)-Z\|}), \|Z - x\| \leq \delta_N\right) + \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \\ &\leq \frac{N_0}{M} \mathbb{P}\left(\frac{f_0(x) - 2L\delta_N}{f_1(x) + L\delta_N} \nu_1(B_{x,\|x-Z\|}) \leq \nu_0(B_{Z,\|x_{(M)}(Z)-Z\|}), \|Z - x\| \leq \delta_N\right) \\ &\quad + \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \leq \frac{N_0}{M} \mathbb{P}\left(\frac{f_0(x) - 2L\delta_N}{f_1(x) + L\delta_N} U \leq U_{(M)}\right) + \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right). \end{aligned} \tag{C.3.29}$$

For the second term in (C.3.29), since $M/\log N_0 \rightarrow \infty$, for any $\gamma > 0$,

$$\frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) = \frac{N_0}{M} \mathbb{P}\left(\text{Bin}\left(N_0, 2\frac{M}{N_0}\right) \leq M\right) \leq \frac{N_0}{M} N_0^{-(1-\log 2)M/\log N_0} \prec N_0^{-\gamma}. \tag{C.3.30}$$

For the first term in (C.3.29), proceeding as (C.3.5), we obtain

$$\frac{N_0}{M} \mathbb{P} \left(\frac{f_0(x) - 2L\delta_N}{f_1(x) + L\delta_N} U \leq U_{(M)} \right) \leq \frac{f_1(x) + L\delta_N}{f_0(x) - 2L\delta_N} \frac{N_0}{N_0 + 1}.$$

Then we obtain

$$\mathbb{E}[\widehat{r}_M(x)] \leq \frac{f_1(x) + L\delta_N}{f_0(x) - 2L\delta_N} \frac{N_0}{N_0 + 1} + o(N_0^{-\gamma}). \quad (\text{C.3.31})$$

Lower bound. Proceeding similar to (C.3.8), we have

$$\begin{aligned} \mathbb{E}[\widehat{r}_M(x)] &= \frac{N_0}{M} \mathbb{E}[\nu_1(A_M(x))] = \frac{N_0}{M} \mathbb{P}(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})) \\ &\geq \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq 2\frac{M}{N_0}\right) \\ &= \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq 2\frac{M}{N_0}, \|Z - x\| \leq \delta_N\right) \\ &\geq \frac{N_0}{M} \mathbb{P}\left(\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} \nu_1(B_{x, \|x - Z\|}) \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq 2\frac{M}{N_0}, \|Z - x\| \leq \delta_N\right) \\ &= \frac{N_0}{M} \mathbb{P}\left(\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} \nu_1(B_{x, \|x - Z\|}) \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|}) \leq 2\frac{M}{N_0}\right) \\ &\geq \frac{N_0}{M} \mathbb{P}\left(\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} U \leq U_{(M)}\right) - \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \\ &= \frac{f_1(x) - L\delta_N}{f_0(x) + 2L\delta_N} \int_0^{\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} \frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt - \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right). \end{aligned}$$

Consider the first term. If $\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} \geq 1$, then

$$\frac{f_1(x) - L\delta_N}{f_0(x) + 2L\delta_N} \int_0^{\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} \frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt = \frac{f_1(x) - L\delta_N}{f_0(x) + 2L\delta_N} \frac{N_0}{N_0 + 1}.$$

If $\frac{f_0(x) + 2L\delta_N}{f_1(x) - L\delta_N} < 1$, using the Chernoff bound, for any $\gamma > 0$,

$$\begin{aligned} &\int_{\frac{f_L}{f_U} \frac{N_0}{M}}^{\frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt \leq \int_{\frac{f_L}{f_U} \frac{N_0}{M}}^{\frac{N_0}{M}} \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} t\right) dt \\ &\leq \left[1 - \frac{f_L}{f_U}\right] \frac{N_0}{M} \mathbb{P}\left(U_{(M)} \geq \frac{f_L}{f_U}\right) \\ &\leq \left[1 - \frac{f_L}{f_U}\right] \frac{N_0}{M} \exp\left[M - \frac{f_L}{f_U} N_0 - M \log M + M \log\left(\frac{f_L}{f_U} N_0\right)\right] \prec N_0^{-\gamma}. \end{aligned}$$

The last step is due to $M \log N_0/N_0 \rightarrow 0$. Recalling (C.3.30), we then obtain

$$\mathbb{E}[\widehat{r}_M(x)] \geq \frac{f_1(x) - L\delta_N}{f_0(x) + 2L\delta_N} \frac{N_0}{N_0 + 1} - o(N_0^{-\gamma}). \quad (\text{C.3.32})$$

Combining (C.3.31) and (C.3.32), and taking N_0 large enough so that $L\delta_N \leq f_U \wedge (f_L/4)$, we obtain

$$\begin{aligned} \left| \mathbb{E}[\widehat{r}_M(x)] - r(x) \right| &\leq \left| \frac{f_1(x) + L\delta_N}{f_0(x) - 2L\delta_N} \frac{N_0}{N_0 + 1} - \frac{f_1(x)}{f_0(x)} \right| \vee \left| \frac{f_1(x) - L\delta_N}{f_0(x) + 2L\delta_N} \frac{N_0}{N_0 + 1} - \frac{f_1(x)}{f_0(x)} \right| \\ &+ o(N_0^{-\gamma}) \leq \frac{f_0(x)L\delta_N + 2f_1(x)L\delta_N}{f_0(x)(f_0(x) - 2L\delta_N)} + \frac{1}{N_0 + 1} \frac{f_1(x) + L\delta_N}{f_0(x) - 2L\delta_N} + o(N_0^{-\gamma}) \\ &\leq \left(\frac{2}{f_L} + \frac{4f_U}{f_L^2} \right) L\delta_N + \frac{4f_U}{f_L} \frac{1}{N_0 + 1} + o(N_0^{-\gamma}). \end{aligned}$$

By the selection of δ_N and that the right hand side does not depend on x , we complete the proof for this case.

Case II. $f_1(x) \leq L\delta_N$. The upper bound (C.3.31) in Case I still holds for this case. Accordingly, taking N_0 large enough so that $L\delta_N \leq f_L/4$, we have

$$\begin{aligned} \left| \mathbb{E}[\widehat{r}_M(x)] - r(x) \right| &\leq \mathbb{E}[\widehat{r}_M(x)] + r(x) \\ &\leq \frac{f_1(x) + L\delta_N}{f_0(x) - 2L\delta_N} \frac{N_0}{N_0 + 1} + \frac{f_1(x)}{f_0(x)} + o(N_0^{-\gamma}) \leq \frac{4}{f_L} L\delta_N + \frac{1}{f_L} L\delta_N + o(N_0^{-\gamma}). \end{aligned}$$

We thus complete the whole proof. $\square$

Proof of Theorem 3.8.3(ii). By the law of total variance,

$$\text{Var} [\widehat{r}_M(x)] = \mathbb{E} \left[\text{Var} [\widehat{r}_M(x) \mid \mathbf{X}] \right] + \text{Var} \left[\mathbb{E} [\widehat{r}_M(x) \mid \mathbf{X}] \right]. \quad (\text{C.3.33})$$

For the first term in (C.3.33), let Z be a copy drawn from ν_1 independently of the data. Then, since $[Z_j]_{j=1}^{N_1}$ are i.i.d,

$$\begin{aligned} \mathbb{E} \left[\text{Var} [\widehat{r}_M(x) \mid \mathbf{X}] \right] &= \mathbb{E} \left[\text{Var} \left[\frac{N_0}{N_1 M} K_M(x) \mid \mathbf{X} \right] \right] \\ &= \left(\frac{N_0}{N_1 M} \right)^2 \mathbb{E} \left[\text{Var} \left[\sum_{j=1}^{N_1} \mathbf{1}(Z_j \in A_M(x)) \mid \mathbf{X} \right] \right] = \frac{N_0^2}{N_1 M^2} \mathbb{E} \left[\text{Var} [\mathbf{1}(Z \in A_M(x)) \mid \mathbf{X}] \right] \\ &= \frac{N_0^2}{N_1 M^2} \mathbb{E} \left[\nu_1(A_M(x)) - \nu_1^2(A_M(x)) \right] \leq \frac{N_0^2}{N_1 M^2} \mathbb{E} [\nu_1(A_M(x))] \\ &= \frac{N_0}{N_1 M} \mathbb{E} [\widehat{r}_M(x)] \lesssim C \frac{N_0}{N_1 M}, \end{aligned} \quad (\text{C.3.34})$$

where $C > 0$ is a constant only depending on f_L, f_U . The last step is due to (C.3.31).

For the second term in (C.3.33), notice that

$$\text{Var} \left[\mathbb{E} \left[\hat{r}_M(x) \mid \mathbf{X} \right] \right] = \text{Var} \left[\mathbb{E} \left[\frac{N_0}{N_1 M} K_M(x) \mid \mathbf{X} \right] \right] = \left(\frac{N_0}{M} \right)^2 \text{Var} \left[\nu_1(A_M(x)) \right].$$

Recalling that $W = \nu_0(B_{Z, \|x-Z\|})$, we have the following lemma about the density of W near 0.

Lemma C.3.3. *Denote the density of W by f_W . Then for any $(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)$,*

$$f_W(0) = r(x).$$

Furthermore, for any $\epsilon > 0$ and N_0 sufficiently large, we have for all $0 \leq w \leq 2M/N_0$,

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, \delta, L, d)} f_W(w) \leq (1 + \epsilon) \frac{f_U}{f_L}.$$

Due to Lemma C.3.3, we can take N_0 sufficiently large so that for any $0 \leq w \leq 2M/N_0$,

$$\sup_{(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, \delta, L, d)} f_W(w) \leq 2 \frac{f_U}{f_L}.$$

Let $Z, \tilde{Z}$ be two independent copies from ν_1 that are further independent of the data. Let $W = \nu_0(B_{Z, \|x-Z\|})$ and $\tilde{W} = \nu_0(B_{\tilde{Z}, \|x-\tilde{Z}\|})$. Let $V = \nu_0(B_{Z, \|x_{(M)}(Z)-Z\|})$ and $\tilde{V} = \nu_0(B_{\tilde{Z}, \|x_{(M)}(\tilde{Z})-\tilde{Z}\|})$. We then have

$$\begin{aligned} \text{Var} \left[\nu_1(A_M(x)) \right] &= \mathbb{E} \left[\nu_1^2(A_M(x)) \right] - \left(\mathbb{E} \left[\nu_1(A_M(x)) \right] \right)^2 \\ &= \mathbb{P} \left(Z \in A_M(x), \tilde{Z} \in A_M(x) \right) - \mathbb{P} \left(Z \in A_M(x) \right) \mathbb{P} \left(\tilde{Z} \in A_M(x) \right) \\ &= \mathbb{P} \left(W \leq V, \tilde{W} \leq \tilde{V} \right) - \mathbb{P} \left(W \leq V \right) \mathbb{P} \left(\tilde{W} \leq \tilde{V} \right). \end{aligned}$$

Due to the independence between Z and $\tilde{Z}$, W and $\tilde{W}$ are independent. Notice that $V \mid Z$ have the same distribution as $U_{(M)}$ for any $Z \in \mathbb{R}^d$, then V and Z are independent, so are $\tilde{V}$ and $\tilde{Z}$.

Let's expand the variance further as

$$\begin{aligned} &\text{Var} \left[\nu_1(A_M(x)) \right] \\ &= \left[\mathbb{P} \left(W \leq V, \tilde{W} \leq \tilde{V}, W \leq 2 \frac{M}{N_0}, \tilde{W} \leq 2 \frac{M}{N_0} \right) \right. \\ &\quad \left. - \mathbb{P} \left(W \leq V, W \leq 2 \frac{M}{N_0} \right) \mathbb{P} \left(\tilde{W} \leq \tilde{V}, \tilde{W} \leq 2 \frac{M}{N_0} \right) \right] \\ &\quad + \left[\mathbb{P} \left(W \leq V, \tilde{W} \leq \tilde{V} \right) - \mathbb{P} \left(W \leq V, \tilde{W} \leq \tilde{V}, W \leq 2 \frac{M}{N_0}, \tilde{W} \leq 2 \frac{M}{N_0} \right) \right] \end{aligned}$$

$$- \left[\mathbb{P}\left(W \leq V\right) \mathbb{P}\left(\widetilde{W} \leq \widetilde{V}\right) - \mathbb{P}\left(W \leq V, W \leq 2\frac{M}{N_0}\right) \mathbb{P}\left(\widetilde{W} \leq \widetilde{V}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right]. \quad (\text{C.3.35})$$

For the first term in (C.3.35), we have the following lemma.

Lemma C.3.4. *We have*

$$\begin{aligned} & \left(\frac{N_0}{M}\right)^2 \left[\mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}, W \leq 2\frac{M}{N_0}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right. \\ & \quad \left. - \mathbb{P}\left(W \leq V, W \leq 2\frac{M}{N_0}\right) \mathbb{P}\left(\widetilde{W} \leq \widetilde{V}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right] \leq C \frac{1}{M}, \end{aligned}$$

where $C > 0$ is a constant only depending on f_L, f_U .

For the second term in (C.3.35),

$$\begin{aligned} & \mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}\right) - \mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}, W \leq 2\frac{M}{N_0}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \\ & \leq \mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}, W > 2\frac{M}{N_0}\right) + \mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}, \widetilde{W} > 2\frac{M}{N_0}\right) \\ & \leq \mathbb{P}\left(V > 2\frac{M}{N_0}\right) + \mathbb{P}\left(\widetilde{V} > 2\frac{M}{N_0}\right) = 2\mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right). \end{aligned}$$

Using the Chernoff bound and $M/\log N_0 \rightarrow \infty$, for any $\gamma > 0$,

$$\left(\frac{N_0}{M}\right)^2 \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \leq \left(\frac{N_0}{M}\right)^2 \exp\left[-(1 - \log 2)M\right] \prec N_0^{-\gamma},$$

We then have

$$\begin{aligned} & \left(\frac{N_0}{M}\right)^2 \left[\mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}\right) - \mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}, W \leq 2\frac{M}{N_0}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right] \\ & \leq 2\left(\frac{N_0}{M}\right)^2 \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \prec N_0^{-\gamma}. \end{aligned} \quad (\text{C.3.36})$$

For the third term in (C.3.35), we can check

$$\left[\mathbb{P}\left(W \leq V\right) \mathbb{P}\left(\widetilde{W} \leq \widetilde{V}\right) - \mathbb{P}\left(W \leq V, W \leq 2\frac{M}{N_0}\right) \mathbb{P}\left(\widetilde{W} \leq \widetilde{V}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right] \geq 0.$$

Plugging Lemma C.3.4 and (C.3.36) into (C.3.35) by taking $\gamma > 1$, we obtain

$$\left(\frac{N_0}{M}\right)^2 \text{Var}\left[\nu_1(A_M(x))\right] \lesssim C \frac{1}{M}, \quad (\text{C.3.37})$$

where $C > 0$ is a constant only depending on f_L, f_U .

Plugging (C.3.34) and (C.3.37) into (C.3.33) completes the proof. $\square$

C.3.6 Proof of Proposition 3.8.1

Proof of Proposition 3.8.1. We take ν_0 and ν_1 to share the same support, and assume x to be the origin of $\mathbb{R}^d$ without loss of generality.

When $N_1 \lesssim N_0$, we take ν_0 to be the uniform distribution with density f_L on $[-f_L^{-1/d}/2, f_L^{-1/d}/2]^d$. Then the MSE is lower bounded by the density estimation over Lipchitz class with N_1 samples.

When $N_0 \lesssim N_1$, we take ν_1 to be the uniform distribution with density f_U on $[-f_U^{-1/d}/2, f_U^{-1/d}/2]^d$. Notice that $1/f_0$ is also local Lipchitz from the lower boundness condition and local Lipchitz condition on f_0 . Then the MSE is lower bounded by the density estimation over Lipchitz class with N_0 samples.

We then complete the proof by combining the above two lower bounds and then using the famous minimax lower bound in Lipschitz density estimation (Tsybakov, 2009, Exercise 2.8), $\square$

C.3.7 Proof of Theorem 3.8.4

Proof of Theorem 3.8.4. We only have to prove the first claim as the second is trivial.

Take $\delta_N = (\frac{4}{f_L V_d})^{1/d} (\frac{M}{N_0})^{1/d}$ as in the proof of Theorem 3.8.3(i) and $\delta'_N = (\frac{2}{af_L V_d})^{1/d} (\frac{M}{N_0})^{1/d}$. For any $x \in \mathbb{R}^d$, denote the distance of x to the boundary of S_1 by $\Delta(x)$, i.e., $\Delta(x) = \inf_{z \in \partial S_1} \|z - x\|$.

Depending on the position of x and the value of $\Delta(x)$, we separate the proof into three cases.

Case I. $x \in S_1$ and $\Delta(x) > 2\delta_N$. In this case, since $\Delta(x) > 2\delta_N$, for any $\|z - x\| \leq \delta_N$, we have $B_{z, \|z-x\|} \subset S_1$. From the smoothness conditions on f_0 and f_1 , similar to the proof of Theorem 3.8.3, we have

$$\begin{aligned} & \mathbb{E} \left[\int_{\mathbb{R}^d} \left| \hat{r}_M(x) - r(x) \right| f_0(x) \mathbf{1}(x \in S_1, \Delta(x) > 2\delta_N) dx \right] \\ & \leq \int_{\mathbb{R}^d} \left(\mathbb{E} \left[\hat{r}_M(x) - r(x) \right]^2 \right)^{1/2} f_0(x) \mathbf{1}(x \in S_1, \Delta(x) > 2\delta_N) dx \\ & \leq C \left[\left(\frac{M}{N_0} \right)^{1/d} + \left(\frac{1}{M} \right)^{1/2} + \left(\frac{N_0}{MN_1} \right)^{1/2} \right] \int_{\mathbb{R}^d} f_0(x) \mathbf{1}(x \in S_1, \Delta(x) > 2\delta_N) dx \\ & \leq C \left[\left(\frac{M}{N_0} \right)^{1/d} + \left(\frac{1}{M} \right)^{1/2} + \left(\frac{N_0}{MN_1} \right)^{1/2} \right], \end{aligned} \tag{C.3.38}$$

where the constant $C > 0$ only depends on f_L, f_U, L, d .

Case II. $x \in S_0 \setminus S_1$ and $\Delta(x) > \delta'_N$. In this case, $r(x) = 0$ and for any $z \in S_1$,

$$\nu_0(B_{z, \|z-x\|}) \geq f_L \lambda(B_{z, \|z-x\|} \cap S_0) \geq af_L \lambda(B_{z, \|z-x\|}) > af_L V_d \delta_N^{d'} \geq 2 \frac{M}{N_0}.$$

Then for any $\gamma > 0$,

$$\mathbb{E} \left[\left| \hat{r}_M(x) - r(x) \right| \right] = \mathbb{E} \left[\hat{r}_M(x) \right] = \frac{N_0}{M} \mathbb{E} [\nu_1(A_M(x))]$$

$$= \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})\right) \leq \frac{N_0}{M} \mathbb{P}\left(U_{(M)} > 2\frac{M}{N_0}\right) \prec N_0^{-\gamma}.$$

We then obtain

$$\begin{aligned} & \mathbb{E}\left[\int_{\mathbb{R}^d} \left|\hat{r}_M(x) - r(x)\right| f_0(x) \mathbb{1}(x \notin S_1, \Delta(x) > \delta'_N) dx\right] \\ & \prec N_0^{-\gamma} \int_{\mathbb{R}^d} f_0(x) \mathbb{1}(x \in S_0 \setminus S_1, \Delta(x) > \delta'_N) dx \leq N_0^{-\gamma}. \end{aligned} \quad (\text{C.3.39})$$

Case III. $x \in S_0$ and $\Delta(x) \leq (2\delta_N) \vee \delta'_N$. In this case, for any $z \in S_1$,

$$\nu_0(B_{z, \|z-x\|}) \geq f_L \lambda(B_{z, \|z-x\|} \cap S_0) \geq a f_L \lambda(B_{z, \|z-x\|}) \geq \frac{a f_L}{f_U} \nu_1(B_{x, \|z-x\|}).$$

Accordingly,

$$\begin{aligned} & \mathbb{E}\left[\left|\hat{r}_M(x) - r(x)\right|\right] \leq \mathbb{E}\left[\hat{r}_M(x)\right] + r(x) = \frac{N_0}{M} \mathbb{P}\left(W \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})\right) + r(x) \\ & \leq \frac{N_0}{M} \mathbb{P}\left(\frac{a f_L}{f_U} \nu_1(B_{x, \|x-Z\|}) \leq \nu_0(B_{Z, \|\mathcal{X}_{(M)}(Z) - Z\|})\right) + r(x) \\ & \leq \frac{N_0}{M} \mathbb{P}\left(\frac{a f_L}{f_U} U \leq U_{(M)}\right) + r(x) = \frac{f_U}{a f_L} (1 + o(1)) + \frac{f_U}{f_L}. \end{aligned}$$

From the definition of δ_N, δ'_N and $M/N_0 \rightarrow 0$, we have $\delta_N, \delta'_N \rightarrow 0$ as $N_0 \rightarrow \infty$. Since the surface area of S_1 is bounded by H , we have $\lambda(\{x : \Delta(x) \leq (2\delta_N) \vee \delta'_N\}) \lesssim H\{(2\delta_N) \vee \delta'_N\}$. Then we obtain

$$\begin{aligned} & \mathbb{E}\left[\int_{\mathbb{R}^d} \left|\hat{r}_M(x) - r(x)\right| f_0(x) \mathbb{1}(\Delta(x) \leq (2\delta_N) \vee \delta'_N) dx\right] \\ & \leq \left(\frac{f_U}{a f_L} (1 + o(1)) + \frac{f_U}{f_L}\right) \int_{\mathbb{R}^d} f_0(x) \mathbb{1}(\Delta(x) \leq (2\delta_N) \vee \delta'_N) dx \\ & \leq \left(\frac{f_U}{a f_L} (1 + o(1)) + \frac{f_U}{f_L}\right) f_U \lambda(\{x : \Delta(x) \leq (2\delta_N) \vee \delta'_N\}) \\ & \lesssim \left(\frac{f_U}{a f_L} + \frac{f_U}{f_L}\right) f_U H (\delta_N + \delta'_N) \leq C \left(\frac{M}{N_0}\right)^{1/d}, \end{aligned} \quad (\text{C.3.40})$$

where the constant $C > 0$ only depends on f_L, f_U, a, H, d .

Combining (C.3.38), (C.3.39), (C.3.40) completes the proof. $\square$

C.3.8 Proof of Proposition 3.8.2

Proof of Proposition 3.8.2. We take ν_0 and ν_1 to be of the same support.

When $N_1 \lesssim N_0$, we take ν_0 to be the uniform distribution with density f_L on $[-f_L^{-1/d}/2, f_L^{-1/d}/2]^d$. Then the L_1 risk is lower bounded by the L_1 risk over support of density estimation over Lipchitz class with N_1 samples.

When $N_0 \lesssim N_1$, we take ν_1 to be the uniform distribution with density f_U on $[-f_U^{-1/d}/2, f_U^{-1/d}/2]^d$. Notice $1/f_0$ is also Lipchitz from the lower boundness condition and Lipchitz condition on f_0 . From the lower boundness condition on f_0 , the L_1 risk is lower bounded by the L_1 risk over support of density estimation over Lipchitz class with N_0 samples.

We then complete the proof by combining the above two lower bounds and then using then the minimax lower bound of L_1 risk for density estimation over Lipchitz class (Zhao and Lai, 2022, Theorem 1). $\square$

C.4 Proofs of the results in Section 3.9

C.4.1 Proof of Lemma 3.9.1

Proof of Lemma 3.9.1. For any $x \in \mathbb{X}$, define $\sigma_\omega^2(x) = \mathbb{E}[U_\omega^2 | X = x] = \mathbb{E}[(Y(\omega) - \mu_\omega(X))^2 | X = x]$ for $\omega \in \{0, 1\}$. Let

$$V^\tau = \mathbb{E}[\mu_1(X) - \mu_0(X) - \tau]^2 \quad \text{and} \quad V^E = \frac{1}{n} \sum_{i=1}^n \left(1 + \frac{K_M(i)}{M}\right)^2 \sigma_{D_i}^2(X_i).$$

From the standard central limit theorem (Billingsley, 2008, Theorem 27.1), we have

$$\sqrt{n}(\bar{\tau}(\mathbf{X}) - \tau) \xrightarrow{d} N(0, V^\tau). \quad (\text{C.4.1})$$

Let $E_{M,i} = (2D_i - 1)(1 + K_M(i)/M)\epsilon_i$ for any $i \in \llbracket n \rrbracket$. Conditional on $\mathbf{X}, \mathbf{D}$, $[E_{M,i}]_{i=1}^n$ are independent. Notice that $\mathbb{E}[E_{M,i} | \mathbf{X}, \mathbf{D}] = 0$ and $\sum_{i=1}^n \text{Var}[E_{M,i} | \mathbf{X}, \mathbf{D}] = nV^E$. To apply the Lindeberg-Feller central limit theorem (Billingsley, 2008, Theorem 27.2), it suffices to verify that: for a given $(\mathbf{X}, \mathbf{D})$,

$$\frac{1}{nV^E} \sum_{i=1}^n \mathbb{E}\left[\left(E_{M,i}\right)^2 \mathbf{1}\left(|E_{M,i}| > \delta\sqrt{nV^E}\right) \mid \mathbf{X}, \mathbf{D}\right] \rightarrow 0,$$

for all $\delta > 0$.

Let $C_\sigma = \sup_{x \in \mathbb{X}, \omega \in \{0,1\}} \{\mathbb{E}[|U_\omega|^{2+\kappa} | X = x] \vee \mathbb{E}[U_\omega^2 | X = x]\} < \infty$. Let $p_1 = 1 + \kappa/2$ and p_2 be the constant such that $p_1^{-1} + p_2^{-1} = 1$. By Hölder's inequality and Markov's inequality,

$$\begin{aligned} & \frac{1}{nV^E} \sum_{i=1}^n \mathbb{E}\left[\left(E_{M,i}\right)^2 \mathbf{1}\left(|E_{M,i}| > \delta\sqrt{nV^E}\right) \mid \mathbf{X}, \mathbf{D}\right] \\ & \leq \frac{1}{nV^E} \sum_{i=1}^n \left(\mathbb{E}\left[|E_{M,i}|^{2+\kappa} \mid \mathbf{X}, \mathbf{D}\right]\right)^{1/p_1} \left(\mathbb{P}\left(|E_{M,i}| > \delta\sqrt{nV^E} \mid \mathbf{X}, \mathbf{D}\right)\right)^{1/p_2} \\ & \leq \frac{1}{nV^E} \sum_{i=1}^n \left(\mathbb{E}\left[|E_{M,i}|^{2+\kappa} \mid \mathbf{X}, \mathbf{D}\right]\right)^{1/p_1} \left(\frac{1}{\delta^2 nV^E} \mathbb{E}\left[\left(E_{M,i}\right)^2 \mid \mathbf{X}, \mathbf{D}\right]\right)^{1/p_2} \end{aligned}$$

$$\leq \frac{C_\sigma}{nV^E} \left(\frac{1}{\delta^2 nV^E} \right)^{1/p_2} \sum_{i=1}^n \left(1 + \frac{K_M(i)}{M} \right)^{2(1+1/p_2)}.$$

Notice that $E[1 + K_M(i)/M]^{2(1+1/p_2)} < \infty$ from Theorem 3.8.2. Let $c_\sigma = \inf_{x \in \mathbb{X}, \omega \in \{0,1\}} E[U_\omega^2 | X = x] > 0$. From the definition of V^E , we have $V^E \geq c_\sigma$ for almost all $\mathbf{X}, \mathbf{D}$. Then

$$E \left[\frac{1}{nV^E} \sum_{i=1}^n E \left[\left(E_{M,i} \right)^2 \mathbf{1} \left(|E_{M,i}| > \delta \sqrt{nV^E} \right) \mid \mathbf{X}, \mathbf{D} \right] \right] = O(n^{-1/p_2}) = o(1).$$

We thus obtain

$$\frac{1}{nV^E} \sum_{i=1}^n E \left[\left(E_{M,i} \right)^2 \mathbf{1} \left(|E_{M,i}| > \delta \sqrt{nV^E} \right) \mid \mathbf{X}, \mathbf{D} \right] = o_P(1).$$

Applying the Lindeberg-Feller central limit theorem then yields

$$\sqrt{n}(V^E)^{-1/2} E_M = (nV^E)^{-1/2} \sum_{i=1}^n E_{M,i} \xrightarrow{d} N(0, 1). \quad (\text{C.4.2})$$

Noticing that $\sqrt{n}(\bar{\tau}(\mathbf{X}) - \tau)$ and $\sqrt{n}(V^E)^{-1/2} E_M$ are asymptotically independent, leveraging the same argument as made in [Abadie and Imbens \(2006, Proof of Theorem 4, Page 267\)](#) and then combining (C.4.1) and (C.4.2) reaches

$$\sqrt{n}(V^\tau + V^E)^{-1/2} (\bar{\tau}(\mathbf{X}) + E_M - \tau) \xrightarrow{d} N(0, 1). \quad (\text{C.4.3})$$

We decompose V^E as:

$$\begin{aligned} V^E &= \frac{1}{n} \sum_{i=1, D_i=1}^n \left(1 + \frac{K_M(i)}{M} \right)^2 \sigma_1^2(X_i) + \frac{1}{n} \sum_{i=1, D_i=0}^n \left(1 + \frac{K_M(i)}{M} \right)^2 \sigma_0^2(X_i) \\ &= \left[\frac{1}{n} \sum_{i=1, D_i=1}^n \left(\frac{1}{e(X_i)} \right)^2 \sigma_1^2(X_i) + \frac{1}{n} \sum_{i=1, D_i=0}^n \left(\frac{1}{1 - e(X_i)} \right)^2 \sigma_0^2(X_i) \right] \\ &\quad + \frac{1}{n} \sum_{i=1, D_i=1}^n \left[\left(1 + \frac{K_M(i)}{M} \right)^2 - \left(\frac{1}{e(X_i)} \right)^2 \right] \sigma_1^2(X_i) \\ &\quad + \frac{1}{n} \sum_{i=1, D_i=0}^n \left[\left(\frac{1}{1 - e(X_i)} \right)^2 - \left(1 + \frac{K_M(i)}{M} \right)^2 \right] \sigma_0^2(X_i). \end{aligned} \quad (\text{C.4.4})$$

For the first term in (C.4.4), notice that $[(X_i, D_i, Y_i)]_{i=1}^n$ are i.i.d. and $E[D_i(e(X_i))^{-2} \sigma_1^2(X_i)], E[(1 - D_i)(1 - e(X_i))^{-2} \sigma_0^2(X_i)] < \infty$. Using the weak law of large numbers, we have

$$\frac{1}{n} \sum_{i=1, D_i=1}^n \left(\frac{1}{e(X_i)} \right)^2 \sigma_1^2(X_i) + \frac{1}{n} \sum_{i=1, D_i=0}^n \left(\frac{1}{1 - e(X_i)} \right)^2 \sigma_0^2(X_i) \xrightarrow{P} E \left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right].$$

For the second term in (C.4.4), using the Cauchy–Schwarz inequality,

$$\begin{aligned} & \mathbb{E} \left| \frac{1}{n} \sum_{i=1, D_i=1}^n \left[\left(1 + \frac{K_M(i)}{M} \right)^2 - \left(\frac{1}{e(X_i)} \right)^2 \right] \sigma_1^2(X_i) \right| \\ & \leq C_\sigma \mathbb{E} \left[D_i \left| \left(1 + \frac{K_M(i)}{M} \right)^2 - \left(\frac{1}{e(X_i)} \right)^2 \right| \right] = C_\sigma \mathbb{E} \left[D_i \mathbb{E} \left[\left| \left(1 + \frac{K_M(i)}{M} \right)^2 - \left(\frac{1}{e(X_i)} \right)^2 \right| \mid \mathbf{D} \right] \right] \\ & \leq C_\sigma \mathbb{E} \left[D_i \left(\mathbb{E} \left[\left(\frac{K_M(i)}{M} - \frac{1 - e(X_i)}{e(X_i)} \right)^2 \mid \mathbf{D} \right] \mathbb{E} \left[\left(2 + \frac{K_M(i)}{M} + \frac{1 - e(X_i)}{e(X_i)} \right)^2 \mid \mathbf{D} \right] \right)^{1/2} \right] = o(1), \end{aligned}$$

where the last step is due to Theorem 3.8.2. Then we obtain

$$\frac{1}{n} \sum_{i=1, D_i=1}^n \left[\left(1 + \frac{K_M(i)}{M} \right)^2 - \left(\frac{1}{e(X_i)} \right)^2 \right] \sigma_1^2(X_i) \xrightarrow{\mathbf{p}} 0.$$

For the third term in (C.4.4), we can establish in the same way that

$$\frac{1}{n} \sum_{i=1, D_i=0}^n \left[\left(\frac{1}{1 - e(X_i)} \right)^2 - \left(1 + \frac{K_M(i)}{M} \right)^2 \right] \sigma_0^2(X_i) \xrightarrow{\mathbf{p}} 0.$$

Then from (C.4.4),

$$V^E \xrightarrow{\mathbf{p}} \mathbb{E} \left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right].$$

By (C.4.3), Slutsky's lemma (van der Vaart, 1998, Theorem 2.8), and the definition of σ^2 , we complete the proof. $\square$

C.4.2 Proof of Lemma 3.9.2

Proof of Lemma 3.9.2. From Assumption 3.8.1 and Assumption 3.4.1, let $R = \text{diam}(\mathbb{X}) < \infty$ and $f_L = \inf_{x \in \mathbb{X}, \omega \in \{0,1\}} f_\omega(x) > 0$. For any $x \in \mathbb{X}$, $\omega \in \{0,1\}$ and $u \leq R$, from Assumption 3.8.1, $\nu_\omega(B_{x,u} \cap \mathbb{X}) \geq f_L \lambda(B_{x,u} \cap \mathbb{X}) \geq f_L a \lambda(B_{x,u}) = f_L a V_d u^d$, where V_d is the Lebesgue measure of the unit ball on $\mathbb{R}^d$.

Let $c_0 = f_L a V_d$. For any $i \in \llbracket n \rrbracket$, $x \in \mathbb{X}$, $M \leq n_{1-D_i}$, if $0 \leq u \leq R n_{1-D_i}^{1/d}$, we have

$$\begin{aligned} & \mathbb{P} \left(\left\| X_j - X_i \right\| \geq u n_{1-D_i}^{-1/d} \mid \mathbf{D}, X_i = x, j = j_M(i) \right) \\ & \leq \mathbb{P} \left(\text{Bin} \left(n_{1-D_i}, \nu_{1-D_i} \left(B_{x, u n_{1-D_i}^{-1/d}} \cap \mathbb{X} \right) \right) \leq M \mid \mathbf{D} \right) \\ & \leq \mathbb{P} \left(\text{Bin} \left(n_{1-D_i}, c_0 u^d n_{1-D_i}^{-1} \right) \leq M \mid \mathbf{D} \right). \end{aligned}$$

Using the Chernoff bound, if $M < c_0 u^d$, then

$$\mathbb{P} \left(\text{Bin} \left(n_{1-D_i}, c_0 u^d n_{1-D_i}^{-1} \right) \leq M \mid \mathbf{D} \right) \leq \exp \left(M - c_0 u^d + M \log \left(\frac{c_0 u^d}{M} \right) \right).$$

Notice that the above upper bound does not depend on x . We then obtain

$$\begin{aligned} & \mathbb{P}\left(\left\|X_j - X_i\right\| \geq un_{1-D_i}^{-1/d} \mid \mathbf{D}, j = j_M(i)\right) \\ & \leq \mathbb{1}\left(M < c_0 u^d\right) \exp\left(M - c_0 u^d + M \log\left(\frac{c_0 u^d}{M}\right)\right) + \mathbb{1}\left(M \geq c_0 u^d\right). \end{aligned}$$

On the other hand, if $u > Rn_{1-D_i}^{1/d}$, then the probability is zero from the definition of R . Accordingly, the above bound holds for any $u \geq 0$.

For any $i \in \llbracket n \rrbracket$, we thus have

$$\begin{aligned} n_{1-D_i}^{p/d} \mathbb{E}\left[\left\|U_{M,i}\right\|^p \mid \mathbf{D}\right] &= p \int_0^\infty \mathbb{P}\left(\left\|X_j - X_i\right\| \geq un_{1-D_i}^{-1/d} \mid \mathbf{D}, j = j_M(i)\right) u^{p-1} du \\ &\leq p \int_0^\infty \left[\mathbb{1}\left(M < c_0 u^d\right) \exp\left(M - c_0 u^d + M \log\left(\frac{c_0 u^d}{M}\right)\right) + \mathbb{1}\left(M \geq c_0 u^d\right)\right] u^{p-1} du \\ &= pc_0^{-p/d} d^{-1} \left[\int_M^\infty \left(\frac{e}{M}\right)^M t^{M+\frac{p}{d}-1} e^{-t} dt + \int_0^M t^{\frac{p}{d}-1} dt \right], \end{aligned} \quad (\text{C.4.5})$$

where the last step is through taking $t = c_0 u^d$.

For the first term in (C.4.5), from Stirling's formula and $M \rightarrow \infty$,

$$\int_M^\infty \left(\frac{e}{M}\right)^M t^{M+\frac{p}{d}-1} e^{-t} dt \leq \int_0^\infty \left(\frac{e}{M}\right)^M t^{M+\frac{p}{d}-1} e^{-t} dt \sim \sqrt{2\pi} M^{\frac{p}{d}-\frac{1}{2}},$$

where $\sim$ means asymptotic convergence.

For the second term in (C.4.5), $\int_0^M t^{\frac{p}{d}-1} dt = \frac{d}{p} M^{\frac{p}{d}}$. Combining the above two terms then complete the proof. $\square$

C.4.3 Proof of Lemma 3.9.3

Proof of Lemma 3.9.3. We decompose $B_M - \widehat{B}_M$ as

$$\begin{aligned} & |B_M - \widehat{B}_M| \\ &= \left| \frac{1}{n} \sum_{i=1}^n (2D_i - 1) \left[\frac{1}{M} \sum_{m=1}^M \left(\mu_{1-D_i}(X_i) - \mu_{1-D_i}(X_{j_m(i)}) - \widehat{\mu}_{1-D_i}(X_i) + \widehat{\mu}_{1-D_i}(X_{j_m(i)}) \right) \right] \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \max_{m \in \llbracket M \rrbracket} \left| \mu_{1-D_i}(X_i) - \mu_{1-D_i}(X_{j_m(i)}) - \widehat{\mu}_{1-D_i}(X_i) + \widehat{\mu}_{1-D_i}(X_{j_m(i)}) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \max_{m \in \llbracket M \rrbracket, \omega \in \{0,1\}} \left| \mu_\omega(X_i) - \mu_\omega(X_{j_m(i)}) - \widehat{\mu}_\omega(X_i) + \widehat{\mu}_\omega(X_{j_m(i)}) \right|. \end{aligned} \quad (\text{C.4.6})$$

Let $k = \lfloor d/2 \rfloor + 1$. For any $\omega \in \{0, 1\}$, by Taylor expansion to k -th order,

$$\left| \mu_\omega(X_{j_m(i)}) - \mu_\omega(X_i) - \sum_{\ell=1}^{k-1} \frac{1}{\ell!} \sum_{t \in \Lambda_\ell} \partial^t \mu_\omega(X_i) U_{m,i}^t \right| \leq \max_{t \in \Lambda_k} \|\partial^t \mu_\omega\|_\infty \frac{1}{k!} \sum_{t \in \Lambda_k} \|U_{m,i}\|^k. \quad (\text{C.4.7})$$

In the same way,

$$\left| \widehat{\mu}_\omega(X_{j_m(i)}) - \widehat{\mu}_\omega(X_i) - \sum_{\ell=1}^{k-1} \frac{1}{\ell!} \sum_{t \in \Lambda_\ell} \partial^t \widehat{\mu}_\omega(X_i) U_{m,i}^t \right| \leq \max_{t \in \Lambda_k} \|\partial^t \widehat{\mu}_\omega\|_\infty \frac{1}{k!} \sum_{t \in \Lambda_k} \|U_{m,i}\|^k. \quad (\text{C.4.8})$$

We also have

$$\left| \sum_{\ell=1}^{k-1} \frac{1}{\ell!} \sum_{t \in \Lambda_\ell} (\partial^t \widehat{\mu}_\omega(X_i) - \partial^t \mu_\omega(X_i)) U_{m,i}^t \right| \leq \sum_{\ell=1}^{k-1} \max_{t \in \Lambda_\ell} \|\partial^t \widehat{\mu}_\omega - \partial^t \mu_\omega\|_\infty \frac{1}{\ell!} \sum_{t \in \Lambda_\ell} \|U_{m,i}\|^\ell. \quad (\text{C.4.9})$$

Notice that $\|U_{M,i}\| = \max_{m \in \llbracket M \rrbracket} \|U_{m,i}\|$ for any $i \in \llbracket n \rrbracket, \omega \in \{0, 1\}$. Then for any $\omega \in \{0, 1\}$, plugging (C.4.7), (C.4.8), (C.4.9) into (C.4.6), we obtain

$$\begin{aligned} |B_M - \widehat{B}_M| &\lesssim \left(\max_{\omega \in \{0,1\}} \max_{t \in \Lambda_k} \|\partial^t \mu_\omega\|_\infty + \max_{\omega \in \{0,1\}} \max_{t \in \Lambda_k} \|\partial^t \widehat{\mu}_\omega\|_\infty \right) \left(\frac{1}{n} \sum_{i=1}^n \|U_{M,i}\|^k \right) \\ &\quad + \sum_{\ell=1}^{k-1} \left(\max_{\omega \in \{0,1\}} \max_{t \in \Lambda_\ell} \|\partial^t \widehat{\mu}_\omega - \partial^t \mu_\omega\|_\infty \right) \left(\frac{1}{n} \sum_{i=1}^n \|U_{M,i}\|^\ell \right). \end{aligned}$$

From Lemma 3.9.2, all moments of $(n/M)^{p/d} \|U_{M,i}\|^p$ are bounded. Then for any positive integer p , using Markov's inequality, we have

$$\frac{1}{n} \sum_{i=1}^n \|U_{M,i}\|^p = O_P\left(\left(\frac{M}{n}\right)^{p/d}\right).$$

By Assumption 3.4.4 and Assumption 3.4.5, we then obtain

$$\begin{aligned} B_M - \widehat{B}_M &= O_P(1) O_P\left(\left(\frac{M}{n}\right)^{k/d}\right) + \max_{\ell \in \llbracket k-1 \rrbracket} O_P(n^{-\gamma_\ell}) O_P\left(\left(\frac{M}{n}\right)^{\ell/d}\right) \\ &= O_P\left(\left(\frac{M}{n}\right)^{k/d}\right) + \max_{\ell \in \llbracket k-1 \rrbracket} O_P\left(n^{-\gamma_\ell} \left(\frac{M}{n}\right)^{\ell/d}\right). \end{aligned}$$

The proof is thus complete by noticing the definition of γ and $M \prec n^\gamma$. □

C.5 Proofs of the results in Appendix

C.5.1 Proof of Lemma C.3.1

Proof of Lemma C.3.1. The first inequality is directly from the definition of Lebesgue points. The second inequality follows by

$$\left| \frac{\nu(B_{z, \|z-x\|})}{\lambda(B_{z, \|z-x\|})} - f(x) \right| \leq \frac{1}{\lambda(B_{z, \|z-x\|})} \int_{B_{z, \|z-x\|}} |f(y) - f(x)| dy$$

$$\begin{aligned}
&\leq \frac{1}{\lambda(B_{z,\|z-x\|})} \int_{B_{x,2\|z-x\|}} |f(y) - f(x)| dy = \frac{\lambda(B_{x,2\|z-x\|})}{\lambda(B_{z,\|z-x\|})} \frac{1}{\lambda(B_{x,2\|z-x\|})} \int_{B_{x,2\|z-x\|}} |f(y) - f(x)| dy \\
&= 2^d \frac{1}{\lambda(B_{x,2\|z-x\|})} \int_{B_{x,2\|z-x\|}} |f(y) - f(x)| dy,
\end{aligned}$$

and then the definition of Lebesgue points. $\square$

C.5.2 Proof of Lemma C.3.3

Proof of Lemma C.3.3. Fix any $(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)$.

We first prove the first claim. First consider $f_1(x) > 0$. For any $\epsilon > 0$, there exists $\delta' > 0$ such that for any $z \in \mathbb{R}^d$ satisfying $\|z - x\| \leq 2\delta'$, we have $|f_0(z) - f_0(x)| \leq \epsilon f_0(x)$ and $|f_1(z) - f_1(x)| \leq \epsilon f_1(x)$ from the local Lipschitz assumption. We take $w > 0$ sufficiently small such that $w < (1 - \epsilon)f_0(x)\lambda(B_{0,\delta'})$. Then $W \leq w$ implies $\|x - Z\| \leq \delta'$. Then for $w > 0$ sufficiently small,

$$P(W \leq w) = P(W \leq w, \|x - Z\| \leq \delta') \leq P\left(\frac{1 - \epsilon f_0(x)}{1 + \epsilon f_1(x)} \nu_1(B_{x,\|x-Z\|}) \leq w\right) = \frac{1 + \epsilon f_1(x)}{1 - \epsilon f_0(x)} w,$$

and

$$\begin{aligned}
P(W \leq w) &= P(W \leq w, \|x - Z\| \leq \delta') \geq P\left(\frac{1 + \epsilon f_0(x)}{1 - \epsilon f_1(x)} \nu_1(B_{x,\|x-Z\|}) \leq w, \|x - Z\| \leq \delta'\right) \\
&= P\left(\frac{1 + \epsilon f_0(x)}{1 - \epsilon f_1(x)} \nu_1(B_{x,\|x-Z\|}) \leq w\right) = \frac{1 - \epsilon f_1(x)}{1 + \epsilon f_0(x)} w.
\end{aligned}$$

Then we have

$$\frac{1 - \epsilon f_1(x)}{1 + \epsilon f_0(x)} \leq \liminf_{w \rightarrow 0} w^{-1} P(W \leq w) \leq \limsup_{w \rightarrow 0} w^{-1} P(W \leq w) \leq \frac{1 + \epsilon f_1(x)}{1 - \epsilon f_0(x)}.$$

Since ϵ is arbitrary, we obtain

$$f_W(0) = \lim_{w \rightarrow 0} w^{-1} P(W \leq w) = \frac{f_1(x)}{f_0(x)} = r(x).$$

The case for $f_1(x) = 0$ can be established in the same way. This completes the proof of the first claim.

For the second claim, for any $0 < \epsilon < f_L$, there exists $\delta' > 0$ such that for any $z \in \mathbb{R}^d$ satisfying $\|z - x\| \leq 2\delta'$, we have $|f_0(z) - f_0(x)| \leq \epsilon$ and $|f_1(z) - f_1(x)| \leq \epsilon$ from the local Lipschitz assumption. We take N_0 sufficiently large such that $2\frac{M}{N_0} < (f_L - \epsilon)\lambda(B_{0,\delta'})$. Then for any $0 < w \leq 2\frac{M}{N_0}$, we have $w < (f_L - \epsilon)\lambda(B_{0,\delta'})$. We take $t > 0$ such that $w + t < (f_L - \epsilon)\lambda(B_{0,\delta'})$. Then for any $(\nu_0, \nu_1) \in \mathcal{P}_{x,p}(f_L, f_U, L, d, \delta)$,

$$P(w \leq W \leq w + t) = \nu_1\left(\left\{z \in \mathbb{R}^d : \nu_0(B_{z,\|x-z\|}) \in [w, w + t]\right\}\right)$$

$$\leq \frac{f_1(x) + \epsilon}{f_0(x) - \epsilon} \nu_0 \left(\left\{ z \in \mathbb{R}^d : \nu_0(B_{z, \|x-z\|}) \in [w, w+t] \right\} \right).$$

Notice that f_0 is lower bounded by f_L . Then for N_0 sufficiently large,

$$\limsup_{t \rightarrow 0} t^{-1} \mathbb{P}(w \leq W \leq w+t) \leq \frac{f_1(x) + \epsilon}{f_0(x) - \epsilon} (1 + \epsilon).$$

This then completes the proof. $\square$

C.5.3 Proof of Lemma C.3.4

Proof of Lemma C.3.4. Due to the i.i.d.-ness of Z and $\tilde{Z}$,

$$\begin{aligned} & \left(\frac{N_0}{M} \right)^2 \left[\mathbb{P} \left(W \leq V, \widetilde{W} \leq \widetilde{V}, W \leq 2 \frac{M}{N_0}, \widetilde{W} \leq 2 \frac{M}{N_0} \right) \right. \\ & \quad \left. - \mathbb{P} \left(W \leq V, W \leq 2 \frac{M}{N_0} \right) \mathbb{P} \left(\widetilde{W} \leq \widetilde{V}, \widetilde{W} \leq 2 \frac{M}{N_0} \right) \right] \\ &= \left(\frac{N_0}{M} \right)^2 \int_0^{2 \frac{M}{N_0}} \int_0^{2 \frac{M}{N_0}} \left[\mathbb{P}(V \geq w_1, \widetilde{V} \geq w_2) - \mathbb{P}(V \geq w_1) \mathbb{P}(\widetilde{V} \geq w_2) \right] f_W(w_1) f_W(w_2) dw_1 dw_2 \\ &\leq 4 \left(\frac{f_U}{f_L} \right)^2 \left(\frac{N_0}{M} \right)^2 \int_0^{2 \frac{M}{N_0}} \int_0^{2 \frac{M}{N_0}} \left| \mathbb{P}(V \geq w_1, \widetilde{V} \geq w_2) - \mathbb{P}(V \geq w_1) \mathbb{P}(\widetilde{V} \geq w_2) \right| dw_1 dw_2 \\ &= 4 \left(\frac{f_U}{f_L} \right)^2 \int_{-1}^1 \int_{-1}^1 \left| \mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1), \widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right) \right. \\ & \quad \left. - \mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1) \right) \mathbb{P} \left(\widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right) \right| dt_1 dt_2, \end{aligned}$$

where the last step is from taking $w_1 = \frac{M}{N_0} (1+t_1)$ and $w_2 = \frac{M}{N_0} (1+t_2)$.

Let

$$S(t_1, t_2) = \left| \mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1), \widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right) - \mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1) \right) \mathbb{P} \left(\widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right) \right|.$$

If $t_1 \geq t_2 \geq 0$, $S(t_1, t_2) \leq \mathbb{P}(V \geq \frac{M}{N_0} (1+t_1)) = \mathbb{P}(U_{(M)} \geq \frac{M}{N_0} (1+t_1))$. If $t_2 \geq t_1 \geq 0$, $S(t_1, t_2) \leq \mathbb{P}(\widetilde{V} \geq \frac{M}{N_0} (1+t_2)) = \mathbb{P}(U_{(M)} \geq \frac{M}{N_0} (1+t_2))$. Then for $t_1, t_2 \geq 0$,

$$S(t_1, t_2) \leq \mathbb{P} \left(U_{(M)} \geq \frac{M}{N_0} (1+t_1 \vee t_2) \right).$$

If $t_1 \leq t_2 \leq 0$ and $\mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1), \widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right) \geq \mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1) \right) \mathbb{P} \left(\widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right)$,

$$S(t_1, t_2) \leq \mathbb{P} \left(\widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right) - \mathbb{P} \left(V \geq \frac{M}{N_0} (1+t_1) \right) \mathbb{P} \left(\widetilde{V} \geq \frac{M}{N_0} (1+t_2) \right)$$

$$\begin{aligned}
&= \mathbb{P}\left(V \leq \frac{M}{N_0}(1+t_1)\right) \mathbb{P}\left(\tilde{V} \geq \frac{M}{N_0}(1+t_2)\right) \\
&\leq \mathbb{P}\left(V \leq \frac{M}{N_0}(1+t_1)\right) = \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0}(1+t_1)\right).
\end{aligned}$$

If $t_1 \leq t_2 \leq 0$ and $\mathbb{P}\left(V \geq \frac{M}{N_0}(1+t_1), \tilde{V} \geq \frac{M}{N_0}(1+t_2)\right) \leq \mathbb{P}\left(V \geq \frac{M}{N_0}(1+t_1)\right) \mathbb{P}\left(\tilde{V} \geq \frac{M}{N_0}(1+t_2)\right)$,

$$\begin{aligned}
S(t_1, t_2) &\leq \mathbb{P}\left(\tilde{V} \geq \frac{M}{N_0}(1+t_2)\right) - \mathbb{P}\left(V \geq \frac{M}{N_0}(1+t_1), \tilde{V} \geq \frac{M}{N_0}(1+t_2)\right) \\
&= \mathbb{P}\left(V \leq \frac{M}{N_0}(1+t_1), \tilde{V} \geq \frac{M}{N_0}(1+t_2)\right) \leq \mathbb{P}\left(V \leq \frac{M}{N_0}(1+t_1)\right) = \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0}(1+t_1)\right).
\end{aligned}$$

If $t_2 \leq t_1 \leq 0$, we can establish in the same way that

$$S(t_1, t_2) \leq \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0}(1+t_2)\right).$$

Then for $t_1, t_2 \leq 0$,

$$S(t_1, t_2) \leq \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0}(1+t_1 \wedge t_2)\right).$$

For $t_1 \geq 0 \geq t_2$, if $t_1 + t_2 \geq 0$, $S(t_1, t_2) \leq \mathbb{P}(U_{(M)} \geq \frac{M}{N_0}(1+t_1))$, and if $t_1 + t_2 \leq 0$, $S(t_1, t_2) \leq \mathbb{P}(U_{(M)} \leq \frac{M}{N_0}(1+t_2))$. Then

$$\begin{aligned}
&\left(\frac{N_0}{M}\right)^2 \left[\mathbb{P}\left(W \leq V, \tilde{W} \leq \tilde{V}, W \leq 2\frac{M}{N_0}, \tilde{W} \leq 2\frac{M}{N_0}\right) \right. \\
&\quad \left. - \mathbb{P}\left(W \leq V, W \leq 2\frac{M}{N_0}\right) \mathbb{P}\left(\tilde{W} \leq \tilde{V}, \tilde{W} \leq 2\frac{M}{N_0}\right) \right] \leq 4\left(\frac{f_U}{f_L}\right)^2 \int_{-1}^1 \int_{-1}^1 S(t_1, t_2) dt_1 dt_2 \\
&= 4\left(\frac{f_U}{f_L}\right)^2 \left[\int_0^1 \int_0^1 S(t_1, t_2) dt_1 dt_2 + \int_{-1}^0 \int_{-1}^0 S(t_1, t_2) dt_1 dt_2 + 2 \int_0^1 \int_{-1}^0 S(t_1, t_2) dt_1 dt_2 \right],
\end{aligned} \tag{C.5.1}$$

where the last step is from the symmetry of $S(t_1, t_2)$.

For the first term in (C.5.1), by the symmetry of $S(t_1, t_2)$ and the Chernoff bound,

$$\begin{aligned}
&\int_0^1 \int_0^1 S(t_1, t_2) dt_1 dt_2 \leq \int_0^\infty \int_0^\infty S(t_1, t_2) dt_1 dt_2 = 2 \int_0^\infty \int_0^\infty S(t_1, t_2) \mathbf{1}(t_1 \geq t_2) dt_1 dt_2 \\
&\leq 2 \int_0^\infty \int_0^\infty \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0}(1+t_1 \vee t_2)\right) \mathbf{1}(t_1 \geq t_2) dt_1 dt_2 = 2 \int_0^\infty t \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0}(1+t)\right) dt \\
&\leq 2 \int_0^\infty t(1+t)^M e^{-Mt} dt.
\end{aligned}$$

Notice that since $M \rightarrow \infty$, by Stirling's approximation,

$$\int_0^\infty t(1+t)^M e^{-Mt} dt = \frac{1}{M} + \frac{e^M}{M} \int_1^\infty t^M e^{-Mt} dt \leq \frac{1}{M}(1+o(1)). \tag{C.5.2}$$

We then obtain

$$\int_0^1 \int_0^1 S(t_1, t_2) dt_1 dt_2 \leq \frac{2}{M} (1 + o(1)). \quad (\text{C.5.3})$$

For the second term in (C.5.1),

$$\begin{aligned} \int_{-1}^0 \int_{-1}^0 S(t_1, t_2) dt_1 dt_2 &= 2 \int_{-1}^0 \int_{-1}^0 S(t_1, t_2) \mathbf{1}(t_1 \leq t_2) dt_1 dt_2 \\ &\leq 2 \int_{-1}^0 \int_{-1}^0 \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0} (1 + t_1 \wedge t_2)\right) \mathbf{1}(t_1 \leq t_2) dt_1 dt_2 \\ &= 2 \int_0^1 t \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0} (1 - t)\right) dt \leq 2 \int_0^1 t (1 - t)^M e^{Mt} dt. \end{aligned}$$

Notice that

$$\int_0^1 t (1 - t)^M e^{Mt} dt \leq \frac{1}{M}. \quad (\text{C.5.4})$$

We then obtain

$$\int_{-1}^0 \int_{-1}^0 S(t_1, t_2) dt_1 dt_2 \leq \frac{2}{M}. \quad (\text{C.5.5})$$

For the third term in (C.5.1),

$$\begin{aligned} &\int_0^1 \int_{-1}^0 S(t_1, t_2) dt_1 dt_2 \\ &= \int_0^1 \int_{-t_1}^0 \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} (1 + t_1)\right) dt_1 dt_2 + \int_0^1 \int_{-1}^{-t_1} \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0} (1 + t_2)\right) dt_1 dt_2 \\ &= \int_0^1 t \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} (1 + t)\right) dt + \int_{-1}^0 (-t) \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0} (1 + t)\right) dt \\ &\leq \int_0^\infty t \mathbb{P}\left(U_{(M)} \geq \frac{M}{N_0} (1 + t)\right) dt + \int_{-1}^0 (-t) \mathbb{P}\left(U_{(M)} \leq \frac{M}{N_0} (1 + t)\right) dt \\ &\leq \frac{1}{M} (1 + o(1)) + \frac{1}{M} = \frac{2}{M} (1 + o(1)), \end{aligned}$$

where the last step is from (C.5.2) and (C.5.4).

We then obtain

$$\int_0^1 \int_{-1}^0 S(t_1, t_2) dt_1 dt_2 \leq \frac{2}{M} (1 + o(1)). \quad (\text{C.5.6})$$

Plugging (C.5.3), (C.5.5), (C.5.6) into (C.5.1) yields

$$\begin{aligned} & \left(\frac{N_0}{M}\right)^2 \left[\mathbb{P}\left(W \leq V, \widetilde{W} \leq \widetilde{V}, W \leq 2\frac{M}{N_0}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right. \\ & \quad \left. - \mathbb{P}\left(W \leq V, W \leq 2\frac{M}{N_0}\right) \mathbb{P}\left(\widetilde{W} \leq \widetilde{V}, \widetilde{W} \leq 2\frac{M}{N_0}\right) \right] \leq 32 \left(\frac{f_U}{f_L}\right)^2 \frac{1}{M} (1 + o(1)), \end{aligned}$$

and thus completes the proof. $\square$