NON-ASYMPTOTIC APPROXIMATION OF SYMMETRIC STATISTICS: APPLICATIONS IN ECONOMETRICS

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL OF BUSINESS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DAVID M. RITZWOLLER

JUNE 2026

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Guido Imbens, Primary Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Joseph Romano, Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Han Hong

Approved for the Stanford University Committee on Graduate Studies.

Kenneth Goodson, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Abstract

This dissertation develops non-asymptotic approximations to various symmetric statistics that arise in econometric analysis. Across the three chapters, the procedures under consideration can be expressed, exactly or to leading order, as averages of statistics formed from partitions, subsets, or pairs of observations drawn from a sample of measurements of economic variables. The essential difficulty is that a single observation may enter many of the terms in these averages, inducing complex dependence across the component statistics. The common mathematical strategy is to obtain tractable approximations by exploiting symmetries in the statistics of interest over rearrangements of the observations.

Preface

Although the chapters in this dissertation differ in substantive focus, they share a common mathematical structure. Each chapter studies an econometric procedure whose behavior is determined by an average over many overlapping functions of the observations in a fixed sample. This overlap induces complex dependence across the summands. As a consequence, standard methods for approximating independent averages do not apply directly.

The common strategy is to replace independence with symmetry. In each chapter, the statistic under consideration is invariant to a relabeling or rearrangement of the observations. That invariance turns the dependence induced by the reuse of observations into a source of structure. The dissertation develops non-asymptotic approximations that exploit this structure, and then uses those approximations to answer econometric questions about reproducibility, learning, and inference.

Chapter 1

The first chapter, coauthored with Joseph P. Romano, studies the residual, or post-data, variability induced by cross-splitting. In cross-fit and cross-validated methods, the researcher does not observe a statistic determined by the data alone. The reported output also depends on a random partition of the sample. When many partitions give nearly the same answer, this distinction is harmless. When they do not, the split itself can affect empirical conclusions.

We first show that this residual randomness can be economically and statistically meaningful. In examples from the applied economics literature, different random splits change the statistical significance of a cross-fit treatment-effect estimate, the interpretation of a model selected by cross-validation, and the features of treatment-targeting rules learned from cross-fit nuisance-parameter estimates.

We then propose a method for removing this randomness. The procedure takes as inputs a tolerance and a nominal error probability. It sequentially averages the statistic over randomly drawn partitions of the sample and estimates the remaining split-induced variation. The procedure stops once this estimated variation is small enough that two independent runs of the same algorithm would differ by more than the tolerance with probability approximately equal to the nominal error rate. Thus, by setting the bound and error rate to be sufficiently small, sample-split statistics aggregated with the procedure are reproducible.

The analysis begins from a common representation of cross-split statistics. We consider averages of the form

$$a(\mathsf{R}_{g,k}, D) = \frac{1}{g}\sum_{h=1}^{g} a(\mathsf{r}_h, D) \tag{P.1}$$

where

$$a(\mathsf{r}, D) = \mathcal{A}(\{T(\mathsf{s}_j, D)\}_{j=1}^{k}) \quad \text{and} \quad T(\mathsf{s}, D) = \Psi(D_{\mathsf{s}}, \hat{\eta}(D_{\tilde{\mathsf{s}}}))\ . \tag{P.2}$$

Here, $D = (D_i)_{i=1}^{n}$ is an i.i.d. sample, $\mathsf{R}_{g,k} = (\mathsf{r}_h)_{h=1}^{g}$ is a collection of k-fold partitions $\mathsf{r}_h = (\mathsf{s}_{j,h})_{j=1}^{k}$ of $[n] = \{1, \ldots, n\}$, and $D_{\mathsf{s}} = (D_i)_{i\in\mathsf{s}}$ denotes the observations indexed by s. The complement of s is $\tilde{\mathsf{s}} = [n] \setminus \mathsf{s}$, the map $\hat{\eta}(\cdot)$ is an estimator of a nuisance parameter, and $\Psi(\cdot)$ and $\mathcal{A}(\cdot)$ are known functions.

Our target is the conditional expectation

$$\overline{a}(D) = \mathbb{E}[a(\mathsf{r}, D) \mid D]\ . \tag{P.3}$$

The quantity (P.3) is the average value the statistic would take over the distribution of random k-fold partitions. The question is how rapidly the computable average $a(\mathsf{R}_{g,k}, D)$ approaches this target as the number of repeated cross-splits grows.

The essential difficulty is that the summands in (P.1) and (P.2) are not independent, both withing and across partitions. We handle this dependence using the method of exchangeable pairs (Stein, 1986). The key observation is that a random collection of partitions can be generated from a random collection of permutations. We construct an exchangeable pair by selecting one of these permutations and swapping two of its components, adapting a coupling argument due to Chatterjee (2005, 2007).

Under regularity conditions, this construction yields the conditional approximation

$$P\left\{|a(\mathsf{R}_{g,k}, D) - \overline{a}\,(D)\,| \leq \sqrt{\frac{1}{n}\frac{1}{k}\frac{1}{g}\frac{\log(\varepsilon^{-1})}{\delta}} \mid D\right\} \geq 1 - \varepsilon \tag{P.4}$$

with probability at least $1-\delta$ over the draw of the data. The bound shows how the residual randomness decreases with the sample size, the number of folds, and the number of repeated cross-splits. We use this approximation to characterize how the computation needed to achieve a given bound on residual randomness and the accuracy of the nominal error rate depend on the desired error tolerance and number of folds used for cross-splitting.

Chapter 2

The second chapter, coauthored with Vasilis Syrgkanis, studies a classical problem in a non-classical regime. A U-statistic is an average of a function evaluated on each of the subsamples of a dataset of a given size; that is, statistics of the form

$$U_{n,b}(u) = \binom{n}{b}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} u(D_{\mathsf{s}}) \,, \tag{P.5}$$

where $\mathcal{S}_{n,b}$ collects all of the subsets of $[n]$ of size b, the vector $D = (D_i)_{i=1}^n$ collects an independent sample, and D_{s} again gives the subset of the sample with indices in s. Many modern non-parametric regression estimators have this structure, including random forest regression. Classical theory is sharp when the size b is fixed or the dimension d is small. The chapter asks what remains true when both b and d are allowed to be non-negligible relative to the sample size n.

The canonical approach for studying the large-sample behavior of statistics with this structure, due to Hoeffding (1948), is based on the observation that the projection of the average (P.5) onto the space of statistics that decompose additively into the contributions of individual observations takes the form

$$\frac{b}{n} \sum_{i=1}^n u^{(1)}(D_i) \,, \quad \text{where} \quad u^{(1)}(D) = \mathbb{E}[u(D_{[b]}) \mid D_1 = D] - \mathbb{E}[u(D_{[b]})] \,. \tag{P.6}$$

This is referred to as the Hájek projection of the U-statistic (P.5), following Hájek (1968). If the residual from this projection is of lower order, then the analysis of the U-statistic reduces to the analysis of an average of independent random variables. In the classical scalar fixed-order setting, Hoeffding (1948) shows that, under regularity conditions,

$$\sqrt{\frac{n}{\sigma_b^2 b^2}} (U_{n,b}(u) - \mathbb{E}[u(D_{[b]})]) \xrightarrow{d} \mathsf{N}(0,1) \,, \quad \text{where} \quad \sigma_b^2 = \mathrm{Var}(u^{(1)}(D_i)) \,. \tag{P.7}$$

See Serfling (1980), Lee (1990), and van der Vaart (2000) for textbook treatments.

We make this projection argument non-asymptotic and order-explicit in a high-dimensional, growing-order regime. Suppose $u = (u_1, \ldots, u_d)$ and each component satisfies the Orlicz-norm bound $\|u_j(D_{[b]})\|_{\psi_1} \leq \phi$. If the order b satisfies $b \log(dn) \leq cn$, we show that

$$\left\| U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n} \sum_{i=1}^{n} u^{(1)}(D_i) \right\|_{\infty} \leq C \frac{b}{n} \phi \log^2(dn) \tag{P.8}$$

with probability at least $1 - C/nd$. The bound is unimprovable up to the polynomial factor in the logarithmic term.

The linearization (P.8) has two main consequences. First, it yields as a corollary an order-explicit Bernstein-type large-deviation bound whose leading term matches the scaling suggested by the asymptotic approximation (P.7). Second, it can be used to derive an order-explicit high-dimensional central limit theorem that applies throughout the regime $b \log^5(dn) = o(n)$. These results extend the available high-dimensional theory for U-statistics to cases where the order and dimension of the statistic are large.

The structure of the proof departs from existing arguments. In the scalar case, one can combine the orthogonality of the Hoeffding decomposition with Chebyshev-type inequalities. This is not enough for high-dimensional approximation, where exponential tails are needed. Existing high-dimensional arguments rely on decoupling and symmetrization methods developed in De la Peña and Montgomery-Smith (1995) and Sherman (1994) that have poor dependence the order b. Our main technical contribution is an order-explicit one-sided decoupling inequality for U-statistics. After decoupling, we combine a symmetrization argument due to Sherman (1994) with a specialization of a moment inequality from Giné et al. (2000). This yields exponential bounds for higher-order Hoeffding terms with the required dependence on both b and d.

The final part of the chapter applies these approximations to inference for nonparametric regression estimators built from subsampled kernels, including subsampled nearest-neighbor regression and generalized random forest regression. Under regularity conditions, these estimators are well-approximated by U-statistics of order b with d-dimensional kernels. At the same time, consistency of the regression estimators requires the subsample size to satisfy $n^{p/(p+c)} = o(b)$, where p is the dimension of the covariates. This places the relevant regime outside the range $b = o(n^{1/3})$ covered by the available high-dimensional central limit theorems, while scalar results that allow for large orders b cannot handle growing dimensions d. The chapter's approximation closes this gap, allowing us to establish the consistency of a class of resampling-based simultaneous confidence intervals.

Chapter 3

The econometric problem considered in the third chapter is motivated by the fact that the indirect effects of economic actions are often multifaceted. For instance, one firm's investment in research and development may influence another firm's value through both product market competition and communication among researchers (Arrow, 1962; Dasgupta and Stiglitz, 1980; Griliches, 1992; Bloom et al., 2013). Likewise, productivity changes in one city may affect output in others through the exchange of goods, labor, and ideas (Marshall, 1890; Ellison et al., 2010; Allen and Arkolakis, 2014; Monte et al., 2018). As a result, empirical strategies that relate spillover effects to single measures of economic proximity risk conflating distinct and potentially countervailing mechanisms.

Concretely, the chapter focuses on the interpretation and statistical analysis of a common regression specification used by empirical economists to characterize how the indirect effects of economic actions vary with measures of economic proximity. In many applications with this structure, empirical researchers consider regressions specifications of the form

$$Y_i = \alpha + \theta \cdot \Delta_i + \varepsilon_i, \quad \text{where} \quad \Delta_i = \sum_{j \neq i} D_{i,j} W_j \, . \tag{P.9}$$

Here, Y_i is the outcome of unit i, W_j is the treatment, shock, or policy variable associated with unit j, and $D_{i,j}$ is a measure of the economic proximity of i and j. I begin by showing that, under regularity conditions, the coefficient θ in (P.9) has a clear nonparametric interpretation. In particular, the coefficient recovers a convex average of the mixed partial derivatives

$$\frac{\partial^2}{\partial \delta \, \partial w} \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w] \, . \tag{P.10}$$

That is, the regression (P.9) measures how the association between one unit's outcome and another unit's treatment correlates with the proximity of the two units, marginalizing over the remainder of the population.

This interpretation is useful, but only descriptive. Suppose that the effects of changes in firm-level investment in research and development W_j on an outcome Y_i are larger for pairs of firms i, j that are technologically similar, and technological similarity is positively correlated with geographic proximity $D_{i,j}$, then the parameter (P.10) does not recover the causal relationship between geographic proximity and productivity spillover intensity. This distinction matters for policy. Estimates of the role of geographic proximity are essential for evaluating policies change the spatial organization of economic activity, such as research parks, transportation investments, and special economic zones.

The remainder of the chapter develops a framework for estimating causal relationships of this form. I introduce a class of estimands, that I call spillover proximity gradients, that describe how the effect of unit j's treatment on unit i's outcome changes under an intervention that alters a chosen measure of proximity between the two units. I then show how to adapt approaches to nonparametric causal inference used in settings without interference to this context. I use strategies based on selection on observables as a running example, because this best describes the relevant empirical literature.

Formally, let X_i and $H_{i,j}$ denote collections of unit-level and pair-level covariates, respectively. The proposed procedure first forms

$$\hat{W}_j^\star = W_j - \hat{\pi}_n(X_j), \quad \hat{D}_{i,j}^\star = D_{i,j} - \hat{\gamma}_n(H_{i,j}), \quad \text{and} \quad \hat{\Delta}_i^\star = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star \,. \tag{P.11}$$

It then estimates the coefficient in the residualized specification

$$Y_i = \alpha + \theta \cdot \hat{\Delta}_i^\star + \varepsilon_i \,. \tag{P.12}$$

The residualization of W_j accounts for selection into treatment using unit-level covariates. The residualization of $D_{i,j}$ removes variation in the proximity measure of interest that is explained by observed pair-level characteristics. Under identifying conditions, the population coefficient in (P.12) recovers a convex average of spillover proximity gradients.

My statistical analysis of this procedure allows for rich forms of interference. Each unit's outcome is modeled as an anonymous, or permutation-invariant, function of the variables that characterize other units. Let the state variable Z_i collect the treatment and any observed and unobserved characteristics of unit i. The maintained structure can be written as

$$Y_i = F\left(Z_i, \{Z_j : j \neq i\}\right) \,, \tag{P.13}$$

where the function F is invariant to permutations of the multi-set of other units' states. This permutation-invariant representation permits a Hoeffding-type decomposition

$$Y_i = \mathbb{E}[Y_i \mid Z_i] + \sum_{j \neq i} (\mathbb{E}[Y_i \mid Z_i, Z_j] - \mathbb{E}[Y_i \mid Z_i]) + R_i \,. \tag{P.14}$$

The first term captures the component of the outcome determined by unit i's own state. The summation captures the contribution of other units one at a time. The remainder R_i collects higher-order interactions among several other units. The chapter controls this remainder by imposing smoothness restrictions on F and a local structure on interference.

The local structure is expressed through an unobserved latent space. Interference can be substantial only between units that are close in this latent space. Let ρ_n denote the fraction of units that are latent neighbors of a typical unit. I focus attention on the asymptotic sequence that allows $n\rho_n$ to diverge while ρ_n converges to zero. The observed proximity measure $D_{i,j}$ need not be the latent distance that organizes interference. This is essential for settings where spillovers are plausibly mediated by social, technological, or economic relationships that are only partially observed.

After substituting the decomposition (P.14) into the the coefficient in (P.12), the leading term can be written as a degenerate U-statistic of order three. This is the main technical distinction from conventional regression asymptotics. I use this representation to show that

$$\hat{\theta}_n^\star = \theta_n^\star + O_p(\rho_n^{1/2}) \ , \tag{P.15}$$

where $\theta_n^\star$ denotes the pseudo-parameter identified by the residualized regression. I show that the rate $O_p(\rho_n^{1/2})$ is unimprovable. The limiting distribution is nonstandard for the same reason. Degenerate U-statistics do not generally converge to Gaussian limits. Their limits can instead be represented as weighted sums of centered chi-square variables, with weights determined by the eigenvalues of the relevant kernel. I show that, under the same regularity conditions need for consistency, these relevant eigenvalues become sufficiently diffuse to converge to a Gaussian limit.

The final part of the chapter develops an inference procedure that is adapted to this structure. Standard clustered standard errors require a partition of units into approximately independent groups. Kernel methods require interference to decay with an observed distance. Neither approach is well suited to a settings where interference may be unobserved. The proposed alternative uses the fact that the residualized treatment variables enter the kernel of the leading U-statistic linearly and are centered. Resampling their signs preserves the variance contribution of this part of the kernel.

Operationally, the procedure computes residuals from (P.12), draws independent Rademacher signs $V_1, \ldots, V_n$, and forms the sign-flipped exposure

$$\hat{\Delta}_i^\star(V) = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star V_j \ . \tag{P.16}$$

For each draw of the signs, it regresses the empirical residuals on $\hat{\Delta}_i^\star(V)$. If $\hat{\phi}_n^\star(V)$ denotes the resulting coefficient, then the variance estimate is given by

$$\hat{\sigma}_n^2 = \frac{2}{B} \sum_{b=1}^{B} \left(\hat{\phi}_n^\star(V^{(b)}) \right)^2 \, . \tag{P.17}$$

The resulting confidence intervals are conservative (and exact in some non-pathological cases) for the pseudo-parameter targeted by the residualized regression. They do not require the researcher to assume that all relevant spillover channels are measured.

The chapter concludes by applying the method to settings in which spillovers are plausibly mediated by multiple correlated measures of proximity. The main application revisits productivity spillovers across firms using data from Bloom et al. (2013). Additional applications study spillovers across locations. These exercises return to the economic motivation for the chapter. Adjusting for correlated channels can materially change the estimated relationship between a chosen proximity measure and spillover intensity. The contribution of the chapter is to make such comparisons interpretable and inferentially tractable when spillovers are heterogeneous, nonlinear, and partly unobserved.

Proofs and discussion supporting Chapters 1 to 3 are given in Appendices A to C, respectively.

Acknowledgments

At many junctures in my life and education, I have been fortunate to receive opportunities, mentorship, and support, without which the development and completion of this dissertation would not have been possible. Although there have been too many of these moments to name, or, I am sure, to recognize, I will mention some that have asserted a particularly prominent role in my reflections as I end my time in graduate school.

My career in academic economics began in my sophomore year at Brandeis University, when Blake LeBaron answered my cold email inquiring about a research assistantship. He taught me many foundational things, about research, economics, and academia, and was extremely generous with his time and support. I am extraordinarily grateful for this, and remain particularly influenced by his insistence on keeping computational considerations at the fore in discussions of both econometric methodology and economic behavior.

I had an incredible stroke of luck when Matthew Gentzkow and Jesse Shapiro accepted my application for a pre-doctoral research fellowship at Stanford. Jesse taught me about rigor, in organization, in reasoning, in revision, and in writing. Matt taught me about taste, in how questions are posed and in how methods are matched to audiences, and about effective identification of pressure points in arguments. Their standard is the one that I will be striving, and failing, to emulate for the remainder of my career.

My academic identity as an econometrician crystallized when Joseph Romano agreed to supervise and collaborate on a study of statistical tests for the hot hand fallacy. In this highly formative project, and in our subsequent collaborations, I learned how to present and defend my ideas, about how to build clearly and honestly on the history of statistical thought, about how to reason about uncertainty with integrity and, most of all, about the joy of giving mathematical explanations for observable statistical phenomena. More than this still, I learned from Joe's example the importance of kindness and humility and the value of living a balanced life.

I was fortunate to join the Stanford econometrics group at a time of incredible growth, afforded, I think, to the leadership and dynamism of my advisor Guido Imbens. I have heard that the most important part of graduate school is the accumulation of voices in your head, asserting their perspective on your ideas. Guido's voice is particularly resonant to me in this respect. I learned from Guido to be ambitious, to value transparency and prioritize simplicity, to be epistemologically positive and empirical, and to confront new ideas and technology with a balanced optimism. It is clear to me that Guido's patience and generosity in managing students and junior scholars has enabled the econometrics and causal inference communities at Stanford, and at large, to flourish.

The econometrics and statistics groups at Stanford were extraordinarily welcoming and motivating. I am enormously thankful to Han Hong for his close engagement with my work, and for his continued warmth and encouragement. Brad Ross and Kevin Chen provided incredible support and friendship during the COVID-19 pandemic. Jann Speiss and Lihua Lei set very compelling examples, and were amazingly generous in their support. I am indebted to my coauthors—Thomas DiCiccio, Azeem Shaikh, Sabri Eyuboglu, Vasilis Syrgkanis, Stefano DellaVigna, and Woojin Kim—for their incredible energy and creativity. Classes and reading groups with Stefan Wager, Andrea Montanari, Sourav Chaterjee, David Donoho, and, particularly, John Duchi were instrumental in my training.

I am thankful to the contemporary students in the econometrics group—Michael Pollman, Jonas Metzger, Evan Munro, Lea Bottmer, Jason Weitze, Amar Venugopal, Sarah Vicol, Janelle Nelson, Alex Almeida, and Gustavo Schlemper—for establishing a very dynamic and supportive culture. My neighbors in the GSB economics cubicles—Cody Cook, Zane Kashner, Joey Feffer, Filip Tokarski, and Ben Barrett—provided a lively and engaging intellectual environment. On my first day as a pre-doc, Frank Yang taught me how to download RStudio. I thank him for this, and for the hundreds of steps, big and small, he's shown me how to take since then.

I am most grateful to my family, who have provided endless patience, encouragement, and support toward the completion of this dissertation. I thank my father Michael for teaching me to value science and the pursuit of knowledge. I thank my mother Debra for teaching me to be tenacious and skeptical. I thank my sister Dawn for her grit and unwavering positivity. And I thank my wife Maya for finding plants for me to hold to the sun, good things great, and bad things funny.

David Ritzwoller
Stanford, CA
May 2026

Contents

Abstract v

Preface vi

Acknowledgments xiv

1 Reproducible Aggregation of Sample-Split Statistics 1

1.1 Introduction 1

1.2 The Residual Randomness of Sample-Split Statistics 4

1.3 Reproducible Aggregation 13

1.4 Theoretical Analysis 21

1.5 Recommendations for Practice 31

2 Order-Explicit Linearization of High-Dimensional U-Statistics 35

2.1 Introduction 35

2.2 Proof of Theorem 2.1.1 40

2.3 Simultaneous Inference for Subsampled Kernel Regression 47

3 Regression Adjustments for Disentangling Spillover Effects 63

3.1 Introduction 64

3.2 A Nonparametric Decomposition of Spillover Regression 70

3.3 Spillover Proximity Gradients 78

3.4 Identification and Estimation 85

3.5 Consistency 93

3.6 Inference 101

3.7 Empirical Application 108

3.8 Conclusion 111

A Supplemental Appendices for Chapter 1 **114**
A.1 Data and Simulations . . . 114
A.2 Auxiliary Results and Discussion . . . 121
A.3 Proofs for Results Stated in Section 1.3 . . . 134
A.4 General Non-Asymptotic Results . . . 138
A.5 Proofs for Lemmas Stated in Appendix A.4 . . . 156

B Supplemental Appendices for Chapter 2 **188**
B.1 Proofs for Results Supporting Theorem 2.1.1 . . . 188
B.2 An Abstract Bound on Coverage Error . . . 204
B.3 Proof of Theorem 2.3.1 . . . 230
B.4 Additional Results and Discussion . . . 249
B.5 Proofs Supporting Appendix B.4 . . . 259
B.6 Supplemental Figures . . . 269

C Supplemental Appendices for Chapter 3 **275**
C.1 Proofs for Decomposition Theorems . . . 275
C.2 Proofs for Identification Results . . . 291
C.3 Proofs for Asymptotic Results . . . 294
C.4 Proofs for Lemmas Supporting Appendix C.3 . . . 312
C.5 Additional Results and Discussion . . . 362
C.6 Data and Additional Applications . . . 383

Chapter 1

Reproducible Aggregation of Sample-Split Statistics

with Joseph P. Romano

Statistical inference is often simplified by sample-splitting. This simplification comes at the cost of the introduction of randomness not native to the data. We propose a simple procedure for sequentially aggregating statistics constructed with multiple splits of the same sample. The user specifies a bound and a nominal error rate. If the procedure is implemented twice on the same data, the nominal error rate approximates the chance that the results differ by more than the bound. We illustrate the application of the procedure to several widely applied econometric methods.

1.1. Introduction

Sample-splitting is ubiquitous in modern econometric theory. Routine statistical tasks—model selection, dimension reduction, nuisance parameter estimation—can be implemented on a randomly selected subsample of a data set, without contaminating the validity of a statistical inference produced on its complement. This principle underlies the widely applied practices of cross-validation for predictive risk estimation (Stone, 1974; Arlot and Celisse, 2010) and cross-fitting for adaptive estimation of semiparametric models (Bickel, 1982; Schick, 1986; Chernozhukov et al., 2018), among a growing set of additional applications.

For a fixed data set, statistics constructed with sample-splitting are not deterministic. Two researchers can compute the same statistic on the same data and obtain different values.

Researchers are incentivized to report significant results. If there is scope to materially alter the statistics that they report through the choice of the split of their sample, should this choice be left to chance?

This paper makes two contributions. First, we show that many widely applied sample-split econometric methods exhibit significant residual randomness. We give examples from the applied economics literature where the randomness induced by sample-splitting determines the statistical significance of a treatment effect estimated with cross-fitting, the interpretation of a model selected with cross-validation, and the qualitative features of treatment targeting rules learned from cross-fit nuisance parameter estimates.

Second, and accordingly, we propose an efficient method for removing the residual randomness from sample-split statistics. The procedure takes as input a bound and an error rate. The statistic of interest is sequentially aggregated over randomly drawn splits of the sample. The procedure is stopped after an estimate of the residual variation of the aggregate statistic falls below a pre-determined threshold. If the procedure were run twice, we show that the chance that the outputs differ by more than the bound is well-approximated by the error rate. That is, by setting the bound and error rate to be sufficiently small, sample-split statistics aggregated with the procedure are reproducible.

We begin, in Section 1.2, by discussing several widely applied sample-split econometric methods and demonstrating that, for each, the randomness induced by sample-splitting can substantively effect results. In Section 1.3, we propose an efficient method for sequentially aggregating sample-split statistics that ensures that the residual randomness is small. We illustrate that, in each of our examples, the proposed method stabilizes results—ensuring reproducibility at a minimal computational expense. We establish that the procedure is valid, in a particular asymptotic sense, without imposing any restrictions on the data generating process or statistic of interest. Similarly, we show that, for a large class of applications, sample-split statistics aggregated with the procedure maintain (or, improve upon) unconditional statistical guarantees. That is, reproducibly aggregated sample-split statistics are still consistent and associated inferences are still valid.

To implement the procedure, a user must make several choices that may affect performance, including the specification of a suitable bound and error rate. Additionally, in most applications, the procedure is applied to stabilize a statistic that is itself constructed with cross-splitting. In these cases, users must also specify how many folds to use for cross-splitting.

To shed light on these choices, in Section 1.4, we give an analysis of the performance of the procedure under a set of simplifying conditions. We give two main results. First, the computation needed to achieve a given bound on residual randomness is very sensitive to the desired error tolerance, but is insensitive to the number of folds that are used for cross-splitting. Second, and on the other hand, the accuracy of the nominal error rate of the procedure deteriorates as the number of folds used for cross-splitting increases. We conclude, in Section 3.8, by synthesizing these results into a set of concrete recommendations for practice. We emphasize simple rules-of-thumb for choosing suitable error tolerances and approaches to cross-splitting.

The main theoretical challenge posed by this analysis is the accommodation of the dependence between statistics computed on cross-splits of a sample. We address this through an application of the method of exchangeable pairs (Stein, 1986; Ross, 2011; Chen et al., 2011). To construct an appropriate exchangeable pair for our problem, we develop a novel application of a coupling argument due to Chatterjee (2005), that may be of independent interest.

1.1.1 Related Literature

The procedure studied in this paper is applicable to a large variety of sample-split statistical methods. In Section 1.2, we give a selective review of various sample-split methods that are frequently used in applied economics. There are many, additional, sample-split methods, proposed in the statistics literature, that have the potential to be useful in econometric applications. Generic sample-split methods for testing statistical hypotheses are studied in Guo and Romano (2017), DiCiccio et al. (2020), and Wasserman et al. (2020). Additional applications include sample-split procedures for selective inference (Rinaldo et al., 2019), inference on high-dimensional linear models (Meinshausen and Bühlmann, 2010), conformal and predictive inference (Lei et al., 2018), and knockoff tests of conditional independence (Barber and Candès, 2015).

Chernozhukov et al. (2018) and Chernozhukov et al. (2023) advocate for the aggregation of estimators and p-values computed with sample splitting, over a pre-determined number of splits, in the context of applications related to semiparametric estimation and characterization of treatment effect heterogeneity, respectively. We second these recommendations and contribute a general purpose method that provides a statistical guarantee that residual randomness has been controlled up to a specified level of error, at a minimal computational cost.

Our setting is related to a large literature that studies methods for constructing confidence intervals for cross-validated estimates of generalization error. Several examples include Dietterich (1998), Nadeau and Bengio (1999), Lei (2020), Bayle et al. (2020), Austern and Zhou (2020), and Bates et al. (2023). By contrast, we are interested in the randomness conditional on the data. Formally, our non-asymptotic results are most similar to the Berry-Esseen bounds given in Austern and Zhou (2020), who study the unconditional normal approximation of statistics similar to the those considered in Section 1.4. Our bounds apply under weaker conditions and are substantially simpler.

Some of our results build on a literature studying the role of algorithmic stability in the accuracy of cross-validation (Kale et al., 2011; Kumar et al., 2013). These papers are related to a broader literature that derives generalization bounds for stable algorithms, originating with Bousquet and Elisseeff (2002). Some of the concentration inequalities that we derive can be compared to the results of Cornec (2010) and Abou-Moustafa and Szepesvári (2019). Again, the setting we study is different and our conditions, and resultant bounds, are substantially simpler.

Although our emphasis is on statistics constructed with sample-splitting, the algorithmic and formal methods studied in this paper are potentially applicable to randomized algorithms more generally (Motwani and Raghavan, 1995). See Beran and Millar (1987) for a classical analysis of the asymptotics of randomized tests and estimators.

1.2. The Residual Randomness of Sample-Split Statistics

Sample-splitting has proliferated as a useful subroutine for simplifying various tasks associated with modern statistical inference. The high-level situation is as follows. Consider a researcher who observes the independent data $D = (D_i)_{i=1}^n$ and wishes to report the statistic

$$\Psi(D, \eta) \,, \tag{1.2.1}$$

where η is some unknown nuisance parameter.

For example, each observation D_i could contain a measurement of an outcome Y_i, a treatment W_i, and a vector collecting a large number of covariates X_i. The researcher could be interested in measuring the effect of W_i on Y_i. To ensure that their estimate is not unnecessarily imprecise, they might like to include controls for only the subset of covariates that are correlated with the outcome. Here, formally, the nuisance parameter η collects the indices of the subset of "relevant" controls and the statistic $\Psi(D, \eta)$ denotes a treatment effect estimate associated with, say, a regression of the outcome on the treatment with controls for covariates with indices in η.

In practice, the researcher cannot compute the statistic (1.2.1), as the nuisance parameter η is unknown. Often, however, an estimator $\hat{\eta}(D)$ is available and so it may be tempting to report the feasible statistic

$$\Psi(D, \hat{\eta}(D)) \ . \tag{1.2.2}$$

In the example, the estimator $\hat{\eta}(D)$ could collect the indices of the covariates whose absolute sample correlation with the outcome is greater than some pre-determined threshold.

Statistical inferences that, counterfactually, would be appropriate if the infeasible statistic (1.2.1) were available will not necessarily be valid if they were instead based on the feasible statistic (1.2.2). In particular, the process of computing the nuisance parameter estimate $\hat{\eta}(D)$ might produce a confounding effect. In the example, screening control variables according to their correlation with the outcome produces a bias in the resultant treatment effect estimate.

Sample-splitting solves this problem.[1] Let s denote a randomly drawn subset of the numbers $[n] = \{1, \ldots, n\}$ of size b and let $\tilde{\mathsf{s}}$ denote its complement. Sample-split statistics take the form

$$T(\mathsf{s}, D) = \Psi(D_{\mathsf{s}}, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \ , \tag{1.2.3}$$

where the quantities $D_{\mathsf{s}} = (D_i)_{i \in \mathsf{s}}$ and $D_{\tilde{\mathsf{s}}} = (D_i)_{i \in \tilde{\mathsf{s}}}$ collect the data with indices in s and $\tilde{\mathsf{s}}$, respectively. Splitting the data D into two independent subsamples $(D_{\mathsf{s}}, D_{\tilde{\mathsf{s}}})$ prevents the construction of nuisance parameter estimates from contaminating statistical inferences that the researcher may wish to make with the feasible statistic (1.2.3).

There are two, well-known, practical issues with this approach. First, by splitting the data, statistical precision may be meaningfully reduced. Second, the statistic (1.2.3) is random through both the data D and the choice of the random subset s. That is, if the same sample-split statistic were computed on the same data by two different researchers, and the statistic was sensitive to the choice of the random subset s, then the researchers could report meaningfully different results.

To address these concerns, researchers often aggregate several replications of sample-split statistics through cross-splitting (Stone, 1974; Schick, 1986). In particular, researchers

[1] Of course, under certain conditions involving the sparsity of the covariance between the treatment, outcome, and controls, "double post selection" of control variables with a Lasso regression (Tibshirani, 1996) can produce consistent treatment effect estimates (Belloni et al., 2014). However, consistent estimates can be obtained under weaker conditions through a closely related approach based on sample-splitting (Chernozhukov et al., 2018; Belloni et al., 2012).

typically report aggregate statistics of the form

$$a(\mathsf{r}, D) = \frac{1}{k}\sum_{j=1}^{k} T(\mathsf{s}_j, D) \ , \tag{1.2.4}$$

where the quantity $\mathsf{r} = (\mathsf{s}_j)_{j=1}^{k}$ denotes a random k-fold partition of $[n]$, i.e., a collection of k mutually exclusive sets whose union is equal to $[n]$. By reusing subsamples of the data for computation of both the statistic of interest and the estimation of nuisance parameters, cross-split statistics mitigate potential losses in statistical precision.

In this section, we demonstrate, in several real applications from applied economics, that the second concern—the residual randomness generated by sample-splitting—can substantively affect results. This residual variability is, often, not resolved by cross-splitting. Consequently, in Section 1.3, we propose an approach to aggregating sample-split statistics that controls the scope of residual randomness at a minimal computational cost. Revisiting the applications, we demonstrate that sample-split statistics aggregated through this procedure are reproducible.

1.2.1 Cross-Validation

The most prevalent instance of sample-splitting in applied economics is the use of cross-validation for model selection. Often, in this setting, the data D_i consist of a measurement of an outcome Y_i and a vector X_i collecting measurements of p covariates. Interest is in choosing a parsimonious subset of the covariates that, together, best predicts the outcome.

Lasso regression is a standard approach to this problem (Tibshirani, 1996; Hastie et al., 2015). The Lasso coefficient is the solution to the regularized least-squares regression

$$\hat{\eta}_\lambda(D) = \underset{\eta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta^\top X_i)^2 + \lambda \sum_{j=1}^{p} |\eta_j| \right\} \ , \tag{1.2.5}$$

where λ is some tuning parameter chosen by the user. Penalization of the ℓ_1-norm of the coefficient η encourages sparsity. That is, often, many elements of $\hat{\eta}_\lambda(D)$ are exactly equal to zero. The set of covariates associated with non-zero coefficients are referred to as the model "selected" by the Lasso. The selected model is sensitive to the choice of the tuning parameter λ. If λ is sufficiently large, no covariates are selected. If λ is sufficiently small,

all covariates are selected.[2]

Cross-validation is a widely applied and, perhaps, uncontroversial approach for choosing tuning parameters. Here, cross-validation entails assigning λ the value that minimizes the cross-split estimate of the out-of-sample mean-squared error

$$a_\lambda(\mathsf{r}, D) = \frac{1}{k}\sum_{j=1}^{k} T_\lambda(\mathsf{s}_j, D)\ , \quad \text{where} \quad T_\lambda(\mathsf{s}, D) = \frac{1}{|\mathsf{s}|}\sum_{i\in\mathsf{s}}(Y_i - \hat{\eta}_\lambda(D_{\bar{\mathsf{s}}})^\top X_i)^2 \tag{1.2.6}$$

and, as before, $\mathsf{r} = (\mathsf{s}_j)_{j=1}^k$ is a random k-fold partition of $[n]$. By measuring error out-of-sample, i.e., in the independent subsample D_s, the cross-validated risk estimate (1.2.6) avoids any "over-fitting" bias induced by the estimation of the coefficient $\hat{\eta}_\lambda(D_{\bar{\mathsf{s}}})$. This approach is widely applied throughout various subfields of applied economics.[3]

The risk estimate (1.2.6) is random both through the data D and through the choice of the k-fold partition r. That is, the choice of λ, and thereby, the selected model, have the potential to change for different choices of the random collection r. To evaluate the scope of this sensitivity, we consider data from Casey et al. (2021a,b). Casey et al. (2021b) study a large-scale experiment, implemented in Sierra Leone, in which a randomly selected subset of parliamentary elections were preceded by direct vote, party-specific primaries. They select covariates to include in various, downstream, econometric analyses with the cross-validated Lasso. In their setting, the outcome Y_i is the vote share, in a poll of party officials, for each of 390 candidates. The vector X_i collects measurements of 48 characteristics for each candidate.

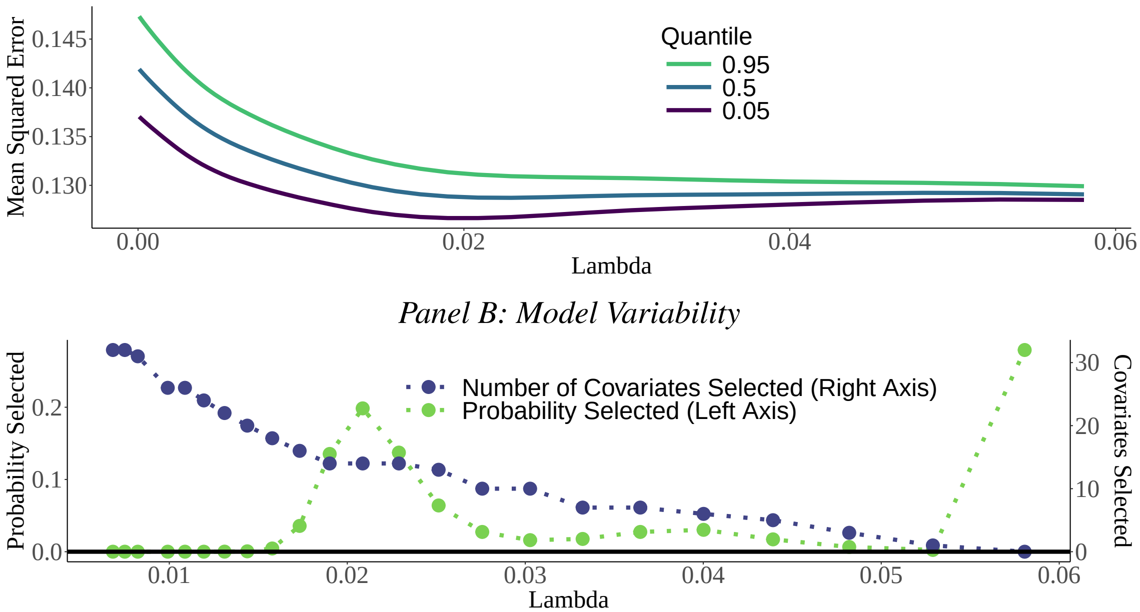
Panel A of Figure 1.1 displays quantiles, across random draws of the 10-fold partition r, of the cross-validated estimate of the mean-squared error (1.2.6) over a grid of values of λ. The curve associated with the 0.05th quantile has a minimum around $\lambda = 0.02$. By contrast, the curve associated with the 0.95th quantile is monotonically decreasing. This induces instability in the value of λ chosen by cross-validation. Panel B of Figure 1.1 displays, in light green, the probability—again, across random 10-fold partitions r—that each value of

[2]For sufficiently small λ, a covariate is selected so long as it is not collinear with other covariates and its sample correlation with the outcome is not exactly equal to zero.

[3]See, for example, applications to Crime (Mastrobuoni, 2020; Arnold et al., 2020), Development (Casey et al., 2021b; Blattman et al., 2024; Sadka et al., 2024), Economic Theory (Fudenberg and Liang, 2019), Education (Ellison and Pathak, 2021), Environment (Deryugina et al., 2019; Cicala, 2022), Finance (Koijen et al., 2024), Health (Abaluck et al., 2016; Cooper et al., 2020), Industrial Organization (Kelly et al., 2023; Dubé and Misra, 2023), Innovation (Chen et al., 2021; Myers and Lanahan, 2022), Labor (Muendler and Becker, 2010; Card et al., 2020; Adermon et al., 2021; Derenoncourt, 2022), Market Design (Agarwal et al., 2019), Macroeconomics (Hansen et al., 2018), and Political Economy (Gentzkow et al., 2019; Cantoni and Pons, 2022).

Figure 1.1: Cross-Validation

Panel A: Mean-Squared Error Quantiles



Notes: Figure 1.1 measures the residual randomness of the cross-validated Lasso, implemented in data from Casey et al. (2021a,b). Panel A displays quantiles of the 10-fold cross-validated estimate of the mean-squared error (1.2.6) of the lasso regression (1.2.5) over a grid of values of λ. In Panel B, the probabilities that each value of λ minimize the cross-validated risk estimate are displayed with light green dots, relative to the left y-axis. The number of covariates selected at each value of λ are displayed with dark blue dots, relative to the right y-axis. See Appendix A.1.1 for further details.

λ is selected, i.e., minimizes the cross-validated estimate of the mean-squared error. The number of covariates in the model associated with each value of λ is displayed in dark blue. The distribution of the selection probabilities has two maxima—both are associated with probabilities greater than 0.2. One entails selecting a model with 14 covariates. The other entails selecting a model with zero covariates. The random choice of the partition r substantively affects the size of the selected model.[4]

1.2.2 Cross-Fitting

Meaningful residual randomness is not particular to cross-validated risk estimation. A second class of sample-split methods, seeing increasing use in applied economics, are characterized by the application of "cross-fitting" to accommodate—or characterize—treatment effect

[4]Casey et al. (2021b) take measures to stabilize their results. In particular, they select covariates that are selected in more than 200 of 400 randomly drawn 10-fold partitions r. We use this setting as an example, in part, because a serious attempt was made to address residual randomness. This is not commonplace in the literature surveyed in Footnote 3 (a similar strategy is used in, e.g., Hansen et al. (2018), however).

heterogeneity. Often, in this case, the data D_i consist of an outcome Y_i, a binary treatment W_i, and a vector of covariates X_i. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes induced by the treatment W_i. As before, let s denote a random a subset of $[n]$, with complement $\tilde{\mathsf{s}}$.

The basic idea underlying this class of methods is to use the data from the units i in $\tilde{\mathsf{s}}$ to predict the treatment effects $Y_i(1) - Y_i(0)$ for the units i in s. For example, a machine learning algorithm can be applied to the data $D_{\tilde{\mathsf{s}}}$ to construct estimates of the conditional expectation

$$\mu_w(x) = \mathbb{E}[Y_i \mid X_i = x, W_i = w] \quad \text{and} \quad \rho(x) = P\{W_i = w \mid X_i = 1\} \tag{1.2.7}$$

for each w in $\{0, 1\}$. Collect these estimates into $\hat{\eta}(D_{\tilde{\mathsf{s}}}) = (\hat{\mu}_1, \hat{\mu}_0)$. Predictions of the treatment effects for the units i in s can be constructed through the sample-split statistic

$$\psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\ . \tag{1.2.8}$$

These predictions can then be used to estimate average treatment effects and characterize treatment effect heterogeneity, among other related problems. The point is, by splitting the data, these "second-stage" analyses are not confounded by the construction of the "first-stage" estimators.

"Double Machine Learning" (DML) estimates of average treatment effects are a leading example of a method that takes this structure (Chernozhukov et al., 2018). Here, the sample-split statistics (1.2.8) are aggregated with cross-splitting, through

$$a(\mathsf{r}, D) = \frac{1}{k}\sum_{j=1}^{k} T(\mathsf{s}_j, D)\ , \quad \text{where} \quad T(\mathsf{s}, D) = \frac{1}{|\mathsf{s}|}\sum_{i \in \mathsf{s}} \psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{1.2.9}$$

and, again, r is a random k-fold partition of $[n]$. An appropriate standard error for (1.2.9) is itself given by the cross-split statistic

$$\mathsf{se}(\mathsf{r}, D) = \frac{1}{n}\sqrt{\sum_{j=1}^{k}\sum_{i \in \mathsf{s}_j} \big(\psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}_j})) - a(\mathsf{r}, D)\big)^2}\ . \tag{1.2.10}$$

Chernozhukov et al. (2018) show that an asymptotically efficient test of the null hypothesis that an average effect treatment effect is less than zero can be constructed by comparing (1.2.9) to the critical value $\mathsf{cv}_\alpha(\mathsf{r}, D) = z_{1-\alpha} \cdot \mathsf{se}(\mathsf{r}, D)$, where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution.[5] DML estimators have been increasingly used in applied economics, as they place weaker restrictions on treatment effect heterogeneity than, say, methods based on linear regression.[6]

The DML estimator (1.2.9) and standard error (1.2.10) are, again, random through both the data D and the k-fold partition r. To evaluate the extent of resultant variability, we consider data from Chakravorty et al. (2024a,b). Chakravorty et al. (2024a) use DML to study the effect of a program, implemented in two Indian states, involving the provision of information concerning prospective jobs to vocational trainees, on employment outcomes. Here, the binary outcome Y_i indicates employment five months after completing training for each of 890 trainees placed into jobs, W_i denotes assignment to the program, and X_i collects measurements of 77 pre-treatment covariates.

Panel A of Figure 1.2 displays a heat map of the joint distribution of the 5-fold cross-fit estimator (1.2.9) and the associated critical value $\mathsf{cv}_\alpha(\mathsf{r}, D)$ over random draws of the partition r.[7] A black line is placed at the threshold where the estimate is equal to the critical value. The residual variability in the estimate is large relative to estimates of the sampling variability, and is sufficient to determine the purported statistical significance. Concretely, the difference between the 0.05th and 0.95th quantiles of the distribution of the estimator (0.094 and 0.131, respectively) is equal to 68% of the median of the distribution of the standard error (0.054).[8]

The same behavior is exhibited in related applications that use cross-fitting in more complicated ways. In these cases, measures to stabilize results are more common, although there is little guidance on how to best operationalize this stabilization. To illustrate this, we

[5]This result does not apply to estimates of the form (1.2.8), but rather to estimators analogous to (1.2.9) constructed with the "Neyman Orthogonal" or "Doubly Robust" moment

$$\psi(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) = (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) + \left(\frac{W_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\rho}(X_i)} - \frac{(1 - W_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\rho}(X_i)}\right) ,$$

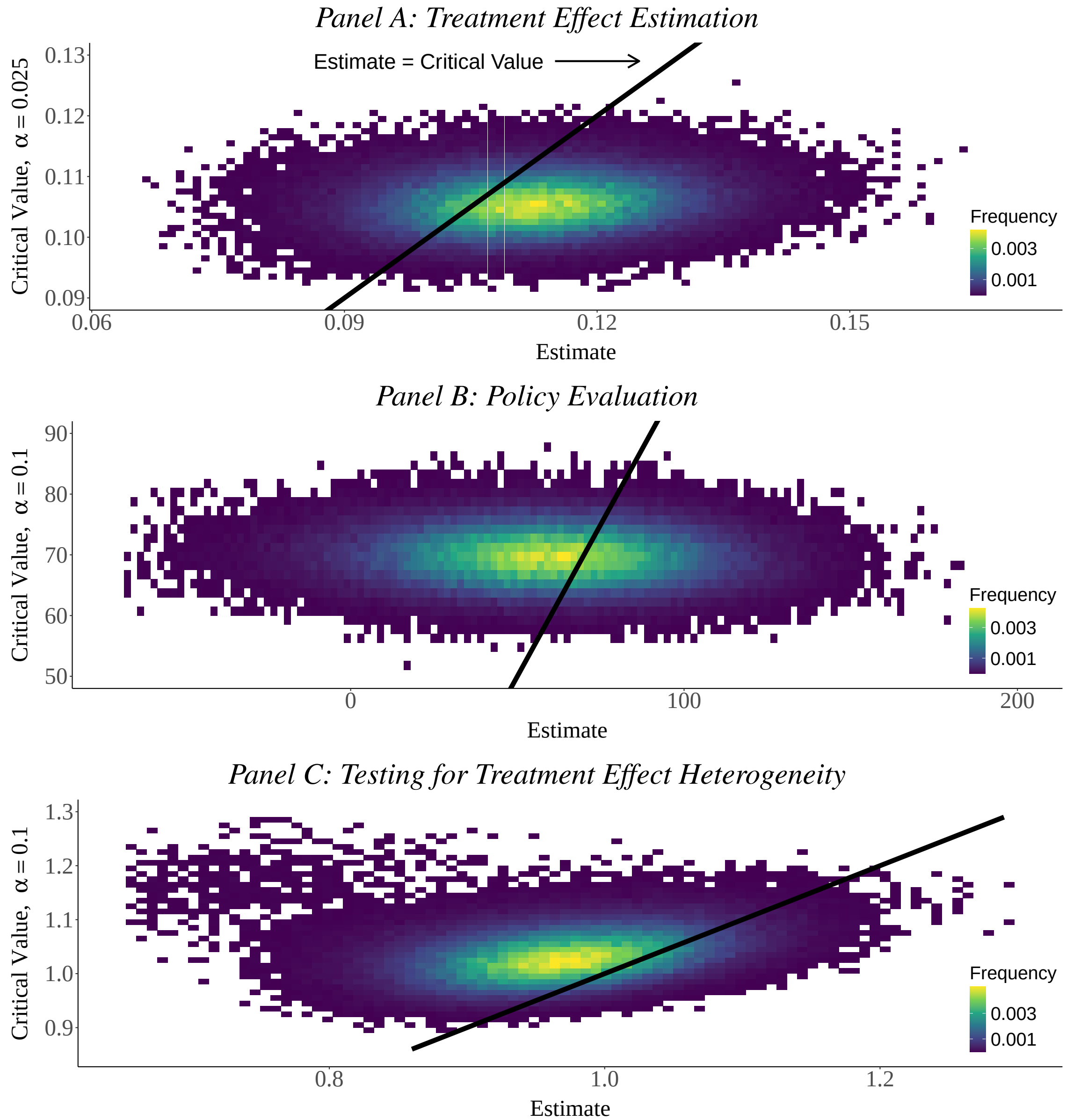
where the nuisance parameter now includes a estimate of the propensity score $\rho(x) = P\{W_i = 1 \mid X_i = x\}$. An estimator of this form is used by Chakravorty et al. (2024a), but not by Haushofer et al. (2022) or Beaman et al. (2023b).

[6]See, for example, applications in Okunogbe and Pouliquen (2022), Beraja et al. (2023), Covert and Sweeney (2023), Delfino (2024), Farronato et al. (2024).

[7]We follow the replication package associated with Chakravorty et al. (2024a). Nuisance parameters are estimated with random forests using the "Ranger" R package (Wright and Ziegler, 2017). Estimates and standard errors are constructed using the "DoubleML" R package (Bach et al., 2024). See Appendix A.1.2 for further details.

[8]Some papers take measures to address analogous residual randomness (see e.g., Covert and Sweeney (2023)). In fact, Chernozhukov et al. (2018) suggest taking the median over several partitions r. Most implementations of DML in standard statistical software allow the user to aggregate estimates, although this aggregation does not occur by default (e.g., Bach et al., 2024 in R or Ahrens et al., 2024 in STATA).

Figure 1.2: Cross-Fitting



Notes: Figure 1.2 displays discretized heat maps quantifying the residual randomness of three estimators constructed with cross-fitting. All three panels give the joint distribution of an estimator and an associated critical value. In each case, a black line has been placed at the threshold where the estimator is equal to the critical value. Panel A displays the distribution of the DML estimate (1.2.9) cross 5-fold cross-splits using data from Chakravorty et al. (2024a,b). Panel B displays the difference between treatment effects estimates for the "most impacted" and "most deprived" individuals, following Haushofer et al. (2022), using data from Egger et al. (2022a,b), again across 5-fold cross-splits. Panel C displays the distribution of the estimate associated with a test of treatment effect heterogeneity using data from Beaman et al. (2023a,b). Here, each estimate and critical value is computed by averaging over 250 independently drawn half-samples s. See Appendix A.1 for further details on the construction of each panel.

consider examples from Haushofer et al. (2022) and Beaman et al. (2023b).[9] Haushofer et al. (2022) use data from a randomized cash transfer implemented in Kenya. These data were originally considered in Egger et al. (2022a,b). Beaman et al. (2023a,b) use data from an experiment concerning agricultural lending in Mali. Both papers use sample-splitting to construct estimates, of the form (1.2.8), of the treatment effects $Y_i(1) - Y_i(0)$ for each of the units i in s. Haushofer et al. (2022) additionally construct sample-split estimates of the untreated outcome $Y_i(0)$ for each of the units in s.

Haushofer et al. (2022) identify the 50% of units in s that have the largest predicted treatment effect as well as the 50% of units that have the smallest predicted untreated outcome. They refer to these groups as the "most impacted" and "most deprived," respectively. They estimate the difference in the average treatment effect for the two groups, and construct a standard error, and associated critical value, for this difference with the bootstrap. See Appendix A.1.3 for details. Panel B of Figure 1.2 displays a heat map of the joint distribution of the 5-fold cross-fit estimate of the difference between the treatment effect estimates for the two groups, and the associated critical value, over random draws of the partition r.[10] The residual randomness is considerable. Here, the difference between the 0.05th and 0.95th quantiles of the distribution of the estimator (37.82 and 109.69, respectively) is equal to 190% of the median of the distribution of the standard error (54.44). To stabilize these results, Haushofer et al. (2022) average over 400 draws of the 5-fold partition r. Due to this increased computation, they only report confidence intervals for their main results. The methods developed in this paper ensure that stabilization of sample-split statistics, in this way, occurs at a minimal computational expense.

Likewise, Beaman et al. (2023b) implement a test of treatment effect heterogeneity proposed by Chernozhukov et al. (2023). In particular, in data for units in s, the outcome is regressed on the treatment, the treatment effect estimate, and an interaction between the treatment and the treatment effect estimate. The idea is that, if treatment effect estimates

[9] Additional, related, methods that uses cross-fitting to estimate and evaluate treatment targeting rules are proposed by Athey and Wager (2021) and Yadlowsky et al. (2024).

[10] We were not able to access a replication package associated with Haushofer et al. (2022), and so implement a simplified version of the exercise considered in that paper using data from the replication package associated with Egger et al. (2022b). Treatment effects, and untreated outcomes, are estimated with random forests using the "GRF" R package (Athey et al., 2019). Analogous estimates reported in Haushofer et al. (2022) are statistically significant. We emphasize that, as we implement only a simplified version of the exercise conducted in Haushofer et al. (2022), our estimates should only be interpreted an an illustration of the scope of residual randomness in analyses of this type, rather than as a substantive characterization of underlying treatment effect heterogeneity. In particular, we make no attempt to reweigh observations according to their sampling probabilities and appear to be using a different measure of the time between the administration of the experiment and the measurement of post-treatment outcomes.

are well-calibrated, then the coefficient on the interaction should be statistically greater than zero. Chernozhukov et al. (2023) recommend aggregating p-values associated with the coefficients on the interaction over 250 replications of this sample-split test. Panel C of Figure 1.2 displays a heat map measuring the joint distribution of the average coefficient and critical value associated with this test.[11] That is, each coefficient and critical value is computed by taking the average over 250 sample-splits. Despite this aggregation, the residual randomness remains meaningful.[12] The difference between the 0.05th and 0.95th quantiles of the distribution of the estimator (0.86 and 1.08, respectively) is equal to 28% of the median of the distribution of the standard error (0.80). Perhaps motivated by this instability, Beaman et al. (2023b) aggregate over 1000 sample-splits.

Two themes emerge from these examples. First, sample-splitting is a versatile tool for simplifying various tasks associated with modern statistical inference. Second, the residual randomness induced by sample splitting is often large and can affect the substantive interpretation of results. In the next section, we provide a general-purpose method for aggregating sample-split statistics that ensures that residual randomness is controlled at a minimal computational cost.

1.3. Reproducible Aggregation

We propose a sequential method for aggregating sample-split statistics. Our objective is to ensure that the auxiliary randomness induced by sample-splitting is small. To introduce the method, we require some additional notation. The set $\mathcal{S}_{n,b}$ consists of all subsets of $[n] = \{1, \ldots, n\}$ of size b. In turn, the set $\mathcal{R}_{n,k,b}$ contains all collections of k mutually exclusive elements of $\mathcal{S}_{n,b}$. That is, if $n = k \cdot b$, then the set $\mathcal{R}_{n,k,b}$ collects all partitions of $[n]$ into k mutually exclusive sets of size b. We refer to the elements of the set $\mathcal{R}_{n,k,b}$ as cross-splits. Throughout, the quantity $\mathsf{R}_{g,k} = (\mathsf{r}_i)_{i=1}^{g}$ collects g cross-splits, each given by $\mathsf{r}_i = (\mathsf{s}_{i,j})_{j=1}^{k}$. Unless otherwise specified, the elements of $\mathsf{R}_{g,k}$ are random, sampled independently and uniformly from the collection $\mathcal{R}_{n,k,b}$.

Suppose that we are interested in some sample-split statistic $T(\mathsf{s}, D)$. We study the

[11] We follow the details of the replication package associated with Beaman et al. (2023b). Nuisance parameters are estimated with random forests using the "GRF" R package (Athey et al., 2019). Further details are given in Appendix A.1.4.

[12] Chernozhukov et al. (2023) recommend aggregating p-values the median to ensure robustness to outliers. In practice, if aggregation of a sample-split statistic with a median, rather than a mean, makes a material difference, we strongly encourage researcher to investigate why some sample-splits generate extreme (or highly skewed) estimates (e.g., outliers in the underlying data or poor overlap in an intervention).

construction of aggregate statistics of the form

$$a(\mathsf{R}_{g,k}, D) = \frac{1}{g}\sum_{i=1}^{g} a(\mathsf{r}_{\mathsf{i}}, D)\ , \quad \text{where} \quad a(\mathsf{r}_{\mathsf{i}}, D) = \mathcal{A}(\{T(\mathsf{s}_{i,j}, D)\}_{j=1}^{k}) \tag{1.3.1}$$

and the function $\mathcal{A}(\cdot)$ aggregates the statistics $T(\mathsf{s}_{i,j}, D)$ across the cross-split r_i. To simplify exposition, for the time being, we will restrict attention to the case that the statistic $a(\mathsf{R}_{g,k}, D)$ is real-valued. This is natural, if, for example, the statistic $a(\mathsf{r}, D)$ is a treatment effect estimate or p-value constructed with cross-fitting. Later, in our application to cross-validated risk estimation, we treat vector-valued sample-split statistics, e.g., cross-validated risk estimates queried at a vector of values of a penalization parameter.

Our task is to formulate a method for choosing the number of cross-splits g to ensure that the residual variability of the aggregate statistic (1.3.1) is small. We formalize this objective as follows.

Definition 1.3.1 (Reproducible Aggregation)**.** Let the sequences $\{\mathsf{r}_i\}_{i=1}^{\infty}$ and $\{\mathsf{r}'_i\}_{i=1}^{\infty}$ be drawn independently and uniformly, conditional on the data D, from the collection of cross-splits $\mathcal{R}_{n,k,b}$. Define $\mathsf{R}_{g,k} = \{\mathsf{r}_g\}_{i=1}^{g}$ and $\mathsf{R}'_{g,k} = \{\mathsf{r}'_g\}_{i=1}^{g}$ for each integer g. Suppose that the integers $\hat{g}$ and $\hat{g}'$ are independent and identically distributed, again conditional on the data D. We say that the aggregate statistic $a(\mathsf{R}_{\hat{g},k}, D)$ is (ξ, β)-reproducible if

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} \leq \beta \tag{1.3.2}$$

almost surely.

Definition 1.3.1 is motivated by the following thought experiment. Suppose that the data D are given to two researchers. Each researcher is tasked with producing an estimate of the form (1.3.1). They generate the collections of splits, $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$, independently using the same, potentially data-dependent, procedure. That is, the two collections of splits are independent and identically distributed, conditional on the data D. If the estimate $a(\mathsf{R}_{\hat{g},k}, D)$ is (ξ, β)-reproducible, then the probability that the two researchers' estimates differ by more than ξ is less than β.

Of course, ensuring that estimates are reproducible, in the sense of Definition 1.3.1, does not preclude deceptive behavior. Researchers might compute a reproducibility aggregated statistic many times, until a desirable estimate is obtained. However, reporting reproducible statistics greatly increases the cost of searches of this form, as the scope of residual randomness has been reduced.

Algorithm 1: Anscombe-Chow-Robbins Aggregation

Input: Data D, tolerance ξ, error rate β, collection size k, split size b, initialization g_{init}

Set $g \leftarrow g_{\mathsf{init}}$

Draw r_1, ..., $\mathsf{r}_{g_{\mathsf{init}}}$ independently and uniformly from $\mathcal{R}_{n,k,b}$. Collect $\mathsf{R}_{g_{\mathsf{init}},k} = (\mathsf{r}_j)_{j=1}^{g_{\mathsf{init}}}$.

while $\hat{v}(\mathsf{R}_{g,k}, D) > \mathsf{cv}(\xi, \beta)$ **do**

  Set $g \leftarrow g + 1$

  Draw r_g uniformly from $\mathcal{R}_{n,k,b}$. Collect $\mathsf{R}_{g,k} = (\mathsf{R}_{g-1,k}, \mathsf{r}_g)$.

end

Set $\hat{g} \leftarrow g$

return $a(\mathsf{R}_{\hat{g},k}, D)$

Notes: Algorithm 1 gives a method for sequentially aggregating sample-split statistics. The critical value $\mathsf{cv}(\xi, \beta)$ is defined in display (1.3.5).

We propose a sequential method for constructing reproducible sample-split statistics. The proposal is based on the fixed-length sequential confidence intervals of Anscombe (1952) and Chow and Robbins (1965). The procedure works by repeatedly drawing a cross-split r_g uniformly from $\mathcal{R}_{n,k,b}$, appending the cross-split to the collection $\mathsf{R}_{g,k} = (\mathsf{R}_{g-1,k}, \mathsf{r}_g)$, and estimating the conditional variance

$$v_{g,k}\,(D) = \mathrm{Var}(a(\mathsf{R}_{g,k}, D) \mid D) \tag{1.3.3}$$

with the plug-in estimator

$$\hat{v}\,(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{g-1}\sum_{i=1}^{g}(a(\mathsf{r}_i, D) - a(\mathsf{R}_{g,k}, D))^2 \tag{1.3.4}$$

until the condition

$$\hat{v}\,(\mathsf{R}_{g,k}, D) \leq \mathsf{cv}(\xi, \beta) = \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \tag{1.3.5}$$

is satisfied, where z_α denotes the α quantile of the standard normal distribution. In particular, let $g_{\mathsf{init}} \geq 2$ denote some “burn-in” period chosen by the user. The number of cross-splits $\hat{g}$ chosen by the procedure is the smallest value of g, greater than g_{init}, such that (1.3.5) is satisfied. The procedure is summarized in Algorithm 1.

In Section 1.3.1, we show that Algorithm 1 is applicable, off-the-shelf, to a large class of problems. In particular, without imposing any conditions on the data generating process or the statistic of interest, we show that sample-split statistics aggregated with Algorithm 1 are reproducible, in a particular asymptotic sense. Moreover, we show that, in many applications, statistics aggregated with Algorithm 1 maintain, or improve on, various unconditional statistical guarantees. We then revisit, in Section 1.3.2, the examples considered in Section 1.2. We show that, in each case, an application of Algorithm 1 produces a reproducible, stabilized, estimate.

1.3.1 Asymptotic Validity

The following theorem establishes that statistics aggregated with Algorithm 1 are reproducible, in a particular asymptotic sense. The proof is closely related to the arguments of Anscombe (1952) and Chow and Robbins (1965) and is given in Appendix A.3. See e.g., Theorem 3.1 of Gut (2009) for a textbook treatment.

Theorem 1.3.1. *Suppose that the conditional variance* $\mathrm{Var}(a(\mathsf{r}, D) \mid D)$ *is strictly positive, almost surely, where* r *denotes a random collection drawn uniformly from* $\mathcal{R}_{n,k,b}$. *If the collections* $\mathsf{R}_{\hat{g},k}$ *and* $\mathsf{R}'_{\hat{g}',k}$ *are independently obtained using Algorithm 1, then*

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} \overset{a.s}{\to} \beta \quad \textit{as} \quad \xi \to 0\,. \tag{1.3.6}$$

Two aspects of Theorem 1.3.1 are worthy of emphasis. First, no restrictions on the data generating process or the statistic of interest are imposed. For instance, the data D do not necessarily need to be i.i.d. and the statistic $a(\mathsf{r}, D)$ can be random though quantities other than just r and D. The result, thus, provides a strong assurance that Algorithm 1 can be applied widely.

Second, in the statement (1.3.6), asymptotics are taken as $\xi \to 0$, with all other quantities, including the sample size n, fixed. This asymptotic framework is somewhat opaque, or at least nonstandard, as the parameter ξ is a choice variable. The operational interpretation is that, if ξ is chosen to be sufficiently small—at a point negligible relative to, say, the conditional variance $v_{1,k}(D)$—then the nominal reproducibility error β is accurate. In Section 1.4, by imposing some simplifying restrictions, we give a set of non-asymptotic results that make the dependence of the performance of Algorithm 1 on the choices of ξ, β, and the statistic $a(\mathsf{r}, D)$ more transparent.

The intuition underlying Theorem 1.3.1 is straightforward. The result follows directly from the observation that the cross-splits r_i are independent and identically distributed conditional on the data D. Thus, for large values of g, the variance estimator $\hat{v}\left(\mathsf{R}_{g,k}, D\right)$ will be close to the conditional variance $v_{g,k}(D)$. As a consequence, if ξ is sufficiently small, then the number of cross-splits $\hat{g}$ chosen by the procedure will be close to the "oracle" stopping time

$$g^\star = \underset{g \geq g_{\mathsf{init}}}{\arg\min} \left\{ \mathsf{cv}(\xi, \beta) = \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \geq v_{g,k}(D) \right\} \tag{1.3.7}$$

$$= \underset{g \geq g_{\mathsf{init}}}{\arg\min} \left\{ \xi \geq z_{1-\beta/2}\sqrt{\frac{2v_{1,k}(D)}{g}} \right\} , \tag{1.3.8}$$

where we have used the fact that $v_{g,k}(D) = g^{-1}v_{1,k}(D)$ in writing (1.3.8). This, then, ensures that the aggregate statistic is approximately (ξ, β)-reproducible, as

$$P\left\{\left|a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)\right| \geq \xi \mid D\right\}$$
$$= P\left\{\left|\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(\frac{1}{g^\star}\sum_{i=1}^{g^\star} a(\mathsf{r}_i, D) - a(\mathsf{r}'_i, D)\right)\right| \geq z_{1-\beta/2} \mid D\right\} \overset{\text{a.s}}{\to} \beta$$

as $\xi \to 0$, where the limit follows from the central limit theorem (as $g^\star \to \infty$ as $\xi \to 0$).[13]

In most cases, sample-split statistics are used because they have some desirable property. For example, a sample-split estimator $a(\mathsf{r}, D)$ might be asymptotically normal or might perform well in terms of some loss function. The following result—roughly, a sequential version of Jensen's inequality—can be applied to show that, in many situations, statistics aggregated with Algorithm 1 inherit these properties. Chernozhukov et al. (2023) give a related risk contraction bound for a fixed number of sample splits. The proof is given in Appendix A.3 and follows from an application of the martingale stopping theorem.

Theorem 1.3.2. *If the collection $\mathsf{R}_{\hat{g},k}$ is obtained with Algorithm 1, the cross-split r is a random element of the set $\mathcal{R}_{n,k,b}$, and $f(\cdot)$ is any convex function, then*

$$\mathbb{E}\left[f(a(\mathsf{R}_{\hat{g},k}, D))\right] \leq \mathbb{E}\left[f(a(\mathsf{r}, D))\right] . \tag{1.3.9}$$

Suppose that the estimator $a(\mathsf{r}, D)$ has good performance in terms of some convex loss, e.g., squared or absolute error. Theorem 1.3.2 demonstrates that the aggregate estimator $a(\mathsf{R}_{\hat{g},k}, D)$ will also perform well (and, in fact, may perform better). For example, Chetverikov et al. (2021) demonstrate that the cross-validated lasso has nearly optimal rates of convergence in sparse regression problems. Theorem 1.3.2 shows that all higher-order moments of risk

[13]The discrepancy between $a(\mathsf{R}_{\hat{g},k}, D)$ and $a(\mathsf{R}_{g^\star,k}, D)$ can be shown to be negligible by applying an appropriate maximal inequality—an idea due to Rényi (1957).

estimates obtained with our procedure are smaller than those of risk estimates obtained from a single cross-split. Thus, we expect the same result to hold for tuning parameters chosen with risk estimates aggregated with Algorithm 1.

Alternatively, in many cases, sample-split statistics are shown to be asymptotically normal by demonstrating that the variance of the discrepancy

$$\sqrt{n}\left(a(\mathsf{r}, D) - \frac{1}{n}\sum_{i=1}^{b}\psi(D_i, \eta)\right) \tag{1.3.10}$$

is small, where $\psi(D_i, \eta)$ is some score, or influence, function and η is some unknown nuisance parameter.[14] This is how Chernozhukov et al. (2018) demonstrate that DML estimates of average treatment effects are asymptotically normal, for example. Theorem 1.3.2 demonstrates that the same argument will apply to statistics aggregated with Algorithm 1.

In some cases, substantial residual randomness in a sample-split statistic might cause concern for the validity of associated inferences. For example, for DML estimators, the variance of the term (1.3.10) is bounded from below by the expectation of its variance conditional on the data. So, if the conditional variance is large, then the unconditional variance of (1.3.10) is likely not small. On the other hand, the conditional variance can be reduced through the application of our procedure. In other cases, substantial residual randomness may not be a reason for concern, because inferences are based on estimating a nuisance parameter in one split of a data set, and then conducting inference in the other split by conditioning on this estimate. This is the case for the applications considered in Beaman et al. (2023b) and Haushofer et al. (2022).

There is a large statistical literature on the use and interpretation of average p-values. In general, a level α test can be constructed by comparing an average p-value to $2 \cdot \alpha$ (see e.g. Rüger, 1978; Vovk and Wang, 2020; DiCiccio et al., 2020).[15] Using a result analogous to Theorem 1.3.2, for a pre-determined number of sample-splits g, Chernozhukov et al. (2023) show that, under some high-level conditions, valid p-values aggregated by averaging over sample-splits are also valid p-values. Under the same conditions, Theorem 1.3.2 can be applied to give an analogous result for p-values aggregated with Algorithm 1.

[14] That is, if second term in (1.3.10) is asymptotically normal and the variance of (1.3.10) converges to zero, then the statistic $a(\mathsf{r}, D)$ is also asymptotically normal by Chebychev's inequality and the continuous mapping theorem.

[15] A variety of methodological papers use this observation to construct sample-split hypothesis tests that control the Type I error rate under very general conditions. In Appendix A.2.1, we show that the sample-split hypothesis tests considered in e.g., DiCiccio et al. (2020), Meinshausen et al. (2009), and Wasserman et al. (2020) continue to be valid when they are constructed sequentially with Algorithm 1.

1.3.2 Reproducible Aggregation in Practice

Equipped with Algorithm 1, we return to the examples considered in Section 1.2. We show that, in each case, an appropriate application of the procedure produces a stabilized, reproducible estimate.

1.3.2.1 Cross-Fitting

We begin by treating the three examples that use cross-fitting. The most important choice to make when implementing Algorithm 1 is the specification of the statistic $a(\mathsf{r}, D)$. There are many reasonable choices that one might make here. For example, one could choose to ensure that estimates or standard errors are stable up to a desired level of precision. For the sake of comparability across settings, we opt to stabilize the p-value

$$a(\mathsf{r_i}, D) = 1 - \Phi\left(\frac{\mathsf{est}(\mathsf{r}_i, D)}{\mathsf{se}(\mathsf{r}_i, D)}\right) , \tag{1.3.11}$$

where the function $\Phi(\cdot)$ is the standard normal c.d.f. and the quantities $\mathsf{est}(\mathsf{r}_i, D)$ and $\mathsf{se}(\mathsf{r}_i, D)$ denote the cross-split estimate and standard error that corresponding to the axes of Figure 1.2. In effect, controlling the residual randomness of the average p-value ensures that the residual randomness in the aggregate estimate is small relative to the sampling error.

Table 1.1 summarizes the results. We choose ξ equal to either 0.001 and 0.01, as these are the levels of precision relevant for the determination of statistical significance at levels $\alpha = 0.025$ and $\alpha = 0.10$, respectively. For now, we follow the approaches to cross-splitting taken by all three papers. In the applications to Chakravorty et al. (2024a) and Haushofer et al. (2022), we aggregate over 5-fold cross splits. In the application to Beaman et al. (2023b), the split r_i contains a single half-sample, i.e., subset of $[n]$ of size $n/2$.[16] The estimates displayed in Table 1.1 indicate that all three applications require aggregation over hundreds of cross-splits to ensure reproducibility, in the sense of Definition 1.3.1, at these levels of error tolerance. Moreover, we find that nominal error rate β associated with Algorithm 1, here set to 0.05, is very accurate.

[16]If it is desirable to aggregate the median p-value, Algorithm 1 will continue to apply with a small modification. In particular, the variance estimator (1.3.4) can be replaced by an estimator constructed with the bootstrap. In this case, a result analogous to Theorem 1.3.1 will continue to hold.

Table 1.1: Reproducible Aggregation in Practice: Cross-Fitting

Application	ξ	p-value	Average $\hat{g}$	Reproducibility
Chakravorty et al. (2024a)	0.001	0.980	868.4	0.946
Haushofer et al. (2022)	0.01	0.80	1360.6	0.947
Beaman et al. (2023b)	0.01	0.82	2794.8	0.950

Notes: Table 1.1 summarizes the application of Algorithm 1 to three of the examples considered in Section 1.2. The first column indicates the application. The second column specifies choices for the error tolerance ξ. In each case, we set the reproducibility error rate to $\beta = 0.05$ and the burn-in sample size g_{init} to 10. The third column gives the p-value produced by one-application of the procedure. The fourth column gives the average number of cross-splits $\hat{g}$ drawn in each implementation, taken across 2,000 replications of the aggregation procedure. The fifth column displays an estimate of the true reproducibility probability, i.e., the probability that two independent implementations of Algorithm 1 produce estimates that differ by less than ξ, computed with these replicates.

1.3.2.2 Cross-Validation

We now turn to our application to cross-validated risk estimation. In this setting, there is more ambiguity in how to best apply Algorithm 1. We find that the following approach works well in the data from Casey et al. (2021b). Recall that, in this application, the sample-split statistic $a_\lambda(\mathsf{r}, D)$ denotes the cross-validated estimate of the mean-squared error, queried at a specified value of the regularization parameter λ. We are interested in stabilizing these estimates for each value of λ in an increasing sequence $(\lambda_\ell)_{\ell=1}^p$, where λ_p is the smallest value such that no covariates have non-zero coefficients when estimated using the full data.[17]

To do this, we apply Algorithm 1 to each component of the vector

$$(b_{\lambda_1}(\mathsf{r}, D), \ldots, b_{\lambda_{p-1}}(\mathsf{r}, D)) , \quad \text{where} \quad b_{\lambda_i}(\mathsf{r}, D) = (a_{\lambda_i}(\mathsf{r}, D) - a_{\lambda_p}(\mathsf{r}, D)) . \qquad (1.3.12)$$

The rationale for this is that the relative, rather than absolute, values of the risk estimates are what are relevant for determining where the estimates takes their minimum value. We find that, in practice, consideration of the differences (1.3.12) can substantially reduce the amount of aggregation needed for stabilization.

We implement Algorithm 1, independently, for each component of the vector (1.3.12). In

[17] Throughout, we use the grid $(\lambda_\ell)_{\ell=1}^p$ of values of the regularization parameter λ chosen by default by the "glmnet" R package (Friedman et al., 2021).

other words, we ensure that each component of the risk estimate is marginally reproducible.[18] We let ξ_i denote the error tolerance used for the ith component of the vector (1.3.12). We use a simple approach for determining suitable values for these tolerances. Roughly speaking, we compute the statistic (1.3.12) for each element of a small, initial sample of cross-splits (e.g., $g = 20$). Using these estimates, we find that the level of precision needed to distinguish between the close-to-optimal values of λ is approximately $\xi_i = 10^{-4}$. The error tolerances specified at values of λ that can be determined to be sub-optimal are set to larger values. Further details are given in Appendix A.2.2.

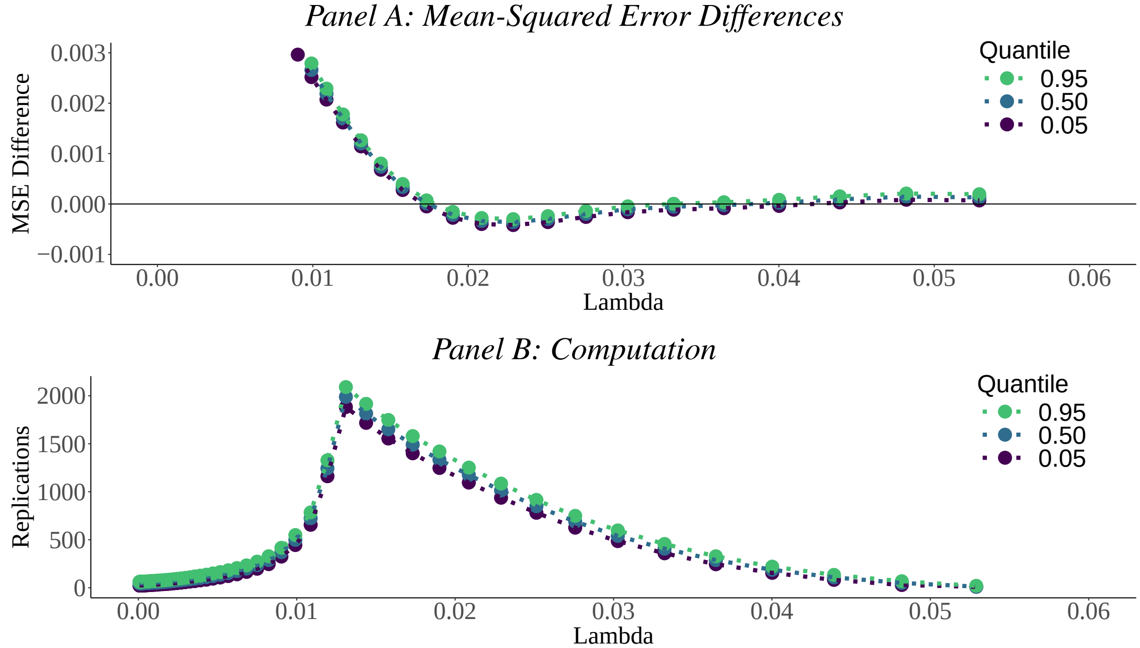
Figure 1.3 gives measurements of the performance of Algorithm 1 for aggregation of the statistic (1.3.12), implemented in the data from Casey et al. (2021b). Panel A displays quantiles of the reproducibly aggregated statistic (1.3.12) across replications of the procedure. The y-axis has been truncated to focus attention on the close-to-optimal values of the regularization parameter. There is little residual randomness. Panel B displays quantiles of the number of 10-fold cross-splits $\hat{g}$ chosen by the procedure at each value of the regularization parameter. Precise estimates at the close-to-optimal values require aggregation over thousands of cross-splits. Observe that less aggregation is used at small, sub-optimal values of λ, as we have used larger error tolerances for these values. We find that, if cross-validated model selection is implemented twice, and, in each case, aggregated at this level of error tolerance, then the same value of λ is selected with probability 0.84 and the same 14 covariates are selected with probability 0.999. In Appendix A.2.2, we show that the nominal reproducability error, again $\beta = 0.05$, is very accurate over the full range of the regularization parameter.

1.4. Theoretical Analysis

The asymptotic results given in Section 1.3 are quite general. In particular, Theorem 1.3.1 holds in the absence of any restrictions on the data generating process or statistic under consideration. It is worth asking, however, whether these results confer a clear statistical understanding. At least two issues arise. First, Theorem 1.3.1 relies entirely on the fact that, in Algorithm 1, successive cross-splits are sampled independently. That is, we have said

[18] In principle, it is straightforward to ensure that the components of the vector (1.3.12) are simultaneously reproducible. For example, simultaneous reproducibility at level β can be obtained by ensuring that each component is reproducible at level $\beta/(p-1)$, mimicking the standard Bonferonni adjustment. More sophisticated schemes, based on the bootstrap, say, are also applicable. In practice, ensuring simultaneous reproducibility can substantially increase the required computation and tends not to lead to materially different risk estimates.

Figure 1.3: Reproducible Aggregation in Practice: Cross-Validation



Notes: Figure 1.3 displays the performance of Algorithm 1 in the application to cross-validated Lasso, implemented in data from Casey et al. (2021b). Algorithm 1 is applied independently to each component of the vector (1.3.12). The error tolerances ξ_i are specified in Appendix A.2.2. We set the nominal reproducibility error to $\beta = 0.05$. Panel A displays quantiles of the statistic (1.3.12) at each value of a grid of values of the regularization parameter λ. The y-axes is truncated to focus attention on large values of λ. We give an un-truncated version in Appendix A.2.2. Panel B displays quantiles of the number of 10-fold cross-splits $\hat{g}$ chosen by the procedure at each value of the regularization parameter.

nothing, yet, about the role of cross-splitting. Second, the interpretation of the asymptotic approximation with $\xi \to 0$ is somewhat opaque. What would be a reasonable value of ξ to choose to ensure that Algorithm 1 is accurate? And how do these choices impact the amount of computation required to implement the procedure?

In this section, we give a non-asymptotic description of the performance of Algorithm 1. This is accomplished by placing some simplifying restrictions on the statistic of interest. The more specialized analysis that follows is aimed at providing qualitative and quantitative intuition for how the computational cost and reproducibility error of statistics aggregated with Algorithm 1 depend on the choices of k and ξ. Proofs for results stated in this section are given in Appendix A.4.

1.4.1 Symmetry, Linearity, and Stability

We impose a set of simplifying restrictions. Recall that, in general, we are considering the aggregation of cross-split statistics of the form

$$a(\mathsf{r}, D) = \mathcal{A}(\{T(\mathsf{s}_j, D)\}_{j=1}^{k}) \ , \quad \text{where} \quad T\left(\mathsf{s}, D\right) = \Psi\left(D_{\mathsf{s}}, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)\right) \tag{1.4.1}$$

is a sample-split statistic and the function $\hat{\eta}(\cdot)$ is an estimator of an unknown nuisance parameter η. In the main text, we restrict attention to the case that $n = k \cdot b$, i.e., where each element r in $\mathcal{R}_{n,k,b}$ is a complete partition of $[n]$ into k sets of size b. Each of the results given here will follow directly from more general results stated in Appendix A.4, where this restriction is not imposed.

First, we assume that the sample-split statistic under consideration is symmetric and deterministic in each part of a split sample.

Assumption 1.4.1 (Symmetry and Determinism)**.** *For all sets s in $\mathcal{S}_{n,b}$ and data D, the statistic $T\left(\mathsf{s}, D\right)$ is deterministic and invariant to permutations of the data with indices in s and $\tilde{\mathsf{s}}$, respectively.*

The intention of Assumption 1.4.1 is to restrict the residual randomness under consideration to the randomness introduced by sample-splitting. This holds in cases where $\hat{\eta}(\cdot)$ is deterministic, e.g., when $\hat{\eta}(\cdot)$ is a coefficient vector determined by a regularized regression or in the applications to hypothesis testing considered by DiCiccio et al. (2020) or Wasserman et al. (2020). Assumption 1.4.1 rules out procedures where the estimator $\hat{\eta}(\cdot)$ is random conditional on the data. This excludes settings where, e.g., $\hat{\eta}(\cdot)$ is estimated with stochastic gradient descent, bagging or subsampling, or is itself constructed with data splitting.

Second, we assume that the aggregation function $\mathcal{A}(\cdot)$ is an average and that the statistic $T\left(\mathsf{s}, D\right)$ is linearly separable in the first part of the split sample.

Assumption 1.4.2 (Linearity)**.** *For all cross-splits $\mathsf{r} = (\mathsf{s}_j)_{j=1}^{k}$ in $\mathsf{R}_{n,k,b}$, the cross-split statistic $a(\mathsf{r}, D)$ can be represented by*

$$a(\mathsf{r}, D) = \frac{1}{k} \sum_{i \in \mathsf{s}} T\left(\mathsf{s}_j, D\right) \ . \tag{1.4.2}$$

Moreover, for all sets s in $\mathcal{S}_{n,b}$, the sample-split statistic $T\left(\mathsf{s}, D\right)$ can be represented by

$$T\left(\mathsf{s}, D\right) = \frac{1}{b} \sum_{i \in \mathsf{s}} \psi(D_i, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) \tag{1.4.3}$$

for some function $\psi(\cdot,\cdot)$.

We make these restrictions to ease exposition. Assumption B.2.3 is satisfied if, for example, the statistic under consideration is a cross-fit treatment effect or cross-validated mean-squared error estimate. In principle, Assumption B.2.3 rules out some applications of interest. In practice, so long as the linear representations (1.4.2) and (1.4.3) hold up to a suitable degree of approximation, the qualitative and quantitative predictions of the results that follow will continue to hold.[19] In particular, we show below that the predictions of our results play out in the data from Chakravorty et al. (2024a), where the statistic $T(\mathsf{s}, D)$ is a p-value of the form (1.3.11).

The remaining assumptions, and our ensuing results, are expressed in terms of two objects that measure the sensitivity of the statistic under consideration to perturbations of the data and of the splits, respectively. We refer to these objects as stabilities. They are defined as follows.

Definition 1.4.1 (Sample Stability)**.** Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$ and let i be an arbitrary element of s. Let D' denote an independent and identical copy of the data D. For each $\mathsf{q} \subseteq [n]$, let $\tilde{D}^{(\mathsf{q})}$ be constructed by replacing D_j with D'_j in D for each j in q. Let q be a randomly selected subset of $\tilde{\mathsf{s}}$ of cardinality q. We refer to the quantity

$$\sigma^{(r,q)} = \mathbb{E}\left[\left|\psi(D_i, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(\mathsf{q})}_{\tilde{\mathsf{s}}}))\right|^r\right] \tag{1.4.4}$$

as the (r,q)-order sample stability.

Definition 1.4.2 (Split Stability)**.** We refer to the quantity

$$\zeta^{(r)} = \mathbb{E}\left[\max_{\mathsf{s},\mathsf{s}'\in \mathsf{S}_{n,b}} \left(T\left(\mathsf{s}, D\right) - T\left(\mathsf{s}', D\right)\right)^r\right]. \tag{1.4.5}$$

as the rth-order split stability.

We restrict attention to statistics whose sample stabilities decay in a suitable way with the sample size n. Throughout, we say $x \lesssim y$ if there exists a universal constant C such that $x \leq Cy$.

Assumption 1.4.3 (Sample Stability Decay)**.** *The* (r,q)*-order sample stability satisfies the bound*

$$\sigma^{(r,q)} \lesssim \left(\frac{\sqrt{q}}{n-b}\right)^r \tag{1.4.6}$$

[19] Extensions of our results to cases where the function $\mathcal{A}(\cdot)$ or the statistic $T(\mathsf{s}, D)$ satisfy component-wise Lipschitz or bounded differences conditions, say, are feasible, and will exhibit the same qualitative behavior.

uniformly for each q in $[b]$ and r in $\{2,4\}$.

We call a statistic sample stable if it satisfies Assumption 1.4.3.[20] Many sample-split statistics of interest are sample stable. In Appendix A.2.3, we show that statistics satisfying Assumption B.2.3 are sample stable if the nuisance parameter estimator $\hat{\eta}(\cdot)$ is an empirical risk minimizer of a, potentially regularized, strictly convex loss. There is a large literature that gives analogous bounds for other standard machine learning estimators, including bagged or subsampled estimators, like random forests (Chen et al., 2022; Ritzwoller and Syrgkanis, 2024), ensemble estimators (Elisseeff et al., 2005), and estimators computed with stochastic gradient descent (Hardt et al., 2016).[21]

Nevertheless, sample-stability should be viewed as a strong assumption, that is only applicable to highly regular estimators. Below, we show that the predictions that follow from the imposition of sample-stability, concerning the qualitative behavior of the residual randomness, play out in the applications to Casey et al. (2021b) and Chakravorty et al. (2024a). These predications may not have the same quality in settings that use less well-behaved nuisance parameter estimators.

The split stability $\zeta^{(r)}$ is a less frequently studied object. We will only require that it is finite for r equal to 4 or 8, depending on the setting. This is a weak restriction that will hold, for example, if the statistic $T(\mathsf{s}, D)$ is bounded.

1.4.2 Computation and Concentration

Algorithm 1 entails sequentially computing the statistic $a(\mathsf{R}_{g,k}, D)$, as g increases, until a stopping criteria is satisfied. How much computation should we expect to do? And how does this quantity depend on the parameters k and ξ?

Recall from Section 1.3.1 that we should expect the total number of splits $\hat{m} = \hat{g} \cdot k$ used by Algorithm 1 to be close to the "oracle" quantity

$$m^{\star} = g^{\star} \cdot k \approx 2k \cdot v_{1,k}(D) \left(\frac{z_{1-\beta/2}}{\xi}\right)^2 , \tag{1.4.7}$$

where $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ denotes the conditional variance of the statistic computed

[20] The $(2,1)$-order sample stability $\sigma^{(2,1)}$ is a widely studied object in the statistical learning literature, where it is referred to as mean-square stability (see e.g., Bousquet and Elisseeff, 2002; Kale et al., 2011; Kumar et al., 2013).

[21] Chen et al. (2022) show that, under some regularity conditions, sample-splitting is unnecessary for the consistency and asymptotic normality of DML estimators, if a condition related to, but partially stronger than, Assumption 1.4.3 is satisfied. We comment on the relationship between Assumption 1.4.3 and the conditions considered in Chen et al. (2022) in Appendix A.2.3.

using a single cross-split. Two aspects of the expression (1.4.7) are worth highlighting. First, the total number of splits depends on the error tolerance ξ through the factor ξ^{-2}. In other words, in order to reduce the reproducibility error by a factor of 10, e.g., to move from $\xi = 0.1$ to $\xi = 0.01$, the total number of splits must be increased by a factor of 100. Second, the total number of splits depends on the parameter k through the factor $k \cdot v_{1,k}(D)$. How should we expect this quantity to scale with k?

We answer this question with the following result, which characterizes the rate of convergence of the statistic $a(\mathsf{R}_{g,k}, D)$ around its conditional mean, given by

$$\overline{a}\,(D) = \mathbb{E}\left[a(\mathsf{R}_{g,k}, D) \mid D\right] = \mathbb{E}\left[T(\mathsf{s}_{i,j}, D) \mid D\right] \tag{1.4.8}$$

under Assumption B.2.3. The nonstandard aspect of this result is that we account for the dependence in the summands in $a(\mathsf{R}_{g,k}, D)$ across cross-splits, i.e., the dependence induced by cross-fitting. This is accomplished by applying a coupling argument due to Chatterjee (2005, 2007).[22]

Theorem 1.4.1. *Suppose that Assumptions 1.4.1, 1.4.3, and B.2.3 hold, the data D are independently and identically distributed, and $n = k \cdot b$. If the 4th-order split stability $\zeta^{(4)}$ is finite, then for each $\delta > 0$, the inequality*

$$v_{g,k}(D) = \mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\,(D)\right)^2 \mid D\right] \lesssim \frac{1}{\delta}\frac{b-1}{n^2}\frac{1}{g}\,. \tag{1.4.9}$$

holds with probability greater than $1 - \delta$ as D varies.

The left-hand side of the inequality (1.4.9) is random through the data D. Theorem 1.4.1 says that, if $\mathcal{F}$ is the event that the inequality (1.4.9) holds, then $P\{\mathcal{F}\} > 1 - \delta$ unconditionally. This bound results from an application of Markov's inequality, at one point in the proof, to bound a complicated, data-dependent term with a term that depends on the sample stability.

[22] Theorem 1.4.1 is closely related to the unconditional variance bounds given in Kale et al. (2011) and Kumar et al. (2013), who give bounds with the same dependence on k for cross-validated risk estimation. Our result follows from a different method of argument. In particular, Theorem 1.4.1 is a corollary of a more general result, presented in Appendix A.4, that gives an analogous large deviations bound. In particular, we show that, for each $\varepsilon > 1$, the bound

$$P\left\{|a(\mathsf{R}_{g,k}, D) - \overline{a}(D)| \leq \sqrt{\frac{b-1}{n^2}\frac{1}{g}\frac{\log(\varepsilon^{-1})}{\delta}} \mid D\right\} \geq 1 - \varepsilon$$

holds with probability greater than $1 - \delta$ as D varies. That is, the rate of convergence suggested by Theorem 1.4.1 holds for all higher-order moments as well. This large deviations bound is applied repeatedly to establish the result given in the following subsection.

This strategy—bounding the conditional quantities of interest with unconditional quantities—is helpful because the resultant unconditional object, the sample stability, is tractable and has been characterized in many settings of interest.

Theorem 1.4.1 demonstrates that the variance $v_{g,k}(D)$ converges to zero at the rate

$$\frac{1}{n}\frac{b-1}{n}\frac{1}{g} \leq \frac{1}{n}\frac{1}{k}\frac{1}{g}\,. \tag{1.4.10}$$

Consequently, for a fixed total number of splits $m = g \cdot k$, the rate of convergence is proportional to $(mn)^{-1}$. That is, the conditional randomness of aggregate statistics constructed with m sample-splits concentrates like averages of m i.i.d. random variables, despite the dependence across cross-splits. Observe, also, that if $b = 1$, corresponding to "jackknife" or "leave-one-out" sample-splitting, then there is no residual randomness and the left-hand-side of (1.4.10) collapses to zero.

Plugging the bound (1.4.9) into the expression (1.4.7), we find that

$$m^{\star} \approx \frac{1}{n}\left(\frac{z_{1-\beta/2}}{\xi}\right)^2\,, \tag{1.4.11}$$

In other words, Theorem 1.4.1 implies that—in contrast to the error tolerance ξ—the number of cross-folds k does not affect the required computation.[23] On the other hand, all else equal, less aggregation is needed to remove residual randomness from settings with larger sample sizes n.

Theorem 1.4.1 plays out empirically. Figure 1.4 displays measurements of the conditional variance $v_{1,k}(D)$ in dark blue, as k varies, for our applications to Casey et al. (2021b) and Chakravorty et al. (2024a). For the application to Casey et al. (2021b), we use the cross-validated estimate of the mean-squared error at the value of λ that minimizes the curves displayed in Panel A of Figure 1.3. For the application to Chakravorty et al. (2024a), we use the p-value (1.3.11), associated with the cross-fit estimate of the average treatment effect. If the variance bound (1.4.9) is accurate, then the approximation

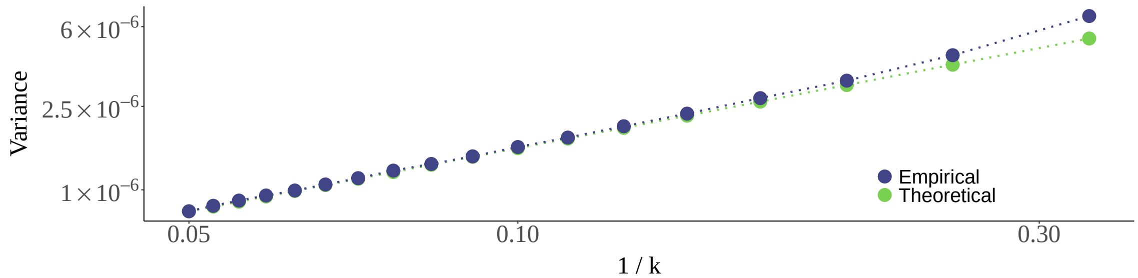
$$v_{1,k}(D) \approx \left(\frac{k'}{k}\right) v_{1,k'}(D) \tag{1.4.12}$$

should hold for each pair k, k'. To test this, we display estimates of right-hand-side of

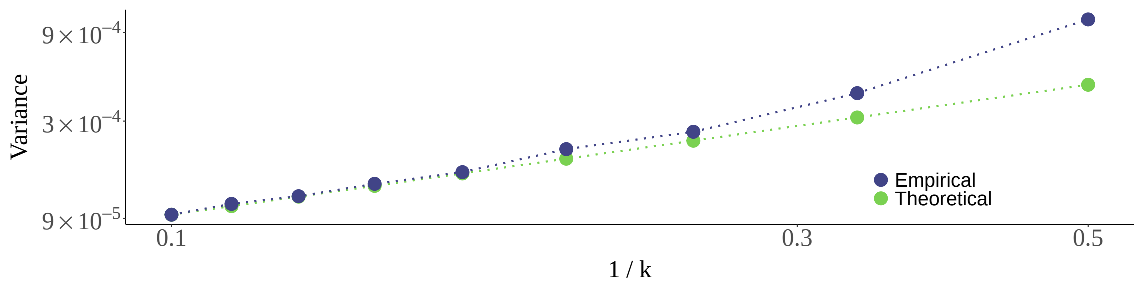
[23]Throughout our discussion, we assume that the computation involved in computing the estimator using one split is invariant to the fold size b. This should hold to a good degree of approximation for small values of k.

Figure 1.4: Concentration

Panel A: Casey et al. (2021b)



Panel B: Chakravorty et al. (2024a)



Notes: Figure 1.4 illustrates the concentration of the residual randomness of various cross-split statistics with the number of cross-splits k, using data from Casey et al. (2021b) and Chakravorty et al. (2024a). For the application to Casey et al. (2021b), we use the cross-validated estimate of the mean-squared error at the value of λ that minimizes the curves displayed in Panel A of Figure 1.3. For the application to Chakravorty et al. (2024a), we use the p-value of the form (1.3.11), associated with the cross-fit estimate of the average treatment effect. The x-axes give $1/k$. The y-axes give measurements of the conditional variance of each statistic. Both the axes are displayed on a logarithmic scale, base 10. The theoretical prediction (1.4.12), based on Theorem 1.4.1, is given in light green.

(1.4.12) in light green, where k' is the largest value of k considered in each sub-figure. In each case, the approximation is remarkably accurate.[24]

1.4.3 Reproducibility

Algorithm 1 confers a guarantee. In particular, statistics aggregated with Algorithm 1 should be interpreted as being reproducible, up to an error tolerance ξ, with probability greater than β. How accurate is this guarantee? In particular, how does the accuracy of the nominal reproducibility error β depend on the choice parameters ξ and k? These questions are answered by the following Berry-Esseen type bound on the accuracy of the nominal

[24] It is worth emphasizing that neither Assumption 1.4.1 nor Assumption B.2.3 holds for the application to Chakravorty et al. (2024a). That is, in this case, nuisance parameters are estimated with random forests, which are random conditional on the data, and the p-value (1.3.11) does not admit an exact representation of the form (1.4.3). Rather, in this setting, both of these assumptions are good approximations (so long as the number of trees used to construct the random forests is sufficiently large), and so the predictions of Theorem 1.4.1 hold.

reproducibility of Algorithm 1.

Theorem 1.4.2. *Suppose that the collections $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are independently obtained using Algorithm 1. If Assumptions 1.4.1, 1.4.3, and B.2.3 hold, the conditional variance $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ is strictly positive, almost surely, the data D are independent and identically distributed, and the eighth-order split stability $\zeta^{(8)}$ is finite, then for all sufficiently small ξ, the inequality*

$$\left| P\left\{ \left| a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D) \right| \geq \xi \mid D \right\} - \beta \right| \lesssim \frac{1}{\delta^{3/4}} \frac{1}{kn} \left(\frac{1}{v_{1,k}(D)} \right)^{5/4} \left(\frac{\xi}{z_{1-\beta/2}} \right)^{1/2}, \tag{1.4.13}$$

holds with probability greater than $1 - \delta$ as D varies, where in writing (1.4.13), we have omitted a multiplicative term that converges to zero logarithmically as ξ decreases to zero.

Theorem 1.4.2 describes the settings under which Algorithm 1 is accurate. To unpack this result, observe that the variance bound (1.4.9) gives

$$\frac{1}{kn} \left(\frac{1}{v_{1,k}(D)} \right)^{5/4} \left(\frac{\xi}{z_{1-\beta/2}} \right)^{1/2} \gtrsim \frac{k^{1/4}}{n} \left(\frac{\xi}{z_{1-\beta/2}} \right)^{1/2}. \tag{1.4.14}$$

This suggests that the performance of Algorithm 1 improves as k and ξ decrease and as n increases.[25] These predictions hold empirically. Panel A of Figure 1.5 displays estimates of the reproducibility error in the application to Casey et al. (2021b) as k and ξ vary. An analogous figure for the application to Chakravorty et al. (2024a) is displayed in Appendix A.1.5. As predicted, over most of the range of ξ, the reproducibility error is increasing as k increases.[26]

Likewise, Panel B displays estimates of the average number of cross-splits $\hat{g}$ used in Algorithm 1, at each value of ξ and k. Again, as predicted, the total number of splits used by the procedure stays constant as k varies. To see this, observe that the average value of $\hat{g}$ is roughly 10 times smaller for $k = 20$ than for $k = 2$, as predicted by the approximation (1.4.11). Comparing Panels A and B, observe that, once the average value of $\hat{g}$ is larger than approximately 500, the reproducibility error is close to the nominal error rate β.

[25] The dependence of the bound (1.4.13) on ξ is sharp, at least up to the logarithmic factor. This follows from general results concerning randomly stopped sums given in Landers and Rogge (1976, 1988).

[26] Note, however, that for large values of ξ, the reproducibility error decreases as k increases. This is due to early stopping. That is, if the variance $v_{1,k}(D)$ is large relative to ξ, then there is an increased chance that $\hat{g}$ stops immediately after the burn-in period, i.e., $\hat{g} = g_{\mathsf{init}}$.

If the total number of splits required to ensure reproducibility at a given error tolerance ξ is constant as k varies, then why does the performance of Algorithm 1 decrease with k? As part of the proof of Theorem 1.4.2, we show that, for each $\varepsilon > 0$, the bound

$$P\left\{\left|\frac{\hat{g}}{g^\star} - 1\right| \lesssim \frac{1}{n}\frac{1}{k}\left(\frac{1}{v_{1,k}(D)}\right)^{3/2} \frac{\xi}{z_{1-\beta/2}}\sqrt{\frac{\log(\varepsilon^{-1})}{\delta}} \mid D\right\} \geq 1 - \varepsilon \tag{1.4.15}$$

holds with probability greater than $1 - \delta$, as D varies. Again the variance bound (1.4.9) gives

$$\frac{1}{n}\frac{1}{k}\left(\frac{1}{v_{1,k}(D)}\right)^{3/2} \frac{\xi}{z_{1-\beta/2}} \gtrsim \frac{k^{1/2}}{n}\frac{\xi}{z_{1-\beta/2}}\,. \tag{1.4.16}$$

Written differently, as k increases the discrepancy between the realized and oracle number of cross-splits, $|\hat{g}/g^\star - 1|$, increases, reducing the accuracy of the nominal error rate. Roughly speaking, this happens because the quality of the estimator $\hat{v}_{g,k}(D)$ for the conditional variance $v_{g,k}(D)$ depends only on the number of cross-splits g.[27] That is, for a given number of sample-splits $\hat{m} = k \cdot \hat{g}$, Algorithm 1 is most accurate when $\hat{g}$ is large, as the variance estimate $\hat{v}_{\hat{g},k}(D)$ is more precise.

As before, these predictions play out in practice. Panel C of Figure 1.5 displays estimates of the 0.05th and 0.95th quantiles of the distribution of the discrepancy $\hat{g}/g^\star - 1$ as k and ξ vary. Over most of the range of ξ the discrepancy between $\hat{g}$ and $g^\star$ is increasing in k. At small values of k and large values of ξ, there is an increased chance that $\hat{g}$ stops immediately after the burn-in period, i.e., $\hat{g} = g_{\mathsf{init}}$.

It is worth pausing to note that these results do not support eschewing cross-splitting altogether, i.e., setting k equal to one and aggregating over independent splits, as we have restricted attention to the case that $n = k \cdot b$. In Appendix A.4, we give analogous results that relax this assumption and show that cross-splitting, i.e., setting $n = k \cdot b$, reduces residual randomness at a faster rate than independent splitting, i.e., setting k equal to one. In other words, all else equal, Algorithm 1 performs best, in the sense that the nominal reproducibility error is most accurate, when k is equal to 2 and b is equal to $b/2$.

To summarize, the computation needed to achieve a desired bound on residual randomness is highly sensitive to the error tolerance ξ, but is not affected by the number of cross-folds

[27] Similarly, in Appendix A.4, we show that the quality of a normal approximation to $a(\mathsf{R}_{g,k}, D)$ depends only on g, although this discrepancy is not the leading term in the reproducibility error. The close approximation exhibited in Figure 1.4 suggests that is may be reasonable to estimate $v_{1,k}(D)$ by taking the sample variance both across and within cross-splits, i.e., computing the sample variance across all m sample-splits. Although this worth further consideration, asymptotic validity, i.e., Theorem 1.3.1, would not hold at the same level of generality.

k. On the other hand, the accuracy of the nominal error rate of Algorithm 1 decreases as the number of cross-folds k increases. If the error bound ξ is chosen to be suitably small, such that the realized number of cross-splits $\hat{g}$ is greater than roughly 500, then the nominal reproducibility probability tends to be quite accurate, and insensitive to changes in the number of cross-splits.

1.5. Recommendations for Practice

Sample-splitting is a helpful tool for simplifying many widely encountered problems in applied econometrics. This simplification comes at the cost of the introduction of residual randomness. We have shown, in several applications, that this residual randomness is large enough to substantively affect results. To address this, we have proposed a simple procedure, summarized in Algorithm 1, for removing the auxiliary randomness from sample-split statistics. The procedure takes as input a bound and an error rate. We have shown that, if the procedure were run twice, the chance that the results differ by more than the bound is well-approximated by the error rate.

We conclude, in this section, by detailing several recommendations for how to best implement Algorithm 1 in practice. The most important choice to make is the specification of the sample-split statistic of interest. For example, suppose that we are interested in estimating an average treatment effect, or other causal contrast, using an estimator based on sample-splitting. Denote this quantity by $\mathsf{est}(\mathsf{r}, D)$ and let $\mathsf{se}(\mathsf{r}, D)$ denote an associated, potentially sample-split, standard error estimate. There are several reasonable choices that one might make. In this case, we recommend applying Algorithm 1 to sequentially aggregate the p-value

$$a(\mathsf{r}, D) = 1 - \Phi\left(\frac{\mathsf{est}(\mathsf{r}, D)}{\mathsf{se}(\mathsf{r}, D)}\right) \tag{1.5.1}$$

associated with a test that the contrast of interest is greater than zero, where $\Phi(\cdot)$ denotes the standard normal c.d.f. This approach has the benefit of benchmarking residual randomness with an estimate of the sampling error.

In some settings, it may be of interest to report the scope of residual randomness for alternative quantities associated with statistics that have aggregated with Algorithm 1. For example, suppose that we have applied Algorithm 1 to stabilize the p-value (1.5.1), and obtain $a(\mathsf{R}_{\hat{g},k}, D)$. We may also wish to report the associated estimate $\mathsf{est}(\mathsf{R}_{\hat{g},k}, D)$. In this case, we recommend reporting and interpreting the standard error

$$\sqrt{\frac{1}{\hat{g}}\frac{1}{\hat{g}-1}\sum_{i=1}^{\hat{g}}(\mathsf{est}(\mathsf{r}_i, D) - \mathsf{est}(\mathsf{R}_{\hat{g},k}, D))^2} \tag{1.5.2}$$

in the usual way. That is, the standard error (1.5.2) estimates the residual randomness of the estimate $\mathsf{est}(\mathsf{R}_{\hat{g},k}, D)$, conditional on the data. If this is too large, the error tolerance ξ should be decreased.

The optimal choice of statistic in settings where cross-validation is used for model selection and risk estimation is less immediate. In this case, we recommend applying Algorithm 1 to each element of the statistic (1.3.12), i.e., the relative values of the risk estimates.

The second most important choice to make is the choice of the error tolerance ξ. In some cases, an appropriate choice is clear. For example, it may be desirable to ensure that a p-value is reproducible at the level relevant for the determination of statistical significance at level 0.10 or 0.025, e.g., setting $\xi = 0.01$ or $\xi = 0.001$. In other cases, like in applications to cross-validated model selection, this choice is less clear. We propose a procedure for making this choice in that setting in Appendix A.2.2.

The error tolerance ξ should be set to a value that is sufficiently small so that the nominal reproducibility error is accurate. On the other hand, ξ should not be set so small that the computation associated with implementing Algorithm 1 becomes infeasible. To ensure that the chosen value of ξ satisfies these constraints, we recommend computing the statistic of interest for each of a small, initial sample of cross-splits (e.g., g equal to 20 or 30). Let $\hat{v}_{1,k}(D)$ denote an estimate of the conditional variance $v_{1,k}(D)$ computed using this sample. An estimate of the total number of cross-splits needed to implement Algorithm 1 at a specified level of ξ can then be obtained by

$$g(\xi) = 2\hat{v}_{1,k}(D)\left(\frac{z_{1-\beta/2}}{\xi}\right)^2 . \tag{1.5.3}$$

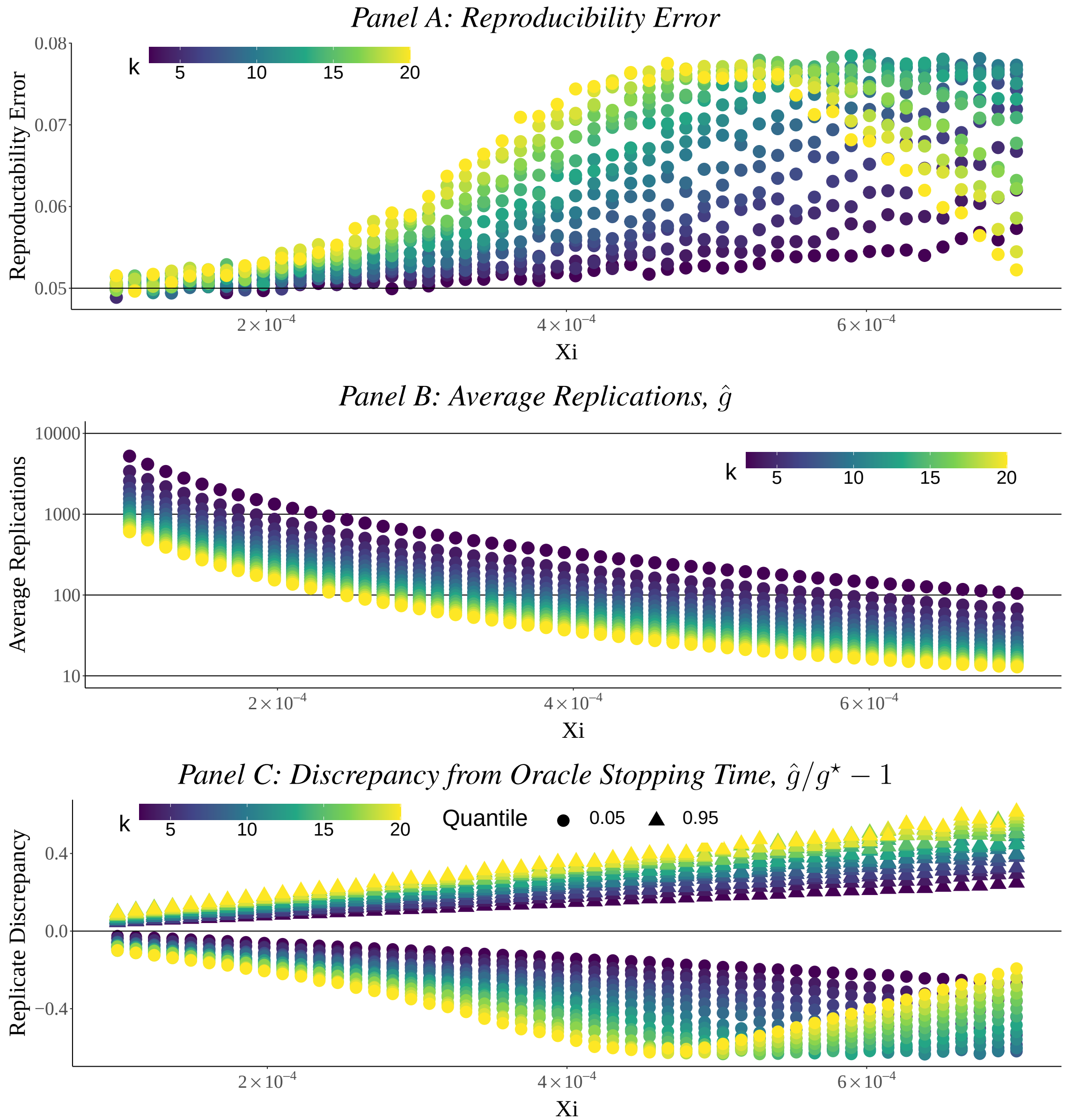
Motivated by the numerical results reported in Section 1.4, we recommend choosing a value of ξ such that this estimate is greater than 500.

In some cases, the estimate (1.5.3) may be too large at practically relevant values of ξ to be computationally feasible. Unfortunately, as we have shown in Section 1.4, changing the number of cross-folds k will not address this issue, and it may be best to consider alternative choices of nuisance parameter estimators. In this situation, one might consider using "leave-one-out" or "jackknife" sample-splitting, i.e., setting $k = n$, which exhibits no residual randomness. This is not an omnibus fix, however. For example, proofs of the asymptotic normality of DML estimates of average treatment effects require that k is small relative to n (Chernozhukov et al., 2018). Similarly, leave-one-out cross-validation is not necessarily optimal for model selection (see e.g., Shao (1993) for early discussion of this point). Chetverikov (2024) gives a review of various, alternative, approaches to tuning parameter selection.

If the number of cross-folds k is set too high, the realized number of cross-splits $\hat{g}$ may be too small and Algorithm 1 may perform poorly. But, so long as the user ensures that the error tolerance ξ is sufficiently small such that the number of cross-splits $\hat{g}$ tends to be large (e.g., greater than 500, say), small changes in the number of folds k should not have adverse effects. Thus, this choice should be made on substantive grounds, that will depend on the application. In practice, the conventional choices of k equal to 2, 5, or 10 should work well in most settings.

We have had relatively little to say about the choice of the nominal reproducibility error β. We have found that setting $\beta = 0.05$ is suitable for most applications. Finally, although it has not played a central role in our theoretical analysis, it is important to choose the burn-in period g_{init} to be suitably large to ensure that the variance estimate $\hat{v}_{g_{\mathsf{init}},k}(D)$ is reasonably accurate, otherwise Algorithm 1 will tend to stop too early. We recommend setting g_{init} to be equal to at least 10.

Figure 1.5: Performance in Application to Casey et al. (2021b)



Notes: Figure 1.5 displays measurements of the performance of Algorithm 1 on the data from Casey et al. (2021b). Panel A displays measurements of the reproducibility error, $P\{|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)| \geq \xi \mid D\}$, as ξ and k vary. A solid horizontal line is displayed at the nominal error rate $\beta = 0.05$. Panel B displays measurements of the average number of replications $\hat{g}$ s ξ and k vary. The y-axis is displayed with a log scale, base 10. Solid horizontal lines are placed at each exponential factor of 10. Panel C displays measurements of the 0.05th and 0.95th quantiles of the discrepancy $\hat{g}/g^{\star} - 1$ as k and ξ vary. Further details on the construction of this figure are given in Appendix A.1.5.

Chapter 2

Order-Explicit Linearization of High-Dimensional U-Statistics

with Vasilis Syrgkanis

We give an order-explicit large deviation bound for the difference between a high-dimensional U-statistic and its Hájek projection. In particular, we show that any U-statistic of order b on n observations, with a d-dimensional kernel whose coordinates have ψ_1-Orlicz norm at most ϕ, has a maximum deviation from its Hájek projection of order $O_p(\phi b n^{-1} \log^2(dn))$. The proof relies on the development of novel order-explicit moment inequalities for higher-order Hoeffding components. We show that this rate is unimprovable, up to the polynomial factor on the logarithmic term. As corollaries, we obtain new Bernstein-type concentration and Gaussian approximation results for high-dimensional U-statistics. We apply these results to establish the consistency of a set of resampling-based simultaneous confidence intervals built around a class of nonparametric regression estimators constructed with subsampled kernels. This class encompasses several forms of random forest regression, including Generalized Random Forests.

2.1. Introduction

Let $\mathbf{D}_n = (D_i)_{i=1}^n$ collect independent observations from a distribution P on a set $\mathcal{D}$. For each symmetric, d-dimensional kernel function $u : \mathcal{D}^b \to \mathbb{R}^d$, define the b-order U-statistic

$$U_{n,b}(u) = \binom{n}{b}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} u(D_{\mathsf{s}}) \; , \tag{2.1.1}$$

where the set $\mathcal{S}_{n,b}$ collects all of the subsets of $[n] = \{1, \ldots, n\}$ of size b and the quantity D_{s} collects the subset of the observed data $\mathbf{D}_n$ with indices in the set s.

The canonical approach for studying the large-sample behavior of statistics with this structure, due to Hoeffding (1948), is based on the observation that the projection of the average (2.1.1) onto the space of statistics that decompose additively into the contributions of individual observations takes the form

$$\frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i) \ , \quad \text{where} \quad u^{(1)}(D) = \mathbb{E}[u(D_{[b]}) \mid D_1 = D] - \mathbb{E}[u(D_{[b]})] \ . \tag{2.1.2}$$

The average (2.1.2) is known as the Hájek projection of the U-statistic (2.1.1), following Hájek (1968). The essential idea is that, if the residual from this projection can be shown to be of a lower stochastic order, then the problem reduces to the analysis of an average of i.i.d. random variables. In particular, under classical regularity conditions, Hoeffding (1948) establishes that if the dimension $d = 1$ and the order b is fixed, then

$$\sqrt{\frac{n}{\sigma_b^2 b^2}}(U_{n,b}(u) - \mathbb{E}[u(D_{[b]})]) \xrightarrow{d} \mathsf{N}(0,1) \ , \quad \text{where} \quad \sigma_b^2 = \mathrm{Var}(u^{(1)}(D_i)) \ . \tag{2.1.3}$$

See Serfling (1980), Lee (1990), and van der Vaart (2000) for textbook treatments.

In this paper, we extend this approach to settings where the order b and the dimension d are non-negligible relative to the sample size n. Formally, we obtain the following order-explicit large deviation bound on the residual from the Hájek projection. Throughout, the quantities c and C denote universal positive constants, whose values are allowed to change in each appearance. We write $A \lesssim B$ if $A \le CB$ almost surely.

Theorem 2.1.1. *Consider a symmetric kernel function $u : \mathcal{D}^b \to \mathbb{R}^d$. If the order b satisfies $b\log(dn) \le cn$ and each component $u_j(\cdot)$ of $u(\cdot)$ satisfies*

$$\|u_j(D_{[b]})\|_{\psi_1} \le \phi \ , \tag{2.1.4}$$

then the large deviation bound

$$\left\| U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i) \right\|_{\infty} \lesssim \frac{b}{n}\phi\log^2(dn) \tag{2.1.5}$$

holds with probability greater than $1 - C/nd$.

We give the proof in Section 2.2. Moreover, we show that the bound (2.1.5) is unimprovable up to the polynomial factor on the logarithmic term.

Theorem 2.1.1 generates two corollaries through the application of the idea proposed in Hoeffding (1948), each closing an open problem in the large-sample approximation of high-dimensional U-statistics. The first corollary is the following order-explicit Bernstein-type large deviation bound. The proof follows by combining Theorem 2.1.1 with a standard Bernstein inequality, and is deferred to Appendix B.1.

Corollary 2.1.1. *Define the Hájek projection variance* $\sigma_{b,j}^2 = \mathrm{Var}(u_j^{(1)}(D_i))$ *and set* $\overline{\sigma}_b = \max_{j\in[d]} \sigma_{b,j}$*. If the conditions of Theorem 2.1.1 hold, then the inequality*

$$\left\| U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] \right\|_\infty \lesssim \sqrt{\frac{b^2 \overline{\sigma}_b^2 \log(nd)}{n}} + \phi \frac{b}{n} \log^2(nd) \tag{2.1.6}$$

holds with probability greater than $1 - C/nd$*.*

Large deviation bounds for U-statistics were first given in Hoeffding (1963) (see Lemma A.5 of Song et al. (2019) for a modern statement). However, the leading term in the bound given there takes the form

$$\sqrt{\frac{b\overline{\nu}\log(dn)}{n}} \quad \text{where} \quad \overline{\nu} = \max_{j\in[d]} \mathrm{Var}(u_j(D_{[b]}))\ . \tag{2.1.7}$$

As it can be shown that $b\sigma_{b,j}^2 \leq \mathrm{Var}(u_j(D_{[b]}))$, the quantity (2.1.7) is larger than the leading order term appearing in (2.1.6).[1] Seen differently, through comparison to the asymptotic approximation (2.1.3), the leading term of the non-asymptotic bound (2.1.6) uses the correct normalizing quantity.

The second corollary is the following order-explicit central limit theorem for high-dimensional U-statistics. The proof follows by combining Theorem 2.1.1 with the high-dimensional central limit theorem for i.i.d. sums stated in Chernozhukov et al. (2022), and is deferred to Appendix B.1.[2]

Corollary 2.1.2. *Continue the notation introduced in Theorem 2.1.1. Define* $\underline{\sigma}_b = \min_{j\in[d]} \sigma_{b,j}$*.*

[1] In part motivated by the deficiency of the Hoeffding (1963) bound, Arcones and Giné (1993), Arcones (1995), Giné et al. (2000), Adamczak (2008), and Peel et al. (2010) establish a series of refined large deviation bounds for high-dimensional U-statistics that use appropriate normalizing factors (among many other related results). However, the constants used to express these bounds depend implicitly on the order b. Maurer (2019) and Minsker (2023) give large deviation bounds for U-statistics that can converge when the order is increasing, but place much stronger assumptions on the regularity and smoothness of the kernel under consideration.

[2] The boundedness condition (2.1.8) can be relaxed to the assumption that $\mathbb{E}[(u_j^{(1)}(D_i))^4] \leq \phi^2\sigma_{b,j}^2$ for each index j.

If the order b satisfies $b\log(dn) \leq cn$ and the kernel $u(\cdot)$ satisfies the bound

$$\|u(D_{[b]}) - \mathbb{E}[u(D_{[b]})]\|_\infty \leq \phi \tag{2.1.8}$$

almost surely, then the inequality

$$\sup_{\mathsf{R}\in\mathcal{R}} \left| P\left\{ \sqrt{\frac{n}{b^2}} \Sigma_b^{-1/2}(U_{n,b} - \mathbb{E}[u(D_{[b]})]) \in \mathsf{R} \right\} - P\left\{ \Sigma_b^{-1/2} Z \in \mathsf{R} \right\} \right| \lesssim \left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n} \right)^{1/4} \tag{2.1.9}$$

holds, where $\Sigma_b = \mathsf{Diag}(\sigma_{b,1}^2, \ldots, \sigma_{b,d}^2)$, Z denotes a centered Gaussian random vector in $\mathbb{R}^d$ with the same covariance matrix as $u^{(1)}(D_i)$, and $\mathcal{R}$ denotes the set of hyper-rectangles in $\mathbb{R}^d$.

The Hoeffding (1948) univariate central limit (2.1.3) was extended only recently to the regime where b is non-negligible relative to n. Under the condition that $b^{-1} \lesssim \sigma_b^2$, DiCiccio and Romano (2022) give a result that applies to the regime $b = o(n^{1/2})$. Wager and Athey (2018) and Peng et al. (2022) strengthen this result to the full regime $b = o(n)$. The results developed in each of these papers apply only to the case that $d = 1$. In parallel, Chen (2018), Chen and Kato (2019), and Song et al. (2019) give results that apply to the case where d can be large relative to n. Of these papers, only Song et al. (2019) gives results with explicit dependence on the order b. However, their results are only applicable to the regime $b^2\underline{\sigma}_b^{-2} = o(n)$. By contrast, in settings where $b^{-1} \lesssim \underline{\sigma}_b^2$, the central limit theorem (2.1.9) applies to the full regime $b\log^5(dn) = o(n)$.

These improvements are the consequence of a set of new order-explicit moment inequalities for higher-order components of Hoeffding decompositions. In the scalar case, existing arguments rely on pairing the total orthogonality of the Hoeffding decomposition with Chebyshev-type bounds. This approach is insufficient for the high-dimensional case, which requires exponential tail bounds. Existing high-dimensional results instead build on the decoupling and symmetrization arguments developed in De la Peña and Montgomery-Smith (1995) and Sherman (1994) and surveyed in De la Pena and Giné (1999). However, available decoupling bounds incur poor dependence on the order b, which becomes prohibitive when b is non-negligible.

Our main technical innovation is a new order-explicit one-sided decoupling bound for U-statistics that is sufficient to resolve this tension. We show that the quantity that results from this bound can be controlled by combining a symmetrization argument due to Sherman

(1994) with a specialization of a moment inequality from Giné et al. (2000). This approach yields exponential tail bounds for higher-order Hoeffding terms that have the correct scaling in the regime where both b and d are large.

The broader scope of Theorem 2.1.1 can be essential for the analysis of modern statistical methods that aggregate algorithms computed on subsamples of a larger data set. To illustrate this point, in Section 2.3, we apply Theorem 2.1.1 to establish the consistency of an approach for constructing resampling-based simultaneous confidence intervals around nonparametric regression estimators based on subsampled kernels, including subsampled nearest-neighbor regression (Fix and Hodges, 1989; Khosravi et al., 2019; Demirkaya et al., 2024) and random forest regression (Breiman, 2001; Mentch and Hooker, 2016; Wager and Athey, 2018).

Our analysis again follows the agenda set out by Hoeffding (1948), and makes essential use of Theorem 2.1.1. We first show that, under regularity conditions, the estimators under consideration are well-approximated by U-statistics of order b with d-dimensional kernels. However, these estimators are only consistent if the subsample size b satisfies $n^{p/(p+c)} = o(b)$, where p is the dimension of the available covariates. This places the relevant regime outside of the range $b = o(n^{1/3})$ covered by the available high-dimensional central limit theorems, while scalar results that allow for larger values of b cannot account for growth in the kernel dimension d. Theorem 2.1.1 closes this gap by showing that the higher-order remainders are negligible at the subsample sizes required for consistency in high dimensions.

Our treatment of this application builds on Oprescu et al. (2019), who augment the Generalized Random Forests framework proposed by Athey et al. (2019) to study pointwise inference for subsampled random forest estimators with nuisance parameters. Our results are comparable to the bounds on the accuracy of Gaussian multiplier bootstrap confidence regions for nonparametric regression and Z-estimation given in Chernozhukov et al. (2014) and Belloni et al. (2018). We generalize these results, in the sense that we treat inference for conditional Z-estimators whose score function is potentially unknown to the researcher.

Additional applications of large order U-statistics to split-sample testing, robust estimation, and conditional density estimation are reviewed in DiCiccio and Romano (2022), Minsker (2023), and Song et al. (2019). Omitted proofs are given in Appendices B.1 to B.3. Appendices B.4 to B.6 give additional results and discussion that will be introduced at appropriate points throughout the paper.

2.2. Proof of Theorem 2.1.1

In this section, we outline the proof of Theorem 2.1.1. Throughout the proof and supporting Lemmas, we maintain $\mathbb{E}[u(D_{[b]})] = 0$ without loss. The result is obtained in two steps. First, we decompose the difference between the U-statistic under consideration and its Hájek projection through a Hoeffding expansion (Hoeffding, 1948; Efron and Stein, 1981). We adopt the following formulation of Hoeffding's expansion based on the use of stochastic differences, stated as Part (i) of the following Lemma (see e.g., Lachièze-Rey and Peccati (2017)). Part (ii) gives the well-known specialization of this result to U-statistics (see e.g., Lemma 5.1.5 A of Serfling (1980)).

Lemma 2.2.1 (Hoeffding's Expansion)**.** *Fix a deterministic function $f : \mathcal{D}^m \to \mathbb{R}$.*
(i) *Let $D'_{[m]}$ denote an independent copy of $D_{[m]}$. Construct $D^{(i)}_{[m]}$ by replacing D_i with D'_i in $D_{[m]}$, leaving all other entries unchanged. For each index i, define the difference operator*

$$\nabla_i f(D_{[m]}) = f(D^{(i)}_{[m]}) - f(D_{[m]}) \ . \tag{2.2.1}$$

If the moment $\mathbb{E}[f(D_{[m]})^2]$ is finite, then the representation

$$f(D_{[m]}) = \mathbb{E}[f(D_{[m]})] + \sum_{\ell=1}^{m} \sum_{\mathsf{s} \in \mathcal{S}_{m,\ell}} f^{(\ell)}(D_{\mathsf{s}}) \ , \quad \text{where} \tag{2.2.2}$$
$$f^{(\ell)}(D_{i_1}, ..., D_{i_\ell}) = (-1)^{\ell} \mathbb{E}[\nabla_{i_1} \cdots \nabla_{i_\ell} f(D_{[m]}) \mid D_{i_1}, ..., D_{i_\ell}] \ ,$$

holds, the non-constant summands appearing in (C.4.2) are mean-zero, and all 2^m summands are uncorrelated.
(ii) *For each integer $\ell \leq n$, define the operator*

$$U_{n,\ell}(h) = \binom{n}{\ell}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,\ell}} h(D_{\mathsf{s}}) \tag{2.2.3}$$

on the set of measurable functions $h : \mathcal{D}^\ell \to \mathbb{R}$. If the statistic $f(D_{[m]})$ is mean-zero and has a finite variance and the function $f(\cdot)$ is symmetric in its arguments, then the decomposition

$$U_{n,m}(f) = \sum_{\ell=1}^{m} \binom{m}{\ell} U_{n,\ell}(f^{(\ell)}) \tag{2.2.4}$$

holds and all m summands appearing in (2.2.4) are uncorrelated.

In particular, specialized to our setting, we obtain the decomposition

$$U_{n,b}(u) - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i) = \sum_{m=2}^{b} \binom{b}{m} U_{n,m}(u^{(m)}) \tag{2.2.5}$$

by applying Part (ii) of Lemma C.4.1.

Second, and more substantively, we bound each of the summands of the decomposition (2.2.5) by applying the following new higher-order moment bound. The essential novelty of the bound is that the dependence on the order m is kept explicit. The proof, given in the following subsection, develops a new one-sided decoupling bound for U-statistics of an arbitrary order, which avoids introducing order-dependent constants present in previous work (see, for comparison, De la Peña and Montgomery-Smith (1995) and Chapter 3.1 of De la Pena and Giné (1999)). The argument then proceeds by combining symmetrization, Bonami's inequality, and a Rosenthal-type decomposition due to Giné et al. (2000) to control the resulting quantity. The final bound is obtained through the application of a new estimate for higher-order moments of conditional expectations of Hoeffding components, derived from a further application of Hoeffding orthogonality.

Lemma 2.2.2. *Fix an integer $1 \leq m \leq \min\{b, n/2\}$. If the function $f : \mathcal{D}^b \to \mathbb{R}$ is symmetric in its arguments and satisfies the bound*

$$\|f(D_{[b]})\|_{\psi_1} \leq \phi \,, \tag{2.2.6}$$

then the higher-order moment inequality

$$\mathbb{E}\left[\left|\binom{b}{m} U_{n,m}(f^{(m)})\right|^q\right]^{1/q} \leq \phi q \left(\frac{Cqb}{n}\right)^{m/2} \tag{2.2.7}$$

holds for each $2 \vee 2\log(n) \leq q \leq cn/b$.

In particular, by setting $q = 2 \vee 2\log(nd)$ and the constant c sufficiently small, the assumption $b\log(nd) \leq cn$ guarantees that $q \leq cn/b$ and $b \leq n/2$, so that Lemma 2.2.2 applies. Thus, for each integer $m \in \{2, \ldots, b\}$ and each component $u_j(\cdot)$ of the kernel $u(\cdot)$, Markov's inequality yields

$$P\left\{\left|\binom{b}{m} U_{n,m}(u_j^{(m)})\right| \geq 2e\phi \log(nd) \left(\frac{Cb\log(nd)}{n}\right)^{m/2}\right\}$$

$$
\begin{aligned}
&\le \exp(-2\log(nd)) \frac{\mathbb{E}\left[\left|\binom{b}{m} U_{n,m}(u_j^{(m)})\right|^{2\log(nd)}\right]}{\left(\phi \log(nd)\left(\frac{Cb\log(nd)}{n}\right)^{m/2}\right)^{2\log(nd)}} \\
&\le \exp(-2\log(nd)) = \frac{1}{n^2 d^2}\,. \qquad (2.2.8)
\end{aligned}
$$

Moreover, by taking the universal constant c sufficiently small, the condition $b\log(dn) \le cn$ again implies

$$
\phi \log(dn) \sum_{m=2}^{b} \left(\frac{Cb\log(dn)}{n}\right)^{m/2} \lesssim \frac{\phi b \log^2(dn)}{n}\,. \qquad (2.2.9)
$$

Consequently, the decomposition (2.2.5) and a union bound imply that

$$
\begin{aligned}
&P\left\{\left\|U_{n,b}(u) - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i)\right\|_\infty \ge C\frac{\phi b \log^2(dn)}{n}\right\} \\
&\le \sum_{j=1}^{d}\sum_{m=2}^{b} P\left\{\left|\binom{b}{m} U_{n,m}(u_j^{(m)})\right| \ge 2e\phi\log(nd)\left(\frac{Cb\log(nd)}{n}\right)^{m/2}\right\} \\
&\le \frac{b-1}{n^2 d} \le \frac{C}{nd}\,, \qquad (2.2.10)
\end{aligned}
$$

completing the proof. ■

Remark 2.2.1. The bound (2.2.10) is unimprovable up to the polynomial factor on the logarithmic term. In particular, in Appendix B.1 we show that, for each $n \ge b \ge 2$ and $d \ge 1$, there exists a symmetric kernel function $u : \mathcal{D}^b \to \mathbb{R}^d$ whose coordinates satisfy the bound (2.1.4) and such that the lower bound

$$
P\left\{\left\|U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i)\right\|_\infty \ge c\frac{\phi b\log(nd)}{n}\right\} \ge \frac{c}{nd}\,. \qquad (2.2.11)
$$

holds. That is, the factor $\phi b/n$ and at least one logarithmic factor in nd are unavoidable. ■

2.2.1 Proof of Lemma 2.2.2

The result is obtained in three steps. Proofs for each sub-lemma are deferred to Appendix B.1. First, we obtain a novel one-sided decoupling bound for the higher-order moments of U-statistics. This is the key innovation in our argument.

Let $\mathsf{R} = (\mathsf{r}_1, \ldots, \mathsf{r}_m, \tilde{\mathsf{r}})$ be a random partition of $[n]$, such that each cell r_i has cardinality $\lfloor n/m \rfloor$ and the remainder cell $\tilde{\mathsf{r}}$ has cardinality $n - m\lfloor n/m \rfloor$. In turn, let $\mathbf{D}_n^{(1)}, \ldots, \mathbf{D}_n^{(m)}$ denote m independent copies of the data $\mathbf{D}_n$ and consider the arbitrary symmetric kernel $g : \mathcal{D}^m \to \mathbb{R}$. Define the identically distributed statistics

$$U_{n,m}(g, \mathsf{R}) = \frac{1}{\lfloor n/m \rfloor^m} \sum_{i_1 \in \mathsf{r}_1} \cdots \sum_{i_m \in \mathsf{r}_m} g(D_{i_1}, \ldots, D_{i_m}) \quad \text{and} \tag{2.2.12}$$

$$\overline{U}_{n,m}(g, \mathsf{R}) = \frac{1}{\lfloor n/m \rfloor^m} \sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)}) \,, \tag{2.2.13}$$

respectively. Observe that the elements of the fixed set $\mathsf{s} \in \mathcal{S}_{n,m}$ are assigned to distinct cells of the partition $\mathsf{r}_1, \ldots, \mathsf{r}_m$ with probability

$$m! \frac{\lfloor n/m \rfloor}{n} \frac{\lfloor n/m \rfloor}{n-1} \cdots \frac{\lfloor n/m \rfloor}{n-m+1} = \lfloor n/m \rfloor^m \binom{n}{m}^{-1} . \tag{2.2.14}$$

Thus, we can evaluate

$$\mathbb{E}\left[U_{n,m}(g, \mathsf{R}) \mid \mathbf{D}_n\right] = \binom{n}{m}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,m}} g(D_{\mathsf{s}}) = U_{n,m}(g) \,. \tag{2.2.15}$$

Consequently, as $q \geq 1$, Jensen's inequality implies that

$$\mathbb{E}\left[|U_{n,m}(g)|^q\right]^{1/q} \leq \mathbb{E}\left[|U_{n,m}(g, \mathsf{R})|^q\right]^{1/q} \,. \tag{2.2.16}$$

But then, as (2.2.12) and (2.2.13) are identically distributed, the bound

$$\mathbb{E}\left[|U_{n,m}(g)|^q\right]^{1/q} \leq \mathbb{E}\left[|\overline{U}_{n,m}(g, \mathsf{R})|^q\right]^{1/q} \tag{2.2.17}$$

also holds.

Remark 2.2.2. It is illustrative to contrast the decoupling bound (2.2.17) with existing results. Many analyses of the large-sample behavior of high-dimensional U-statistics are based on the decoupling bound

$$\mathbb{E}\left[|U_{n,m}(g)|^q\right]^{1/q} \leq h(m) \mathbb{E}\left[\left|\binom{n}{m}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,m}} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)})\right|^q\right]^{1/q} \tag{2.2.18}$$

where $h(m) \asymp m^{\Theta(m^2)}$. This bound was first given by De la Peña and Montgomery-Smith (1995). See Chapter 3.1 of De la Pena and Giné (1999) for a textbook treatment. Although this decoupling is sufficient to derive a bound with the same dependence on n that we obtain in Lemma 2.2.2 through an argument analogous to that pursued below, the explosive growth of the quantity $h(m)$ with m makes the dependence on m dramatically worse. By contrast, the multiplicative factor appearing in the bound (2.2.17) is equal to one.

The bound (2.2.17) is instead reminiscent of the well-known representation

$$U_{n,m}(g) = \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/m \rfloor} g(D_{\mathsf{s}_{\pi,\ell}}) \tag{2.2.19}$$

used in Hoeffding (1948), where $\mathcal{P}_n$ denotes the set of permutations of $[n]$ and $\mathsf{s}_{\pi,\ell} = \{\pi((\ell-1)m+1), \dots, \pi(\ell m)\}$. In particular, Hoeffding (1948) argues that

$$\mathbb{E}\left[|U_{n,m}(g)|^q\right]^{1/q} \leq \mathbb{E}\left[\left|\left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/m \rfloor} g(D_{\mathsf{s}_{\pi,\ell}})\right|^q\right]^{1/q} \tag{2.2.20}$$

by appealing to (2.2.19) and Jensen's inequality, thereby reducing the problem to an i.i.d. sum. The essential aspect of the bound (2.2.17), relative to the bound (2.2.20), is that the dominating quantity exhibits a recursive structure, shared by the term appearing in (2.2.18), that we are able to leverage in the ensuing argument. ■

Second, we apply the symmetrization and multilinear Khintchine inequalities stated in the following Lemma. Part (i) is obtained by an iterated application of a standard symmetrization inequality for sums of independent, centered random variables; see e.g., Section 6.4 of Vershynin (2018). Part (ii) is a direct specialization of the Bonami inequality for homogeneous Rademacher chaos, stated as Theorem 3.2.2 of De la Pena and Giné (1999); see also Sherman (1994).

Lemma 2.2.3. *For each ℓ in $[m]$, let $V^{(\ell)} = (V_i^{(\ell)})_{i=1}^n$ denote a collection of i.i.d. Rademacher variables.*
(i) *If the symmetric function $g : \mathcal{D}^m \to \mathbb{R}$ satisfies the condition*

$$\mathbb{E}[g(D_1, D_2, \dots, D_m) \mid D_2, \dots, D_m] = 0 \tag{2.2.21}$$

almost surely, then the moment inequality

$$\begin{aligned}&\mathbb{E}\left[|\overline{U}_{n,m}(g,\mathsf{R})|^q\right]^{1/q}\\&\leq \frac{2^m}{\lfloor n/m\rfloor^m}\mathbb{E}\left[\left|\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}V_{i_1}^{(1)}\cdots V_{i_m}^{(m)}g(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})\right|^q\right]^{1/q}\end{aligned}\tag{2.2.22}$$

holds for each $q\geq 1$.

(ii) *Fix the deterministic real-valued array* $\{g_{(i_1,\ldots,i_m)}: i_j\in[\lfloor n/m\rfloor]$ *for all* $j\in[m]\}$. *The moment inequality*

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}V_{i_1}^{(1)}\cdots V_{i_m}^{(m)}g_{(i_1,\ldots,i_m)}\right|^q\right]^{1/q}\\&\leq q^{m/2}\left(\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}g^2_{(i_1,\ldots,i_m)}\right)^{1/2}\end{aligned}\tag{2.2.23}$$

holds for each $q>2$.

It is well-known that the Hoeffding projection $f^{(m)}$ is completely degenerate, in the sense of the condition (2.2.21). Thus, Parts (i) and (ii) of Lemma 2.2.3 and the law of iterated expectation imply that

$$\begin{aligned}&\mathbb{E}\left[\left|\binom{b}{m}\overline{U}_{n,m}(f^{(m)},\mathsf{R})\right|^q\right]^{1/q}\\&\leq \frac{2^m}{\lfloor n/m\rfloor^m}\mathbb{E}\left[\left|\binom{b}{m}\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}V_{i_1}^{(1)}\cdots V_{i_m}^{(m)}f^{(m)}(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})\right|^q\right]^{1/q}\\&\leq C^mq^{m/2}\lfloor n/m\rfloor^{-m}\binom{b}{m}\mathbb{E}\left[\left|\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}f^{(m)}(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})^2\right|^{q/2}\right]^{1/q}.\end{aligned}\tag{2.2.24}$$

Hence, by choosing $g=f^{(m)}$ and combining (2.2.17) and (2.2.24), we obtain the inequality

$$\begin{aligned}&\mathbb{E}\left[\left|\binom{b}{m}U_{n,m}(f^{(m)})\right|^q\right]^{1/q}\\&\leq C^mq^{m/2}\lfloor n/m\rfloor^{-m/2}\binom{b}{m}\end{aligned}\tag{2.2.25}$$

$$\cdot\, \mathbb{E}\left[\left|\frac{1}{\lfloor n/m\rfloor^m}\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor} f^{(m)}(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})^2\right|^{q/2}\right]^{1/q},$$

for each $q > 2$.

Finally, we bound the moment

$$\mathbb{E}\left[\left|\frac{1}{\lfloor n/m\rfloor^m}\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor} f^{(m)}(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})^2\right|^{q/2}\right]^{2/q} \tag{2.2.26}$$

through the application of the following Lemma. Part (i) is a specialization of Proposition 2.1 of Giné et al. (2000) to large higher-order moments of symmetric non-negative kernels. The bound controls moments of the form (2.2.26) with a geometrically weighted series of conditional moments obtained by fixing subsets of arguments of the kernel under consideration. Part (ii) then gives a suitable bound for these conditional moments. The result follows by fixing the first t arguments of the kernel f and identifying the corresponding section of $f^{(m)}$ as the $(m-t)$th Hoeffding component of the induced kernel. This representation then facilitates a further application of Hoeffding orthogonality. We are not aware of a direct antecedent for this estimate, although Hoeffding orthogonality has been used in a related way to control the variance of higher-order terms in Hoeffding expansions. See, for instance, Athey et al. (2019) and Peng et al. (2022).

Lemma 2.2.4. ***(i)*** *Let $g : \mathcal{D}^m \to \mathbb{R}_+$ be a symmetric, non-negative kernel. For each integer $0 \le t \le m$, define*

$$\Gamma_{t,q}(g) = \mathbb{E}\left[\left|\mathbb{E}\left[g(D_{[m]}) \mid D_1,\ldots,D_t\right]\right|^q\right]^{1/q}, \tag{2.2.27}$$

where we adopt the convention that $\Gamma_{0,q}(g) = \mathbb{E}[g(D_{[m]})]$. The inequality

$$\mathbb{E}\left[\left|\frac{1}{\lfloor n/m\rfloor^m}\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor} g(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})\right|^{q}\right]^{1/q} \le C^m\sum_{t=0}^{m}\left(\frac{q}{\lfloor n/m\rfloor}\right)^t\Gamma_{t,q}(g)$$

holds for each $q \ge 1 \vee \log(\lfloor n/m\rfloor)$.

(ii) *Let $f : \mathcal{D}^b \to \mathbb{R}$ be a symmetric kernel that satisfies the bound $\|f(D_{[b]})\|_{\psi_1} \le \phi$. Then,* *for each integer $m \le b$ and $0 \le t \le m$, the inequality*

$$\Gamma_{t,q}((f^{(m)})^2) \le \phi^2 (Cq)^2 C^t \binom{b}{m}^{-1} \left(\frac{b}{m}\right)^t \tag{2.2.28}$$

holds for each $q \ge 1$.

In particular, as $\log(n) \le q/2$, Parts (i) and (ii) of Lemma 2.2.4 imply that

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{\lfloor n/m\rfloor^m}\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor} f^{(m)}(D_{i_1}^{(1)},\dots,D_{i_m}^{(m)})^2\right|^{q/2}\right]^{2/q}\\ &\le C^m \sum_{t=0}^{m}\left(\frac{q/2}{\lfloor n/m\rfloor}\right)^t \Gamma_{t,q/2}((f^{(m)})^2) \le C^m\phi^2 q^2 \binom{b}{m}^{-1}\sum_{t=0}^{m}\left(\frac{q/2}{\lfloor n/m\rfloor}\frac{Cb}{m}\right)^t .\end{aligned} \tag{2.2.29}$$

Consequently, as $q \le cn/b$ and $b \le n/2$, we obtain the bound

$$\mathbb{E}\left[\left|\binom{b}{m} U_{n,m}(f^{(m)})\right|^q\right]^{1/q} \le C^m q^{m/2+1}\phi \left\lfloor\frac{n}{m}\right\rfloor^{-m/2}\binom{b}{m}^{1/2} \tag{2.2.30}$$

by plugging (2.2.29) into (2.2.25) and applying the geometric series formula. Hence, as

$$\binom{b}{m} \le \left(\frac{eb}{m}\right)^m \quad \text{and} \quad \left\lfloor\frac{n}{m}\right\rfloor^{-m/2} \le \left(\frac{2m}{n}\right)^{m/2}, \tag{2.2.31}$$

we obtain the bound

$$\mathbb{E}\left[\left|\binom{b}{m} U_{n,m}(f^{(m)})\right|^q\right]^{1/q} \le \phi q \left(\frac{Cqb}{n}\right)^{m/2} \tag{2.2.32}$$

for all $2\log(n) \le q \le cn/b$, as required. ■

2.3. Simultaneous Inference for Subsampled Kernel Regression

In this section, we apply Theorem 2.1.1 to establish the consistency of an approach for constructing resampling-based simultaneous confidence intervals around nonparametric regression estimators based on subsampled kernels. We begin in Appendix A.4.4 by defining and illustrating the confidence region under consideration. In Sections 2.3.2 and 2.3.3 we define subsampled kernel regression and state several regularity conditions. The main result is given in Section 2.3.4. Throughout, for any function $f(x)$ and vector $\boldsymbol{x}^{(d)} = (x^{(j)})_{j=1}^d$, we let $f(\boldsymbol{x}^{(d)})$ denote the vector $(f(x^{(1)}), \ldots, f(x^{(d)}))$.

2.3.1 Construction

The target for inference is the parameter vector $\theta_0(\boldsymbol{x}^{(d)})$, where $\theta_0(x)$ is defined as the unique solution in θ to the conditional moment equation

$$M(x; \theta, g_0) = \mathbb{E}\left[m(D_i; \theta, g_0) \mid X_i = x\right] = 0\,. \tag{2.3.1}$$

Here, X_i is a sub-vector of the observation D_i, $\boldsymbol{x}^{(d)} = (x^{(j)})_{j=1}^d$ is a vector of query points, and g_0 is a nuisance parameter. We focus on estimators $\hat{\theta}_n(x)$ obtained as solutions to the empirical moment equation

$$M_n(x; \theta, \hat{g}_n, \mathbf{D}_n) = \sum_{i=1}^n K(x, X_i) m(D_i; \theta, \hat{g}_n) = 0\,, \tag{2.3.2}$$

where $\hat{g}_n$ is some first-stage estimator of g_0 and $K(\cdot, \cdot)$ is a data-dependent kernel constructed with subsampling.

Example 1. To fix concepts, consider Banerjee et al. (2015), who study the effects of a poverty alleviation program implemented in Ghana.[3] For each individual in their sample, they observe the data $D_i = (Y_i, W_i, Z_i)$, where Y_i is a measurement of total assets taken two years after the implementation of the program, W_i is an indicator denoting assignment to the program, and Z_i is a vector of covariates. A broad aim of the study is to determine the conditions under which recipients of aid experience lasting improvements in welfare. One quantity that can inform this determination is the conditional average treatment effect (CATE)

$$\theta_0(x) = \mathbb{E}_P\left[Y_i(1) - Y_i(0) \mid X_i = x\right]\,, \tag{2.3.3}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes generated by the intervention W_i and X_i is some chosen subvector of Z_i. Many modern approaches to estimating CATEs are premised on the observation that $\theta_0(x)$ is the solution to the conditional moment equation

$$M(x; \theta, g_0) = \mathbb{E}\left[(\mu_1(Z_i) - \mu_0(Z_i)) + \beta(W_i, Z_i)(Y_i - \mu_{W_i}(Z_i)) - \theta \mid X_i = x\right] = 0\,, \tag{2.3.4}$$

[3] Banerjee et al. (2015) study data collected from several similar graduation programs. We focus on the data from their evaluation of the program implemented in Ghana. Appendix B.4.3 gives further details.

where the nuisance parameter g_0 collects the conditional outcome regression and Horvitz-Thompson weight

$$\mu_w(z) = \mathbb{E}_P\left[Y_i \mid W_i = w, Z_i = z\right] \quad \text{and} \quad \beta(w,z) = \frac{w}{\pi(z)} - \frac{1-w}{1-\pi(z)}\ , \tag{2.3.5}$$

for the propensity score $\pi(z) = P\{W_i = 1 \mid Z_i = z\}$. See Semenova and Chernozhukov (2021), Foster and Syrgkanis (2023), and Kennedy (2023), for further discussion.

Figure 2.1 displays CATE estimates for the experiment studied in Banerjee et al. (2015), where the chosen conditioning covariates X_i are pretreatment measurements of monthly consumption and total assets. The nuisance parameter estimate $\hat{g}_n$ and the kernel $K(x, x')$ are constructed with the implementation of random forest regression made available through the "GRF" R package (Athey et al., 2019). The graduation program appears to be most effective for individuals with a high level of baseline consumption and a low level of baseline assets.[4] These results are suggestive of a poverty trap: individuals with an opportunity to increase their assets are able to do so only if they have a high level of baseline consumption (Balboni et al., 2022; Kraay and McKenzie, 2014). ■

We consider a method for assessing the statistical significance of estimates typified by Figure 2.1. In particular, we study a simple procedure for constructing confidence intervals $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$ centered around the solutions to (2.3.2) that satisfy the bound

$$\sup_{P \in \mathbf{P}} \left| P\left\{\theta_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)})\right\} - (1-\alpha)\right| \leq r_{n,d} \tag{2.3.6}$$

for a sequence $r_{n,d}$ that converges to zero in regimes where d may increase more quickly than n and $\mathbf{P}$ is some statistical family that contains the distribution P.[5]

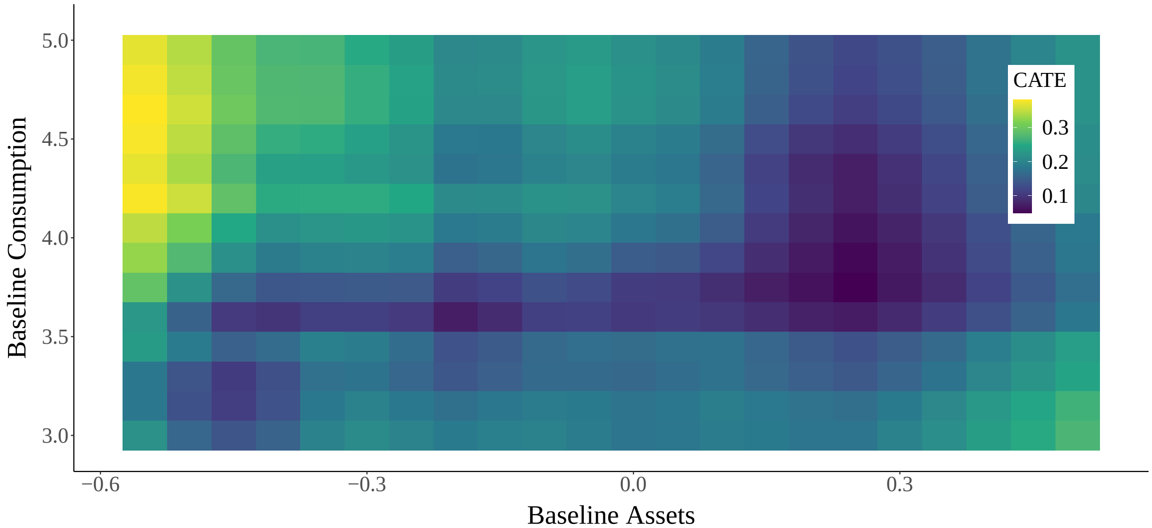
To this end, let h denote a random element of $\mathcal{S}_{n,n/2}$, i.e., a random half-sample of $[n]$ and let $\hat{\theta}_{\mathsf{h}}(\boldsymbol{x}^{(d)})$ denote the vector of the solutions to (2.3.2), with the data $\mathbf{D}_n$ replaced by the half-sample D_{h}. Our construction is premised on approximating the distribution of the root

$$R_n(\boldsymbol{x}^{(d)}) = \hat{\theta}_n(\boldsymbol{x}^{(d)}) - \theta_0(\boldsymbol{x}^{(d)}) \tag{2.3.7}$$

[4]The quartiles of baseline log consumption are 3.33, 3.76, and 4.20. The quartiles of baseline assets are -0.45, -0.71, and 0.03. Panel A of Figure B.3 displays a scatter plot of baseline log consumption and assets.

[5]For the sake of simplicity, we consider only the case that the moment $m(\cdot;\theta,g_0)$ and the parameter θ are scalar-valued. The vector-valued case is relevant, e.g., when estimating CATEs with multiple treatment variables. Our results generalize to the vector-valued case at the cost of additional notation.

Figure 2.1: CATE Estimates



Notes: Figure 2.1 displays a heat map giving CATE estimates for the intervention studied in Banerjee et al. (2015) on post-treatment assets. The color of each rectangle indicates the estimate of the CATE queried at the rectangle's central point. The horizontal and vertical axes display the baseline monthly consumption, normalized to 2014 US dollars and measured in logs base 10, and the baseline value of an index for total assets. CATE estimates are obtained by solving the empirical moment equation (2.3.2) for each value x on an evenly spaced grid on both axes. See Appendix B.4.3 for further details.

with the conditional distribution of the half-sample bootstrap root

$$R_n^*(\boldsymbol{x}^{(d)}) = \hat{\theta}_{\mathsf{h}}(\boldsymbol{x}^{(d)}) - \hat{\theta}_n(\boldsymbol{x}^{(d)}) \ . \tag{2.3.8}$$

The nuisance parameter estimator $\hat{g}_n$ does not need to be re-estimated when computing (2.3.8). A version of the half-sample bootstrap is implemented by default in the GRF R package (Athey et al., 2019). The half-sample bootstrap is an instance of subsampling (Politis and Romano, 1994; Politis et al., 1999).

Let $\hat{\lambda}_{n,j}^2$ denote the variance of $\sqrt{n}R_n^*(x^{(j)})$, conditioned on the data $\mathbf{D}_n$, and let $\hat{\mathsf{cv}}_n(\alpha)$ denote the $1-\alpha$ quantile of the distribution of the studentized statistic

$$\hat{S}_n^*(\boldsymbol{x}^{(d)}) = \sqrt{n}\left\|\hat{\Lambda}_n^{-1/2} R_n^*(\boldsymbol{x}^{(d)})\right\|_\infty \ , \tag{2.3.9}$$

again conditioned on the data $\mathbf{D}_n$. Here, $\hat{\Lambda}_n$ denotes the diagonal matrix with elements $\hat{\lambda}_{n,j}^2$.[6] We consider confidence intervals with the following structure.

[6]The quantities $\hat{\lambda}_{n,j}^2$ and $\hat{\mathsf{cv}}_n(\alpha)$ are easily approximated by resampling the bootstrap root (2.3.8). To simplify exposition, we omit explicit consideration of residual randomness induced by this approximation.

Definition 2.3.1. Define the intervals

$$\hat{\mathcal{C}}(x^{(j)}) = \hat{\theta}_n(x^{(j)}) \pm n^{-1/2}\hat{\lambda}_{n,j}\widehat{\mathsf{cv}}_n(\alpha) \quad \text{for each} \quad j \text{ in } [d] \,. \tag{2.3.10}$$

The level-α half-sample confidence region for $\theta_0(\boldsymbol{x}^{(d)})$ is given by $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$.

Confidence intervals with the same structure, based on different choices of bootstrap root, are studied in, e.g., Chernozhukov et al. (2014) and Belloni et al. (2018).

The essential feature of the bootstrap root (2.3.8) is that it can be computed without knowing anything about the structure of the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$. In particular, approaches based on the Rademacher or Gaussian multiplier bootstrap rely on knowledge of a linear approximation to $\hat{\theta}_n(\boldsymbol{x}^{(d)})$. To gain intuition, suppose that the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$ satisfies the exact linear representation

$$\hat{\theta}_n(\boldsymbol{x}^{(d)}) = \frac{1}{n}\sum_{i=1}^{n} \overline{u}(\boldsymbol{x}^{(d)}, D_i) \tag{2.3.11}$$

for some function $\overline{u}(\cdot,\cdot)$. Let $V_1, \dots, V_n$ denote a collection of random variables, where V_i takes the value 1 if the index i is an element of the random set h used to define the half-sample bootstrap root (2.3.8), and takes the value -1 otherwise. Observe that

$$\begin{aligned} R_n^*(\boldsymbol{x}^{(d)}) = \hat{\theta}_{\mathsf{h}}(\boldsymbol{x}^{(d)}) - \hat{\theta}_n(\boldsymbol{x}^{(d)}) &= \frac{2}{n}\sum_{i=1}^{n} \mathbb{I}\{i \in \mathsf{h}\}\overline{u}(\boldsymbol{x}^{(d)}, D_i) - \frac{1}{n}\sum_{i=1}^{n}\overline{u}(\boldsymbol{x}^{(d)}, D_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} V_i\left(\overline{u}(\boldsymbol{x}^{(d)}, D_i) - \theta_0(\boldsymbol{x}^{(d)})\right) \,. \end{aligned} \tag{2.3.12}$$

The weights V_i are exchangeable Rademacher random variables, i.e., they are uniformly distributed on $\{1, -1\}$. If the weights were fully independent, the representation (2.3.12) reduces to the more familiar Rademacher bootstrap. The representation (2.3.12) is due to Yadlowsky et al. (2023), who draw on a similar observation made in the context of two-sample testing in Chung and Romano (2013).

In practice, many widely applied estimators are not perfectly linearly decomposable. Often, however, estimators do satisfy an approximate linear decomposition, in the sense that the equality (2.3.11) holds with a remainder term of order, say, $o_p(n^{-\gamma})$ for some positive constant γ. As we will see, estimators constructed with subsampled kernels are approximately linear around some unknown function $\overline{u}(\cdot,\cdot)$. If an estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$ is approximately linear, then the subsampled estimate $\hat{\theta}_{\mathsf{h}}(\boldsymbol{x}^{(d)})$ is *immediately* also approximately linear.[7] Thus, for approximately linear statistics, the representation (2.3.12) continues to hold, now with a remainder term of order $o_p(n^{-\gamma})$.

Example 1 (Continued)**.** We now return to the data studied in Banerjee et al. (2015). Figure 2.2 displays upper and lower half-sample confidence bounds for the CATE (2.3.3) on post-treatment assets. The null hypothesis that the CATE is equal to zero is only rejected for individuals with low baseline assets. The lower bounds are meaningfully larger than zero only for individuals with low baseline assets and high baseline consumption. That is, the graduation program has a positive impact on individuals who do not have many assets to begin with, but who do have access to a stable source of consumption. On the other hand, the confidence regions contain zero for individuals who have low baseline consumption or high baseline assets. In Section 3.6.4, we measure the coverage and average length of the confidence region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$ in a simulation calibrated to this setting. ■

2.3.2 Subsampled Kernel Regression

Subsampled kernel regression is a broad class of algorithms for solving regression problems of the form (2.3.2), based on constructing a data-driven kernel function $K(x, x')$ with subsampling. Formally, fix some positive integer r and let $(\mathsf{s}_q)_{q=1}^r$ collect a sequence of subsets of $[n]$ drawn independently and uniformly from $\mathcal{S}_{n,b}$. Let ξ denote some auxiliary source of randomness and let $(\xi_{\mathsf{s}_q})_{q=1}^r$ collect a set of independent random variables with the same distribution as ξ. We study conditional empirical moment estimators of the form (2.3.2), where the kernel function $K(x, x')$ admits the decomposition

$$K(x, X_i) = \frac{1}{r}\sum_{q=1}^{r} \mathbb{I}\{i \in \mathsf{s}_q\}\kappa(x, X_i, D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) \tag{2.3.13}$$

for some known kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi_{\mathsf{s}})$. Several widely applied instances of subsampled kernels are as follows.[8]

Example 2. Subsampled nearest-neighbors regression is a simple example of a kernel with the structure (2.3.13) (Fix and Hodges, 1989). Here, the kernel $\kappa(x, X_{i'}, D_{\mathsf{s}}, \xi_{\mathsf{s}})$ is non-zero if and only if $X_{i'}$ is one of the k closest points to x among the points in the subsample D_{s}.

[7] Approximate linearity does not immediately imply a representation analogous to (2.3.12) for a root constructed with an empirical bootstrap, as approximate linearity would not necessarily hold under the empirical distribution.

[8] The half-sample bootstrap is particularly computationally efficient for estimators constructed with subsampling, as the estimator and the half-sampled estimator can be constructed using the same collection of subsamples. See Section 4.1 of Athey et al. (2019).

Example 3. Random forest regression, introduced by Breiman (2001), is another example of a kernel with the structure (2.3.13). In this case, each pair $(D_{\mathsf{s}}, \xi_{\mathsf{s}})$ generates some partition of $\mathcal{X}$. The kernel $\kappa(x, x', D_{\mathsf{s}}, \xi_{\mathsf{s}})$ is non-zero if and only if x and x' are in the same element of the partition generated by $(D_{\mathsf{s}}, \xi_{\mathsf{s}})$. Often, such partitions are constructed with recursive algorithms, e.g., the "CART" algorithm of Breiman et al. (1984).

We impose the following restrictions on the kernels under consideration.

Assumption 2.3.1 (Honesty and Positive Symmetry)**.**
(i) The kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi)$ is Honest in the sense that

$$\kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) \perp\!\!\!\perp m(D_i; \theta, g) \mid X_i, D_{\mathsf{s}_{-i}} , \tag{2.3.14}$$

where $\perp\!\!\!\perp$ denotes conditional independence and s_{-i} denotes the set $\mathsf{s} \setminus \{i\}$.
(ii) The kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi)$ is positive and satisfies the restriction $\sum_{i \in \mathsf{s}} \kappa(\cdot, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) = 1$ almost surely. Moreover, the conditional expectation $\mathbb{E}\left[\kappa(\cdot, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mid D_{\mathsf{s}}\right]$ is invariant to permutations of the data D_{s}.

The "Honesty" condition stipulated in Part (i) of Assumption 2.3.1 imposes the restriction that any part of the data D_i that can affect the value of the moment $m(D_i; \theta, g)$ cannot affect the value of the kernel $\kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}})$. This condition was introduced in Athey and Imbens (2016). Honesty is often achieved through kernel construction schemes based on sample-splitting; see Wager and Athey (2018) and Athey et al. (2019) for further discussion. Part (ii) of Assumption 2.3.1 imposes several weak regularity conditions.

The following two quantities restrict the "size" and "variability" of the chosen kernel.

Definition 2.3.2 (Shrinkage and Incrementality)**.**
Let s be an arbitrary element of $\mathcal{S}_{n,b}$ and let ℓ be any element of s.
(i) We say that the kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi_{\mathsf{s}})$ has a uniform shrinkage rate ε_b if

$$\sup_{P \in \mathbf{P}} \sup_{j \in [d]} \mathbb{E}\left[\max\left\{\|X_i - x^{(j)}\|_\infty : \kappa(x^{(j)}, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) > 0\right\}\right] \leq \varepsilon_b \, . \tag{2.3.15}$$

(ii) We say that the kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi_{\mathsf{s}})$ is uniformly incremental if

$$\inf_{P \in \mathbf{P}} \inf_{j \in [d]} \operatorname{Var}\left(\mathbb{E}\left[\sum_{i \in \mathsf{s}} \kappa(x^{(j)}, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) m(D_i; \theta, g) \mid D_\ell = D\right]\right) \gtrsim b^{-1} \tag{2.3.16}$$

where D is an independent random variable with distribution P.

The shrinkage rate of a kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi_{\mathsf{s}})$ is analogous to the bandwidth of a classic deterministic kernel. The incrementality restriction ensures that the chosen kernel is not overly dependent on a single data point. Both notions were introduced by Wager and Athey (2018) and have been characterized explicitly for various widely applied subsampled kernel estimators.[9]

Example 2 (Continued)**.** In the case of honest subsampled k-NN regression, Khosravi et al. (2019) show that $\varepsilon_b \lesssim b^{-1/p}$, where p is the intrinsic dimension of the measure of the covariates X_i. See e.g., Kpotufe (2011) for further discussion. In turn, Khosravi et al. (2019) and Peng et al. (2022) show that the kernels associated with both honest and non-honest variants of k-NN regression are incremental, up to logarithmic factors that depend on the dimension of the covariates. ■

Example 3 (Continued)**.** Analogously, for honest random forest regression, under suitable regularity conditions, Wager and Athey (2018) establish that $\varepsilon_b \lesssim b^{-c/p}$, where p is the dimension of the domain of X_i. See e.g., Wager and Athey (2018) and Oprescu et al. (2019) for further discussion. Bounds adaptive to the intrinsic dimension of the measure of X_i are given in Huo et al. (2023). Wager and Athey (2018) and Peng et al. (2022) give simple conditions under which the kernel associated with subsampled, honest, random forest regression is uniformly incremental, again up to dimension-dependent logarithmic factors.■

2.3.3 Moment Restrictions

We restrict attention to conditional moments that satisfy a Neyman orthogonality condition (Chernozhukov et al., 2018). We assume that the nuisance parameter g_0 is a collection of h real-valued functions $g_0 = (g_0^{(k)})_{k=1}^h$, each having domain $\mathcal{Z}$ given by the support of the covariates Z_i. Define the norm

$$\|g - g_0\|_{2,\infty} = \sup_{k \in [h]} \sup_{j \in [d]} \left(\mathbb{E}\left[(g^{(k)}(Z_i) - g_0^{(k)}(Z_i))^2 \mid X_i = x^{(j)} \right] \right)^{1/2} \tag{2.3.17}$$

for any $g = (g^{(k)})_{k=1}^h$. The nuisance parameter g_0 takes values in the space $\mathcal{G}$. Throughout, for a functional F on $\mathcal{F}$, we use the notation

$$\partial_f F(f)[h] = \frac{\mathrm{d}}{\mathrm{d}t} F(f + th)\Big|_{t=0} \quad \text{and} \quad \partial_{f,f} F(f)[h] = \frac{\mathrm{d}^2}{\mathrm{d}t^2} F(f + th)\Big|_{t=0}$$

[9]The terminology "shrinkage" was introduced in Oprescu et al. (2019), and is not intended to connote (explicit) regularization.

to denote first- and second- order directional derivatives, respectively.

Definition 2.3.3 (Local Neyman Orthogonality)**.** We say that a conditional moment $M(\cdot;\theta_0,g_0)$ is uniformly locally Neyman orthogonal if

$$\partial_g M(x^{(j)};\theta_0,g_0)[g-g_0]=0 \tag{2.3.18}$$

for all P in $\mathbf{P}$ and $x^{(j)}$ in $\boldsymbol{x}^{(d)}$.

The use of Neyman orthogonal moments ensures that the bias induced by the estimation of the nuisance parameter g_0 with $\hat{g}_n$ is small (Chernozhukov et al., 2018).

In addition to Neyman orthogonality, we require several smoothness restrictions on the function $m(\cdot;\theta,g)$. In the main text, to ease exposition, we impose the following linearity and boundedness restriction.

Assumption 2.3.2 (Moment Linearity and Boundedness)**.** *The moment function $m(\cdot;\theta,g)$ satisfies the linear representation*

$$m(D_i;\theta,g)=m^{(1)}(D_i;g)\cdot\theta+m^{(2)}(D_i;g) \tag{2.3.19}$$

for some known functions $m^{(1)}(\cdot;g)$ and $m^{(2)}(\cdot;g)$. Moreover, the absolute value of the function $m(\cdot;\theta,g)$ is bounded by the constant $(|\theta|+1)\phi$ for some $\phi\geq 1$ almost surely.

The linearity restriction entailed in Assumption 2.3.2 is inessential and is imposed for the sake of simplicity.[10] Like in Corollary 2.1.2, the boundedness restriction can be weakened to a slightly more involved assumption stated in terms of sub-exponential norms. Again, we impose boundedness to ease exposition.

We maintain the following additional mild smoothness restrictions.

Assumption 2.3.3 (Moment Smoothness)**.**
(i) *The moment $M(\cdot;\theta,g_0)$ is uniformly second order smooth, in the sense that*

$$\sup_{P\in\mathbf{P}}\sup_{j\in[d]}\left|\partial_{g,g}M(x^{(j)};\theta_0,g_0)[g-g_0]\right|\lesssim\|g-g_0\|_{2,\infty}^2 \tag{2.3.20}$$

for each g in $\mathcal{G}$.

[10]In Appendix B.2, we show that moment linearity can be replaced by the high-level assumption that $\hat{\theta}_n(\boldsymbol{x}^{(d)})$ is consistent for $\theta_0(\boldsymbol{x}^{(d)})$. This state of affairs is standard in M-estimation problems (see e.g., Newey and McFadden, 1994). As our running examples use linear moments, and sufficient conditions for the consistency of $\hat{\theta}_n(\boldsymbol{x}^{(d)})$ have been established (see e.g., Assumption 4.1 and Theorem 4.3 of Oprescu et al., 2019)), we omit a detailed consideration of nonlinear moments.

(ii) *The variogram*

$$V(x;g) = \mathbb{E}\left[(m(D_i;\theta_0(x),g) - m(D_i;\theta_0(x),g_0))^2 \mid X = x\right]$$

is uniformly Lipschitz in its first component, in the sense that

$$\sup_{P\in\mathbf{P}}\sup_{g\in\mathcal{G}}\left|V(x;g) - V(x';g)\right| \lesssim \|x - x'\|_\infty \tag{2.3.21}$$

holds for all x and x' in $\mathcal{X}$. Moreover, the variogram satisfies the mean-squared continuity condition

$$\sup_{P\in\mathbf{P}}\sup_{j\in[d]}|V(x^{(j)},g) - V(x^{(j)},g')| \lesssim \|g - g'\|_{2,\infty}^2 \tag{2.3.22}$$

for each g and g' in $\mathcal{G}$.
(iii) *Define the moments*

$$M^{(1)}(x;\theta,g) = \mathbb{E}\left[m^{(1)}(D_i;g) \mid X_i = x\right] \quad \textit{and} \tag{2.3.23}$$

$$M^{(2)}(x;g) = \mathbb{E}\left[m^{(2)}(D_i;g) \mid X_i = x\right]\ , \tag{2.3.24}$$

associated with the functions $m^{(1)}(\cdot;g)$ and $m^{(2)}(\cdot;g)$ introduced in Assumption 2.3.2. Both moments are uniformly Lipschitz in their first component. That is, it holds that

$$\sup_{P\in\mathbf{P}}\sup_{g\in\mathcal{G}}\left|M^{(1)}(x;g) - M^{(1)}(x';g)\right| \lesssim \|x - x'\|_\infty \tag{2.3.25}$$

for all x, x', and θ, and analogously for $M^{(2)}(x;g)$. Moreover, the first moment is uniformly Lipschitz in its second component and bounded from below in the sense that

$$\sup_{P\in\mathbf{P}}\sup_{j\in[d]}\left|M^{(1)}(x^{(j)};\theta_0,g) - M^{(1)}(x^{(j)};g_0)\right| \lesssim \|g - g_0\|_{2,\infty} \quad \textit{and} \tag{2.3.26}$$

$$\inf_{P\in\mathbf{P}}\inf_{j\in[d]}\left|M^{(1)}(x^{(j)};g)\right| \geq c \tag{2.3.27}$$

for each g and θ and some positive constant c.
(iv) *The parameter space Θ is a bounded, convex subset of $\mathbb{R}$. Moreover, the target parameter satisfies $\theta_0(x) \in \Theta$ for every $x \in \mathcal{X}$ and every $P \in \mathbf{P}$ and the estimator $\hat{\theta}_n(x)$ is constructed as a solution over Θ.*

Neyman orthogonal moments satisfying the smoothness restrictions specified in Assumption 2.3.2 and Assumption 2.3.3 are available for many widely considered statistical problems. For example, the moment (2.3.4) is the unique Neyman orthogonal identifying moment for the parameter (2.3.3) (see e.g., Hahn, 1998; Chernozhukov et al., 2018).[11] Additional examples of problems where smooth Neyman orthogonal identifying moments are available include estimation of partially linear regression and partially linear instrumental variable regression (Chernozhukov et al., 2018), local average treatment effects (Tan, 2006; Frölich, 2007), dynamic treatment effects (Lewis and Syrgkanis, 2021), and long-term treatment effects identified by surrogate outcomes (Athey et al., 2020; Chen and Ritzwoller, 2023).

2.3.4 Coverage Accuracy

The following theorem gives a bound on the error in the nominal coverage probability of the confidence regions introduced in Definition 2.3.1. We emphasize that the result is applicable to asymptotic regimes where the dimension of the query-vector $\boldsymbol{x}^{(d)}$ can be exponentially larger than the sample size n.

Theorem 2.3.1. *Suppose that the kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi_{\mathsf{s}})$ satisfies Assumption 2.3.1, has uniform shrinkage rate ε_b, and is uniformly incremental. Additionally, suppose that the Neyman orthogonal moment function $M(\cdot; \theta_0, g_0)$ satisfies Assumption 2.3.2 and Assumption 2.3.3 and that r has been chosen to satisfy $n \leq b\sqrt{r}$. If the estimator $\hat{g}_n$ is statistically independent of the data $\mathbf{D}_n$ and satisfies the moment bound*

$$\sup_{P \in \mathbf{P}} \mathbb{E}\left[\|\hat{g}_n - g_0\|_{2,\infty}^q\right]^{1/q} \leq q \left(\frac{b}{n}\right)^{1/4} \delta_{n,g} \tag{2.3.29}$$

for some sequence $\delta_{n,g}$ and each $q \geq 2 \vee 2\log(dn)$, then the confidence region given in Definition 2.3.1 satisfies

$$\begin{aligned} &\sup_{P \in \mathbf{P}} \left| P\left\{\theta_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)})\right\} - (1-\alpha)\right| \\ &\qquad \lesssim \left(\frac{b\phi^4 \log^8(dn)}{n}\right)^{1/4} + \left(\delta_{n,g}^2 + \sqrt{\frac{n}{b}}\varepsilon_b\right) \log^3(dn)\,. \end{aligned} \tag{2.3.30}$$

[11] For estimation of CATEs, Part (i) of Assumption 2.3.3 is implied by the more refined bound

$$\sup_{P \in \mathbf{P}} \sup_{j \in [d]} \left|\partial_{g,g} M(x^{(j)}; \theta_0, g_0)[g - g_0]\right| \lesssim \|\mu - \mu_0\|_{2,\infty} \|\beta - \beta_0\|_{2,\infty}\,, \tag{2.3.28}$$

where μ and β are defined in (2.3.5). This structure yields the celebrated "Double Robustness" result for estimation of conditional average treatment effects; see Chernozhukov et al. (2018) and Kennedy (2023) for further discussion. We impose the more general condition (2.3.20), as this bound exhibits many of the same features and will hold for a wider variety of problems.

Remark 2.3.1. To expedite exposition, in stating Theorem 2.3.1, we have assumed that the nuisance parameter estimator $\hat{g}_n$ is computed on a separate sample, independent of the data $\mathbf{D}_n$. This can be achieved by randomly splitting the available data into two subsamples. The first can be used to construct the nuisance parameter estimate $\hat{g}_n$; the second to construct the confidence region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$. There are practical issues with this strategy. First, the region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$ might be sensitive to the choice of the split of the data (Ritzwoller and Romano, 2023). Second, by splitting the data, researchers incur a potentially meaningful loss in statistical precision. In Appendix B.4.1 we give additional conditions under which the nuisance parameter estimator $\hat{g}_n$ can be computed using the same data used to construct the region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$.[12] If these conditions are not plausible for a given application, in Appendix B.4.2, we detail a modified version of the confidence region given in Definition 2.3.1 that is constructed with cross-splitting. ■

Remark 2.3.2. The bound (2.3.30) can be interpreted as a bias-variance decomposition. The first term in (2.3.30) results from a Berry-Esseen-type bound, analogous to Corollary 2.1.2, on the accuracy of a normal approximation to (2.3.31). The term involving the kernel shrinkage ε_b is a remnant of a bound on the bias of the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$. ■

Remark 2.3.3. Theorem 2.3.1 follows from an application of an abstract result stated in Appendix B.2. The proof is given in Appendix B.3. The abstract result applies to conditional moment estimators that are not necessarily constructed with subsampled kernels. Rather, the result holds under the high-level assumption that the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$ is approximately linear, with a sufficiently small remainder term, and has sufficiently small bias and stochastic equicontinuity. ■

The proof of Theorem 2.3.1 has three main steps. First, through a standard series of expansions (see e.g., Chernozhukov et al., 2018; Oprescu et al., 2019), we show that the root $R_n(\boldsymbol{x}^{(d)})$ can be approximated by

$$W_{n,b}(\boldsymbol{x}^{(d)}) = \binom{n}{b}^{-1} \sum_{\mathsf{s}\in\mathcal{S}_{n,b}} \mathbb{E}\left[u(\boldsymbol{x}^{(d)}; D_{\mathsf{s}}, \xi_{\mathsf{s}}, \theta_0, g_0) \mid D_{\mathsf{s}}\right] , \quad \text{where} \tag{2.3.31}$$

$$u(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, \theta, g) = -M^{(1)}(x; g_0)^{-1} \sum_{i\in\mathsf{s}} \big(\kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) m(D_i; \theta, g) - \mathbb{E}\big[\kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) m(D_i; \theta, g)\big]\big) ,$$

[12] Our analysis builds on Chen et al. (2022), who give a pointwise analysis of unconditional moment estimators.

with a remainder term given by the second term in (2.3.30). The quantity (2.3.31) can be recognized as a U-statistic of order b. Second, we apply Theorem 2.1.1 to show that this quantity is approximately linear. Written differently, we establish that the root $R_n(\boldsymbol{x}^{(d)})$ satisfies a linear representation of the form (2.3.11), up to a small remainder term. It immediately follows that the bootstrap root $R_n^*(\boldsymbol{x}^{(d)})$ satisfies the linear representation (2.3.12), up to a small remainder term. Finally, we conclude by applying suitable central limit theorems (Chernozhukov et al., 2022) to the linear terms (2.3.11) and (2.3.12).

The broadened scope of Theorem 2.1.1 is an essential input into making Theorem 2.3.1 operational. To see this, recall that for many honest subsampled kernel estimators, the shrinkage rate ε_b satisfies the bound $\varepsilon_b \lesssim b^{-c/p}$ for some small constant c and some integer p measuring the dimension of the covariates X_i. Suppose that

$$b = n^{\gamma_b} \quad \text{and} \quad \|\hat{g}_n - g_0\|_{2,\infty} \lesssim n^{-\gamma_g} \; , \tag{2.3.32}$$

with probability greater than $1 - 1/n$, for some constants γ_b and γ_g between 0 and 1. In this case, ignoring logarithmic factors and other constants, the bound (2.3.30) can be re-expressed as

$$n^{\frac{1}{4}(\gamma_b - 1)} + n^{\frac{1}{2}(1-\gamma_b-4\gamma_g)} + n^{\frac{1}{2}(1-\gamma_b(1+c/p))} \; . \tag{2.3.33}$$

In other words, the confidence region defined in Definition 2.3.1 is consistent if

$$1 \leq \gamma_b + 4\gamma_g \quad \text{and} \quad 1 \leq \gamma_b(1 + c/p) \; , \tag{2.3.34}$$

and so the bound (2.3.30) only converges if

$$\gamma_b \geq \frac{p}{p+c} \; . \tag{2.3.35}$$

Written differently, it is essential to accommodate subsample exponents γ_b very close to one, i.e., the regime $b = o(n)$. This is exactly the range enabled by Theorem 2.1.1.

2.3.5 Simulation

We now measure the performance of the confidence region formulated in Definition 2.3.1. We apply a method for simulation design proposed by Athey et al. (2021). In particular, we calibrate a simulation to the Banerjee et al. (2015) data using a Generative Adversarial Network (GAN) (Goodfellow et al., 2014). In effect, we construct a data generating process that approximates the Banerjee et al. (2015) data, where we know the true value of the CATE $\theta_0(x)$ queried at each value x used to construct the grids displayed in Figures 2.1 and 2.2. Further details are given in Appendix B.4.4.

Measurements of performance are taken as two parameters vary. First, we consider several values of the sample size n. In particular, we consider settings with $n = h \cdot n_0$, for h in $\{1, 2.5, 5, 7.5\}$, where n_0 is the sample size of the Banerjee et al. (2015) data. Second, we vary the proportion b/n. We consider three regimes: $b/n = 0.05$, $b/n = (2/(h+1))0.05$, and $b/n = (1/h)0.05$. Observe that b increases in proportion to n in the first regime and that b is constant as n varies in the third regime. The second regime resides between these two extremes.

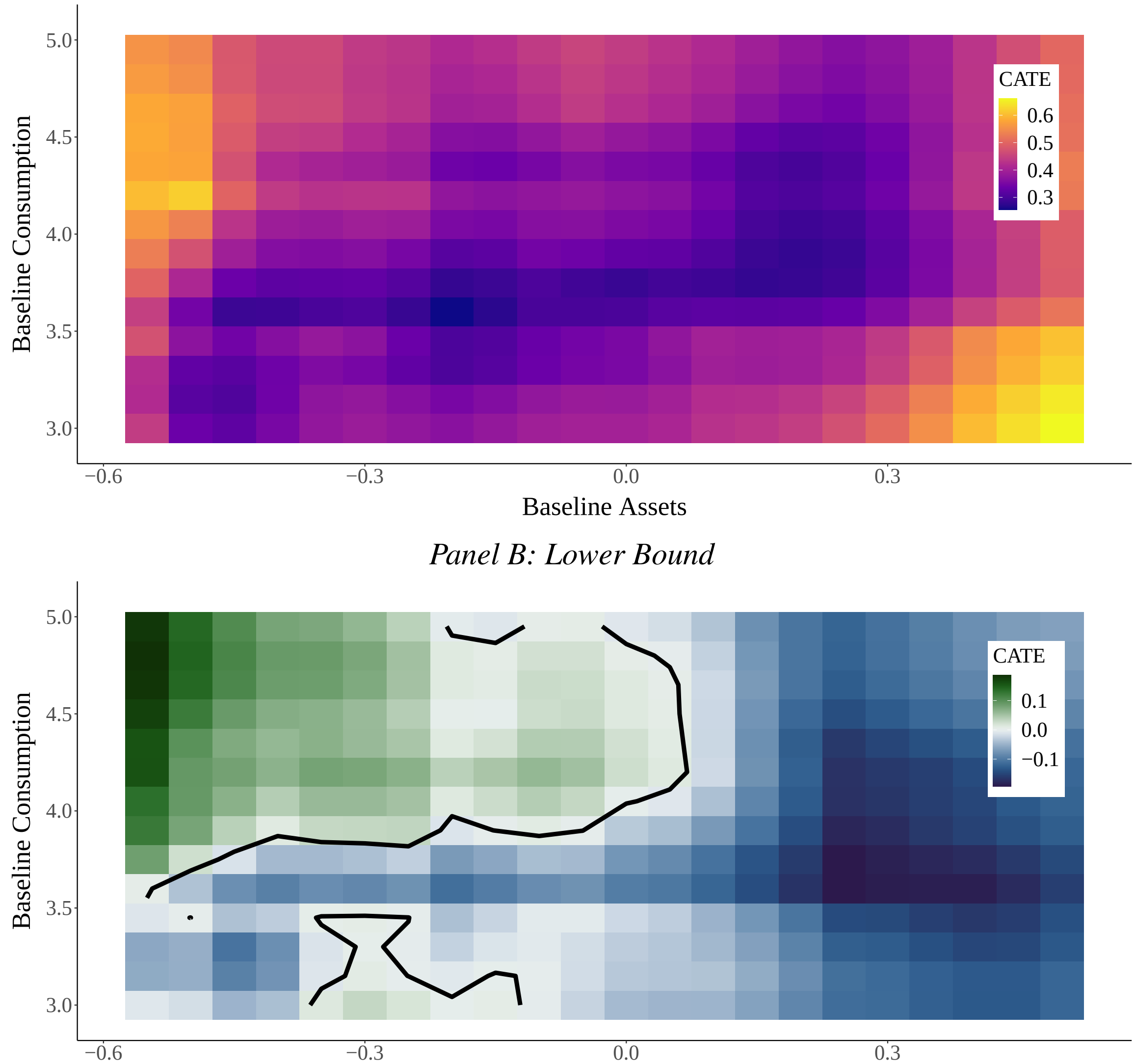
Figure 2.3 displays measurements of the coverage and width of the confidence region formulated in Definition 2.3.1, in addition to measurements of the bias of the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$. Here, we consider the setting where the nuisance parameter estimator $\hat{g}_n$ is computed using the same data used to construct the region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$. The first row displays measurements of the coverage of the confidence region, the coverage of the lower bound (i.e., Panel B of Figure 2.2), and the coverage of the upper bound (i.e., Panel A of Figure 2.2). The nominal level is $\alpha = 0.1$. Throughout, the confidence region is somewhat conservative.

The second row of Figure 2.3 illustrates a bias-variance trade-off with the subsample size b. The first panel displays measurements of the average width of the confidence region. Here, the average is taken over both simulation draws and the query-vector $\mathbf{x}^{(d)}$. The width of the confidence region is increasing in the proportion b/n and is essentially constant if b/n is constant as n increases. By contrast, the second and third panels display the maximum and average bias of the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$, again taken over the query-vector $\mathbf{x}^{(d)}$. The bias is decreasing in the proportion b/n and is essentially constant if b is constant as n varies.

Figure 2.2: Half-Sample Confidence Region

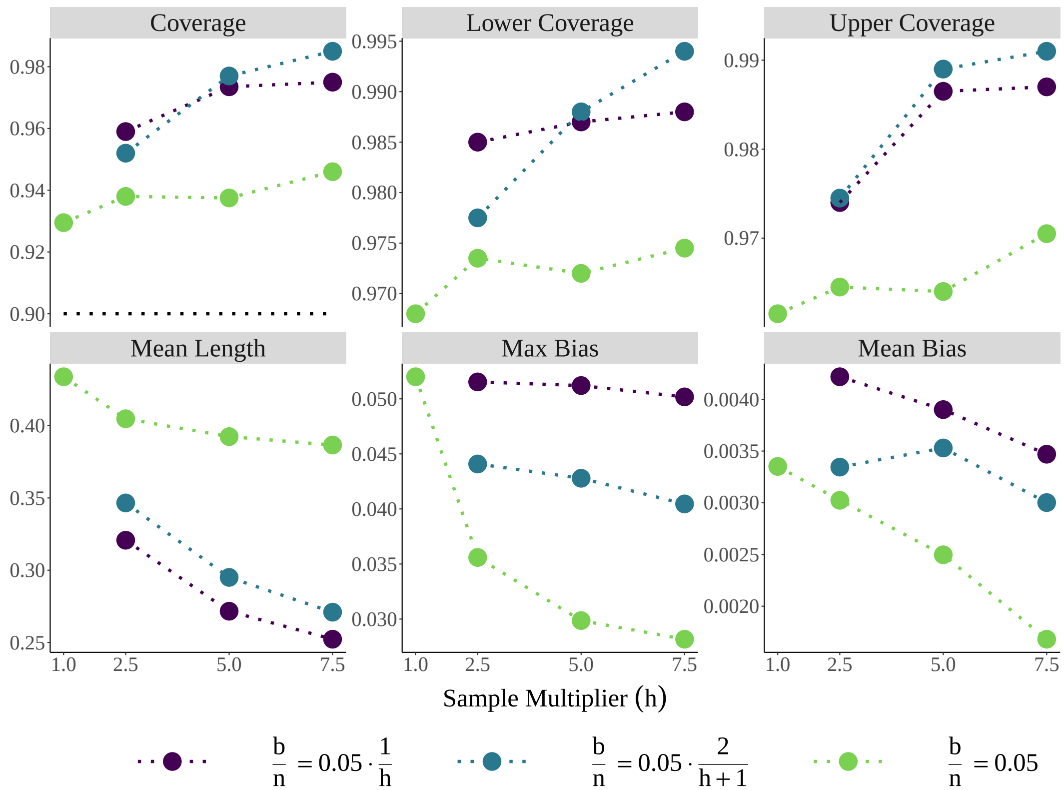
Panel A: Upper Bound

Panel B: Lower Bound



Notes: Figure 2.2 displays heat maps giving half-sample upper and lower confidence bounds for the CATE of the intervention studied in Banerjee et al. (2015) on post-treatment total assets. The confidence bounds are constructed at level $\alpha = 0.1$. The upper and lower bounds are displayed with different color palettes to emphasize the use of different scales. A contour line has been superimposed over the lower bound to demarcate where the bound crosses zero. The axes and estimator are the same as in Figure 2.1.

Figure 2.3: Performance



Notes: Figure 2.3 displays several measurements of the performance of the confidence intervals formulated in Definition 2.3.1 in a simulation calibrated to the Banerjee et al. (2015) data. The nominal level is $\alpha = 0.1$; a horizontal dotted line is displayed at the nominal coverage $1 - \alpha$ in the first panel. The confidence bounds considered are constructed analogously to the confidence bounds displayed in Figure 2.2. The x-axis of each panel is the sample multiplier h. The color of each measurement varies with b/n.

Chapter 3

Regression Adjustments for Disentangling Spillover Effects

Empirical analyses that characterize the mechanisms that mediate spillover effects often do so by relating responses to a treatment, shock, or policy change with a measure of economic proximity, such as geographic distance, technological similarity, trade costs, or migration flows. Typically, such efforts are based on regressions that associate outcomes with proximity-weighted averages of the treatments received by other units. We show that regressions with this structure measure how the association between one unit's outcome and another unit's treatment correlates with the proximity between the two units. We then argue that, if the proximity measure of interest is associated with other channels that mediate spillover effects, causal interpretations of such relationships are susceptible to confounding. For instance, if technologically similar firms tend to be geographically proximate, then a positive association between technological similarity and the intensity of the productivity spillovers between firms might arise spuriously. We give conditions under which the effect of a proximity measure on spillover intensity can instead be recovered by regressing outcomes on averages of other units' treatments, reweighted by residualized versions of the proximity measure under consideration. We show that estimates obtained in this manner achieve the optimal rate of convergence and give a resampling procedure for constructing estimates of the associated uncertainty. We illustrate the proposed method with several empirical applications.

3.1. Introduction

The indirect effects of economic actions are often multifaceted. For instance, one firm's investment in research and development may influence another firm's value through both product market competition and communication among researchers (Arrow, 1962; Dasgupta and Stiglitz, 1980; Griliches, 1992; Bloom et al., 2013). Likewise, productivity changes in one city may affect output in others through the exchange of goods, labor, and ideas (Marshall, 1890; Ellison et al., 2010; Allen and Arkolakis, 2014; Monte et al., 2018).

As a result, aggregate estimates of indirect effects can conflate distinct and potentially countervailing mechanisms. A shock to one location's natural resource endowment, for example, might benefit locations closely connected through labor markets, while harming locations linked through trade in intermediates, as factor reallocation toward the natural-resource sector crowds out manufacturing (i.e., "Dutch Disease"; Krugman, 1987; Allcott and Keniston, 2018). The indirect effects of changes to industrial policy (Greenstone et al., 2010; Giroud et al., 2024) and infrastructure investment (Kline and Moretti, 2014; Donaldson and Hornbeck, 2016), similarly, can depend on how—and how closely—locations are connected. As a consequence, efforts to disentangle the channels through which spillover effects propagate are frequently critical in the design and evaluation of economic policy.

Empirical analyses that characterize the mechanisms that mediate spillover effects often do so by relating responses to a shock or policy change with a chosen measure of economic proximity. For example, Moretti (2004) finds that the spillover effects of increases in human capital on productivity are larger for firms whose investments in research and development are concentrated in more similar technological areas.[1] Analyses with this structure confront an essential ambiguity. The association between spillover intensity and any one proximity measure may be confounded by correlation with other mediating channels. That is, if productivity spillovers are larger for technologically similar firms, and technologically similar firms are likely to be geographically proximate, should we infer that technological

[1]Results with an analogous structure are common. For example, Franklin et al. (2024) show that the spillover effects of localized public investment on wages are larger for locations that are more closely connected in labor markets. Myers and Lanahan (2022) find that the spillovers from venture grants to small firms are larger for firms that are technologically or geographically close. Bernstein et al. (2019) find that the spillover effects of bankruptcy are larger for firms that draw on more similar customer bases. Moscona and Seck (2024) show that the spillover effects of cash transfer programs on consumption are larger for more culturally similar individuals. Miguel and Kremer (2004), Feyrer et al. (2017), Diamond and McQuade (2019), and Egger et al. (2022b) (among many other analogous examples) find that the spillover effects of de-worming programs, shocks to natural resource endowments, low income housing credits, and cash transfer programs, respectively, decrease with the geographic distance.

similarity, itself, affects the intensity of productivity spillovers?[2]

In this paper, we develop an econometric framework for identifying and estimating the effect of a measure of proximity on spillover intensity. To motivate our approach, we begin by giving a nonparametric decomposition, in the spirit of Yitzhaki (1996), Angrist (1998), and Angrist and Krueger (1999), of the implicit estimands of a class of regression specifications typically used in empirical applications that link spillover effects to measures of proximity. In particular, in such settings, researchers often report coefficients obtained from regressions that relate outcomes to proximity-weighted averages of other units' treatments—that is, through specifications similar to

$$Y_i = \alpha + \theta \cdot \Delta_i + \varepsilon_i \ , \quad \text{where} \quad \Delta_i = \sum_{j \neq i} D_{i,j} W_j \ . \tag{3.1.1}$$

Here, Y_i is an outcome, W_j is a treatment, and $D_{i,j}$ is a proximity measure of interest. See, for instance, Miguel and Kremer (2004), Conley and Udry (2010), Bloom et al. (2013), Acemoglu et al. (2016a), Cai and Szeidl (2024), and Franklin et al. (2024). In Section 3.2, we show that, under certain conditions, the coefficient θ on the covariate Δ_i identifies a convex average of the mixed-partial derivatives[3]

$$\frac{\partial^2}{\partial \delta \, \partial w} \, \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w] \ . \tag{3.1.2}$$

Written differently, such regressions measure how the association between one unit's outcome and another unit's treatment *correlates* with the proximity between the two units.[4]

Yet researchers who consider regressions of this kind are, in many instances, fundamentally interested in assigning a causal interpretation to such relationships. That is, often, substantive interest is in measuring the structural relationship between a measure of proximity and spillover intensity. For example, a central question animating the study of economic agglomeration concerns the role of geographic proximity in generating and magnifying spillover effects (Jaffe, 1986; Ellison and Glaeser, 1997; Ellison et al., 2010). To what extent

[2]This indeterminacy was highlighted saliently by Jaffe et al. (1993), who write "The most difficult problem confronted by the effort to test for spillover-localization is the difficulty of separating spillovers from correlations that may be due to a pre-existing pattern of geographic concentration of technologically related activities."

[3]In the result stated in Section 3.2, the conditional expectation in parameter (3.1.2) is replaced by a conditional expectation that effectively holds the "location" of the unit j fixed, and averages over other units i with $D_{i,j} = \delta$. Further notation will be introduced to make the distinction precise.

[4]We allow for the distributions of the treatments and proximity measures to be discrete, continuous, or to contain point masses. For instance, if the treatment variable is valued on $\{0, 1\}$, the parameter (3.1.2) is replaced by the derivative of the difference $\frac{\partial}{\partial \delta} \left(\mathbb{E}[Y_i \mid D_{i,j} = \delta, W_i = 1] - \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_i = 0] \right)$.

is Silicon Valley's outsized success in the production of novel technologies a consequence of the close geographic proximity of innovative firms? Would industrial policies that encouraged greater clustering (e.g., tax credits, investments in transportation and housing infrastructure, etc.) increase the social returns to investment in research and development? This literature has long recognized, however, that productivity spillovers can also propagate in other ways. For example, ideas may move through scientific networks, and so productivity spillovers might thus be determined by the quantity of effort directed to related scientific problems, rather than through geography, per se (Jaffe, 1986; Jaffe et al., 1993).

As a consequence, if the proximity measure of interest is not the only channel that potentially mediates spillovers, then the parameter (3.1.2) does not necessarily measure changes in spillover effects induced by exogenous changes to proximity. That is, if the effects of changes in public sector investment in research and development W_j on the outcome Y_i are larger for pairs of firms i, j that are technologically similar, and technological similarity is positively correlated with geographic proximity $D_{i,j}$, then the parameter (3.1.2) does not recover the causal relationship between geographic proximity and productivity spillover intensity.

We formalize this point in Section 3.3. In particular, we propose a class of parameters, analogous to the descriptive parameter (3.1.2), that reflect the causal relationship between a measure of proximity and spillover intensity. We refer to the elements of this class as "Spillover Proximity Gradients." We show, with an example calibrated to data from Bloom et al. (2013), that when there are several, correlated measures of proximity that mediate spillover effects, unadjusted estimates of the causal relationship between spillover intensity and any one proximity measure can be badly biased.

Accordingly, in Section 3.4, we propose a regression-based approach for recovering convex averages of spillover proximity gradients. This proposal is premised on the observation that, in many applications, researchers observe additional, auxiliary measures of the features and economic proximity of the units in their sample. For instance, Myers and Lanahan (2022) measure how the spillover effects of small business loans depend on firms' technological proximity, and also observe the geographic distance between each pair of firms. Formally, let X_i and $H_{i,j}$ denote vectors of covariates associated with the unit i and the pair of units i, j, respectively, where X_i and X_j can be included as sub-vectors of $H_{i,j}$.

The core methodological idea advanced in this paper is to adjust the regression (3.1.1) by residualizing the treatment and proximity measure of interest with appropriately chosen covariates and auxiliary proximity measures. The proposed method has two steps. First, the user constructs estimates of the conditional expectations

$$\pi_n(x) = \mathbb{E}[W_j \mid X_j = x] \quad \text{and} \quad \gamma_n(h) = \mathbb{E}[D_{i,j} \mid H_{i,j} = h]\ , \tag{3.1.3}$$

respectively. Denote these estimates by $\hat{\pi}_n(\cdot)$ and $\hat{\gamma}_n(\cdot)$. Second, the user solves the regression specification

$$Y_i = \alpha + \theta \cdot \hat{\Delta}_i^\star + \varepsilon_i\ , \quad \text{where} \quad \hat{\Delta}_i^\star = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star \tag{3.1.4}$$

and the variables

$$\hat{W}_j^\star = W_j - \hat{\pi}_n(X_j) \quad \text{and} \quad \hat{D}_{i,j}^\star = D_{i,j} - \hat{\gamma}_n(H_{i,j}) \tag{3.1.5}$$

denote residualized versions of the treatment and proximity measure of interest.

This two-step estimator accounts for potential endogeneity in both the treatment W_j and the proximity measure $D_{i,j}$ by conditioning on observable covariates. Formally, we assume that the treatment W_i is unconfounded conditional on the covariates X_i, and that the covariates and auxiliary proximity measures $H_{i,j}$ are sufficient to stratify pairs of units in terms of any characteristics that jointly determine spillover intensity and the proximity measure of interest. For instance, in the setting considered in Myers and Lanahan (2022), the latter condition would be satisfied if, on average, firms in the same geographic location have the same response to changes in the productivity of firm j, up to their technological similarity to j.[5] In extensions deferred to the Appendix, we illustrate how to augment the specification (3.1.4) to accommodate settings in which treatment endogeneity is addressed using difference-in-differences or instrumental variable research designs, as well as settings in which instrumental variables are available for the proximity measure of interest.

In Section 3.5, we give conditions under which the estimator (3.1.4) is consistent. We enforce two main restrictions. First, we restrict attention to settings where the estimators $\hat{\pi}_n(\cdot)$ and $\hat{\gamma}_n(\cdot)$ are obtained from well-specified linear regressions. This is analogous to the "saturated specification" condition considered by Angrist (1998). Second, we assume that interference is determined by an underlying, unobserved measure of the proximity of each pair of units in a latent space. That is, we assume that interference can be substantial only among pairs of units that are close in this latent space, and that units are only ever

[5]Roughly speaking, this is the identification strategy considered in Jaffe et al. (1993). It has been applied, implicitly or explicitly, by Bloom et al. (2013) and Myers and Lanahan (2022), among many others.

proximate in observed measures if they are proximate in the latent measure. The converse, for the latter, need not hold. We focus our consideration on asymptotic regimes in which the number of units that are proximate in the latent space to each unit increases, so long as the fraction of such units converges to zero. The main novelty of this framework is that we allow for interference to be mediated by proximity in both observed and unobserved characteristics, rather than through the proximity measure of interest alone.

We show that, in such settings, coefficients obtained from regression specifications of the form (3.1.4) converge at a rate strictly slower than $O_p(n^{-1/2})$. In particular, the rate of convergence achieved by the class of estimators under consideration decreases as the extent of cross-unit interference increases. Nevertheless, we establish that this sub-parametric rate is optimal, up to constant factors, in a minimax sense.

In Section 3.6, we detail a procedure for quantifying the uncertainty associated with estimates obtained from regressions with the structure (3.1.4). Most approaches to inference in settings with interference take two routes. In the first, units are placed into disjoint groups, and clustered standard errors can be constructed under the assumption that interference across groups is sufficiently small (Leung, 2023). In the second, interference is assumed to be bounded above by a decreasing function of an observed proximity measure, and inferences can be constructed with kernel-based methods (Kojevnikov, 2021; Leung, 2022a).

Neither approach is applicable to our setting. In particular, as we allow for interference to be mediated by unobserved characteristics, units cannot necessarily be arranged into groups or otherwise organized around an observed proximity measure. Instead, we show that conservative estimates of uncertainty can be obtained via a procedure based on repeatedly computing the coefficient from a regression specification analogous to (3.1.4), where the outcomes have been replaced by the empirical residuals and the signs of the residualized treatments have been flipped at random. The implementation of the procedure does not require the choice of any bandwidths or tuning parameters, and places no a priori restrictions on the structure of the interference across units. This approach is closely related to the variance estimators proposed in Adao et al. (2019) and the randomization inference-based methods considered by Borusyak and Hull (2023). Relative to these procedures, we show that the proposed method is asymptotically conservative for tests concerning the value of the parameter targeted by the estimator (3.1.4) and is applicable to settings where spillovers are non-linear, heterogeneous, and potentially mediated by unobserved channels.

In Section 3.7, we illustrate the application of the proposed methodology to measure the relationship between productivity spillover intensity and technological and product market proximity, using data from Bloom et al. (2013). In line with the results reported in Bloom et al. (2013), we find that adjusting estimates of the relationship between spillover intensity and a given measure of proximity for correlated channels that also mediate spillovers meaningfully changes spillover effect estimates. We give two additional applications, using data from Hornbeck and Moretti (2024) and Franklin et al. (2024), in Appendix C.6.

This paper contributes to an extensive literature that considers the identification and estimation of spillover effects. Much of this research builds on the "exposure mapping" and "effective treatment" frameworks developed in Hudgens and Halloran (2008), Manski (2013), Toulis and Kao (2013), and Aronow and Samii (2017), which parameterize outcomes with correctly specified, low-dimensional summaries of the treatments received by other units. By contrast, and joining Leung (2022a), Sävje (2024), and Menzel (2025), we consider parameters that can be defined without placing such restrictions and do not assume that the complete structure of interference is known a priori (or that information sufficient to recover this structure is observed).[6]

The key distinction, in our asymptotic analysis, is that we assume that interference is determined by proximity in both the observed and unobserved characteristics of the population under consideration. The assumptions that we use to express this structure are closely related to the "latent surface models" considered by, for example, Hoff et al. (2002), McCormick and Zheng (2015), and Breza et al. (2020). We differ, in that our primary aim is not to recover the latent surface. Instead, we use the geometric structure associated with the latent surface as device in our theoretical analysis for organizing the influence of unobserved interference.

The main contribution of this paper is a method for characterizing the *causal* relationship between a measure of proximity and spillover intensity. This objective is related to, but distinct from, the descriptive problems considered in Pollmann (2023) and Wang et al. (2025), who aim to characterize the average spillover effect between pairs of units that are separated by a given distance (see also Munro et al. (2025) and Li and Wager (2022), who consider related problems). Put differently, this paper is distinguished by the fact that its explicit objective is to estimate the *effect* of interventions that affect the proximity between pairs of units.

We contribute, as well, to a growing literature that considers the estimation of spillover

[6] See also Gao and Ding (2023), who apply the framework developed in Leung (2022a) to regression-based estimators.

effects in observational data (see e.g., Leung and Loupos 2025; Auerbach and Tabord-Meehan 2025).[7] In this context, the main novelty of our analysis is that we consider models where measures of proximity are endogenous, in the sense that pairs of units that are more proximate in a particular measure can be more likely to be proximate in other observed or unobserved factors that might also mediate spillover effects.[8]

Finally, we contribute to a broader literature concerning the interpretation and statistical analysis of regression specifications where the covariates of interest take the form of weighted averages (Adao et al., 2019; Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022b; Borusyak and Hull, 2023). In this literature, this paper is distinguished by the fact that we do not assume that the regressions under consideration are well-specified, and study inference for a class of parameters that can be defined nonparametrically.

Section 3.8 concludes. Proofs for results stated in Sections 3.2 to 3.4 are given in Appendices C.1 and C.2. Proofs for results stated in Sections 3.5 and 3.6, and all supporting lemmas, are given in Appendices C.3 and C.4, respectively. Appendices C.5 and C.6 give additional results and discussion that will be introduced at appropriate points throughout the paper.

3.2. A Nonparametric Decomposition of Spillover Regression

We consider settings with the following structure. There are n units. Each unit i in $[n] = \{1, \ldots, n\}$ is associated with a real-valued treatment W_i and a real-valued outcome Y_i. In turn, each pair of units i, j is associated with a variable $D_{i,j}$ that measures geographic or economic proximity. The main interest is in characterizing the relationship between the proximity measure $D_{i,j}$ and the spillover effect of the treatment W_j on the outcome Y_i. Settings with this structure occur throughout applied economics.[9]

---

[7] A distinct line of this literature develops methods for accommodating network interference in treatment selection, primarily through an assumption that a well-specified exposure mapping is unconfounded, conditional on a vector of covariates (see e.g., Veitch et al., 2019; Forastiere et al., 2021; Ogburn et al., 2024; Emmenegger et al., 2025; Auerbach and Tabord-Meehan, 2025). Leung and Loupos (2025) consider a more general model, where treatments are jointly determined, conditional on an observed network and covariates. In each case, the networks that mediate selection into treatment and spillovers between units are taken to be fixed and observed. Moreover, in contrast to the regression-based methods at the center of this paper, explicit specification, or estimation, of the network dependence in the treatments is required.

[8] We note that we use the term “endogenous networks” in a different sense than it is used in Gao (2024), who studies experimental settings where treatments can affect measures of proximity.

[9] To simplify exposition, we have ruled out applications where treatments and outcomes are measured for different units; see, e.g., Kovak (2013), Autor et al. (2013), and Adao et al. (2019), where treatments are assigned to industries and outcomes are measured at geographic units of aggregation. We treat this extension in Appendix C.5.1.

Example 4 (Research and Development)**.** Firm level investments in R&D affect the market outcomes of other firms (Arrow, 1962; Griliches, 1992). Bloom et al. (2013) measure how "knowledge spillovers" vary with "technological proximity." The main hypothesis is that a firm's innovative activity has larger effects on the productivity of firms that operate in similar technological areas. In this application, the treatment W_j is the stock of R&D for firm j, the outcome Y_i is the market value of firm i, and the variable $D_{i,j}$ denotes technological proximity, which, following Jaffe (1986), is measured by the uncentered correlation (or "cosine similarity") of patent shares across USPTO technology classes.[10] This measure is widely adopted, see e.g., Ellison and Glaeser (1997), Moretti (2004), and Ellison et al. (2010). ■

Example 5 (Productivity)**.** Using city-level data from the U.S. Census, Hornbeck and Moretti (2024) argue that the indirect effect of a change in a cities' manufacturing productivity on the wages in another city depends on the propensity to migrate between the two cities. That is, productivity growth increases labor demand in city j, encouraging migration from city i, which reduces labor supply. Here, the treatment W_j denotes total factor productivity growth, the outcome Y_i denotes growth in average earnings, and the proximity measure $D_{i,j}$ denotes the pretreatment probability that an individual that migrates from city i relocates to city j. ■

Example 6 (Public Works)**.** Similarly, Franklin et al. (2024) study the spillover effects of a public works program implemented in Addis Ababa, Ethiopia. The program provided guaranteed public work to targeted households and was randomized across neighborhoods. In this setting, the treatment W_j indicates that neighborhood j has been assigned to the program and the outcome Y_i denotes the average earnings in neighborhood i. Franklin et al. (2024) argue that the indirect effect of the program on earnings depends on commuting flows. That is, individuals who participate in the program no longer commute to other neighborhoods, reducing private sector labor supply. In our notation, they posit that the effect of the program W_j on earnings Y_i is increasing in the pretreatment probability $D_{i,j}$ that a worker who works in neighborhood i lives in neighborhood j. ■

3.2.1 Spillover Regression

In the examples outlined above, the main interest is in characterizing the effect of economic proximity on spillover intensity. What, precisely, does this mean? To give empirical content

[10] Bloom et al. (2013) consider several alternative measures of technological proximity, largely based on reweighting or rescaling the inner product between patent shares. We focus our discussion on their baseline measure for the sake of simplicity.

to an answer to this question, we begin our analysis by considering the interpretation of regression specifications often used in applications with this structure. In particular, in such settings, researchers typically report coefficients obtained from regression specifications similar to

$$Y_i = \alpha + \theta \cdot \Delta_i + \varepsilon_i \ , \quad \text{where} \quad \Delta_i = \sum_{j \neq i} D_{i,j} W_j \tag{3.2.1}$$

measures the "exposure" of unit i to the treatments W_j through the proximity measure $D_{i,j}$.[11]

The interpretation of coefficients obtained in this way varies widely. In some cases, measures of proximity are observed directly in data and the specification (3.2.1) is posed without a strict model-based justification; see, for instance, Bloom et al. (2013), Cai et al. (2015), Fafchamps and Quinn (2018), Huber (2018), Zacchia (2020), Arque-Castells and Spulber (2022), Lerche (2025), and Bergeaud et al. (2025). In others, either the specification (3.2.1) or the measure $D_{i,j}$ are derived from (or approximate the solution to) models of spatial equilibrium; see e.g., Redding and Venables (2004), Hanson (2005), Conley and Udry (2010), Ahlfeldt et al. (2015), Acemoglu et al. (2016a), Acemoglu et al. (2016b), Rotemberg (2019), Helm (2020), Borusyak et al. (2022a), Franklin et al. (2024), Cai and Szeidl (2024), and Adao et al. (2025).

The sentiment adopted in much of this literature is that such estimates recover structural quantities—parameters defined and justified in terms of the primitives and assumptions of an economic, or parametric, model, but that are not necessarily nonparametrically meaningful. In this paper, we advance an alternative perspective. We begin by giving a result, in the spirit of Yitzhaki (1996), Angrist (1998), and Angrist and Krueger (1999), that shows that, under certain conditions, the regression (3.2.1) identifies a weighted average of particular mixed-partial derivatives of the conditional expectation of the outcomes. This result will suggest that the coefficient $\hat{\theta}_n$ on the covariate Δ_i in the specification (3.2.1) can be interpreted as

[11] Often, researchers also include the variable W_i as a covariate, and report the associated coefficient as a measure of the "direct effect." In other cases, the proximity measure of interest enters through some monotonic function $Q(D_{i,j})$. For example, when $D_{i,j}$ is a function of geographic distance, researchers often set $Q(\cdot)$ as a binary indicator, or vector of binary indicators, that $D_{i,j}$ is an element of some pre-defined interval; see, e.g, Miguel and Kremer (2004), Egger et al. (2022b), Myers and Lanahan (2022), and Muralidharan et al. (2023). Similarly, in settings where the specification (3.2.1) has been derived from a model of spatial equilibrium, the function $Q(\cdot)$ often depends on various elasticities fixed at values chosen by referencing pre-existing literature; see e.g., Kovak (2013), Autor et al. (2013), Donaldson and Hornbeck (2016), and Adao et al. (2025). We treat several of these extensions in Appendix C.5.1, but focus our consideration on the specification (3.2.1) in the main text for the sake of parsimony.

targeting a particular nonparametric, causal estimand, whose identification and estimation are the main subjects of this paper.

3.2.2 Baseline Model

Our objective is to give a nonparametric interpretation for coefficients obtained from regression specifications of the form (3.2.1). To do this, we impose several baseline conditions on the joint distribution of the treatments W_j and proximity measures $D_{i,j}$ across units. First, we posit that the proximity measure $D_{i,j}$ is generated by a function of some, potentially unobserved, features of the units i and j.

Assumption 3.2.1 (Latent Factor Structure)**.** *There exists a function $D(\cdot,\cdot)$, valued on the non-negative, real-valued set $\mathcal{D}$, and potentially unobserved coordinates $(S_i)_{i=1}^n$ on the set $\mathcal{S}$ such that the representation*

$$D_{i,j} = D(S_i, S_j) \ , \tag{3.2.2}$$

holds for each pair of units i,j.

If, for instance, the variable $D_{i,j}$ is a function of the geographic distance between units i and j, then S_i can be taken to denote the geographic coordinates of unit i. Likewise, in Example 4, technological proximity is given by the uncentered correlation of the shares of each firm's patents that belong to each USPTO technology class. In more abstract settings, like in Examples 5 and 6, the factors can be viewed more expansively. That is, if $D_{i,j}$ measures the probability of migrating between cities i and j, then the coordinate S_i could, for example, collect city i's geographic coordinates, average wage, and average amenity value.

The main restriction is that the proximity measure $D_{i,j}$ is not determined by any features inherent to the *pair* i, j. In Appendix C.5.1 we show that the result stated in this section can be extended to settings where $D_{i,j}$ is determined by the coordinates S_i, S_j as well as i.i.d., pair-specific errors. This includes, for instance, the latent position and graphon network models considered by, e.g., Hoff et al. (2002), Breza et al. (2020), and Li and Wager (2022). Nevertheless, ruling out pair-specific errors greatly simplifies the exposition of the identification and statistical results given in subsequent sections.

Second, we assume that the treatments are i.i.d., centered, and orthogonal to the latent factors.

Assumption 3.2.2 (Exogeneity)**.** *The replicates $(S_i, W_i)_{i=1}^n$ are i.i.d. and satisfy the condition*

$$\mathbb{E}[W_i \mid S_i] = 0 \tag{3.2.3}$$

almost surely.

This assumption permits dependence in the outcomes across units, as would arise from spillover effects, but rules out cross-sectional dependence in the treatments. We impose this restriction because regression specifications of the form (3.2.6) are not suited to distinguish spillover effects "mediated by" the proximity measure $D_{i,j}$ from correlation between the treatments W_i and W_j.[12] This is a version of the identification problem studied in Manski (1993).

Assumption 3.2.2 also rules out settings where the mean of the treatment varies with S_i. Equivalent assumptions, in closely related models, are used in Adao et al. (2019), Borusyak et al. (2022b), and Borusyak and Hull (2023). In Section 3.4, we relax this aspect of Assumption 3.2.2 to hold after conditioning on available covariates. We have also stipulated that the treatments are centered. This is equivalent to the main methodological recommendation of Borusyak and Hull (2023) and is increasingly adopted in practice.

Third, we require that the association between the treatment and outcome of two distinct units is relatively insensitive to changes in their proximity to a third unit. To state this assumption, we introduce some useful notation. For a real-valued function f defined on the set $\mathcal{A}$ in $\mathbb{R}$, we use the short-hand $\partial_a f(a')$ to denote the derivative of $f(a)$ evaluated at a'. If the set $\mathcal{A}$ is connected, this is defined in the usual way. If the set $\mathcal{A}$ is not connected, then, following a convention set out in Kolesár and Plagborg-Møller (2024), we define $\partial_a f(a)$ on the convex hull of $\mathcal{A}$ by extending f via linear interpolation. That is, if $\mathcal{A} = \{0, 1\}$, then $\partial_a f(1/2) = f(1) - f(0)$. We use the notation $\partial^2_{a,b} f(a, b)$ for the analogous mixed-partial derivative.

Assumption 3.2.3 (Limited Second-Order Interference)**.** *For distinct units i, j, and k, it holds that*

$$\partial^2_{\delta,w}\, \mathbb{E}[Y_i \mid D_{k,j} = \delta, W_j = w, S_j = s] = o(n^{-1/2}) \tag{3.2.4}$$

uniformly over each δ, w, and s in their respective domains.

[12] This choice of emphasis is common in models of network interference (see, e.g., Leung 2020; Auerbach 2022; Li and Wager 2022). The settings considered in Forastiere et al. (2021), Pollmann (2023), Ogburn et al. (2024), Emmenegger et al. (2025), and Leung and Loupos (2025) are exceptions. However, the approaches to estimation taken in these cases are substantially more complex, and require explicit specification (or, in the case of Leung and Loupos (2025), estimation) of the dependence across treatments.

Assumption 3.2.3 is relatively mild, ruling out only situations where changes to the coordinate S_k can make a non-negligible difference on the relationship between the distinct units i and j. Finally, we require two mild regularity conditions concerning the smoothness of the conditional expectation of the outcome and the scaling of the proximity measure $D_{i,j}$. We defer the statement of these assumptions to Appendix C.1.

3.2.3 Decomposition

The following Theorem expresses the population solution to the regression (3.2.1) as a convex average of the mixed-partial derivatives of the conditional expectation

$$\mu(\delta, w \mid S_j) = \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w, S_j] \ , \tag{3.2.5}$$

defined for each δ and w in their respective domains. Let

$$L_n(\theta) = \min_{\alpha} \sum_{i=1}^{n} (Y_i - \alpha - \theta \cdot \Delta_i)^2 \tag{3.2.6}$$

denote the loss function associated with the regression specification (3.2.1).

Theorem 3.2.1. *Assume that the treatments and proximity measures have support contained in intervals $\overline{\mathcal{W}}$ and $\overline{\mathcal{D}}$ of constant length and that the outcomes are identically distributed. Under Assumptions 3.2.1 to 3.2.3, and two regularity conditions stated in Appendix C.1, the parameter $\overline{\theta}_n$ that minimizes the risk $\mathbb{E}[L_n(\theta)]$ admits the representation*

$$\overline{\theta}_n = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(\delta, w \mid S_j)\, \partial^2_{\delta,w} \mu(\delta, w \mid S_j)\Big]\, \mathrm{d}w\, \mathrm{d}\delta + o(n^{-1/2}) \ , \tag{3.2.7}$$

where the weights

$$\lambda(\delta, w \mid S_j) = \frac{\mathrm{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j} \mid S_j)\, \mathrm{Cov}\,(\mathbb{I}\{W_j \geq w\}, W_j \mid S_j)}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid S_j)\, \mathrm{Var}(W_j \mid S_j)]} \tag{3.2.8}$$

are convex, in the sense that

$$\lambda(\delta, w \mid S_j) \geq 0 \quad \textit{and} \quad \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_j)]\, \mathrm{d}w\, \mathrm{d}\delta = 1 \ , \tag{3.2.9}$$

respectively. That is, the weights (3.2.8) are positive almost surely and integrate to one in expectation.

Remark 3.2.1. Theorem 3.2.1 applies to settings where distributions of the treatments and proximity measures are discrete, have continuous support, or have point masses. For instance, if the treatments W_j are valued on $\{0,1\}$ and $\mathrm{Var}(W_j \mid S_j)$ is constant almost surely, then the representation (3.2.7) can be re-expressed as the weighted average derivative

$$\begin{aligned}\overline{\theta}_n = \int_{\mathcal{D}} \lambda(\delta \mid S_j)\, \partial_\delta (\mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = 1, S_j] \\ - \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = 0, S_j])\, \mathrm{d}\delta + o(n^{-1/2})\,, \\ \text{where} \quad \lambda(\delta \mid S_j) = \frac{\mathrm{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j} \mid S_j)}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid S_j)]}\,, \end{aligned} \tag{3.2.10}$$

by the convention set out prior to the statement of Assumption 3.2.3. Likewise, if, additionally, the proximity measure $D_{i,j}$ is supported on $\{0,1\}$, and $\mathrm{Var}(D_{i,j} \mid S_j)$ is constant almost surely, then (3.2.7) takes the form of the double-difference

$$\begin{aligned}\overline{\theta}_n = (\mathbb{E}[Y_i \mid D_{i,j} = 1, W_j = 1] - \mathbb{E}[Y_i \mid D_{i,j} = 1, W_j = 0]) \\ - (\mathbb{E}[Y_i \mid D_{i,j} = 0, W_j = 1] - \mathbb{E}[Y_i \mid D_{i,j} = 0, W_j = 0]) + o(n^{-1/2})\,,\end{aligned} \tag{3.2.11}$$

by the same convention. ■

Remark 3.2.2. Observe that the minimization over the intercept α is included as part of the definition of the loss function (3.2.6), and so occurs within the expectation used to define the parameter $\overline{\theta}_n$. This arrangement is nonstandard, but is essential to obtain the representation (3.2.7). In particular, by the Frisch-Waugh-Lovell Theorem, we can write

$$\begin{aligned}\mathbb{E}[L_n(\theta)] = \mathbb{E}\left[\sum_{i=1}^{n} (Y_i - \theta \cdot \tilde{\Delta}_i)^2\right]\,, \\ \text{where} \quad \tilde{\Delta}_i = \sum_{j \neq i} \left(D_{i,j} - \frac{1}{n}\sum_{k \neq j} D_{k,j}\right) W_j - \frac{1}{n}\sum_{k \neq i} D_{k,i} W_i\,,\end{aligned} \tag{3.2.12}$$

and so, by the Projection Theorem, the parameter $\overline{\theta}_n$ admits the representation

$$\overline{\theta}_n = \frac{\sum_{i=1}^{n} \mathbb{E}[Y_i \tilde{\Delta}_i]}{\sum_{i=1}^{n} \mathbb{E}[(\tilde{\Delta}_i)^2]}\,. \tag{3.2.13}$$

Here, by including the construction of the intercept within the expectation, the proximity measures $D_{i,j}$ are effectively centered by the averages $\frac{1}{n}\sum_{k \neq j} D_{k,j}$ (up to a negligible error term introduced by the second term in (3.2.12)). That is, the proximity measures

$D_{i,j}$ are *implicitly* centered conditional on the coordinate S_j. The error in the average $\frac{1}{n}\sum_{k\neq j} D_{k,j}$ about the expectation $\mathbb{E}[D_{k,j} \mid S_j]$ contributes the leading term of the error in the approximation (3.2.7), and is controlled by Assumption 3.2.3. This implicit centering causes each of the terms that appear in the representation (3.2.7) to condition on the coordinate S_j, which has important implications for the identification arguments pursued below. ■

This result can be best understood through comparison to the simpler setting where an outcome Y_i is regressed on a real-valued treatment W_i and a vector of covariates X_i—that is, through the specification

$$Y_i = \alpha + \theta \cdot W_i + \gamma^\top X_i + \varepsilon_i \ . \tag{3.2.14}$$

Building on results due to Yitzhaki (1996), Angrist (1998), Angrist and Krueger (1999), Kolesár and Plagborg-Møller (2024) show that, under conditions analogous to those used in Theorem 3.2.1, if the expectation of W_i conditional on the covariates X_i is linear, then the population coefficient on W_i admits the representation

$$\int_{\mathcal{W}} \mathbb{E}\Big[\lambda(w \mid X_i)\, \partial_w \mathbb{E}[Y_i \mid W_i = w, X_i]\Big]\, \mathrm{d}w \ , \tag{3.2.15}$$
$$\text{where} \quad \lambda(w \mid X_i) = \frac{\mathrm{Cov}(\mathbb{I}\{W_i \geq w\}, W_i \mid X_i)}{\mathbb{E}[\mathrm{Var}(W_i \mid X_i)]} \ .$$

Here, as before, the weight function $\lambda(w \mid X_i)$ is positive almost surely and integrates to one in expectation. Written differently, linear regression coefficients identify convex averages of the change in the outcome associated with incremental changes in the treatment.

Theorem 3.2.1 shows that, under certain conditions, an analogous, nonparametric interpretation can be given to coefficients obtained from spillover regressions.[13] That is, regression specifications of the form (3.2.1) identify convex averages of the change in the dependence of one unit's outcome on another unit's treatment associated with changes in the proximity of the two units.[14] This decomposition is purely descriptive. That is, we have placed no restrictions on the causal relationship between the treatments, proximity measures,

[13] Vazquez-Bare (2023) consider estimators similar to (3.2.1) in settings where there are a collection of groups, units within the same group receive the same treatment, and the proximity measure $D_{i,j}$ is binary and indicates that i and j are elements of the same group. He finds that the resultant estimate does not have a coherent causal interpretation. By contrast, we show that, if treatments are independent and identically distributed, the implicit estimand of the specification under consideration permits a clean representation.

[14] As it can be shown that $\int\int \lambda(\delta, w \mid S_j)\, \mathrm{d}w\, \mathrm{d}\delta = \mathrm{Var}(D_{i,j} \mid S_j)\, \mathrm{Var}(W_j \mid S_j)$, units j that have larger conditional variances $\mathrm{Var}(W_j \mid S_j)$ and $\mathrm{Var}(D_{i,j} \mid S_j)$ tend to receive larger weights.

and outcomes, and are only describing the joint distribution of their realized values.[15]

Linear regression coefficients invite causal interpretations, as measuring as the *effect* of W_i on Y_i. So too does the representation (3.2.7), as measuring the *effect* of the proximity measure $D_{i,j}$ on the intensity of the spillover effect of W_j on Y_i. In the remainder of the paper, we define causal parameters that encode this structure, state conditions under which they are identified, and develop procedures for augmenting the regression specification (3.2.6) to operationalize these conditions.

3.3. Spillover Proximity Gradients

In many settings in applied economics, the primary, substantive interest is in characterizing the causal relationship between a measure of proximity and the spillover effects of a shock, policy, or action. In particular, often, spillover effects are potentially mediated by several distinct economic phenomena, whose extent and intensity have different implications for welfare and policy. Recovering the structural relationship between responses to a treatment any one proximity measure, then, requires disentangling these channels.

Example 4 (Continued)**.** Consider, for instance, the problem of interest in Bloom et al. (2013). Aggregate estimates of the indirect effects of one firm's investment in R&D on other firms' values risk conflating two distinct mechanisms. The first mechanism gives rise to the core market failure in markets for innovation (Arrow, 1962). One firm's investment in R&D may reduce the cost of R&D for technologically similar firms. Such cost reductions are socially valuable, but are rarely internalized by the innovating firm. Thus, when this mechanism exists—and, in particular, when it is quantitatively meaningful—there is a risk of persistent underinvestment in R&D. The second mechanism concerns product-market interactions: investments in R&D by one firm may reduce sales for firms marketing substitutes and increase sales for firms marketing complements (Dasgupta and Stiglitz, 1980). The welfare and policy implications of the second channel have little overlap with the first, in the sense that the underlying economic phenomena are fundamentally different, and depend largely on market structure (Spence, 1984).

To estimate the effect of technological proximity on R&D spillover intensity, it is necessary to account for the potentially countervailing effect of product market rivalry.

[15] The orthogonality condition stipulated by Assumption 3.2.2 is best motivated by the assumption that the treatments are randomly assigned. If this condition did not hold, the part of the weight $\lambda(\delta, w \mid S_j)$ associated with $D_{i,j}$ would vary with W_j, and would no longer necessarily integrate to one. We relax this assumption, by introducing covariates, in Section 3.4.

To operationalize this idea, Bloom et al. (2013) measure the "product market proximity" of firms i and j by computing the uncentered correlation $G_{i,j}$ of the share of each firms' sales across four-digit industry codes. Panel A of Figure 3.1 displays features of the joint distribution of technological proximity $D_{i,j}$ and product market proximity $G_{i,j}$ across the sample of firms considered by Bloom et al. (2013).[16] Technological proximity increases substantially and systematically with product market proximity. ■

Example 5 (Continued)**.** In Hornbeck and Moretti (2024), interest centers on characterizing how the spillover effect of total factor productivity growth on wages varies with migration. Of course, other notions of economic proximity might also mediate indirect effects. For instance, suppose that we measure the proportion of output produced in city i that is purchased by firms in city j. Denote this measure by $G_{i,j}$. An increase in city i's productivity may increase demand for intermediate inputs produced by city j, increasing city j's labor demand in proportion to $G_{i,j}$. ■

Example 6 (Continued)**.** Franklin et al. (2024) measure how the indirect effects of a public works program depend on commuting flows. There are, again, other channels that might also mediate spillover effects. For example, as the program consists of publicly subsidized employment in small-scale activities that improve shared amenities (e.g., garbage disposal, greening of public spaces), businesses that are geographically close to the treated neighborhoods may experience increased demand, and so spillover effects might, just as well, be driven by geographic distance $G_{i,j}$. ■

In this section, we propose a class of parameters, which we refer to as spillover proximity gradients, that express the causal relationship between a measure of proximity and spillover intensity. We show that, when there are several, correlated measures of proximity that could potentially mediate spillover effects, unadjusted estimates of the causal relationship between spillover intensity and any one measure can be badly biased.

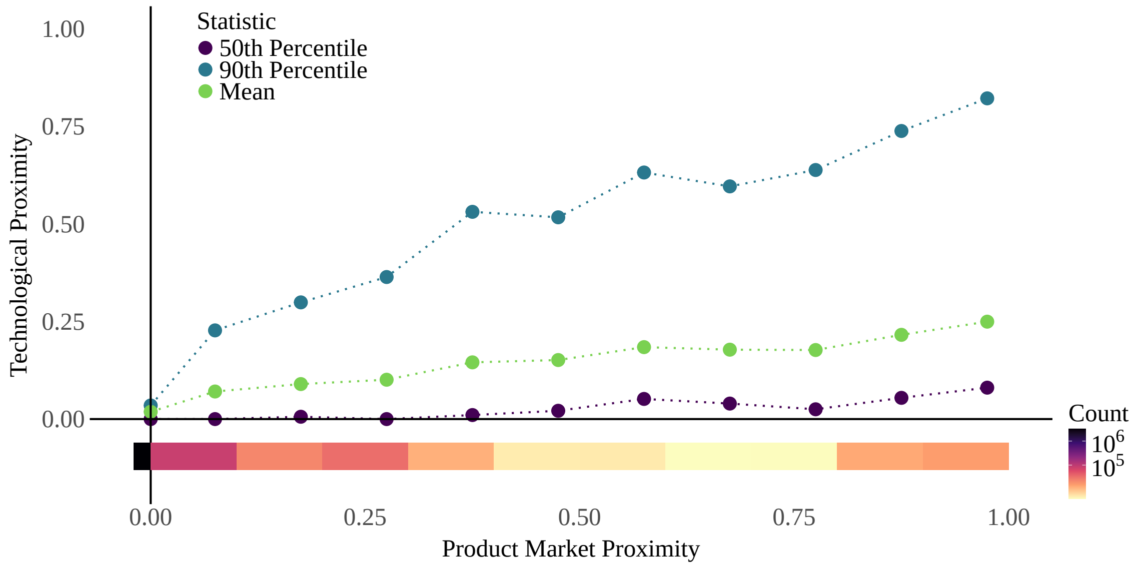
3.3.1 Potential Outcomes

Our interest is in estimating the effect of a measure of proximity on spillover intensity. To pose this objective formally, we must prescribe the potential value of each unit's outcome under counterfactual interventions to each other unit's treatment and proximity.

To this end, we impose a condition that enables us to intervene on the proximity measure of interest $D_{i,j}$, while holding constant any unobserved features of other units. We assume

[16] We detail our treatment of these data, and display analogous figures for Examples 5 and 6, in Appendix C.6.

Figure 3.1: Correlation in Measures of Economic Proximity



Notes: Figure 3.1 displays features of the joint distribution of technological and product market proximity across a sample of publicly traded U.S. firms, using data from the replication packages associated with Bloom et al. (2013) and Lucking et al. (2019). The measures are computed using the shares of patents filed across USPTO technology classes and shares of sales across four digit industry codes from 1990 to 1995, respectively. Further details concerning the treatment of these data are given in Appendix C.6. The x-axis measures product market proximity and has been partitioned into bins. The means, 50th percentiles, and 90th percentiles of technological proximity within each bin are displayed using green, purple, and blue dots and are measured relative to the y-axis. A heatmap measuring the number of pairs of units in each bin is displayed below both the x-axis.

---

that each unit is associated with an additional, potentially unobserved, variable U_i. The variable U_i should be thought of as a unit-specific residual or latent factor. We refer to the variable

$$Z_i = (W_i, S_i, U_i) \tag{3.3.1}$$

as the *state* of unit i and let Z_{-i} collect the state of all units other than i.

Assumption 3.3.1 (Anonymous Interference)**.** *Outcomes are generated through the structural equation*

$$Y_i = F(Z_i, Z_{-i}) \, , \tag{3.3.2}$$

where the function $F(\cdot, \cdot)$ is symmetric in the components of its second argument.

The state Z_i should be interpreted as collecting all of the features of the unit i that can impact any unit's outcome. Like Assumption 3.2.1, the main restriction is that the variable Y_i is not determined by any factors inherent to the pair i, j. That is, Assumption 3.3.1 ensures that the influence of unit j on unit i is "anonymous," in the sense that unit i's outcome depends only on the value of Z_j and not in any way on j's "identity."

Assumption 3.3.1 reduces the problem of defining potential outcomes associated with interventions to the proximity of another unit into the specification of choices for counterfactual values for the coordinate S_j. Any counterfactual coordinate consistent with the intervention $D_{i,j} = \delta$ is necessarily an element of the set

$$\mathcal{S}^{(i)}(\delta) = \{s : D_n(S_i, s) = \delta\} \ . \tag{3.3.3}$$

The convex combination

$$\begin{aligned} S_j^{(i)}(\delta) &= S_j + \alpha_j^{(i)}(\delta)(S_i - S_j) \ , \quad \text{where} \\ \alpha_j^{(i)}(\delta) &= \arg\min_{\alpha} \{|\alpha| : S_j + \alpha(S_i - S_j) \in \mathcal{S}^{(i)}(\delta)\} \ , \end{aligned} \tag{3.3.4}$$

is a natural choice. If this choice is uniquely-defined, we can express the potential outcome

$$Y_{i,j}(\delta, w) = F(Z_i, Z_j(w, S_j^{(i)}(\delta)), Z_{-i,j}) \ , \quad \text{where} \quad Z_j(w,s) = (w, s, U_j) \tag{3.3.5}$$

denotes the counterfactual state of unit j that would be realized if the coordinate S_j were set to s and the treatment W_j were set to w. Here, the vector $Z_{-i,j}$ collects each unit's state, excluding the units i and j.

Unless stated otherwise, for the remainder of the paper, we use (3.3.5) to denote the potential outcome for unit i associated with the intervention that sets the treatment W_j to w and the proximity $D_{i,j}$ to δ. Thus, it remains to enforce structure that ensures that the choice (3.3.4) can be uniquely defined. To do this, we impose the following strengthening of Assumption 3.2.1.

Assumption 3.3.2 (Monotone Factor Structure)**.** *There exists a norm $\|\cdot\|$ on $\mathcal{S}$ such that*

$$D_{i,j} = D(S_i, S_j) = \overline{D}(\|S_i - S_j\|) \tag{3.3.6}$$

for some function $\overline{D}(\cdot)$ that is continuous and non-increasing.

It can be shown that, under Assumption 3.3.2, so long as the set (3.3.3) is non-empty, the choice (3.3.4) is uniquely defined.[17] Assumption 3.3.2 holds in settings where the proximity measure of interest is a decreasing function of geographic distance, such as pure iceberg trade costs. Panel A of Figure 3.2 illustrates the construction (3.3.4) in such a setting. Assumption 3.3.2 also holds in Example 4.

Example 4 (Continued)**.** In Bloom et al. (2013), technological proximity is given by

$$D_{i,j} = \frac{\langle S_i, S_j \rangle}{\|S_i\|_2 \|S_j\|_2} = 1 - \frac{1}{2} \left\| \frac{S_i}{\|S_i\|_2} - \frac{S_j}{\|S_j\|_2} \right\|_2^2 , \tag{3.3.7}$$

where S_i denotes the share of firm i's patents in each USPTO technology class and $\|\cdot\|_2$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean norm and inner-product. Thus, Assumption 3.3.2 is satisfied. Panel B of Figure 3.2 illustrates a simplified version of the construction (3.3.4) for this case. ■

In Appendix C.5.2, we detail several approaches for extending Assumption 3.3.2 to settings where proximity measures are functions of many features of the units under consideration. For instance, commuting and migration probabilities, like those considered in Examples 5 and 6, are often modeled as being functions of geographic distance, as well as wages and amenity values (Monte et al., 2018). In essence, we address such settings by assuming that the measure $D_{i,j}$ is monotone in some feature of the unit j and conditioning on other information sufficient to isolate the dependence of $D_{i,j}$ on this feature.

3.3.2 Spillover Proximity Gradients

We give a framework for identifying and estimating averages of the variables

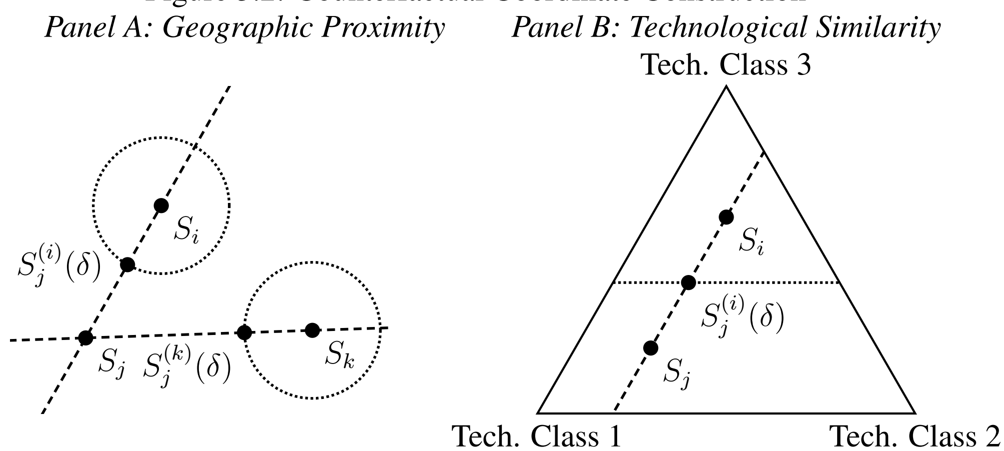
$$\partial^2_{\delta,w} Y_{i,j}(\delta, w) , \tag{3.3.8}$$

measuring the incremental change in the spillover effect of W_j on Y_i as $D_{i,j}$ is increased. We refer to the variable (3.3.8) as a "spillover proximity gradient." Spillover proximity gradients are analogous to the descriptive parameters

$$\partial^2_{\delta,w}\, \mu(\delta, w \mid S_j) = \partial^2_{\delta,w} \mathbb{E}\left[Y_{i,j}(\delta, w) \mid W_j = w, D_{i,j} = \delta, S_j\right] \tag{3.3.9}$$

[17] In Appendix C.5.2, we extend Assumption 3.3.2 to accommodate setting where $\overline{D}(\cdot)$ has discontinuities. For example, in some settings, the proximity measure $D_{i,j}$ might take the form of an indicator that the distance $\|S_i - S_j\|$ is less than or equal to a specified radius. This can be addressed analogously to formalizations of regression discontinuity designs, by defining the potential outcome (3.3.5) as the limiting value as S_j approaches the boundary of $\mathcal{S}^{(i)}(0)$ along the line $S_j + \alpha(S_i - S_j)$. Assumption 3.3.2 can also be extended without loss to settings where $D_{i,j} = \overline{D}(S_i, \|S_i - S_j\|)$, so long as $\overline{D}(S_i, \cdot)$ remains monotone.

Figure 3.2: Counterfactual Coordinate Construction



Notes: Figure 3.2 illustrates the construction of the counterfactual coordinate $S_j^{(i)}(\delta)$ in two examples. In Panel A, the variables S_i measure geographic coordinates and the proximity measure of interest is a function of Euclidean distance. Panel B gives a simplified version of the setting consider by Bloom et al. (2013). Here, the coordinates S_i take values on a simplex and measure the proportion of each firm's patents filed in each of three technology classes. The proximity measure of interest is given by the inner product $D_{i,j} = \langle S_i, S_j \rangle$. In both cases, the level sets (3.3.3) are displayed with dotted lines. The values of the points $S_j + \alpha(S_i - S_j)$, as α varies, are displayed with dashed lines. The points (3.3.4), where these lines intersect, are marked with dots.

---

identified by the regression (3.2.6). However, unless the potential outcome inside the conditional expectation (3.3.9) is independent of the treatment W_j and proximity measure $D_{i,j}$, conditional on the factor S_j, the average

$$\mathbb{E}[\partial^2_{\delta,w} Y_{i,j}(\delta, w) \mid S_j] \tag{3.3.10}$$

and the parameter (3.3.9) will be, in general, different. In practice, this difference can be large.

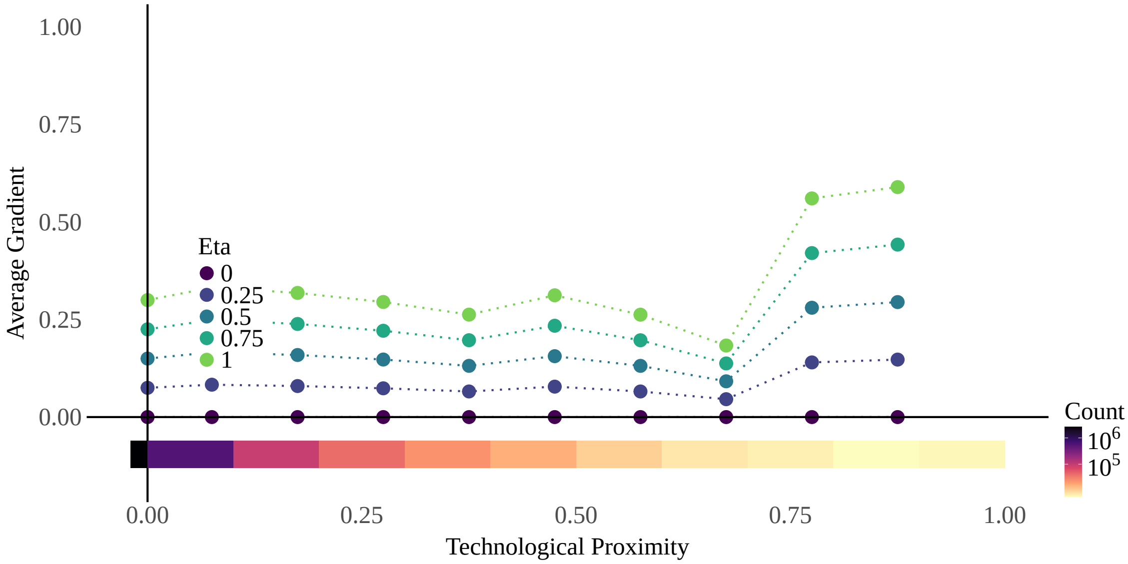
Example 4 (Continued)**.** Suppose, for the sake of illustration, that the outcomes in Bloom et al. (2013) were generated by the specification

$$Y_i = \sum_{j \neq i} (\theta \cdot D_{i,j} - \eta \cdot G_{i,j}) W_j \ , \tag{3.3.11}$$

where we recall that $D_{i,j}$ and $G_{i,j}$ denote the technological and product market proximity between the firms i and j, respectively.[18] Assume also, for the sake of simplicity, that the

[18] In our notation, product market proximity is generated by $G_{i,j} = G(U_i, U_j)$. The product market proximity measure $G_{i,j}$ is given by the $G_{i,j} = \langle R_i, R_j \rangle / \|R_i\|_2 \|R_j\|_2$, where R_i denotes the share of firm i sales across four digit industry codes. Thus, the variable R_i is a component of the variable U_i.

Figure 3.3: Confounding in a Calibrated Example



Notes: Figure 3.3 displays estimates of the mixed-partial derivative (3.3.12), under the calibrated model (3.3.11), averaged over bins of the support of technological proximity. We estimate $\partial_\delta\, \mathbb{E}[G_{i,j} \mid D_{i,j} = \delta]$ at each bin center with a local-linear regression using a triangular kernel with bandwidth 0.1. We set the coefficient $\theta = 1$. Each series corresponds to a specified value of the coefficient η, which parametrizes the extent of the confounding with product market proximity. A heatmap measuring the number of pairs of units in each bin are displayed below the x-axis.

---

treatments W_j are independent of variables S_j and U_j. As the magnitude of the coefficient η increases in proportion to the coefficient θ, the confounding effect of product market proximity increases. To see this, Figure 3.3 displays estimates of the value of the quantity

$$\partial^2_{\delta,w}\mathbb{E}[Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, W_j = w] = \theta - \eta \cdot \partial_\delta\, \mathbb{E}[G_{i,j} \mid D_{i,j} = \delta] \tag{3.3.12}$$

averaged over bins of the support of technological proximity using the sample from Bloom et al. (2013). We set $\theta = 1$ and vary the value of η. By contrast, the average spillover proximity gradient satisfies

$$\mathbb{E}[\partial^2_{\delta,w} Y_{i,j}(\delta, w)] = \theta \tag{3.3.13}$$

irrespective of the coefficient η. Thus, in this sense and in this calibrated example, estimates of the effect of technological proximity on productivity spillover intensity that do not account for the potential that product market proximity affects spillover intensity can be severely negatively biased. ■

It is worth pausing to emphasize what spillover proximity gradients do and do not measure. First, averages of spillover proximity gradients are related to the "average indirect effect" and "marginal spatial effect,"

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\partial_w Y_{i,j}(D_{i,j},w) \quad \text{and} \quad \mathbb{E}[\partial_w Y_{i,j}(D_{i,j},w) \mid D_{i,j}=\delta]\ , \tag{3.3.14}$$

studied by Hu et al. (2022); Munro et al. (2025); Li and Wager (2022) and Pollmann (2023); Wang et al. (2025), respectively. The main difference, in both cases, is that we explicitly consider interventions to the proximity measure $D_{i,j}$. Second, spillover proximity gradients are related to the "average exposure contrasts," considered by, e.g., Aronow and Samii 2017, Sävje et al. 2021, Auerbach and Tabord-Meehan 2025, and Leung and Loupos 2025, where the main interest is in intervening on a low dimensional summary of the treatments and connections of "neighboring" units. Here, the low dimensional summary is often chosen to parameterize (or approximate) the global structure of interference across outcomes. By contrast, our aim is estimate average changes in one unit's outcome associated with changes to the treatment and proximity of another unit, while remaining agnostic, a priori, toward the global structure of interference.

Finally, in some cases, coefficients obtained from regressions of the form (3.2.1) have been interpreted as estimates of "total effects"; that is, parameters measuring the effect of changing the treatment of all units, in unison. Theorem 3.2.1 indicates that, without placing further structure on the data at hand, estimators based on specifications of the form (3.2.6) target a more modest quantity, again relating to the effect of changing the proximity and treatment of only one unit. Indeed, although, as we will see, average spillover proximity gradients can be estimated nonparametrically, recovering richer objects demands the imposition of further structure, either on the distribution of the treatments or on how how interference aggregates across units.[19]

3.4. Identification and Estimation

We propose a method for identifying and estimating averages of spillover proximity gradients. The core idea is to adjust the regressions considered in Section 3.2 by residualizing the treatment and proximity measure of interest with appropriately chosen covariates and auxiliary proximity measures. We give conditions that rationalize estimation strategies with

[19]There are many papers in the network interference literature that aim to estimate total effects. In most cases, this is enabled by restricting attention to settings where the realized treatments are clustered (see e.g., Leung 2022b; Faridani and Niehaus 2022; Viviano et al. 2023) or by restricting the ways in which interference aggregates across units (Munro, 2025).

this structure.

3.4.1 Regression Adjustment

Any approach for recovering averages of spillover proximity gradients must account for potential endogeneity in both the treatments and proximity measures of interest. We consider strategies that address these two sources of confounding by separately residualizing the two associated variables.

To this end, we assume that we additionally observe the vectors of covariates X_i and auxiliary proximity measures $G_{i,j}$. For instance, in Example 4, the covariates X_i might denote features of the firm i, like lagged values of R&D expenditure, and the proximity measure $G_{i,j}$ might denote the product market proximity and geographic distance between the firms i and j. Define the vector $H_{i,j} = (X_i, G_{i,j}, X_j)$. We impose the following convention to fit these measurements into the model proposed in Section 3.3. Let the vector

$$\overline{S}_i = (S_i, U_i) \tag{3.4.1}$$

collect the two latent factors associated with unit i.

Assumption 3.4.1. *There exists a function $H(\cdot,\cdot)$ such that $H_{i,j} = H(\overline{S}_i, \overline{S}_j)$.*

Assumption 3.4.1 stipulates that the covariates $H_{i,j}$ are functions of the unit-specific factors $\overline{S}_i$ and $\overline{S}_j$. This condition rules out covariates that summarize features of units other than i and j.

The proposed method has two steps. First, the user constructs estimates of the conditional expectations

$$\pi_n(x) = \mathbb{E}[W_j \mid X_j = x] \quad \text{and} \quad \gamma_n(h) = \mathbb{E}[D_{i,j} \mid H_{i,j} = h]\ , \tag{3.4.2}$$

respectively. Denote these estimates by $\hat{\pi}_n(\cdot)$ and $\hat{\gamma}_n(\cdot)$. Second, the user computes the coefficients from the regression specification

$$Y_i = \alpha + \theta \cdot \hat{\Delta}_i^\star + \varepsilon_i\ , \quad \text{where} \quad \hat{\Delta}_i^\star = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star\ . \tag{3.4.3}$$

Here, the variables

$$\hat{W}_i^\star = W_i - \hat{\pi}_n(X_j) \quad \text{and} \quad \hat{D}_{i,j}^\star = D_{i,j} - \hat{\gamma}_n(H_{i,j}) \tag{3.4.4}$$

denote residualized versions of the treatment and proximity measure of interest.[20] In this section, we give conditions under which endogeneity in the treatments and proximity measures is addressed by the use of the residualized variables (3.4.4).

Remark 3.4.1. The estimator obtained by solving the residualized regression (3.4.3) differs from the types of regressions usually encountered in empirical practice. In particular, specifications of the form

$$Y_i = \alpha + \theta \cdot \Delta_i + \xi \cdot \Gamma_i + \varepsilon_i \ , \quad \text{where} \quad \Delta_i = \sum_{j \neq i} D_{i,j} W_j \quad \text{and} \quad \Gamma_i = \sum_{j \neq i} G_{i,j} W_j \ , \tag{3.4.5}$$

are considered by Conley and Udry (2010), Bloom et al. (2013), Acemoglu et al. (2016a), and Lerche (2025), among many others. (Here, for the sake of simplicity, we have omitted consideration of the covariates X_i.) In contrast to cross-sectional regressions, the residualized spillover regression (3.4.3) and the "long" spillover regression (3.4.5) are not numerically equivalent and do not identify the same parameter, even when the expectation of $D_{i,j}$ conditional on $G_{i,j}$ is linear. In Appendix C.5.3, we show that, under this linearity condition, the two regressions identify the same parameter only if treatments are homoskedastic. ■

We focus the majority of our discussion on addressing endogeneity in measures of proximity, as our approaches for handling endogeneity in treatments are standard. In particular, in the main text, we assume that treatments are as-good-as randomly assigned conditional on the covariates X_j. In Appendix C.5.4, we show that this condition can be replaced with analogous assumptions based on parallel trends or instrumental variable identification strategies by making appropriate changes to the structure of the estimator (3.4.3).

Assumption 3.4.2 (Conditional Ignorability in Treatments)**.**
The replicates $(W_i, \overline{S}_i)_{i=1}^n$ are i.i.d. and satisfy the condition

$$\mathbb{E}[W_j \mid X_j, \overline{S}_j] = \mathbb{E}[W_j \mid X_j] \tag{3.4.6}$$

[20] In practice, different sets of covariates X_i can be used to residualize the treatments and proximity measures. In particular, in settings where the coordinates S_i are observed, they can be used to residualize the treatments, but should not be used (without being coarsened in some way) to residualize the proximity measures, as otherwise there would be no variability in the residuals $\hat{D}^\star_{i,j}$. We assume that the same set of covariates are used in both cases, for the sake of simplicity.

almost surely.

Assumption 3.4.2 is a direct generalization of Assumption 3.2.2. Like Assumption 3.2.2, cross-sectional dependence in the treatments has been ruled out.[21] In this case, however, the means of the treatments are allowed to vary with the latent factor, so long as this heterogeneity can be explained by the observed covariates X_j.

3.4.2 Decomposition, Redux

To discipline our discussion, we first state an extension to Theorem 3.2.1. In particular, we consider the infeasible objective function

$$L_n^\star(\theta) = \min_\alpha \sum_{i=1}^n (Y_i - \alpha - \theta \cdot \Delta_i^\star)^2 \ , \quad \text{where} \quad \Delta_i^\star = \sum_{j \neq i} D_{i,j}^\star W_j^\star \tag{3.4.7}$$

is constructed using the oracle residuals

$$W_i^\star = W_i - \pi_n(X_i) \quad \text{and} \quad D_{i,j}^\star = D_{i,j} - \gamma_n(H_{i,j}) \ . \tag{3.4.8}$$

We give conditions under which regressions of the form (3.4.7) identify convex averages of mixed-partial derivatives of the conditional expectation

$$\mu(\delta, w \mid H_{i,j}, \overline{S}_j) = \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j] \ . \tag{3.4.9}$$

In Section 3.5, in turn, we give conditions under which the coefficient obtained from the regression (3.4.3) is consistent for the parameter that minimizes the expectation of the objective function (3.4.7).

We require a final condition that justifies our approach to residualizing the proximity measure. In particular, we assume that any unobserved heterogeneity in the relationship between the proximity measure and the covariates is additively separable.

Assumption 3.4.3 (Additively Separable Heterogeneity)**.** *There exists a function $\psi_n(\cdot)$ such*

[21] Peer effects in treatment selection are considered by Forastiere et al. (2021), Ogburn et al. (2024), and Emmenegger et al. (2025), who study models where treatment is as-good-as randomly assigned conditional on a low-dimensional vector collecting features of neighboring units. Leung and Loupos (2025) consider a more general model where treatments are jointly determined, conditional on an observed network and covariates. In each case, in contrast to the regression-based approaches at the center of this paper, explicit specification, or estimation, of the cross-sectional dependence in treatments and network dependence across outcomes is required.

that

$$\mathbb{E}[D_{i,j} \mid H_{i,j}, S_j] = \gamma_n(H_{i,j}) + \psi_n(S_j) \tag{3.4.10}$$

almost surely.

This condition allows us to residualize the proximity measures $D_{i,j}$ with an estimate of the expectation $\gamma_n(H_{i,j}) = \mathbb{E}[D_{i,j} \mid H_{i,j}]$, rather than with an estimate of the expectation $\mathbb{E}[D_{i,j} \mid H_{i,j}, S_j]$. The former can be obtained using the full sample, whereas the latter would need to be re-estimated for each unit j.[22] In particular, recall from the discussion in Remark 3.2.2 that regressions with the structure (3.2.6) effectively center the proximity measures $D_{i,j}$ conditional on the latent factors S_j. Regressions with the structure (3.4.7) center the residualized proximity measures $D^\star_{i,j}$ analogously. Thus, Assumption 3.4.3 ensures that, by centering the residual $D^\star_{i,j}$ conditional on the latent factor S_j, we effectively center the proximity measure $D_{i,j}$, conditional on both the covariates $H_{i,j}$ and the latent factor S_j.

The following Theorem gives a direct generalization of Theorem 3.2.1. As before, we require several moment conditions. We additionally impose a mild generalization of Assumption 3.2.3. In each case, we now condition on the covariates $H_{i,j}$. Again, we defer the statement of these conditions to Appendix C.1.

Theorem 3.4.1. *Assume that the treatments and proximity measures have support contained in intervals $\overline{\mathcal{W}}$ and $\overline{\mathcal{D}}$ of constant length and that the outcomes are identically distributed. Under Assumptions 3.2.1 and 3.4.1 to 3.4.3, and two regularity conditions stated in Appendix C.1, the parameter*

$$\overline{\theta}^\star_n = \arg\min_\theta \mathbb{E}[L^\star_n(\theta)] \tag{3.4.11}$$

admits the representation

$$\begin{aligned} \overline{\theta}^\star_n &= \theta^\star_n + o(n^{-1/2})\,, \quad \textit{where} \\ \theta^\star_n &= \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j)\, \partial^2_{\delta,w}\, \mu(\delta, w \mid H_{i,j}, \overline{S}_j)\Big]\, \mathrm{d}w\, \mathrm{d}\delta\,, \end{aligned} \tag{3.4.12}$$

[22] If Assumption 3.4.3 is unlikely to be accurate, then the results stated in this section will hold, unchanged, if an estimate of $\mathbb{E}[D_{i,j} \mid H_{i,j}, S_j]$ is used to residualize the proximity measure $D_{i,j}$, in place of the conditional expectation $\gamma_n(H_{i,j})$. Such an estimate can be obtained by regressing $D_{i,j}$ on $H_{i,j}$ separately for each unit j. In our empirical applications, we prefer to impose Assumption 3.4.3 because such unit-by-unit estimates are less precise and are more computationally expensive.

and the weights

$$\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) = \frac{\mathrm{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j} \mid H_{i,j}, S_j)\,\mathrm{Cov}(\mathbb{I}\{W_j \geq w\}, W_j \mid \overline{S}_j)}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid H_{i,j}, S_j)\,\mathrm{Var}(W_j \mid \overline{S}_j)]} \tag{3.4.13}$$

are convex, in the sense that they are positive almost surely and integrate to one in expectation.

Corollary 3.4.1. *Assume that the conditions of Theorem 3.4.1 hold. If, additionally, Assumption 3.3.1 holds and the potential outcomes $(\delta, w) \mapsto Y_{i,j}(\delta, w)$ have continuous first and second partial derivatives, then the parameter $\theta_n^\star$, defined in (3.4.12), can be re-expressed as*

$$\theta_n^\star = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j)\, \partial_\delta\, \mathbb{E}[\partial_w\, Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j]\Big]\, \mathrm{d}w\, \mathrm{d}\delta\ , \tag{3.4.14}$$

where the weight function $\lambda(\delta, w \mid H_{i,j}, \overline{S}_j)$ is defined in (3.4.13).

Theorem 3.4.1 is entirely analogous to Theorem 3.2.1, except that every quantity now additionally conditions on the covariates $H_{i,j}$ and the latent factor $\overline{S}_j$.[23] Corollary 3.4.1, says that, if additionally Assumption 3.3.1 holds, then regressions of the form (3.4.7) identify convex averages of the parameter

$$\partial_\delta\, \mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j]\ . \tag{3.4.15}$$

As a consequence, if the covariates $H_{i,j}$ and latent coordinate $\overline{S}_j$ are sufficient to hold constant features of the unit i that might covary with both its proximity $D_{i,j}$ to unit j and its response to j's treatment, then the regression specification (3.4.7) recovers convex averages of spillover proximity gradients.

3.4.3 Conditional Ignorability

Our baseline approach for addressing endogeneity in measures of proximity centers on the following conditional ignorability assumption.

[23] As it can be shown that $\int\int \lambda(\delta, w \mid H_{i,j}, \overline{S}_j)\, \mathrm{d}w\, \mathrm{d}\delta = \mathrm{Var}(D_{i,j} \mid H_{i,j}, S_j)\,\mathrm{Var}(W_j \mid \overline{S}_j)$, the weighted average (3.4.17) tends to place larger weight on the pairs of units i, j associated with larger values of $\mathrm{Var}(D_{i,j} \mid H_{i,j}, S_j)$. This echos classical results from Angrist (1998) and Angrist and Krueger (1999) associated with contexts without interference.

Assumption 3.4.4 (Conditional Ignorability in Proximity)**.** *The potential spillover effect $\partial_w Y_{i,j}(\delta, w)$ associated with an intervention to the treatment of unit j on the outcome of unit i is mean independent of the realized proximity measure $D_{i,j}$, conditional on the covariates $H_{i,j}$ and the latent factor $\overline{S}_j$; that is*

$$\mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j] = \mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid H_{i,j}, \overline{S}_j] \tag{3.4.16}$$

almost surely for each δ, w in the support of $D_{i,j}$ and W_j.

Under Assumption 3.4.4, the parameter $\theta_n^\star$, targeted by regressions with the structure (3.4.7), identifies a convex average of spillover proximity gradients.

Corollary 3.4.2. *Assume that the conditions of Corollary 3.4.1 hold. If, additionally, Assumptions 3.3.2 and 3.4.4 hold, then the parameter $\theta_n^\star$, defined in* (3.4.12)*, admits the representation*

$$\theta_n^\star = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \partial^2_{\delta,w} Y_{i,j}(\delta, w)\Big] \, \mathrm{d}w \, \mathrm{d}\delta \ , \tag{3.4.17}$$

where the weight function $\lambda(\delta, w \mid H_{i,j}, \overline{S}_j)$ is defined in (3.4.13).

Example 4 (Continued)**.** In the context of Bloom et al. (2013), the condition (3.4.16) stipulates that, if we fix a firm j, once we account for any observable features X_i and, e.g., the product market proximity $G_{i,j}$, then technological proximity $D_{i,j}$ doesn't vary systematically with any other quantities that might affect how firm i's value responds to R&D by firm j. This condition is satisfied, for instance, in the model (3.3.11) used to generate Figure 3.3. Bloom et al. (2013) use language suggestive of this assumption, arguing that their identification rests on "substantial independent variation in the two measures," namely product market and technological proximity.[24] ■

Assumption 3.4.4 asks that the covariates $H_{i,j}$ stratify units i in terms of the characteristics that influence the intensity of the spillover effects they receive from j's treatment, up to the realized proximity measure $D_{i,j}$. To visualize this restriction, return to Panel A of Figure 3.2. The figure displays three coordinates S_i, S_j, and S_k. Assume that i and k have the same covariates and the same relation to j in terms of the auxiliary proximity measures, that is $G_{i,j} = G_{k,j}$. Assumption 3.4.4 implies that the difference in the proximity measures $D_{i,j}$

[24] Bloom et al. (2013) support this assertion with a series of illustrative examples, writing e.g., "In our sample period, IBM was technologically close to Intel, Motorola, and Apple ... However, while IBM is close to Apple in product market space, ... [it] is not close to Intel and Motorola."

and $D_{k,j}$ cannot be correlated with any differences in the characteristics of the units i and k that influence the intensity of the spillover effects that they receive from the unit j. So, for instance, if all else equal, units that are further north tend to receive larger spillovers from unit j (perhaps, because they are closer in terms of some unobserved characteristic), and also tend to be closer to S_j, then, unless the covariates $H_{i,j}$ and $\overline{S}_j$ fix unit i's north-south coordinate, Assumption 3.4.4 will not hold.

As in other contexts in causal inference, conditional ignorability is a strong assumption. That is, in any particular setting, it may not be too difficult to hypothesize further channels that might also mediate spillover effects. For instance, Lychagin et al. (2016), incorporate geography into the empirical framework developed in Bloom et al. (2013), arguing that technological and geographic proximity are highly correlated and that R&D spillovers are potentially mediated by geographic proximity (see also Arque-Castells and Spulber (2022) and Zacchia (2020) who study how R&D spillovers vary with voluntary technology transfers and patent co-authorship networks, respectively). Nevertheless, we center our consideration on conditional ignorability because it represents the simplest departure from the assumption that spillovers are entirely mediated by the proximity measure of interest that enables identification, and because, in our view, it best reflects the approaches to disentangling spillover effects that have been taken in empirical practice. For instance, Lerche (2025) measures how the spillover effects of investment tax credits vary with input-output linkages, but controls for proximity in labor market flows. See also Conley and Udry (2010), Acemoglu et al. (2016a), Rotemberg (2019), Helm (2020), and Cai and Szeidl (2024) who consider similar problems in other settings.

Moreover, conditioning on observables is an essential building block for more nuanced identification strategies. Concretely, in Appendix C.5.5, we demonstrate how to adapt our framework to treat settings where suitable instrumental variables are available. For example, Fafchamps and Quinn (2018) conduct an experiment where social connections between managers of African manufacturing firms are exogenously formed. They then measure how business practices propagate through the exogenous linkages. Roughly speaking, under restrictions resembling standard conditions for the validity of linear two-stage least-squares estimators (Imbens and Angrist, 1994; Angrist et al., 2000), the residualized proximity measures can be replaced in the estimator (3.4.3) with their values predicted on the basis of the instrument.

3.5. CONSISTENCY

In this section, we give conditions under which the class of estimators proposed in Section 3.4 are consistent for convex averages of spillover proximity gradients. We focus our consideration on settings where each unit's outcome can be potentially affected by the treatments, locations, and unobserved characteristics of many other units. We show that, in such settings, coefficients obtained from regression specifications of the form (3.4.3) converge at a rate strictly slower than $O_p(n^{-1/2})$. In particular, the rate of convergence achieved by the estimator under consideration decreases as the extent of cross-unit interference increases. Nevertheless, we establish that, under our regularity conditions, this sub-parametric rate is optimal, in a minimax sense.

Throughout, the quantities c and C denote universal, positive constants, whose values are allowed to change in each appearance. For two sequences $f(n)$ and $g(n)$, the relations $f(n) \lesssim g(n)$ and $f(n) \asymp g(n)$ indicate that $f(n) \le Cg(n)$ and $cg(n) \le f(n) \le Cg(n)$ for all sufficiently large n, respectively.

3.5.1 Linear Residualization

To facilitate these results, and in keeping with our emphasis on regression-based estimation procedures, we restrict attention to settings where the estimators $\hat{\pi}_n$ and $\hat{\gamma}_n$ are obtained from least-squares regressions. In particular, with an abuse of notation, we consider the estimator

$$\hat{\theta}_n^\star = \arg\min_\theta \min_\alpha \sum_{i=1}^n (Y_i - \alpha - \theta \cdot \hat{\Delta}_i^\star)^2 \,, \quad \text{where} \quad \hat{\Delta}_i^\star = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star \tag{3.5.1}$$

has been constructed using the empirical residuals

$$\hat{W}_i^\star = W_i - \hat{\pi}_n^\top X_i \quad \text{and} \quad \hat{D}_{i,j}^\star = D_{i,j} - \hat{\gamma}_n^\top H_{i,j} \tag{3.5.2}$$

centered with the linear regression coefficients

$$\hat{\pi}_n = \arg\min_\pi \sum_{i=1}^n (W_i - \pi^\top X_i)^2 \quad \text{and} \quad \hat{\gamma}_n = \arg\min_\gamma \sum_{i=1}^n \sum_{j \neq i}^n (D_{i,j} - \gamma^\top H_{i,j})^2 \,, \tag{3.5.3}$$

respectively. Here, we adopt the convention that both X_i and $H_{i,j}$ contain intercepts.

Moreover, we assume that the estimators (3.5.3) are well-specified.

Assumption 3.5.1 (Fixed Dimension, Linear Nuisance Parameters). *The nuisance parameters $\pi_n(x)$ and $\gamma_n(h)$ are linear in their arguments and the covariate vector $H_{i,j}$ has a fixed dimension.*

The linearity enforced by Assumption 3.5.1 plays an essential role in our arguments. In settings without interference, the use of more generic nuisance parameter estimators is increasingly common. Often, such approaches are enabled by the use of sample-splitting (Chernozhukov et al., 2018). Sample-splitting is not generally appropriate for our setting, because the structure of the cross-sectional dependence across units is not known a priori.[25] As a practical matter, the covariate vector $H_{i,j}$ can be constructed by taking a basis expansion or binning the observable covariates, so long as the number of components remains small.

3.5.2 Regularity Conditions

Our asymptotic analysis centers on the idea that cross-unit interference is non-negligible only for pairs of units that are "proximate," either through the measure of interest or through other observed or unobserved characteristics. To give structure to this idea, we assume that there exists a symmetric, binary-valued function

$$A_{i,j} = A_n(\overline{S}_i, \overline{S}_j) = A_n(\overline{S}_j, \overline{S}_i) \in \{0, 1\} \tag{3.5.4}$$

of the latent factors that indicates whether the units i and j are proximate. The latent proximity indicator $A_{i,j}$ is used to encode two relationships: interference arises only among proximate pairs of units, and the measure of interest $D_{i,j}$ is nonzero only for such pairs. For the latter, the converse is not necessarily true. We express these assumptions formally as follows.

Assumption 3.5.2 (Latent Spatial Structure). *Define the sets*

$$L_n(S_j) = \{s : A_n(s, \overline{S}_j) = 1\} \quad \text{and} \quad L_n^{(1)}(\overline{S}_i, \overline{S}_j) = L_n(\overline{S}_i) \cap L_n(\overline{S}_j) \tag{3.5.5}$$

for each pair of latent factors $\overline{S}_i$, $\overline{S}_j$. There exists a sequence ρ_n, with $\rho_n \to 0$ and $n\rho_n \to \infty$,

[25] Leung and Loupos (2025) consider a more complicated nuisance parameter estimator in a setting with interference. Their argument is not directly applicable to our setting, as our data, nuisance parameters, and estimator have different structures.

that satisfies

$$P\{\overline{S}_j \in L_n(\overline{S}_i) \mid \overline{S}_i\} \asymp \rho_n \quad \textit{and} \quad P\{P\{\overline{S}_k \in L_n^{(1)}(\overline{S}_i, \overline{S}_j) \mid \overline{S}_i, \overline{S}_j\} > 0\} = O(\rho_n) \ , \tag{3.5.6}$$

uniformly almost surely, such that:

(i) *Outcomes are nearly mean independent of the states of distant units, in the sense that*

$$\textit{if} \quad A_{i,j} = 0, \quad \textit{then} \quad \mathbb{E}[Y_i \mid Z_j, Z_i] - \mathbb{E}[Y_i \mid Z_i] = o(\rho_n^{1/2}) \tag{3.5.7}$$

uniformly almost surely.

(ii) *The proximity measure $D_{i,j}$ is equal to zero for distant units and is non-negligible, with a non-vanishing probability, for proximate units, in the sense that*

$$P\{D_{i,j} = 0 \mid A_{i,j} = 0\} = 1 \quad \textit{and} \quad \mathrm{Var}(D_{i,j} \mid A_{i,j} = 1, S_j) \asymp 1 \tag{3.5.8}$$

hold uniformly almost surely.

(iii) *Covariates $H_{i,j}$ associated with distant units are not predictive of proximity, in the sense that*

$$\textit{if} \quad A_{i,j} = 0 \ , \quad \textit{then} \quad \mathbb{E}[D_{i,j} \mid H_{i,j}, S_j] = O(\rho_n) \ , \tag{3.5.9}$$

uniformly almost surely.

The sequence ρ_n measures the expected fraction of units that are proximate to any given unit j. In the first part of the relation (3.5.6), we have imposed the condition that this fraction can vary across units only up to a constant factor. The conditions $\rho_n \to 0$ and $n\rho_n \to \infty$ restrict our consideration to asymptotic sequences in which the number of units that are proximate to each given unit grows without bound, but remains small relative to the sample size. The second part of the condition (3.5.6) stipulates that the fraction of pairs of units i, j that could feasibly be proximate to a third unit k is also of order $O(\rho_n)$. This restriction should be interpreted as encoding the idea that i and j can only *ever* share a common neighbor (in the sense that $P\{\overline{S}_k \in L_n^{(1)}(\overline{S}_i, \overline{S}_j) \mid \overline{S}_i, \overline{S}_j\}$ has a positive probability) when they themselves are sufficiently "close" in the latent space. This event, that i and j are sufficiently close to feasibly share a neighbor, cannot, in turn, be *much* more likely to occur than the event that i and j are, themselves, neighbors.

For the sake of concreteness, consider the setting where the latent proximity indicators are generated by

$$A_n(\overline{S}_i, \overline{S}_j) = \mathbb{I}\{\|\overline{S}_i - \overline{S}_j\| \leq \phi_n\}\ , \tag{3.5.10}$$

where $\|\cdot\|$ is some norm (or semi-norm) on the latent factors and ϕ_n is a sequence of bandwidths. Networks with the structure (3.5.10) are referred to as "random geometric graphs," and are widely studied in the network interference literature (Penrose, 2003; Leung, 2022a). In this case, the second part of the condition (3.5.6) will be satisfied when the measure of the latent factors is doubling with respect to the norm $\|\cdot\|$, in the sense that $P\{\|\overline{S}_i - \overline{S}_j\| \leq 2\phi_n\} \lesssim P\{\|\overline{S}_i - \overline{S}_j\| \leq \phi_n\}$. Doubling is a common feature of well-behaved measures, including distributions with densities bounded above and below on finite-dimensional Euclidean spaces (Heinonen, 2001). In this sense, the condition (3.5.6) asks that the latent proximity indicator $A_{i,j}$ behaves like its name, that is, it should reflect an underlying measure of proximity.[26]

Part (i) of Assumption 3.5.2 stipulates that only units that are proximate to a given unit can have non-negligible effects on the expected value of its outcome. This restriction is closely related to the "Approximate Neighborhood Interference" framework developed in Leung (2022a), and adopted in, e.g., Leung (2022b) and Leung and Loupos (2025).[27] Crucially, in our setting, interference is parameterized by the latent proximity measure $A_{i,j}$, rather than through the observed proximity measure $D_{i,j}$, per se. That is, we explicitly allow for spillover effects, and other forms of interference, to be mediated by other observed or unobserved characteristics. Like in Leung (2022a), because we have not imposed the restriction that outcomes that are sufficiently far away are independent, our framework accommodates general forms of "endogenous" spillover effects, in the sense of Manski (1993).

In each of our running examples, the proximity measure $D_{i,j}$ is equal to zero for most pairs of units i, j. That is, the measure $D_{i,j}$ is sparse. For instance, in our treatment of the data associated with Example 4, 92% of pairs firms never file patents in the same technology class, and so their technological proximity is equal to zero.[28] Parts (ii) and

[26] In the region $n^{-1/2} \lesssim \rho_n$, the second part of the relation (3.5.6) will fail in sparse graphon models without a geometric structure (like those considered in Li and Wager (2022)). That is, in that context, in general, the size of the intersection of the neighborhoods of two randomly drawn nodes is of order $n\rho_n^2$.

[27] The Approximate Neighborhood Interference framework is more expressive than Assumption 3.5.2, in that, there, consistency is established by measuring how the variability in a unit's outcome changes as one progressively resamples the treatments of all units whose distances exceed an increasing threshold. By contrast, our approach (effectively) fixes a single threshold and measures dependence by resampling the latent factor of just one unit whose distance exceeds that threshold.

[28] In Examples 5 and 6, the measure $D_{i,j}$, which measures the probability of migrating or commuting between the pair of location, is smaller than 0.01 for 93% and 92% pairs of locations, respectively

(iii) of Assumption 3.5.2 relate the indicator $A_{i,j}$ to the sparsity of the proximity measure of interest and its association with the covariates. In particular, Part (ii) requires that the measure $D_{i,j}$ is only ever non-zero for units whose latent factors are proximate. Part (iii), in turn, requires that values of the covariates $H_{i,j}$ that predict large values of $D_{i,j}$ only occur among proximate units. In other words, the binary, latent proximity measure $A_{i,j}$ respects the structure of the observed proximity measures $D_{i,j}$ and $H_{i,j}$, while allowing for interference between units that are only proximate through unobserved characteristics.

Likewise, many of the covariates $H_{i,j}$ that are potentially relevant to our running examples are equal to zero for the majority of pairs of units. To accommodate such situations, we require the following condition concerning the sparsity and correlation of the components of the covariates.

Assumption 3.5.3 (Well-Behaved Covariate Sparsity)**.**
(i) *Each component of the covariate vector $H_{i,j} = (H_{i,j}^{(k)})_{k=1}^p$ is bounded by a constant almost surely. Moreover, each of the components of the covariate vector X_i has variance bounded below by a constant.*
(ii) *For each component $H_{i,j}^{(k)}$ of the covariate vector $H_{i,j}$ there exists a sequence $\kappa_{n,k} \gtrsim \rho_n$ such that*

$$P\{H_{i,j}^{(k)} \neq 0 \mid \overline{S}_j\} \gtrsim \kappa_{n,k}\ , \quad P\{H_{i,j}^{(k)} \neq 0 \mid \overline{S}_i\} \gtrsim \kappa_{n,k}\ , \quad \textit{and} \quad \mathrm{Var}(H_{i,j}^{(k)}) \gtrsim \kappa_{n,k}\ . \tag{3.5.11}$$

The heterogeneity in the sparsity, in turn, is well-behaved, in the sense that

$$\sum_{\ell=1}^{p} |K_{k,\ell}^{-1}| = O(1) \quad \textit{and} \quad \sum_{\ell=1}^{p} \sqrt{\frac{\kappa_{n,k}}{\kappa_{n,\ell}}} |K_{k,\ell}^{-1}| = O(1) \tag{3.5.12}$$

both hold for each k, where $K_{k,\ell} = \mathrm{Corr}(H_{i,j}^{(k)}, H_{i,j}^{(\ell)})$ denotes the k, ℓ component of $K = \mathrm{Corr}(H_{i,j})$.

The restriction (3.5.11) stipulates that none of the covariates in $H_{i,j}$ are strictly more sparse than the proximity measure of interest. This condition is necessary to ensure that the variability in the estimator $\hat{\gamma}_n$ does not, on its own, determine the rate of convergence of the estimator $\hat{\theta}_n^\star$. The first part of the restriction (3.5.12) requires that different covariates are not so strongly collinear that the inverse covariance structure becomes unstable. The second part, in effect, ensures that each component of $\hat{\gamma}_n$ can be estimated at a rate that is determined by its own sparsity, rather than being driven by association with other covariates with different levels of sparsity.

In turn, we require a condition that limits the curvature of the function $F(\cdot)$.

Assumption 3.5.4 (Curvature). *Let the variable Z_j' denote an independent copy of Z_j. Construct $Z^{(j)}$ by replacing Z_j with Z_j' in the collection $Z = (Z_i)_{i=1}^n$. Let $Z^{(j,k)}$ be constructed analogously, by replacing Z_j and Z_k with Z_j' and Z_k'. Define the difference operators*

$$\nabla_j f(Z) = f(Z) - f(Z^{(j)}) \quad \text{and} \quad \nabla^2_{j,k} f(Z) = \nabla_j \nabla_k f(Z) \, , \tag{3.5.13}$$

respectively, for any real-valued function $f(\cdot)$. The bounds

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k} F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}]] = O(n^{-2}\rho_n^{-3}) \, , \\ &\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k} F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 1, A_{i,k} = 0, Z^{(j,k)}]] = O(n^{-2}\rho_n^{-1}) \, , \quad \text{and} \\ &\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k} F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 0, A_{i,k} = 0, Z^{(j,k)}]] = O(n^{-2}\rho_n) \end{aligned} \tag{3.5.14}$$

hold.

The quantity $\nabla^2_{j,k} F(Z_i, Z_{-i})$ characterizes how the association between i's outcome and j's state varies as k's state changes, and can be interpreted like a randomized version of a mixed-partial derivative. The first bound can be thought of as describing the proportion of pairs of units j, k proximate to i for which this curvature can be on the order of a constant. That is, the curvature can be non-zero and large for all pairs of units proximate to i, so long as $\rho_n \lesssim n^{-2/3}$. As the quantity ρ_n becomes larger, this proportion decreases. For instance, when $\rho_n = n^{-1/2}$, each unit will have, on average, n pairs of neighbors. The bound (3.5.14) implies $n^{1/2}$ of these pairs can be associated curvature on the order of a constant. The second two bounds pertain to the case that only j is proximate to i and that neither j nor k are proximate to i, respectively, and, consequently, get progressively stronger. We discuss this condition in further detail in Appendix C.5.6.[29]

Finally, we require a set of additional, mild moment bounds that restrict the variability of the outcomes on certain events. These bounds are satisfied if, for instance, the random variables $\mathbb{E}[Y_i|Z_i, Z_j]$ and $\mathbb{E}[Y_i|Z_i]$ are bounded almost surely. To facilitate exposition, we defer the statement of this assumption to Appendix C.3.

[29]We note that Assumption 3.5.4 is satisfied by models where outcomes are given by structural equations of the form $Y_i = F_i(W_i, \frac{1}{n\rho_n}\Delta_i)$, where $F_i(\cdot,\cdot)$ is some unit-specific function, so long as the second derivative of $F_i(W_i,\cdot)$ does not grow too quickly. Models with this structure are considered in, e.g., Li and Wager (2022).

3.5.3 Consistency

The following Theorem demonstrates that, under the conditions developed in this section, the estimator $\hat{\theta}_n^\star$, obtained by minimizing the objective function (3.4.3), is consistent for the parameter $\theta_n^\star$, characterized in Theorem 3.4.1. The rate of convergence is determined by the sequence ρ_n, introduced in Assumption 3.5.2.

Theorem 3.5.1. *Suppose that the conditions of Theorem 3.4.1 hold. If, additionally, Assumption 3.3.1, Assumptions 3.5.1 to 3.5.4, and an additional regularity condition stated in Appendix C.3 hold, then*

$$\hat{\theta}_n^\star = \theta_n^\star + O_p(\rho_n^{1/2}) \ , \tag{3.5.15}$$

where $\theta_n^\star$ is defined in (3.4.12).

As a consequence, if the conditions of Corollary 3.4.2 are satisfied, the estimator $\hat{\theta}_n^\star$ is consistent for a convex averages of spillover proximity gradients. Observe that none of the conditions needed to establish identification are used to establish consistency. That is, under the assumptions developed in this section, even if the identifying conditions developed in Section 3.4 fail, the estimator $\hat{\theta}_n^\star$ will still consistently recover the descriptive parameter $\theta_n^\star$.

Remark 3.5.1. Theorem 3.5.1 follows by first showing that, under the maintained assumptions, the estimator $\hat{\theta}_n^\star$ can be approximated by the statistic

$$\frac{\sum_{i=1}^n Y_i \tilde{\Delta}_i^\star}{\sum_{i=1}^n (\tilde{\Delta}_i^\star)^2} \ , \quad \text{where} \quad \tilde{\Delta}_i = \sum_{j\neq i} \tilde{D}_{i,j}^\star W_j^\star \quad \text{and} \quad \tilde{D}_{i,j}^\star = D_{i,j} - \mathbb{E}[D_{i,j} \mid H_{i,j}, S_j] \ , \tag{3.5.16}$$

up to an error of order $O_p(n^{-1/2})$. Assumptions 3.5.1 and 3.5.3 are essential for this step. The numerator and denominator of (3.5.16), in turn, satisfy the approximations

$$\frac{1}{n^2}\sum_{i=1}^n Y_i \tilde{\Delta}_i^\star = \mathbb{E}[Y_i \tilde{D}_{i,j}^\star W_j^\star] + O_p(\rho_n^{3/2}) \quad \text{and} \tag{3.5.17}$$

$$\frac{1}{n^2}\sum_{i=1}^n (\tilde{\Delta}_i^\star)^2 = \mathbb{E}[\operatorname{Var}(D_{i,j} \mid H_{i,j}, S_j) \operatorname{Var}(W_j \mid \overline{S}_j)] + O_p(\rho_n^{3/2}) \ , \tag{3.5.18}$$

respectively. Roughly speaking, Theorem 3.5.1 then follows from the facts that the parameter $\theta_n^\star$ can be expressed as the quotient of the expectations in (3.5.17) and (3.5.18) and that the expectation in the approximation to the denominator can be bounded from below by ρ_n.

The most nonstandard aspect of this argument pertains to the approximation (3.5.17). This bound relies on the application of the Hoeffding type decomposition

$$Y_i = \mathbb{E}[Y_i \mid Z_i] + \sum_{j \neq i} (\mathbb{E}[Y_i \mid Z_j, Z_i] - \mathbb{E}[Y_i \mid Z_i]) + R_i \tag{3.5.19}$$

to approximate each outcome. We control the remainder terms R_i by pairing Assumption 3.5.4 with a suitable, higher-order Efron-Stein type inequality, due to Bobkov et al. (2019). After plugging the decomposition (3.5.19) into the numerator of (3.5.16), the leading term can be represented as a degenerate U-statistic of order three (and likewise for the leading term in the denominator). The variability of these quantities can then be controlled through Assumption 3.5.2. ■

3.5.4 Optimality

Theorem 3.5.1 indicates that the rate of convergence of the estimator $\hat{\theta}_n^\star$ is at most $\rho_n^{1/2}$. Recall that, by Assumption 3.5.2, the sequence ρ_n is asymptotically, strictly larger than $1/n$. That is, on the sequences under consideration, each unit has an increasing number of neighbors. Thus, the rate of convergence implied by Theorem 3.5.1 is strictly slower than $n^{-1/2}$, the rate usually achieved by estimates of scalar parameters in settings without interference.

In this section, we show that, despite this, the rate $\rho_n^{1/2}$ is the best that any estimator can achieve, in a minimax sense, in data satisfying our regularity conditions. For the purposes of this subsection, we restrict attention to the setting without covariates X_i or $G_{i,j}$. That is, we are not asking the more complicated question whether our estimator makes efficient use of the information made available through the covariates. Instead, we ask whether the estimator $\hat{\theta}_n^\star$ depends optimally on the extent of cross-unit interference.

To state the result, we require three additional pieces of notation. First, let the set $\mathcal{P}_n(\rho_n)$ denote the class of distributions for the collection of random variables $(Y_i, Z_i)_{i=1}^n$, such that the conditions of Theorem 3.5.1 are satisfied with the sequence ρ_n. Second, we let $\theta_n^\star(P)$ denote the parameter $\theta_n^\star$, evaluated at the distribution P in $\mathcal{P}_n(\rho_n)$. Third, we let the set $\hat{\Theta}$ collect all, potentially randomized, real-valued functions of the observable data $(Y_i, W_i, D_i)_{i=1}^n$, where the vector $D_i = (D_{i,j})_{j \neq i}$ collects the proximity between i and each other unit.[30]

[30]The same result will hold, by an identical argument, if we allow the estimators in the set $\hat{\Theta}$ to use the coordinates S_i.

With this in place, we obtain the following lower bound on the minimax risk

$$\mathfrak{M}_n(\rho_n) = \inf_{\hat{\theta}\in\hat{\Theta}} \sup_{P\in\mathcal{P}_n(\rho_n)} \mathbb{E}_P\left[(\hat{\theta} - \theta_n^\star(P))^2\right]^{1/2}, \tag{3.5.20}$$

where the dependence of the estimators $\hat{\theta}$ on the observable data is left implicit.

Theorem 3.5.2. *It holds that*

$$\mathfrak{M}_n(\rho_n) \gtrsim \rho_n^{1/2}, \tag{3.5.21}$$

for all sufficiently large n.

By Theorem 3.5.2, there are no estimators that converge to the parameter $\theta_n^\star$, in root-mean-squared error, uniformly more quickly, up to constant factors, than the regression-based estimator $\hat{\theta}_n^\star$ proposed in Section 3.4.

Remark 3.5.2. The lower bound (3.5.21) is obtained from an application of the Le Cam (1973) two-point method (see e.g., Wainwright 2019 for a textbook treatment). The argument follows by explicitly constructing two distributions in $\mathcal{P}_n(\rho_n)$ that are close in the total-variation norm, but are associated with diverging values of the parameter $\theta_n^\star(P)$. Our construction is based on partitioning the units into ρ_n^{-1} groups, such that units are only proximate to other units in their group, and specifying the outcomes and latent factors in such a way that units within the same group have exactly the same outcome. In effect, in this construction, there are ρ_n^{-1} i.i.d. observations, giving rise to the minimax rate $\rho_n^{1/2}$ in a natural way.[31] ■

3.6. Inference

Existing approaches to inference in settings with interference take two routes. In some cases, units are partitioned into disjoint groups. If interference across groups is negligible, clustered standard errors can be constructed in the usual way (Leung, 2023). In others, interference is assumed to be bounded above by a decreasing function of an observed proximity measure, and so kernel-based methods are appropriate (see e.g., Kojevnikov 2021; Leung 2022a; Wang et al. 2025).

[31]A similar construction is noted, heuristically, in Li and Wager (2022), in order to suggest that the upper bounds achieved in that paper are unimprovable. In effect, Theorem 3.5.2 formalizes that observation. We note that the upper bounds in Li and Wager (2022), who consider a different but closely related setting, are restricted to the case that $n^{-1/2} \lesssim \rho_n$. By contrast, in our setting, we give matching upper and lower bounds over the full regime $n\rho_n \to \infty$.

Neither approach is directly applicable to our setting. In particular, as we allow interference to be mediated by unobserved characteristics, units cannot necessarily be arranged into groups or otherwise organized around an observed proximity measure. Thus, in this section, we propose an alternative, resampling-based procedure for giving conservative estimates of the uncertainty associated with the estimators proposed in Section 3.4.

3.6.1 Construction

Our proposal is based on resampling the coefficient from a regression specification analogous to (3.5.1), where the outcomes have been replaced by the empirical residuals, in data where the signs of the (residualized) treatments have been flipped at random. Formally, define the empirical residual

$$\hat{\varepsilon}_i = Y_i - \hat{\alpha}_n - \hat{\theta}_n^\star \hat{\Delta}_i^\star \tag{3.6.1}$$

for each unit i in $1, ..., n$, where $\hat{\alpha}_n$ denotes the intercept from the regression (3.5.1) used to construct the estimator $\hat{\theta}_n^\star$. Let the vector $V = (V_i)_{i=1}^n$ collect a sequence of i.i.d. Rademacher random variables. That is, each coordinate V_i is independently, uniformly distributed on $\{-1, 1\}$. Let $\hat{\phi}_n^\star(V)$ denote the coefficient obtained from the regression specification

$$\hat{\phi}_n^\star(V) = \arg\min_{\phi} \min_{\alpha} \sum_{i=1}^n \left(\hat{\varepsilon}_i - \alpha - \phi \cdot \hat{\Delta}_i^\star(V)\right)^2 , \quad \text{where} \quad \hat{\Delta}_i^\star(V) = \sum_{j \neq i} \hat{D}_{i,j}^\star W_j^\star V_j \tag{3.6.2}$$

denotes a version of the treatment exposure $\hat{\Delta}_i^\star$ where the signs of the residualized treatments $\hat{W}_j^\star$ have been flipped at random. We give conditions under which the distribution of the coefficients $\hat{\phi}_n(V)$, over the randomness in the signs V, can be used to construct conservative estimates of the uncertainty associated with the estimator $\hat{\theta}_n^\star$.

In particular, let $V^{(1)}, \ldots, V^{(B)}$ denote B independent copies of the vector V. We show that a conservative, asymptotically level α, confidence interval for the parameter $\theta_n^\star$ can be obtained by reporting the interval

$$\hat{\theta}_n^\star \pm z_{1-\alpha/2}\hat{\sigma}_n , \quad \text{where} \quad \hat{\sigma}_n^2 = \frac{2}{B} \sum_{b=1}^B \left(\hat{\phi}_n^\star(V^{(b)})\right)^2 \tag{3.6.3}$$

and $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of a standard Gaussian distribution. Likewise, the statistic $\hat{\sigma}_n$ can be used to report a conservative standard error. We show that scaling the variance estimator by a factor of 2, in this way, is essential, in the sense that, in some (non-pathological) configurations, the resultant confidence interval is asymptotically exact. This proposal is summarized in Algorithm 2.

The confidence interval (3.6.3) is closely related to both the plug-in methods developed in Adao et al. (2019) and the randomization inference-based procedures considered by Borusyak and Hull (2023). Relative to the former procedure, which has a related structure but it not based on resampling, we show that our proposal is conservative even when the regression (3.5.1) is not well-specified. The latter procedure, in our context, would entail comparing the estimator $\hat{\theta}_n^\star$ to its randomization distribution constructed by repeatedly permuting the residualized treatments.[32] Methods of this form have the appeal of ensuring exact error control for tests of strong null hypotheses in finite samples (Ritzwoller et al., 2024), but, unlike our proposal, will not necessarily control the level of tests of the value of the parameter $\theta_n^\star$.

3.6.2 Gaussian Approximation

We begin by characterizing the limiting distribution of the estimator $\hat{\theta}_n^\star$. To facilitate exposition, We require three additional conditions. We defer the statements of these conditions to Appendix C.3. The first condition is a mild strengthening of the second relation in (3.5.6), stated in Assumption 3.5.2. The second ensures that, when suitably scaled, the variance of the leading term of the estimator $\hat{\theta}_n^\star$ is bounded away from zero. The final restriction is a mild strengthening of the moment bounds needed for Theorem 3.5.1.

Theorem 3.6.1. *Define the population residual and conditional expectation*

$$\varepsilon_i = Y_i - \alpha_n - \theta_n^\star \cdot \Delta_i^\star \quad \textit{and} \quad \tilde{\varepsilon}_{i,j} = \mathbb{E}[\varepsilon_i \mid Z_j, Z_i] - \mathbb{E}[\varepsilon_i \mid Z_i] \ , \tag{3.6.4}$$

where α_n is the mean of Y_i and $\theta_n^\star$ is defined in Theorem 3.4.1. Define the sequence

$$\sigma_n^2 = \frac{1}{2\mathbb{E}[(\tilde{D}_{i,j}^\star W_j^\star)^2]^2} \mathrm{Var}(\mathbb{E}[\tilde{\varepsilon}_{k,i}\tilde{D}_{k,j}^\star \mid Z_i, Z_j]W_j^\star + \mathbb{E}[\tilde{\varepsilon}_{k,j}\tilde{D}_{k,i}^\star \mid Z_j, Z_i]W_i^\star) \ . \tag{3.6.5}$$

Suppose that the conditions of Theorem 3.5.1 hold. If three additional regularity conditions, stated in Appendix C.3, hold, then

$$\sigma_n^{-1}(\hat{\theta}_n^\star - \theta_n^\star) \xrightarrow{d} \mathsf{N}(0,1) \ , \quad \textit{where} \quad \sigma_n \asymp \rho_n^{1/2} \ . \tag{3.6.6}$$

[32]Results similar to those given in this section for a related procedure where the residualized treatments in the regression (3.6.2) are permuted, rather than randomly sign-flipped, will follow from analogous methods, although the details of the supporting arguments would be substantially more involved.

Algorithm 2: Estimation and Inference

Input: Outcome Y_i, treatment W_i, and proximity measure of interest $D_{i,j}$
Covariates X_i and auxiliary proximity measures $G_{i,j}$

Compute the linear regression coefficients

$$(\hat{\alpha}_{w,n}, \hat{\pi}_n) = \arg\min_{\alpha,\pi} \sum_{i=1}^{n} (W_i - \alpha - \pi^\top X_i)^2 \quad \text{and}$$
$$(\hat{\alpha}_{d,n}, \hat{\gamma}_n) = \arg\min_{\alpha,\gamma} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (D_{i,j} - \alpha - \gamma^\top H_{i,j})^2$$

and residuals

$$\hat{W}_i^\star = W_i - \hat{\alpha}_{w,n} - \hat{\pi}_n^\top X_i \quad \text{and} \quad \hat{D}_{i,j}^\star = D_{i,j} - \hat{\alpha}_{d,n} - \hat{\gamma}_n^\top H_{i,j} \ ,$$

where $H_{i,j} = (X_i, G_{i,j}, X_j)$

Compute the linear regression coefficients

$$(\hat{\alpha}_n, \hat{\theta}_n^\star) = \arg\min_{\alpha,\theta} \sum_{i=1}^{n} (Y_i - \alpha - \theta \cdot \hat{\Delta}_i^\star)^2 \ , \quad \text{where} \quad \hat{\Delta}_i^\star = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star \ ,$$

and the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{\alpha}_n - \hat{\theta}_n^\star \hat{\Delta}_i^\star$$

for b *in* $1, \ldots, B$ **do**

Draw a vector $V^{(b)} = (V_i^{(b)})_{i=1}^n$ of i.i.d. Rademacher random variables

Compute the linear regression coefficient

$$\hat{\phi}_n^\star(V^{(b)}) = \arg\min_{\phi} \min_{\alpha} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i - \alpha - \phi \cdot \hat{\Delta}_i(V^{(b)}) \right)^2 \ , \quad \text{where}$$
$$\hat{\Delta}_i^\star(V^{(b)}) = \sum_{j \neq i} \hat{D}_{i,j}^\star \hat{W}_j^\star V_j^{(b)}$$

end

Return the spillover proximity gradient estimate $\hat{\theta}_n^\star$ and conservative variance estimate

$$\hat{\sigma}_n^2 = \frac{2}{B} \sum_{b=1}^{B} \left(\hat{\phi}_n^\star(V^{(b)}) \right)^2$$

Notes: Algorithm 2 details the construction of the residualized estimator of convex averages of spillover proximity gradients, and associated standard error and confidence interval. We recommend taking B on the order of 2000.

Remark 3.6.1. As discussed in Remark 3.5.1, the leading order terms of both the numerator and denominator of the estimator $\hat{\theta}_n^\star$ can be expressed as degenerate U-statistics of order three. In general, the limiting distributions of degenerate U-statistics are non-Gaussian, and instead can be represented as mixtures of chi-squared random variables, where the mixing weights are determined by the eigenvalues of each U-statistic's kernel (see e.g., Serfling 1980 for a textbook treatment). In effect, Theorem 3.6.1 follows by showing that, under our regularity conditions (particularly, Assumptions 3.5.2 and 3.5.4) these eigenvalues become sufficiently diffuse that the mixture of chi-squared variables converges to a Gaussian distribution (see Hall 1984 for discussion of arguments with this structure). Functionally, this step is enabled by a general Berry-Esseen type central limit theorem for degenerate U-statistics, given in Liu et al. (2025). ■

It is worth emphasizing that the leading order term in the variance of the estimator $\hat{\theta}_n^\star$ is determined by the distribution of the residual-like quantity

$$\tilde{\varepsilon}_{i,j} = (\mathbb{E}[Y_i \mid Z_j, Z_i] - \mathbb{E}[Y_i \mid Z_i]) - \theta_n^\star \tilde{D}_{i,j}^\star W_j^\star \ , \tag{3.6.7}$$

defined in (3.6.4). This variable can be interpreted as the contribution of the unit j to unit i's outcome, net any linear dependence on the quantity $\tilde{D}_{i,j}^\star W_j^\star$.

3.6.3 Consistent Inference

The following theorem establishes that the limiting distribution of the regression coefficient $\hat{\phi}_n^\star(V)$, defined in (3.6.2), over the random signs V, is closely related to the limiting, unconditional distribution of the estimator $\hat{\theta}_n^\star$, characterized in Theorem 3.6.1. In particular, the conditional distribution of the sign-flipped coefficient can be used to bound the variance of the unconditional distribution of the baseline estimator.

Theorem 3.6.2. *Define the sequence*

$$\varphi_n^2 = \frac{1}{\mathbb{E}[(\tilde{D}_{i,j}^\star W_j^\star)^2]^2} \operatorname{Var}(\mathbb{E}[\tilde{\varepsilon}_{k,i} \tilde{D}_{k,j}^\star \mid Z_i, Z_j] W_j^\star) \tag{3.6.8}$$

and the cumulative distribution function

$$\hat{R}_n(x) = \frac{1}{2^n} \sum_{v \in \{-1,1\}^n} \mathbb{I}\{\varphi_n^{-1} \hat{\phi}_n^\star(v) \leq x\} \ . \tag{3.6.9}$$

If the conditions of Theorem 3.6.1 hold, then

$$\rho_n \lesssim \sigma_n^2 \leq 2\varphi_n^2 \lesssim \rho_n \quad \text{and} \quad \hat{R}_n(x) \xrightarrow{p} \Phi(x) \tag{3.6.10}$$

for all $x \in \mathbb{R}$, where $\Phi(x)$ denotes the standard normal cumulative distribution function.

As the vector V is uniformly distributed on the set $\{-1, 1\}^n$, the function $\hat{R}_n(\cdot)$ gives the cumulative distribution function, conditional on the observed data, for a suitably scaled version of the coefficient $\hat{\phi}_n^\star(V)$. The statement (3.6.10) implies that both the sampling law of the baseline estimator and the conditional law of the sign-flipped estimator are approximately normal, with scales ordered so that the former is no larger than twice the latter. As a consequence, the variability of the resampling distribution (3.6.9) can be used to provide an upper bound for the sampling variability of the baseline estimator. That is, using twice the resampling variance yields a conservative standard error and the confidence interval (3.6.3) controls the asymptotic level.

Observe that the difference between the variance σ_n^2 of the baseline estimator $\hat{\theta}_n^\star$ and the conditional variance φ_n^2 of the sign-flipped estimator $\hat{\phi}_n^\star(V)$ is given by the covariance

$$\frac{1}{\mathbb{E}[(\tilde{D}_{i,j}^\star W_j^\star)^2]^2} \operatorname{Cov}(\mathbb{E}[\tilde{\varepsilon}_{k,i} \tilde{D}_{k,j}^\star \mid Z_i, Z_j] W_j^\star, \mathbb{E}[\tilde{\varepsilon}_{k,j} \tilde{D}_{k,i}^\star \mid Z_j, Z_i] W_i^\star) \,. \tag{3.6.11}$$

As the two terms appearing in this covariance are identically distributed, the Cauchy-Schwarz inequality implies that the quantity (3.6.11) is bounded above by the variance φ_n^2. This bound yields the factor of two appearing in the relation (3.6.10) and the conservative variance estimate $\hat{\sigma}_n^2$.

The covariance (3.6.11) can be close (or equal) to the variance φ_n^2 in reasonable data generating distributions. The simulation considered in the following subsection gives a concrete illustration of this point. Heuristically, the covariance (3.6.11) is larger when the channels that mediate spillovers from j to i and from i to j are more correlated. That is, if the outcome Y_i depends on the treatment W_j though the (potentially, unobserved) proximity measure $G_{i,j}$, and likewise for Y_j, W_i, and $G_{j,i}$, then the covariance (3.6.11) increases with the correlation between the proximity measures $G_{i,j}$ and $G_{j,i}$.

3.6.4 Simulation

We conclude this section by reporting the results of a simple simulation, designed to illustrate the settings under which the variance estimator $\hat{\sigma}_n^2$ is more, or less, conservative. For the sake of simplicity, we assume that there are no covariates X_i or proximity measures $G_{i,j}$ and that the treatments W_i are uniformly distributed on $\{-1/2, 1/2\}$. In turn, we assume that the coordinates S_i are uniformly distributed on the integers $\mathcal{M}_n = \{0, \ldots, m_n - 1\}$ and the proximity measure of interest is given by $D_{i,j} = \mathbb{I}\{S_i = S_j\}$. In this setting, the sequence ρ_n is given by m_n^{-1}. We take $m_n = \sqrt{n}$ for the sake of concreteness. The factors S_i and U_i are perfectly correlated, in the sense that all units with the same value of S_i will have the same value of U_i. The marginal distribution of the variables U_i is multinomial on $\mathcal{M}_n$, with parameter proportional to

$$p^{(\eta)} = (1-\eta)p^{(0)} + \eta p^{(1)} \ , \tag{3.6.12}$$

where $p^{(0)}$ places mass m_n^{-1} on each coordinate and $p^{(1)}$ places mass m_n^{-1} and $1 - m_n^{-1}$ on the coordinates zero and $m/2$, respectively.[33] Here, the free parameter η will be used to vary the correlation (3.6.11).

Spillovers are mediated by an additional, unobserved proximity measure $G'_{i,j}$, which is determined by the i.i.d. factors U_i through the specification

$$G'_{i,j} = \mathbb{I}\{(S_i - S_j) \bmod m_n = U_i\} \ . \tag{3.6.13}$$

In particular, outcomes are generated by the specification

$$Y_i = \sum_{j \neq i} (\theta_n^\star (D_{i,j} - \rho_n) + G'_{i,j}) W_j \ . \tag{3.6.14}$$

Observe that, when $\eta = 0$, the measure $G'_{i,j}$ is uncorrelated with the measure $G'_{j,i}$, and that when $\eta = 1$ the two quantities are almost surely equal.

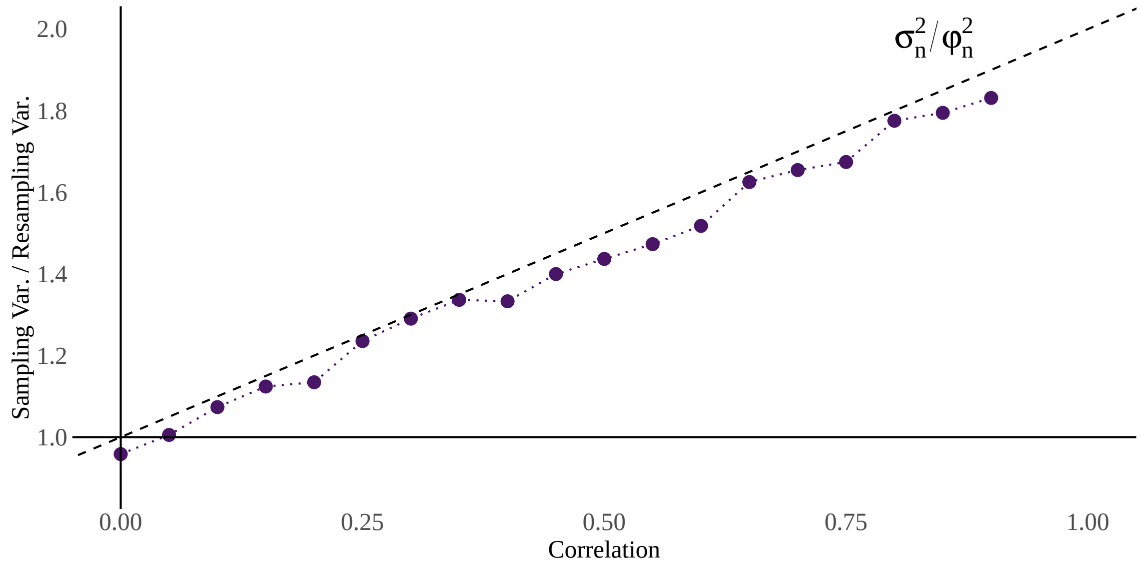
Theorems 3.6.1 and 3.6.2 imply that the ratio of the sampling variance of the estimator $\hat{\theta}_n^\star$ to the conditional variance of the sign-flipped estimator $\hat{\phi}_n^\star(V)$ should be well approximated by sequence σ_n^2/φ_n^2. In this model, it can be shown that the sequence σ_n^2/φ_n^2 is equal to one plus the correlation

$$\mathrm{Corr}(\mathbb{E}[\tilde{\varepsilon}_{k,i}\tilde{D}_{k,j}^\star \mid Z_i, Z_j]W_j^\star, \mathbb{E}[\tilde{\varepsilon}_{k,j}\tilde{D}_{k,i}^\star \mid Z_j, Z_i]W_i^\star) = \eta^2 + O(m_n^{-1}) \ . \tag{3.6.15}$$

This prediction is bourne out in finite-samples. In particular, Figure 3.4 displays estimates of the ratio of the sampling variance of the estimator $\hat{\theta}_n^\star$ to the conditional variance of the

[33]That is, the variables U_i have a multinomial distribution with parameter given by a version of $p^{(\eta)}$ normalized to sum to one.

Figure 3.4: Observed and Predicted Variance Ratios



Notes: Figure 3.4 compares observed and predicted values of the ratio of the sampling variance of the estimator $\hat{\theta}_n^\star$ to the conditional variance of the sign-flipped estimator $\hat{\phi}_n^\star(V)$, in the simulation environment detailed in Section 3.6.4. The observed values of this ratio are displayed with purple dots. The value of this ratio predicted by Theorems 3.6.1 and 3.6.2 is displayed with a black dashed line. The x-axis displays values of the correlation (3.6.15). At each value of this correlation on a grid between 0 and 0.95, we use $5,000$ simulation replications with $B = 500$ random sign vectors $V^{(b)}$ used per replication.

---

sign-flipped estimator $\hat{\phi}_n^\star(V)$, for the case that $n = 1600$ and $m_n = \sqrt{n} = 40$, over a range of values for the correlation (3.6.15). The ratio σ_n^2/φ_n^2 is juxtaposed with a dashed line.

When the parameter η is equal to zero, the conditional variance of the sign-flipped estimator is very close to the sampling variance of estimator $\hat{\theta}_n^\star$. As η increases, the ratio of the two quantiles increases in proportion to the correlation (3.6.15), as predicted by Theorems 3.6.1 and 3.6.2. When η is close to one, two times the conditional variance of the sign-flipped estimator, that is, the variance estimate $\hat{\sigma}_n^2$, provides a good approximation to the sampling variance of the estimator $\hat{\theta}_n^\star$.

3.7. Empirical Application

We now illustrate the application of the methodology developed in this paper to Example 4. In particular, building on the data and empirical framework considered in Bloom et al. (2013), we characterize the relationship between technological similarity, product market similarity, and productivity spillover intensity. For the sake of parsimony, we detail our treatment of these data in Appendix C.6, where we also illustrate applications to Examples 5 and 6.[34]

We consider a sample of publicly traded U.S. firms. Here, the treatment variable W_i is a binary indicator that firm i spent more than 100 million dollars on R&D between between 1996 and 2000. Following Bloom et al. (2013), we consider two proximity measures of interest. First, we set the proximity measure $D_{i,j}$ as the uncentered correlation (or "cosine similarity") between i and j's vectors of patent shares across USPTO technology classes, computed from all patents filed between 1996 and 2000. Second, we set $D_{i,j}$ as the uncentered correlation between i and j's vectors of sales shares across four digit industry codes, again from 1996 to 2000. We consider four choices for the outcome variable Y_i: market value (i.e., Tobin's Q), patent cites, sales, and R&D expenditure, each averaged between 2001 and 2005. We use the logarithm of each outcome variable. Results are displayed in Table 3.1. Panels A and B consider Technological and Product Market Proximity, respectively.

The first row in both panels displays the realized value of the estimator $\hat{\theta}^{\star}_n$, associated with each outcome, when neither the treatments nor the proximity measures of interest have been residualized (but the treatments are centered). Causal interpretations of these estimates are susceptible to two forms of confounding.

First, R&D expenditure might be endogenous. To address this, we residualize the treatments with lagged values of the treatment and outcome variables as well as fixed effects for the technology subcategory where each firm filed the majority of its patents. This is a simplification of the empirical strategy taken in Bloom et al. (2013).[35] The updated coefficients are displayed in the second row of Table 3.1. Second, the proximity measures of interest might be correlated with other factors that mediate spillover effects. To address this, in the third row, we additionally residualize the proximity measure of interest with the other proximity measure. We also consider a specification where we additionally control for geographic proximity. The standard error $\hat{\sigma}_n$, defined in (3.6.3), is reported below each estimate.

These estimates indicate that, across nearly all outcomes, firms that are more technologically similar capture larger R&D spillovers. Together, these findings are consistent with the

[34]We rely on data from the replication package associated with Lucking et al. (2019), who improve the coverage of the data considered in Bloom et al. (2013), both in terms of the number of firms and years that are available.

[35]Bloom et al. (2013) consider a similar panel specification, and so include firm fixed effects. We use a cross-sectional specification for the sake of simplicity. Bloom et al. (2013) also consider an instrument based on tax-induced changes in the user cost of R&D capital, which varies across states. We do not consider this approach, because these changes are highly correlated for pairs of firms whose R&D is concentrated in similar places (i.e., an analogue of Assumption 3.4.2 is violated).

Table 3.1: Application to R&D Spillover Estimation

		Value	Patent Cites	Sales	R&D Exp.
Panel A: Technological Proximity					
Unadjusted		0.052**	0.119**	0.106**	0.252**
		(0.012)	(0.024)	(0.022)	(0.012)
Resid. Treatment		0.201**	0.395**	0.272	0.647**
		(0.087)	(0.156)	(0.188)	(0.100)
Resid. Proximity	Prod. Market Prox.	0.227**	0.384**	0.171	0.636**
		(0.079)	(0.129)	(0.181)	(0.100)
	+ Geographic Prox.	0.231**	0.381**	0.142	0.626**
		(0.074)	(0.131)	(0.183)	(0.104)
Number of firms		1214	852	1294	1276
Panel B: Product Market Proximity					
Unadjusted		0.051	0.068	0.123**	0.248**
		(0.034)	(0.050)	(0.060)	(0.018)
Resid. Treatment		-0.146	0.013	0.626**	0.412
		(0.249)	(0.203)	(0.186)	(0.407)
Resid. Proximity	Technological Prox.	-0.294**	-0.153*	0.435**	-0.135
		(0.142)	(0.088)	(0.157)	(0.237)
	+ Geographic Prox.	-0.283*	-0.163*	0.414**	-0.141
		(0.146)	(0.087)	(0.168)	(0.245)
Number of firms		1370	839	1466	1364

Notes: Table 3.1 displays values of the estimator $\hat{\theta}_n^{\star}$ obtained by minimizing the objective function (3.4.3), using two choices for the proximity measure of interest, a variety of choices for covariates, and four choices for the outcome variable. The standard error $\hat{\sigma}_n$, defined in Algorithm 2, is displayed in parentheses below each estimate. One or two asterisks are placed beside each estimate to denote significance at the 10% and 5% levels, respectively. The estimates are obtained using data from the replication package associated with Lucking et al. (2019). Further details are given in Appendix C.6. In each case, the treatment variable W_i denotes a binary indicator that firm i spent more than one hundred million dollars on R&D between 1996 and 2000. The proximity measures $D_{i,j}$ are the uncentered correlation between patent shares filed between 1991 and 1995 across USPTO technology classes and the uncentered correlation between sales shares between 1991 and 1995 across four digit industry codes, respectively. The outcomes are averaged between 2001 and 2005, and enter in logs. In the first row, neither the treatments nor proximity measures have been residualized. In the second row, the treatments are residualized using lagged values of the treatment and outcome variables as well as fixed effects for the technology subcategory where each firm filed the majority of its patents. In the third and fourth rows, the proximity measure has been residualized using various measures of proximity, in addition to the covariates used to residualize the treatments. The estimators $\hat{\pi}_n(\cdot)$ and $\hat{\gamma}_n(\cdot)$ are both obtained from linear regressions.

idea that investments in R&D reduce the cost of technologically related investment. The exceptional outcome is sales: across each specification, the coefficient is not statistically different from zero. As these specifications control for product market proximity, this is intuitive.

The sign of the relationship between product market proximity and spillover intensity depends on the outcome. Increases in product market proximity are associated with decreased R&D spillovers on market value and patent citations and increased R&D spillovers on sales. The decrease in market value is consistent with the idea that investors interpret the investments of a rival firm as likely to depress long-term firm growth and earnings. The decrease in patent citations is consistent with models of cumulative innovation, in which older technologies are displaced—here, quite literally in citation patterns—by newer variants (Scotchmer, 2004). Increases in sales suggest some short-term strategic complementarity. It is worth cautioning that these estimates aggregate across many industries with heterogeneous market structures.

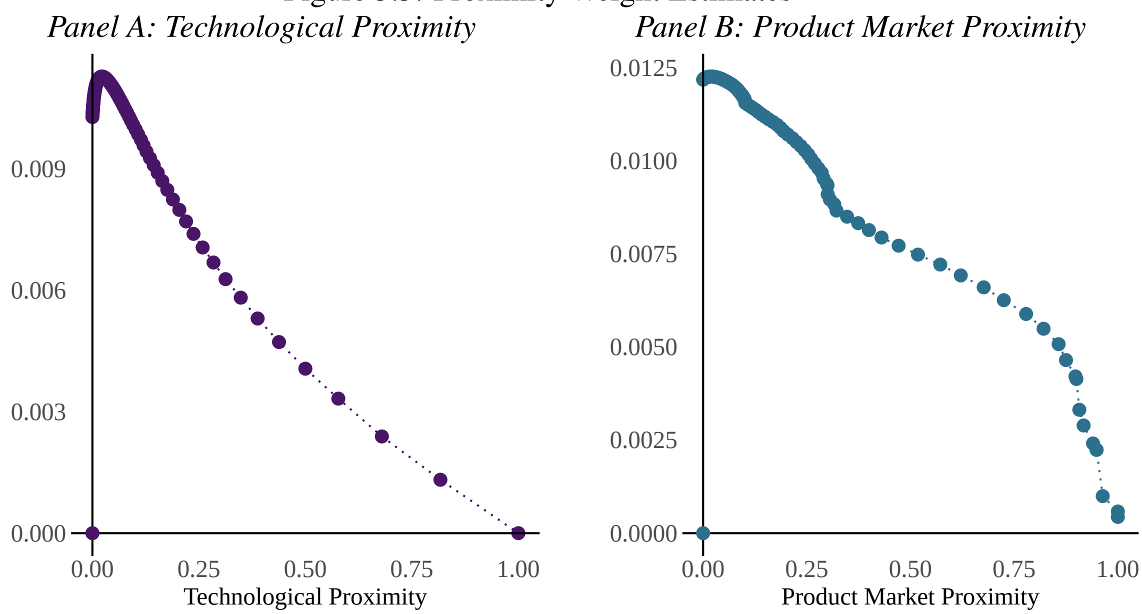
To unpack these results, Figure 3.5 displays estimates of the covariance $\mathrm{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j})$ for values of δ across the support of the two proximity measures of interest. Up to heterogeneity across covariates, this quantity determines the weight function (3.4.13) characterized in Theorem 3.4.1. These estimates indicate that much of the weight in the coefficients displayed in Table 3.1 is placed on pairs of firms whose technological and product market proximities are relatively small.

3.8. Conclusion

The spillover effects of economic actions often propagate through distinct, and potentially countervailing channels. This paper proposes a framework for identifying and estimating the causal relationship between a given measure of economic proximity—such as geographic distance, technological similarity, trade, or migration flows—and the intensity of the spillover effects of a treatment, shock, or policy change. Our consideration centers on widely-applied regressions that relate outcomes to proximity-weighted averages of the treatments assigned to other units. We show that coefficients obtained from such regressions admit a nonparametric interpretation—as targeting convex averages of parameters that measure how the association between one unit's outcome and another unit's treatment correlates with the proximity measure of interest.

The core premise of this paper is that when the proximity measure of interest is associated with other channels that also mediate spillover effects, causal interpretations of

Figure 3.5: Proximity Weight Estimates



Notes: Figure 3.5 displays estimates of the parameter $\text{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j})$ across the support proximity measures of interest for the estimates displayed in Table 3.1. Each series has been normalized so that the points add to one. Panels A and B, display the weight estimates for technological proximity and product market proximity, respectively. In both cases, we measure $\text{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j})$ at each of the percentiles of the distribution of the proximity measure $D_{i,j}$, conditional on $D_{i,j}$ being non-zero.

---

such relationships are susceptible to confounding. For example, Bloom et al. (2013) find that the spillover effects of investments in research and development are larger for pairs of firms whose research is concentrated in more technologically similar areas. If technologically similar firms tend to be geographically close, such a relationship might arise, spuriously, from spillover effects that are determined by geography alone.

The main contribution of this paper is a regression-based method for adjusting such estimates to account for correlated channels that might also transmit spillover effects. In particular, we give conditions under which the effect of a given proximity measure on spillover intensity can be recovered by regressing outcomes on averages of other units' treatments, reweighted by versions of the proximity measure that have been residualized with available auxiliary proximity measures. That is, in the application to Bloom et al. (2013), the propagation of spillovers through geography can be accounted for by residualizing technological similarity with geographic distance. We show that estimates obtained in this way are consistent, and optimal, in a minimax sense, and propose a new resampling based approach for constructing conservative estimates of the associated uncertainty. We illustrate, in data from Bloom et al. (2013), that the proposed adjustments can make a meaningful difference in practice.

Several extensions to the results presented in this paper have the potential to be particularly impactful. First, throughout, we have assumed that treatments are assigned independently across units. Further research should consider the extent to which peer effects in treatment selection can be accommodated. Leung and Loupos (2025) give a general treatment of a related problem. Second, our main approach for addressing endogeneity in measures of proximity is based on a conditional ignorability assumption. We have given a brief treatment of an extension to settings where suitable instrumental variables are available.[36] Further consideration of such settings, both theoretically and empirically, would be useful. Finally, the estimands of interest in this paper are local, in the sense that they relate to the spillover effects of changing the treatment and proximity of only one unit. It would be valuable to consider how to use such estimates as inputs into the recovery of more complicated counterfactuals that require the prescription of greater economic structure.

[36] For instance, Ellison et al. (2010) instrument geographic co-location in the United States with geographic co-location in the United Kingdom. Similarly, Fafchamps and Quinn (2018) implement an experiment that introduces exogenous variation in the social ties across business leaders, and measure the propagation of management practices through these ties. See Appendix C.5.5.

Appendix A

Supplemental Appendices for Chapter 1

with Joseph P. Romano

A.1. Data and Simulations

A.1.1 Casey et al. (2021b)

We obtain the data associated with Casey et al. (2021a,b) from the replication package posted on OpenICPSR.[1] The units of observation are the aspirant political candidates in Sierra Leone's 2018 parliamentary elections. The data that we consider consist of an outcome and 48 covariates. The outcome is the vote share for each aspirant candidate in a poll of party officials. The covariates collect various measurements relating to the candidate's personal, political, and financial characteristics. All data cleaning steps match Casey et al. (2021b).

We consider the exercise summarized in Appendix Table A.4 of Casey et al. (2021b). These authors repeatedly compute 10-fold cross-validated estimates of the risk associated with a regularized regression of the outcome on the collection of covariates, at each value of a grid of penalization parameters. They include both Ridge and Lasso type penalties. For each replication, they compute the covariates that are selected in the regression associated with the penalty parameter having the minimum risk estimate. They display the covariates that are selected in more than 200 of the 400 replications. These covariates are included as controls in various downstream analyses. In our replication of this exercise, to ease exposition, we only include a Lasso type penalty.

[1] The replication package associated with Casey et al. (2021b) is posted at the URL `https://www.openicpsr.org/openicpsr/project/124501/version/V1/view`.

A.1.2 Chakravorty et al. (2024a)

We obtain the data and replication package associated with Chakravorty et al. (2024a,b) directly from the authors.[2] The units of observation are trainees in a job training program implemented in the Indian states of Bihar and Jharkhand. There are 2,488 total trainees, who were divided into batches of approximately 30 trainees. Batches of trainees were randomly assigned to treatment with a program involving the provision of information concerning potential placement jobs. All data cleaning steps match those taken in Chakravorty et al. (2024a) .

We replicate the analysis reported in Column 4 of Table 1 of Chakravorty et al. (2024a). This estimate restricts attention to the 890 trainees that completed the training and were subsequently placed into a job. The outcome variable is an indicator for whether the trainee is still in the job five months after training completion. The covariates collect measurements of 77 attributes related to the demographics, human capital, and expectations of the trainees. These covariates are listed in Table 3.1 of Chakravorty et al. (2024a).

Following Chakravorty et al. (2024a), we estimate the effect of the treatment using the implementation of Double Machine Learning available through the "DoubleML" R package (Bach et al., 2024). We use 5-fold cross-fitting. Nuisance parameters (i.e., outcome regressions and propensity scores) are estimated with Random Forests using the default tuning parameters associated with the "Ranger" R package (Wright and Ziegler, 2017). Standard errors are clustered at the batch level.

A.1.3 Haushofer et al. (2022)

Haushofer et al. (2022) consider data originally studied in Egger et al. (2022a,b). At the time of the preparation of this paper, no replication materials for Haushofer et al. (2022) were publicly available. We replicate a simplified version of the exercise considered in Haushofer et al. (2022), using data obtained from the replication package associated with Egger et al. (2022b).[3] The units of observation are households in three sub-counties of Siaya County, western Kenya. Households were randomly allocated large cash transfers. See Egger et al. (2022b) and Haushofer et al. (2022) for further details on the design of this experiment.

Using the replication data from Egger et al. (2022b,a), we begin with a sample of 5,423

[2]We thank Bhaskar Chakravorty for sharing this material.

[3]The replication package for Egger et al. (2022b) is available at the link https://www.econometricsociety.org/publications/econometrica/2022/11/01/General-Equilibrium-Effects-of-Cash-Transfers-Experimental-Evidence-From-Kenya.

treatment eligible households. We then restrict attention to 4,758 households who were surveyed in both the first and second round of the experiment. We drop any households that are missing measurements of post-treatment assets, consumption, income, or food security, leaving a sample of 4,755 households. These data are merged with pre-treatment measurements of the demographics, assets, food security, and labor market participation for each household, leaving a sample of 4,754 households.

We obtain a cleaned dataset that records 24 measurements associated with each household. The dataset identifies the village that contains each household, whether each household was assigned to treatment, and post-treatment measurements of indices of assets, consumption, income, and food security. The covariates are measurements of the household size, the number of meals eaten the day before the survey, the number of high-protein meals eaten the day before the survey, the time between the administration of the program and the measurement of post-treatment outcomes,[4] and indicators for whether the household contains a widow, contains a female, has children, has children under 3, has children under 6, contains an elderly resident, has livestock, has land, has more than a quarter acre of land, has a radio and tv, has a self-employed resident, and has an employed resident.

We implement a simplified version of the exercise studied in Haushofer et al. (2022). Consider a subset s. Let the complement of s in $[n]$ be denoted by $\tilde{\mathsf{s}}$. Let Y_i denote a measurement of an index quantifying post-treatment consumption, X_i denote a vector of measurements of pre-treatment covariates, and W_i denote assignment to treatment. Let H_i denote the number of occupants of household i. Let D_i collect these observations. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes induced by the treatment W_i. For each household i in s, let

$$\mathtt{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \quad \text{and} \quad \mathtt{uto}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{A.1.1}$$

denote sample-split estimates of the treatment effect $Y_i(1) - Y_i(0)$ and *per-capita* untreated outcome $Y_i(0)/H_i$, respectively. In this case, the nuisance parameters $\hat{\eta}(D_{\tilde{\mathsf{s}}})$ consist of nonparametric estimates of outcome regressions, i.e., the conditional expectations (1.2.7), computed using the data for units i in $\tilde{\mathsf{s}}$.

Following Haushofer et al. (2022), we compute these estimates with Random Forests using the "GRF" R package (Athey et al., 2019).[5] As best as we can tell, there are two

[4]We use the measurement of the time between the administration of the program and the measurement of post-treatment outcomes available as "exptoend" in the dataset "GE_Survey_and_Transfer_Dates.dta" in the replication package for Egger et al. (2022b). This variable appears to differ from the analogous quantity used in Haushofer et al. (2022), whose density is displayed in their Figure A.2.

[5]Following Haushofer et al. (2022), we use the default tuning parameters, but set "sample.fraction" equal to 0.1 and "min.node.size" equal to 10.

differences between our implementation and the results presented in Haushofer et al. (2022). First, we make no attempt to include sample weights that reflect the sampling probabilities. Second, we include the time between treatment and the measurement of post-treatment outcomes as a covariate, rather than implement some form of de-meaning and re-weighing of estimates across time.

Let $\mathsf{Impacted}(\mathsf{s})$ collect the 50% of units in s associated with the largest values of $\mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}}))$. Similarly, let $\mathsf{Deprived}(\mathsf{s})$ collect the 50% of units i in s associated with the smallest values of $\mathsf{uto}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}}))$. Let $\mathsf{Impacted}_w(\mathsf{s})$ collect the subset of $\mathsf{Impacted}(\mathsf{s})$ with $W_i = w$. Define $\mathsf{Deprived}_w(\mathsf{s})$ analogously. Consider the sample-split statistic

$$\begin{aligned} T(\mathsf{s}, D) = &\left(\frac{1}{|\mathsf{Impacted}_1(\mathsf{s})|} \sum_{i \in \mathsf{Impacted}_1(\mathsf{s})} Y_i - \frac{1}{|\mathsf{Impacted}_0(\mathsf{s})|} \sum_{i \in \mathsf{Impacted}_0(\mathsf{s})} Y_i \right) \\ &- \left(\frac{1}{|\mathsf{Deprived}_1(\mathsf{s})|} \sum_{i \in \mathsf{Deprived}_1(\mathsf{s})} Y_i - \frac{1}{|\mathsf{Deprived}_0(\mathsf{s})|} \sum_{i \in \mathsf{Deprived}_0(\mathsf{s})} Y_i \right) . \end{aligned} \tag{A.1.2}$$

That is, the sample-split statistic (A.1.2) is an estimator of the difference in the average treatment effect of the most impacted and most deprived units in s. The statistic (A.1.2) can be aggregated with cross-splitting, through

$$a(\mathsf{r}, D) = \frac{1}{k} \sum_{i=1}^{k} T(\mathsf{s}_i, D) \ , \tag{A.1.3}$$

where $\mathsf{r} = (\mathsf{s}_i)_{i=1}^k$ is a k-fold partition of $[n]$. Following Haushofer et al. (2022), in Section 1.2, we use 5-fold cross-fitting.

Haushofer et al. (2022) say that they construct a standard error for (A.1.3) with the bootstrap. There are a variety of ways that one might do this. In our implementation, we keep the values of the treatment effect and untreated outcome estimates (A.1.1) fixed for each household, and compute a bootstrap standard error for the statistic (A.1.3), re-computing the groups $\mathsf{Impacted}(\mathsf{s})$ and $\mathsf{Deprived}(\mathsf{s})$ in each bootstrap replicate.

A.1.4 Beaman et al. (2023b)

We obtain the data associated with Beaman et al. (2023a,b) from the associated replication package posted on the Econometric Society's webpage.[6] The units of observations are low-income farmers in Mali. The cleaned data consist of observations of 5210 farmers across 198 villages. All data cleaning steps match Beaman et al. (2023b).

Beaman et al. (2023b) consider a two-stage experiment. In the first stage, a microcredit organization randomly offered group-liability loans to women in 88 of the 198 villages. In the second stage, households were randomly offered cash grants. We replicate the analysis reported in Column (4) of Appendix Table VI of Beaman et al. (2023b). In this setting, attention is restricted to the 2,142 households who received a loan. They are interested in testing whether the effects of cash-grants on farm profits are heterogeneous. For this problem, the data D_i consist of a measurement Y_i of the post-transfer profit for farm i, the variable W_i denotes assignment to the cash transfer, and the vector X_i collects a vector of pre-treatment measurements of the physical and financial attributes of each farm. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes induced by the treatment W_i.

Beaman et al. (2023b) implement a sample-split test for treatment effect heterogeneity proposed by Chernozhukov et al. (2023). Let s denote a random half-sample of $[n]$, i.e., a random subset of $[n]$ of size $n/2$. As before $\tilde{\mathsf{s}}$ denotes the complement of s in $[n]$. For each farm i in s, let

$$\mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) \tag{A.1.4}$$

denote sample-split estimates of the treatment effect $Y_i(1) - Y_i(0)$. The nuisance parameters $\hat{\eta}(D_{\tilde{\mathsf{s}}})$ consist of nonparametric estimates of outcome regressions, i.e., the conditional expectations (1.2.7), computed using the data for units i in $\tilde{\mathsf{s}}$. We compute these estimates with Random Forests using the "GRF" R package (Athey et al., 2019). Tuning parameters are chosen with cross-validation in the same manner implemented in Beaman et al. (2023b).

To test for treatment effect heterogeneity, we estimate the coefficients of the linear regression

$$Y_i = \alpha + \tau_1 \cdot W_i + \tau_2 \cdot \mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) + \tau_3 \cdot W_i \cdot \mathsf{te}(D_i, \hat{\eta}(D_{\tilde{\mathsf{s}}})) + \varepsilon_i\,. \tag{A.1.5}$$

The idea is that, if the treatment effect estimates are heterogeneous, then the "true" value

[6] The replication package for Beaman et al. (2023b) is available at https://www.econometricsociety.org/publications/econometrica/2023/09/01/Selection-into-Credit-Markets-Evidence-from-Agriculture-in-Mali

of the interaction coefficient τ_3 should be positive. Let $\hat{\tau}_3(\mathsf{s}, D)$ denote the estimate of τ_3 computed with linear regression in the sample s. Let $\mathsf{se}(\mathsf{s}, D)$ denote the associated standard error, clustered at the village level. Chernozhukov et al. (2023) advocate for computing the median estimate and standard error over 250 half-splits s. Beaman et al. (2023b) use 1000 half-splits.

A.1.5 Simulations

In this section, we give further details concerning the simulations whose results are reported in Section 1.4. For both the applications to Casey et al. (2021b) and Chakravorty et al. (2024a), we sample 100,000 replications of the cross-split statistic of over a range of values of k. In the application to Chakravorty et al. (2024a), we have increased the number of trees used to compute the random forest nuisance parameter estimates, from 100 to 3000, in order to ensure that the residual randomness induced by cross-splitting is greater than the residual randomness induced by the nuisance parameter estimate. We use these replicates to estimate the variance $v_{1,k}(D)$ for each value of k, in addition to the oracle stopping time

$$g^{\star} = 2v_{1,k}(D)\left(\frac{z_{1-\beta/2}}{\xi}\right)^2 , \tag{A.1.6}$$

at each value of k and ξ

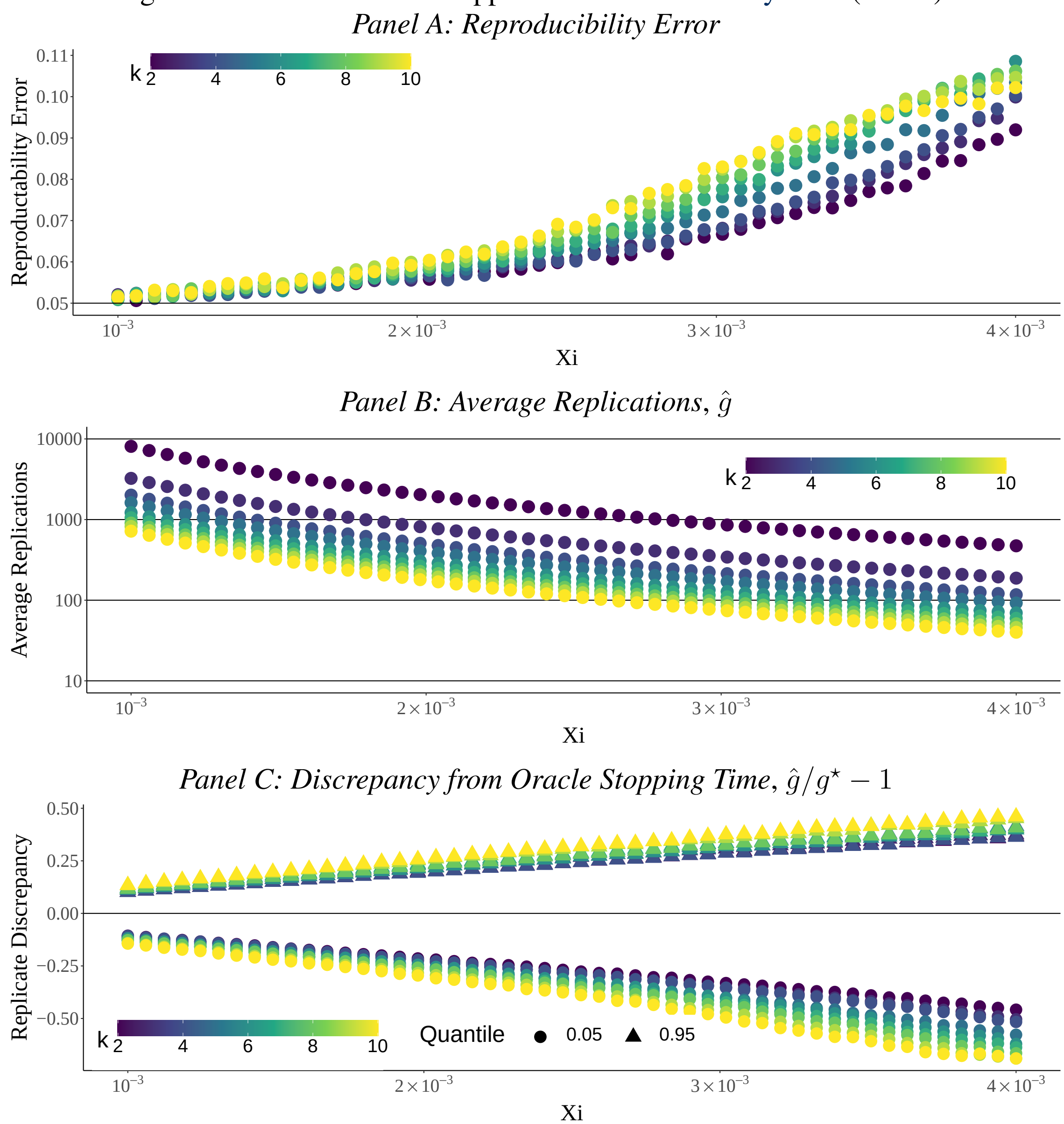
We implement Algorithm 1 100,000 times for each value of k and ξ, for each application, by sampling with replacement from the replicates of 100,000 draws. We use these estimates to estimate the reproducibility error

$$P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} \tag{A.1.7}$$

as well as the distribution of $\hat{g}$ for each k and ξ. These estimates are used to construct Figure 1.5.

Figure A.1 displays measurements analogous to those displayed in Figure 1.5 for the application to Chakravorty et al. (2024a). The accuracy of the reproducibility error, relative to the application to Casey et al. (2021b), is somewhat worse. This is likely a consequence of the fact that the quality of a normal approximation to the conditional distribution of the p-value (1.3.11) is worse than the quality of the normal approximation to the conditional distribution of the mean-squared error estimate. The reproducibility error tends to be most accurate at values of k and ξ where the average number of cross-splits $\hat{g}$ is greater than 500.

Figure A.1: Performance in Application to Chakravorty et al. (2024a)



Notes: Figure A.1 displays measurements of the performance of Algorithm 1 on the data from Chakravorty et al. (2024a). Panel A displays measurements of the reproducibility error, $P\{|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)| \geq \xi \mid D\}$, as ξ and k vary. A solid horizontal line is displayed at the nominal error rate $\beta = 0.05$. Panel B displays measurements of the average number of replications $\hat{g}$ s ξ and k vary. The y-axis is displayed with a log scale, base 10. Solid horizontal lines are placed at each exponential factor of 10. Panel C displays measurements of the 0.05th and 0.95th quantiles of the discrepancy $\hat{g}/g^{\star} - 1$ as k and ξ vary. Further details on the construction of this figure are given in Appendix A.1.5.

A.2. Auxiliary Results and Discussion

A.2.1 Validity of Testing Procedures Based on Multiple Sample-Splitting

In this appendix, we discuss the application of Algorithm 1 to testing procedures based on averaging over multiple splits of the same sample. We show that methods based on both p-values and e-values constructed with sample splitting continue to control the Type I error rate if they are aggregated sequentially with Algorithm 1. Both results follow from the "Exchangeable Markov Inequality" of Ramdas and Manole (2023). As before, let $D = (D_i)_{i=1}^n$ be independent and identically distributed according to a probability distribution P. Interest is in testing the null hypothesis $H_0 : P \in \mathbf{P}$ for some collection of probability distributions $\mathbf{P}$.

A.2.1.1 Methods Based on p-Values

Suppose that we have access to a valid p-value $\hat{p}(\mathsf{s}, D)$. That is, the statistic $\hat{p}(\mathsf{s}, D)$ satisfies

$$P\left\{\hat{p}(\mathsf{s}, D) \leq u \mid D_{\bar{\mathsf{s}}}\right\} \leq u$$

for all u in $(0, 1)$ and P in $\mathbf{P}$. For example, a test-statistic could be chosen using the data in $D_{\bar{\mathsf{s}}}$ and a p-value can be constructed based on this test statistic using the data in D_{s}. For any collection $\mathsf{R}_{g,k}$ in $\mathcal{R}_{n,k,b}$, let

$$a_\delta(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{k}\sum_{i=1}^{g}\sum_{j=1}^{k} \mathbb{I}\left\{\hat{p}(\mathsf{s}_{i,j}, D) \leq \delta\right\} \tag{A.2.1}$$

denote the proportion of p-values that are less than or equal than δ. Rüger (1978), Meinshausen et al. (2009), and DiCiccio et al. (2020) observe that if $\mathsf{R}_{g,k}$ is constructed independently of the data D, then

$$P\left\{a_\delta(\mathsf{R}_{g,k}, D) \geq c\right\} \leq \frac{\mathbb{E}\left[a_\delta(\mathsf{R}_{g,k}, D)\right]}{c} \leq \frac{\delta}{c}$$

by Markov's inequality, for all P in $\mathbf{P}$. Thus, if δ and c are chosen such that $\delta/c = \alpha$, then the test that rejects the null hypothesis H_0 if $a_\delta(\mathsf{R}_{g,k}, D)$ is larger than c has level α.

The following theorem establishes that this test continues to be valid if the collection of sample-splits $\mathsf{R}_{g,k}$ is constructed sequentially with Algorithm 1.

Theorem A.2.1. *If the statistic $a_\delta(\mathsf{R}_{\hat{g},k}, D)$ defined in (A.2.1) is constructed sequentially* *with Algorithm 1, then*

$$P\left\{a_\delta(\mathsf{R}_{\hat{g},k}, D) \geq c\right\} \leq \frac{\delta}{c} \tag{A.2.2}$$

for all P in $\mathbf{P}$.

Proof. We apply the following inequality, due to Ramdas and Manole (2023).

Theorem A.2.2 (Theorem 1.1, Ramdas and Manole (2023))**.** *If $X_1, X_2, \ldots$ form an exchangeable sequence of integrable random variables, then*

$$P\left\{\exists t \geq 1 : \frac{1}{t}\sum_{i=1}^{t} |X_i| \geq 1/a\right\} \leq a\mathbb{E}\left[|X_i|\right] \tag{A.2.3}$$

for any $a > 0$.

Consequently, we have that

$$\begin{aligned} P\left\{a_\delta(\mathsf{R}_{\hat{g},k}, D) \geq c\right\} &\leq P\left\{\exists g \geq 1 : a_\delta(\mathsf{R}_{g,k}, D) \geq c\right\} \\ &\leq \frac{\mathbb{E}\left[\mathbb{I}\left\{\hat{p}(\mathsf{s}_{i,j}, D) \leq \delta\right\}\right]}{c} \leq \frac{\delta}{c} \end{aligned} \tag{A.2.4}$$

by Theorem A.2.2, as required. ■

A.2.1.2 Methods Based on e-Values

Next, we consider settings where we have access to a valid e-value $\hat{e}(\mathsf{s}, D)$ (see Ramdas et al. (2023) for a recent review). That is, the nonnegative statistic $\hat{e}(\mathsf{s}, D)$ satisfies

$$\mathbb{E}_P\left[\hat{e}(\mathsf{s}, D)\right] \leq 1$$

for all P in $\mathbf{P}$. For example, this setting applies to the “Universal Inference” procedure of Wasserman et al. (2020). Here, an estimator $\hat{P}(\tilde{\mathsf{s}})$ of P is formed using the data $D_{\tilde{\mathsf{s}}}$ and is used in the split-likelihood ratio test statistic

$$\hat{e}(\mathsf{s}, D) = \inf_{P \in \mathbf{P}} \prod_{i \in \mathsf{s}} \frac{\mathrm{d}\hat{P}(\tilde{\mathsf{s}})}{\mathrm{d}P}(D_i)\,. \tag{A.2.5}$$

Wasserman et al. (2020) prove that (A.2.5) is an e-value. For any collection $\mathsf{R}_{g,k}$ in $\mathcal{R}_{n,k,b}$, let

$$a(\mathsf{R}_{g,k}, D) = \frac{1}{g}\frac{1}{k}\sum_{i=1}^{g}\sum_{j=1}^{k} \hat{e}(\mathsf{s}_{i,j}, D) \tag{A.2.6}$$

denote an aggregate e-value. See Dunn et al. (2023) and Tse and Davison (2022) for further discussion of aggregate e-values. Observe that

$$P\left\{a(\mathsf{R}_{g,k}, D) \geq 1/\alpha\right\} \leq \alpha\mathbb{E}\left[\hat{e}(\mathsf{s}, D)\right] \leq \alpha \;, \tag{A.2.7}$$

by Markov's inequality, for all P in $\mathbf{P}$. Thus, the test that rejects the null hypothesis H_0 if $a(\mathsf{R}_{g,k}, D)$ is larger than $1/\alpha$ has level α.

We again establish that this test continues to be valid if the collection of sample splits $\mathsf{R}_{g,k}$ is constructed sequentially with Algorithm 1.

Theorem A.2.3. *If the aggregate e-value $a_\delta(\mathsf{R}_{\hat{g},k}, D)$ defined in (A.2.6) is constructed sequentially with Algorithm 1, then*

$$P\left\{a(\mathsf{R}_{\hat{g},k}, D) \geq \frac{1}{\alpha}\right\} \leq \alpha \tag{A.2.8}$$

for all P in $\mathbf{P}$.

Proof. The claim is established with an argument very similar to the proof of Theorem A.2.1. Namely, the Markov inequality used to establish (A.2.7) can then be replaced by Theorem A.2.2, as before. ■

A.2.2 Determining Suitable Levels of Precision for Cross-Validated Risk Estimation

In this section, we give the details supporting our application of Algorithm 1 to cross-validated risk estimation, using data from Casey et al. (2021b). Recall from Section 1.3.2.2, that we apply the procedure, independently, to each component of the vector

$$\left(b_{\lambda_1}(\mathsf{r}, D), \ldots, b_{\lambda_{p-1}}(\mathsf{r}, D)\right) \;, \quad \text{where} \quad b_{\lambda_i}(\mathsf{r}, D) = \left(a_{\lambda_i}(\mathsf{r}, D) - a_{\lambda_p}(\mathsf{r}, D)\right) \,. \tag{A.2.9}$$

Let ξ_i denote the error tolerance used for the ith component of the vector (A.2.9). We use a simple, data-driven procedure for choosing these tolerances.

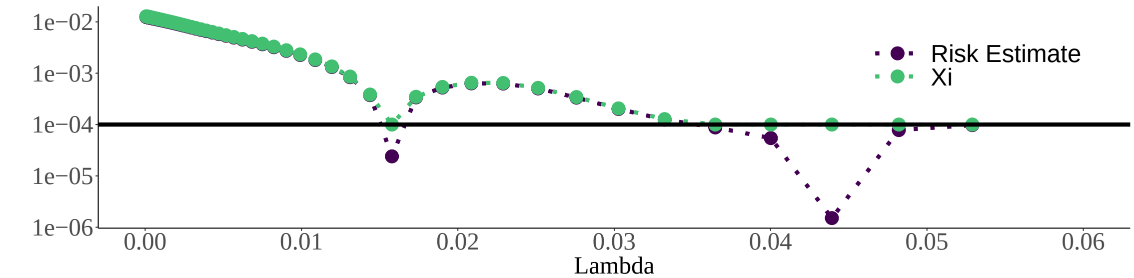
In particular, we draw a small number of cross-splits, i.e., $g = 20$ and compute the aggregated difference

$$b_{\lambda_i}(\mathsf{R}_{g,k}, D) = \frac{1}{g}\sum_{i=1}^{g} b_{\lambda_i}(\mathsf{r}_i, D) \tag{A.2.10}$$

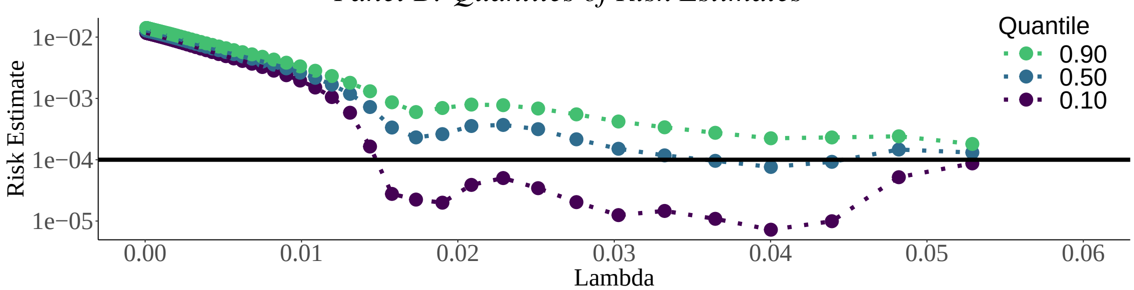
for each i in $1, \ldots, p-1$. Panel A of Figure A.2 displays the absolute value of these estimates

Figure A.2: Determining Error Tolerances for Cross-Validated Risk Estimation

Panel A: One Draw of Risk Estimates, The Choice of ξ_i



Panel B: Quantiles of Risk Estimates



Notes: Figure A.2 displays auxiliary measurements relating to the the performance of Algorithm 1 in the application to cross-validated Lasso, implemented in data from Casey et al. (2021b). Panel A displays the absolute value of the differences (A.2.10) computed on a random sample of $g = 20$ cross-splits, in purple, and the realized values of the error tolerances (A.2.13), in green. Panel B displays the quantiles of the absolute value of the differences (A.2.10), over replications of random samples of $g = 20$ cross-splits. In both cases, the y-axes is displayed in a logarithmic scale.

---

in purple, where the y-axis has been transformed to a logarithmic scale. The absolute value of the differences (A.2.10) are on the order 10^{-4} at the close-to-optimal values of λ, i.e., values of λ above about 0.01. The differences for smaller values of λ are much larger, and, as indicated by Figure 1.1, have larger conditional variances. These qualitative features are not particularly sensitive to residual randomness. Panel B of Figure A.2 displays the quantiles of the difference (A.2.10) across draws of $g = 20$ cross-splits.

These observations suggest the following approach for choosing the tolerances ξ_i. We would like to set $\xi_i = 10^{-4}$ for the large, close-to-optimal values of the regularization parameter, but would be willing to tolerate a larger error tolerance at smaller, sub-optimal values. This is practically important, as achieving the same level of precision for the small values of λ requires aggregation over many more cross-splits (as the conditional variances are much larger). We operationalize this approach in the following way. First, using the sample of $g = 20$ cross-splits that we used to estimate (A.2.10), we compute the p-values

$$q_{\lambda_i}(\mathsf{R}_{g,k}, D) = 1 - \Phi\left(\frac{b_{\lambda_i}(\mathsf{R}_{g,k}, D)}{\sqrt{\hat{v}_{\lambda_i}(\mathsf{R}_{g,k}, D)}}\right) \tag{A.2.11}$$

for each i in $1, \dots, p-1$, where $\hat{v}_{\lambda_i}(\mathsf{R}_{g,k}, D)$ denotes the sample-variance of the statistics $b_{\lambda_i}(\mathsf{r}_i, D)$ across the sample-splits. Let $\overline{\lambda}$ denote the largest value of the regularization parameter such that $q_{\overline{\lambda}}(\mathsf{R}_{g,k}, D) < 0.2$ and let the constant "scale" solve the equality

$$10^{-4} = \frac{b_{\overline{\lambda}}(\mathsf{R}_{g,k}, D)}{\mathsf{scale}} . \tag{A.2.12}$$

We set the error tolerances to

$$\xi_i = \max\{10^{-4}, b_{\lambda_i}(\mathsf{R}_{g,k}, D) \times \mathsf{scale}\} . \tag{A.2.13}$$

In other words, we set $\xi_i = 10^{-4}$ for all i with $q_{\lambda_i}(\mathsf{R}_{g,k}, D) \geq 0.2$ and increase the error tolerance in proportion to $b_{\lambda_i}(\mathsf{R}_{g,k}, D)$ for all other values of i. The realized values of ξ_i are displayed in Panel A of Figure A.2. These error tolerances are used throughout the main text.

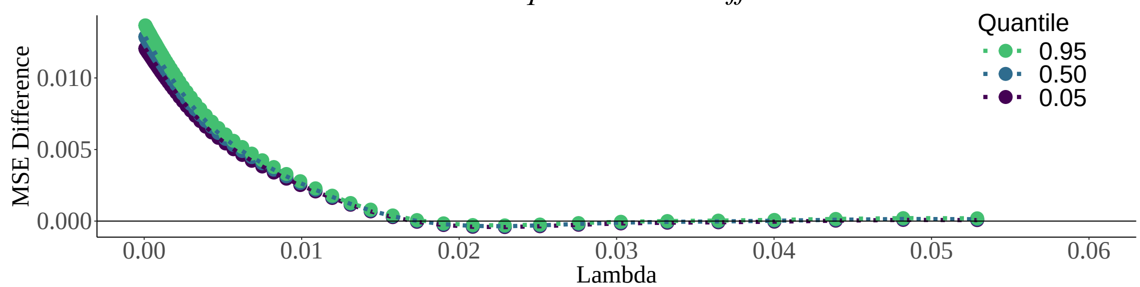
Panel A displays quantiles of the reproducibly aggregated statistic (1.3.12) across replications of the procedure, without truncation of the y-axis. Panel B of Figure A.2 displays estimates of the marginal reproducibility probability at each value of the regularization parameter, computed using $2,000$ replications of the procedure. Over the full range of the regularization parameter, the reproducibility error is very close to the nominal value $\beta = 0.05$.

A.2.3 Sample Stability

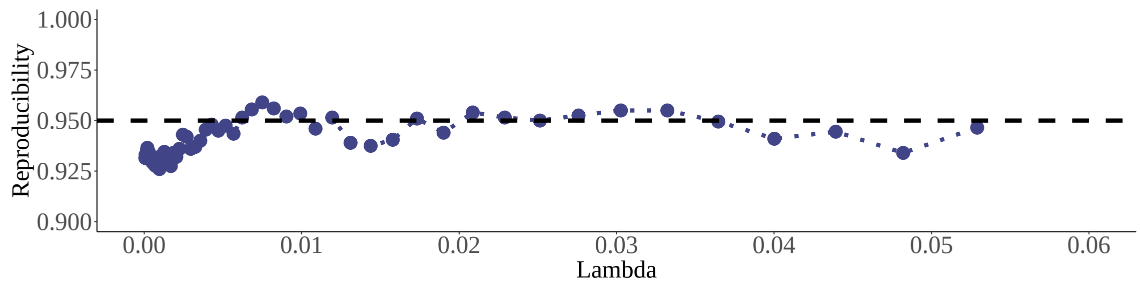
This appendix collects discussion and analysis concerning the (r, q)-sample stability $\sigma^{(r,q)}$ defined in Definition 1.4.1. First, in **??** A.2.3.1, we specialized our consideration to the case that the estimator $\hat{\eta}$ is a regularized empirical risk minimizer. Some closely related arguments are used in proof of Proposition 4 of Austern and Zhou (2020). Second, in **??** A.2.3.2 we compare we compare the stable-stability decay condition Assumption 1.4.3 to the conditions considered by Chen et al. (2022). The proof of the Theorem stated in **??** A.2.3.1 is given in **??** A.2.3.3

Figure A.3: Auxiliary Figures for Application to Cross-Validation

Panel A: Mean-Squared Error Differences



Panel B: Reproducibility



Notes: Figure A.2 displays auxiliary measurements relating to the the performance of Algorithm 1 in the application to cross-validated Lasso, implemented in data from Casey et al. (2021b). Panel A is analogous to Panel A of Figure 1.3, but the y-axis is not truncated. Panel B displays estimates of the marginal reproducibility probability for each value of the regularization parameter, where the nominal reproducibility error is set to $\beta = 0.05$.

A.2.3.1 Specialization to Regularized Empirical Risk Minimizers

Assume that the parameter η is an element of some closed convex space $H \subseteq \mathbb{R}^p$. Consider the estimator

$$\Psi\left(D_{\mathsf{s}}, \eta\right) = \frac{1}{b} \sum_{i \in \mathsf{s}} \psi\left(D_i, \eta\right) \tag{A.2.14}$$

$$\hat{\eta} = \underset{\eta \in H}{\arg\min} \left\{ \frac{1}{n-b} \sum_{i \in \tilde{\mathsf{s}}} \ell\left(D_i, \eta\right) + \lambda_{1,n} \|\eta\|_1 + \lambda_{2,n} \|\eta\|_2 \right\}, \tag{A.2.15}$$

where $\psi\left(\cdot,\cdot\right)$ are $\ell\left(\cdot,\cdot\right)$ functions and $\lambda_{1,n}, \lambda_{2,n} \geq 0$ are penalty parameters.

Let $\nabla \ell\left(d, \eta\right)$ and $\nabla^2 \ell\left(d, \eta\right)$ denote the gradient and Hessian of the function $\eta \mapsto \ell\left(d, \eta\right)$. Let $\nabla_j \ell\left(\cdot,\cdot\right)$ denote the jth component of $\nabla \ell\left(\cdot,\cdot\right)$. Similarly, we write

$$\overline{\nabla}_{\tilde{\mathsf{s}}} \ell\left(D, \eta\right) = \frac{1}{n-b} \sum_{i \in \tilde{\mathsf{s}}} \nabla \ell\left(D_i, \eta\right) \quad \text{and}$$

$$\overline{\nabla}^2_{\tilde{\mathsf{s}}}\ell\left(D,\eta\right)=\frac{1}{n-b}\sum_{i\in\tilde{\mathsf{s}}}\nabla^2\ell\left(D_i,\eta\right)$$

for the empirical averages of the gradients and the Hessian evaluated on the data in the set $\tilde{\mathsf{s}}$.

We impose the following set of restrictions. The first two restrictions concern the curvature of the loss function.

Assumption A.2.1. *The minimum eigenvalue of* $\overline{\nabla}^2_{\tilde{\mathsf{s}}}\ell\left(D,\eta\right)$ *is bounded below by* ρ *almost surely.*

Assumption A.2.2. *The loss function* $\ell\left(\cdot,\cdot\right)$ *is strictly convex and twice continuously differentiable in its second argument.*

Remark A.2.1. Under Assumption A.2.1, if $\rho>0$, then the dimension p of the parameter vector η is less than the number of observations n. To the best of our knowledge, analysis of the stability of ℓ_1 regularized empirical risk minimization in the regime with p larger than n is an open problem. Further analysis of the sample stability in the high-dimensional regime is an interesting direction for further research. ■

The second two restrictions are statistical. The first concerns the expected curvature of the moment function $\psi\left(\cdot,\cdot\right)$. The second is a sub-exponential tail bound on the gradient $\nabla_j\ell\left(\cdot,\cdot\right)$.

Assumption A.2.3. *The function* $\psi\left(\cdot,\cdot\right)$ *satisfies the inequality*

$$\mathbb{E}\left[\left(\psi(D_i,\eta)-\psi(D_i,\eta')\right)^r\right]\lesssim\mathbb{E}\left[\|\hat{\eta}-\hat{\eta}'\|_2^r\right] \tag{A.2.16}$$

for each pair $\eta,\eta'\in H$*.*

Assumption A.2.4. *The* ψ_1*-Orlicz norm bound*

$$\|\nabla_j\ell_j(D_i',\eta)-\nabla_j\ell(D_i,\eta)\|\lesssim\kappa^2 \tag{A.2.17}$$

holds for each pair $\eta\in H$*.*

Remark A.2.2. Assumption A.2.3 is satisfied in many problems of interest. For example, Chen et al. (2022) show that similar conditions hold when $\psi(\cdot,\cdot)$ is given by various moment functions used for semiparametric causal inference. Assumption A.2.4 is equivalent to the assumption that $\nabla_j\ell_j(D_i',\eta)-\nabla\ell_j(D_i,\eta)$ is sub-exponential with parameter κ^2. ■

The following theorem gives a bound on the (r, q)-sample stability.

Theorem A.2.4. *If Assumptions A.2.1, A.2.3, A.2.4, and C.1.1, if* $\rho + \lambda_{2,n} > 0$, *then*

$$\sigma^{(r,q)} \lesssim C_r p \left(\frac{\sqrt{q}}{n-b} \frac{\kappa}{\rho + \lambda_{2,n}} \right)^r + C_r \left(\lambda_{1,n} \frac{p}{\rho + \lambda_{2,n}} \right)^r ,$$

uniformly over all even integers r *and positive integers* $q \leq n - b$, *where* C_r *is some positive constant that depends only on* r.

Remark A.2.3. A necessary and sufficient condition for $\hat{\eta}$ to be consistent for the population risk minimizer associated with (A.2.15) is that $\lambda_{1,n} = o\left((n-b)^{-1}\right)$. See e.g., Knight and Fu (2000). Thus, so long as the penalty $\lambda_{1,n}$ is chosen in this regime, the sample stability $\sigma^{(r,q)}$ will be

$$O\left(\left(\frac{\sqrt{q}\kappa}{n-b} \right)^r \right)$$

up to terms that depend on the dimension of η, as required. ■

A.2.3.2 Comparison to Chen et al. (2022)

In the main text, we restrict attention to sample stable estimators. In a recent article, Chen et al. (2022) show that, under some regularity conditions, if a condition related to, but partially stronger than, sample stability is satisfied, then sample-splitting is unnecessary for the consistency and asymptotic normality of DML estimators. In this section, we comment on the difference between the notion of sample stability used in this article and the related condition used in Chen et al. (2022).

First, for ease of reference, we recall our definition of sample stability. We restrict attention to the case that $q = 1$ and $r = 2$, as this is what is relevant to make the comparison. Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$ and let i be an arbitrary element of s. Let D' denote an independent and identical copy of the data D. For each i in $[n]$, let $\tilde{D}^{(i)}$ be constructed by replacing D_i with D'_i in D. Let I be a randomly selected element of $\tilde{\mathsf{s}}$. In this paper, we refer to the quantity

$$\sigma^{(2,1)} = \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(I)}_{\tilde{\mathsf{s}}})) \right)^2 \right] \tag{A.2.18}$$

as the $(2, 1)$-sample stability. In particular, in the main text, we restrict attention to settings where the bound

$$\sigma^{(2,1)} \lesssim \left(\frac{1}{n-b} \right)^2 \tag{A.2.19}$$

holds.

Chen et al. (2022) consider settings where the nuisance parameter estimator $\hat{\eta}$ is not constructed with sample-splitting. Likewise, they consider moments of the form

$$\tilde{\sigma}^{(2,1)} = \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\,(D)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(i)}))\right)^2\right] , \tag{A.2.20}$$

where i is an arbitrary element of $[n]$. Observe that, in contrast to (A.2.18), the moment (A.2.20) is evaluated using the same observation D_i used to perturb the nuisance parameter estimate—and that the nuisance parameter estimate is evaluated using the complete data. Roughly speaking, Chen et al. (2022) show that, under some regularity conditions, DML estimators are consistent and asymptotically normal if the bound

$$\tilde{\sigma}^{(2,1)} \lesssim n^{-1} \tag{A.2.21}$$

holds for a suitable choice of the function $\psi(\cdot, \cdot)$. Moreover, they show that, under certain conditions on the choice of tuning parameters, nuisance parameter estimators constructed with a bagged 1-nearest-neighbor estimator satisfies this bound.

Although the right-hand-side of the bound (A.2.21) is larger than the bound (A.2.19), control of the moment (A.2.20) can be more difficult to achieve than control of the moment (A.2.18). To see the substantive differences between the conditions (A.2.19) and (A.2.20), we consider an artificial, but illustrative, example adapted from an example given in the introduction to Chernozhukov et al. (2018). Suppose that the observation D_i consists of the real-valued quantities Y_i, W_i, and X_i. Furthermore, suppose that the nuisance parameter $\eta(x)$ denotes the conditional expectation $\mathbb{E}[W_i \mid X_i = x]$ and that

$$\psi(D_i, \eta) = Y_i(W_i - \mathbb{E}[W_i \mid X_i]) = Y_i(W_i - \eta(x)) \, . \tag{A.2.22}$$

Suppose that the nuisance parameter estimator takes the form

$$\hat{\eta}(D_{\tilde{\mathsf{s}}})(x) = \sum_{j \in \tilde{\mathsf{s}}} w_j(x) \hat{\eta}_j(x) \tag{A.2.23}$$

where $\tilde{\mathsf{s}}$ is an arbitrary subset of $[n]$, $\hat{\eta}_j(x)$ is an estimator of the nuisance parameter $\eta(x)$ based on the observation D_j, and the weight $w_j(x)$ again depends on the observation D_j and the set $\tilde{\mathsf{s}}$. We refer the reader to discussion in Section 1 of Chernozhukov et al. (2018) and Section 2 of Chen et al. (2022) for further context on why understanding quantities of the form (A.2.22) is essential for establishing the consistency of DML estimators.[7]

To study the sample stability (A.2.18) in this example, i.e., in the sense considered in this article, we are interested in the difference

$$\psi(D_i, \hat{\eta}\,(D_{\tilde{\mathsf{s}}})) - \psi(D_i, \hat{\eta}(\tilde{D}_{\tilde{\mathsf{s}}}^{(I)})) = Y_i(w_I(X_i)\hat{\eta}_I(X_i) - w_I'(X_i)\hat{\eta}_I'(X_i)) \tag{A.2.24}$$

where $w_I'(X_i)$ and $\hat{\eta}_I'(X_i))$ denote the updated weight and estimator associated with the perturbed data $\tilde{D}_{\tilde{\mathsf{s}}}^{(I)}$. On the other hand, in order to consider the stability (A.2.20), i.e., in the sense of Chen et al. (2022), we are interested in the difference

$$\psi(D_i, \hat{\eta}\,(D)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(i)})) = Y_i(w_i(X_i)\hat{\eta}_i(X_i) - w_i'(X_i)\hat{\eta}_i'(X_i)) \tag{A.2.25}$$

where, as before, $w_i'(X_i)$ and $\hat{\eta}_i'(X_i)$ denote the updated weight and estimator associated with the perturbed data $\tilde{D}^{(i)}$.

The quantities (A.2.24) and (A.2.25) will behave differently if the nuisance parameter estimator exhibits overfitting. In particular, to take a highly stylized example, suppose that the weights satisfy

$$w_i(X_j) = \begin{cases} |\tilde{\mathsf{s}}|^{-1/3}, & i = j, \\ |\tilde{\mathsf{s}}|^{-1}, & \text{otherwise.} \end{cases} \tag{A.2.26}$$

That is, the nuisance parameter estimator (A.2.23) meaningfully up-weights the estimate obtained from the evaluation point X_j, if X_j is part of the training data $\tilde{\mathsf{s}}$, and is otherwise equally balanced. Overfitting to the training data, in a perhaps less stylized form, is a well-known property of some machine learning estimators.

It is easy to see that, in this case, if all other quantities are bounded, we should expect that

$$\begin{aligned} \sigma^{(2,1)} &= \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\,(D_{\tilde{\mathsf{s}}})) - \psi(D_i, \hat{\eta}(\tilde{D}_{\tilde{\mathsf{s}}}^{(I)}))\right)^2\right] \\ &= \mathbb{E}\left[\left(Y_i(w_I(X_i)\hat{\eta}_I(X_i) - w_I'(X_i)\hat{\eta}_I'(X_i))\right)^2\right] \lesssim \left(\frac{1}{|\tilde{\mathsf{s}}|}\right)^2 \end{aligned} \tag{A.2.27}$$

whereas

$$\tilde{\sigma}^{(2,1)} = \mathbb{E}\left[\left(\psi(D_i, \hat{\eta}\,(D)) - \psi(D_i, \hat{\eta}(\tilde{D}^{(i)}))\right)^2\right]$$

[7] In particular, establishing the Stochastic Equicontinuity of a given Neyman Orthogonal moment amounts to bounding scaled averages of quantities of the form (A.2.22).

$$= \mathbb{E}\left[\left(Y_i(w_i(X_i)\hat{\eta}_i(X_i) - w_i'(X_i)\hat{\eta}_i'(X_i))\right)^2\right] \lesssim \left(\frac{1}{|\tilde{\mathsf{s}}|}\right)^{2/3}. \tag{A.2.28}$$

In particular, we can see that all that is needed for us to expect the moment (A.2.27) to be approximately of order $O((n-b)^{-2})$, is that the *randomly selected* weights $w_I(X_i)$ *tend* to be close to, or smaller than, $(n-b)^{-1}$. The moment (A.2.28), by contrast, is highly sensitive to the weight $w_i(X_i)$ placed on the evaluation point X_i.

The main point of this example is that establishing that a nuisance parameter estimator is stable, in the sense of Chen et al. (2022), requires showing that it does not overfit its training data, i.e., that evaluations of points *in-sample* are not overly-determined by a few observations. By contrast, all that is needed to establish that a nuisance parameter estimator is sample stable, in the sense used in this article, is that evaluations of points *out-of-sample* are not overly concentrated on a small number of training samples. In general, we should expect that out-of-sample balance is more likely to occur in practice than in-sample balance.

A.2.3.3 Proof of Theorem A.2.4

By Assumptions A.2.1 and C.1.1, the objective function for the estimator (A.2.15) is strongly convex. Thus, there is a unique solution to (A.2.15) for any data D. Let $\hat{\eta}$ and $\hat{\eta}'$ denote the solutions to (A.2.15) for the data D and $\tilde{D}^{(\mathsf{q})}$ respectively. Let $\partial f(x)$ denote the subgradient set of the function $x \mapsto f(x)$. The Karush-Kuhn-Tucker condition for the program (A.2.15) is given by

$$\overline{\nabla}_{\tilde{\mathsf{s}}}\ell(D,\eta) + \lambda_{2,n}\hat{\eta} + \lambda_{1,n}\hat{z} = 0, \tag{A.2.29}$$

where $\hat{z} \in \partial\|\hat{\eta}\|_1$ is the subgradient associated with the Lasso penalty. Observe that, in this case, $\hat{z} \in \text{sign}(\hat{\eta})$, where we set $\text{sign}(0) = [-1,1]$. Let $\hat{z}$ and $\hat{z}'$ denote the subgradients obtained from D and $\tilde{D}^{(\mathsf{q})}$, respectively. As $\ell(d,\cdot)$ is twice continuously differentiable under Assumption C.1.1 (i), we have that

$$\begin{aligned}
&\overline{\nabla}_{\tilde{\mathsf{s}}}\ell(D,\hat{\eta}') + \lambda_{2,n}\hat{\eta}' + \lambda_{1,n}\hat{z} \\
&= \overline{\nabla}_{\tilde{\mathsf{s}}}\ell(D,\hat{\eta}) + \lambda_{2,n}\hat{\eta} + \lambda_{1,n}\hat{z} + \left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}}\ell(D,\tilde{\eta}) + \lambda_{2,n}I_d\right)(\hat{\eta}-\hat{\eta}') \\
&= \left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}}\ell(D,\tilde{\eta}) + \lambda_{2,n}I_d\right)(\hat{\eta}-\hat{\eta}')
\end{aligned} \tag{A.2.30}$$

for some vector $\tilde{\eta}$, by a Taylor expansion and the optimality condition (A.2.29). On the other hand, we have that

$$\begin{aligned}\overline{\nabla}_{\tilde{\mathsf{s}}}\ell\left(D,\hat{\eta}'\right)+\lambda_{2,n}\hat{\eta}'+\lambda_{1,n}\hat{z} &= \overline{\nabla}_{\tilde{\mathsf{s}}}\ell(\tilde{D}^{(\mathsf{q})},\hat{\eta}')+\lambda_{2,n}\hat{\eta}'+\lambda_{1,n}\hat{z}' \\ &\quad+\frac{1}{n-b}\left(\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\right)+\lambda_{1,n}\left(\hat{z}-\hat{z}'\right) \\ &= \frac{1}{n-b}\left(\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\right)+\lambda_{1,n}\left(\hat{z}-\hat{z}'\right)\end{aligned} \tag{A.2.31}$$

again by the optimality condition (A.2.29). Thus, we have that

$$\begin{aligned}&\left(\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}}\ell\left(D,\tilde{\eta}\right)+\lambda_{2,n}I_p\right)\left(\hat{\eta}-\hat{\eta}'\right) \\ &=\frac{1}{n-b}\left(\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\right)+\lambda_{1,n}\left(\hat{z}-\hat{z}'\right)\end{aligned} \tag{A.2.32}$$

by (A.2.30) and (A.2.31). Observe that the the matrix $\overline{\nabla}^{(2)}_{\tilde{\mathsf{s}}}\ell\left(D,\tilde{\eta}\right)+\lambda_{2,n}I_p$ is invertible by Assumption A.2.1. Thus, we find that

$$\begin{aligned}\hat{\eta}-\hat{\eta}' &= \frac{1}{n-b}\left(\overline{\nabla}^{(2)}\ell\left(D,\tilde{\eta}\right)+\lambda_2 I_p\right)^{-1}\left(\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\right) \\ &\quad+\lambda_{1,n}\left(\overline{\nabla}^{(2)}\ell\left(D,\tilde{\eta}\right)+\lambda_2 I_d\right)^{-1}\left(\hat{z}-\hat{z}'\right)\end{aligned}$$

and that consequently

$$\begin{aligned}\|\hat{\eta}-\hat{\eta}'\|_2 &\lesssim \frac{1}{n-b}\frac{\|\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\|_2}{\rho+\lambda_{2,n}}+\lambda_{1,n}\frac{\|\hat{z}-\hat{z}'\|_2}{\rho+\lambda_{2,n}} \\ &\lesssim \frac{1}{n-b}\frac{\|\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\|_2}{\rho+\lambda_{2,n}}+\lambda_{1,n}\frac{p}{\rho+\lambda_{2,n}}\end{aligned} \tag{A.2.33}$$

by Assumption A.2.1.

Now, observe that

$$\begin{aligned}\sigma^{(r,q)} &= \mathbb{E}\left[\left(\psi(D_i,\hat{\eta})-\psi(D_i,\hat{\eta}')\right)^r\right] \\ &\lesssim \mathbb{E}\left[\|\hat{\eta}-\hat{\eta}'\|_2^r\right]\end{aligned}$$

$$\lesssim \mathbb{E}\left[\left(\frac{1}{n-b}\frac{\|\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\|_2}{\rho+\lambda_{2,n}}+\lambda_{1,n}\frac{p}{\rho+\lambda_{2,n}}\right)^r\right] \tag{A.2.34}$$

$$\lesssim 2^r\left(\frac{1}{n-b}\frac{1}{\rho+\lambda_{2,n}}\right)^r\mathbb{E}\left[\|\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\|_2^r\right]$$
$$+\,2^r\left(\lambda_{1,n}\frac{p}{\rho+\lambda_{2,n}}\right)^r, \tag{A.2.35}$$

where the first inequality follows from Assumption A.2.3, the second inequality follows from (A.2.33), and the third inequality follows from the Binomial Theorem and Cauchy-Schwarz. By Assumption A.2.4, we have that

$$\|\sum_{i\in\mathsf{q}}\nabla_j\ell_j(D_i',\hat{\eta}')-\nabla\ell_j(D_i,\hat{\eta}')\|_{\psi_1}\lesssim q\|\nabla_j\ell_j(D_i',\hat{\eta}')-\nabla\ell_j(D_i,\hat{\eta}')\|_{\psi_1}\lesssim q\kappa\,. \tag{A.2.36}$$

Consequently, we have that

$$\mathbb{E}\left[\|\sum_{i\in\mathsf{q}}\nabla\ell(D_i',\hat{\eta}')-\nabla\ell(D_i,\hat{\eta}')\|_2^r\right]$$
$$\lesssim\sum_{j\in[p]}\mathbb{E}\left[\left|\sum_{i\in\mathsf{q}}\nabla_j\ell_j(D_i',\hat{\eta}')-\nabla\ell_j(D_i,\hat{\eta}')\right|^{r/2}\right] \tag{Jensen}$$
$$\lesssim p((r/2)!)^{r/2}(\kappa q)^{r/2} \tag{A.2.37}$$

where, for the final inequality, we have used the fact that $\|X\|_{r/2}\lesssim((r/2)!)\|X\|_{\psi_1}$ (see e.g., Section 2.2 of Van Der Vaart and Wellner (1996)). Putting the pieces together, we find that

$$\sigma^{(r,q)}\lesssim 2^r((r/2)!)^{r/2}p\left(\frac{q^{1/2}}{n-b}\frac{\kappa^{1/2}}{\rho+\lambda_{2,n}}\right)^r+2^r\left(\lambda_{1,n}\frac{p}{\rho+\lambda_{2,n}}\right)^r, \tag{A.2.38}$$

as required. ■

A.3. Proofs for Results Stated in Section 1.3

A.3.1 Proof of Theorem 1.3.1

For each collection $\mathsf{r} = (\mathsf{s}_j)_{j=1}^k$ in $\mathcal{R}_{n,k,b}$, we write

$$\overline{a}(\mathsf{r}, D) = \frac{1}{k}\sum_{j=1}^{k} T(\mathsf{s}_j, D) - \mathbb{E}\left[T(\mathsf{s}_j, D) \mid D\right].$$

Define the oracle stopping time

$$g^\star = \arg\min_{g \geq 2} \left\{ v_{g,k}(D) \leq \frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \right\}.$$

Observe that $v_{g,k}(D) = (1/g)\cdot v_{1,k}(D)$ as the collections of cross-splits are drawn independently and identically. Moreover, we have that

$$\frac{\hat{v}(\mathsf{R}_{g,k}, D)}{v_{g,k}(D)} = \frac{g\hat{v}(\mathsf{R}_{g,k}, D)}{v_{1,k}(D)} \overset{\text{a.s}}{\to} 1 \tag{A.3.1}$$

as $g \to \infty$, by the strong law of large numbers. Now, observe that

$$\frac{\hat{g}\cdot\hat{v}(\mathsf{R}_{\hat{g},k}, D)}{v_{1,k}(D)} \leq \frac{\hat{g}}{2v_{1,k}(D)}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \leq \frac{\hat{g}\cdot\hat{v}(\mathsf{R}_{\hat{g}-1,k}, D)}{v_{1,k}(D)}, \tag{A.3.2}$$

by definition. Consequently, as $\hat{g} \to \infty$ and

$$g^\star\left(2v_{1,k}(D)\left(\frac{z_{1-\beta/2}}{\xi}\right)^2\right) \to 1$$

as $\xi \to 0$, we have that

$$\hat{g}/g^\star \overset{\text{a.s}}{\to} 1 \tag{A.3.3}$$

as $\xi \to 0$ by (A.3.1) and (A.3.2).

Define the objects

$$U\left(\mathsf{R}_{g,k}, D\right) = \frac{1}{g^\star}\left(\sum_{i=1}^{g}\overline{a}\left(\mathsf{r}_i, D\right) - \sum_{i=1}^{g^\star}\overline{a}\left(\mathsf{r}_i, D\right)\right) \quad \text{and} \tag{A.3.4}$$

$$Q\left(\mathsf{R}_{g,k}, D\right) = \left(1 - \frac{g}{g^\star}\right)\left(a(\mathsf{R}_{g,k}, D) - \mathbb{E}\left[T(\mathsf{s}_j, D) \mid D\right]\right) \tag{A.3.5}$$

Observe that we can decompose

$$\begin{aligned}
&\left(a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right) / \sqrt{2v_{g^\star,k}(D)} \\
&= \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)\right) \\
&+ \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)\right) + \sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(Q(\mathsf{R}_{\hat{g},k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)\right) .
\end{aligned} \tag{A.3.6}$$

First, we have that

$$\sqrt{\frac{g^\star}{2v_{1,k}(D)}} \left(Q(\mathsf{R}_{\hat{g},k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)\right) = o_p(1)$$

almost surely as $\xi \to 0$ by (A.3.3). To handle the terms involving (A.3.4), fix $\varepsilon > 0$ and observe that

$$\begin{aligned}
&P\left\{\left|\sum_{i=1}^{\hat{g}} a(\mathsf{r}_i, D) - \sum_{i=1}^{g^\star} a(\mathsf{r}_i, D)\right| > \varepsilon\sqrt{g^\star} \mid D\right\} \\
&\leq P\left\{\left|\sum_{i=1}^{\hat{g}} a(\mathsf{r}_i, D) - \sum_{i=1}^{g^\star} a(\mathsf{r}_i, D)\right| > \varepsilon\sqrt{g^\star}, \hat{g} \in \left[g^\star\left(1-\varepsilon^3\right), g^\star\left(1+\varepsilon^3\right)\right] \mid D\right\} \\
&+ P\left\{\left|\sum_{i=1}^{\hat{g}} a(\mathsf{r}_i, D) - \sum_{i=1}^{g^\star} a(\mathsf{r}_i, D)\right| > \varepsilon\sqrt{g^\star}, \hat{g}' \notin \left[g^\star\left(1-\varepsilon^3\right), g^\star\left(1+\varepsilon^3\right)\right] \mid D\right\} \\
&\leq P\left\{\max_{g^\star(1-\varepsilon^3)\leq g\leq g^\star}\left|\sum_{i=1}^{g} a(\mathsf{r}_i, D)\right| > \varepsilon\sqrt{g^\star} \mid D\right\} \\
&+ P\left\{\max_{g^\star\leq g\leq(1+\varepsilon^3)g^\star}\left|\sum_{i=1}^{g} a(\mathsf{r}_i, D)\right| > \varepsilon\sqrt{g^\star} \mid D\right\} \\
&+ P\left\{\hat{g} \notin \left[g^\star\left(1-\varepsilon^3\right), g^\star\left(1+\varepsilon^3\right)\right] \mid D\right\} .
\end{aligned} \tag{A.3.7}$$

The third term in (A.3.7) is smaller than ε for all sufficiently small ξ by (A.3.3). To handle the first two terms, observe that

$$P\left\{\max_{g^\star\leq g\leq(1+\varepsilon^3)g^\star}\left|\sum_{i=1}^{g} a(\mathsf{r}_i, D)\right| > \varepsilon\sqrt{\hat{g}} \mid D\right\}$$

$$\leq \frac{1}{\varepsilon^2}\frac{1}{g^\star}\operatorname{Var}\left(\sum_{i=1}^{\lfloor \varepsilon^3 g^\star \rfloor} a(\mathsf{r}_i, D) \mid D\right)$$
$$\leq \varepsilon \operatorname{Var}(a(\mathsf{r}_i, D)),$$

almost surely, where the first inequality follows from Kolmogorov's maximal inequality (see e.g., Proposition 2.3.16 of Dembo (2021)) and the second inequality follows from the fact that the cross-splits r_i are independent. An analogous bound holds for the remaining term. Thus, we have

$$\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)\right) = o_p(1)$$

almost surely as $\xi \to 0$. Hence, we have that

$$\begin{aligned}
&P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| > \xi \mid D\right\} \\
&= P\left\{\left|\sqrt{\frac{g^\star}{2v_{1,k}(D)}}\left(\frac{1}{g^\star}\sum_{i=1}^{g^\star} a(\mathsf{r}_{i,k}, D) - a(\mathsf{r}'_{i,k}, D)\right)\right| > z_{1-\beta/2} \mid D\right\} + o(1) \\
&= 1 - \beta + o(1)
\end{aligned}$$

as $\xi \to 0$ almost surely, by the central limit theorem. ■

A.3.2 Proof of Theorem 1.3.2

Recall that the cross-split r is drawn independently and uniformly from the collection $\mathcal{R}_{n,k,b}$. The result is based on the bound

$$\begin{aligned}
\mathbb{E}\left[f(a(\mathsf{R}_{g,k}, D))\right] &= \mathbb{E}\left[\mathbb{E}\left[f\left(\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} a(\mathsf{r}_i, D)\right) \mid \hat{g}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} f\left(a(\mathsf{r}_i, D)\right) \mid \hat{g}\right]\right] && \text{(Jensen)} \\
&= \mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} f\left(a(\mathsf{r}_i, D)\right)\right] \\
&= \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right)\right] + \mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}} \left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right)\right].
\end{aligned} \tag{A.3.8}$$

In particular, it will suffice to show that the second term in (A.3.8) is weakly negative. To do this, we apply the following optimal stopping theorem.

Theorem A.3.1 (Theorem 5.7.5, Durrett (2019)). *Suppose that X_g is a real-valued supermartingale, that $\mathcal{F}_g$ is the σ-algebra generated by $X_1, \ldots, X_g$, and that there exists some constant B such that*

$$\mathbb{E}\left[|X_{g+1} - X_g| \mid \mathcal{F}_g\right] \leq B \tag{A.3.9}$$

almost surely. If $\hat{g}$ is a stopping time with respect to the filtration $\{\mathcal{F}_g\}_{g=1}^{\infty}$, then $\mathbb{E}\left[X_{\hat{g}}\right] \leq \mathbb{E}\left[X_1\right]$.

Consider the sequence of random variables

$$X_g = \frac{1}{g}\sum_{i=1}^{g}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \tag{A.3.10}$$

Observe that

$$\begin{aligned}
\mathbb{E}\left[X_{g+1} \mid \mathcal{F}_g, D\right] &= \mathbb{E}\left[\frac{1}{g+1}\sum_{i=1}^{g+1}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \mid \mathcal{F}_g, D\right] \\
&= \frac{1}{g+1}\mathbb{E}\left[f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right] \mid \mathcal{F}_g, D\right] \\
&+ \frac{g}{g+1}\frac{1}{g}\sum_{i=1}^{g}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \\
&= \frac{g}{g+1}X_g\ .
\end{aligned} \tag{A.3.11}$$

Consequently, X_g is a supermartingale. To verify the condition (A.3.9), observe that, as the collection $\mathcal{R}_{n,k,b}$ is finite, there exists some positive, finite function $B(D)$ such that

$$f(a(\mathsf{r}, D)) \leq B(D) \tag{A.3.12}$$

almost surely. Thus, we find that

$$\begin{aligned}
\mathbb{E}\left[|X_{g+1} - X_g| \mid \mathcal{F}_g, D\right] &= \mathbb{E}\left[\left|\frac{1}{g+1}\sum_{i=1}^{g+1} f\left(a(\mathsf{r}_i, D)\right) - \frac{1}{g}\sum_{i=1}^{g} f\left(a(\mathsf{r}_i, D)\right)\right| \mid \mathcal{F}_g, D\right] \\
&\leq \frac{1}{g+1}\mathbb{E}\left[|f\left(a(\mathsf{r}, D)\right)|\right] + \frac{1}{g(g+1)}\sum_{i=1}^{g}\mathbb{E}\left[|f\left(a(\mathsf{r}_i, D)\right)|\right] \\
&\leq \frac{2}{g+1}B(D)
\end{aligned} \tag{A.3.13}$$

almost surely. Hence, by Theorem A.3.1, we have

$$\mathbb{E}\left[\frac{1}{\hat{g}}\sum_{i=1}^{\hat{g}}\left(f\left(a(\mathsf{r}_i, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right]\right) \mid D\right]$$
$$\leq \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) - \mathbb{E}\left[f\left(a(\mathsf{r}, D)\right) \mid D\right] \mid D\right] = 0 \tag{A.3.14}$$

almost surely, completing the proof. ■

A.4. General Non-Asymptotic Results

In this appendix, we give a general, non-asymptotic analysis of the performance of Algorithm 1. We begin, in Appendix A.4.1, by studying the concentration and normal approximation of the aggregate statistic $a(\mathsf{R}_{k,g}, D)$. Equipped with these results, in Appendix A.4.2, we characterize the accuracy of the nominal error rate of Algorithm 1. In Appendix A.4.3, we show that these general results imply the results stated in Section 1.4, by restricting attention to sample-stable, cross-split statistics. Proofs are given in Appendices A.4.4 and A.4.5.

A.4.1 Concentration and Normal Approximation

We begin by giving a non-asymptotic characterization of the concentration and normal approximation of the statistic $a(\mathsf{R}_{g,k}, D)$. To state these results, we require a slightly more general definition of sample stability.

Definition A.4.1 (General Sample Stability)**.** Let D' denote an independent and identical copy of the data D. For each $\mathsf{q} \subseteq [n]$, let $\tilde{D}^{(\mathsf{q})}$ be constructed by replacing D_i with D_i' in D for each i in q. Fix a set $\mathsf{s} \subseteq \mathcal{S}_{n,b}$. Let i be an arbitrary element of s. Let q be a randomly selected subset of $\tilde{\mathsf{s}}$ of cardinality q. We refer to the quantities

$$\sigma_{\mathsf{valid}}^{(r)} = \mathbb{E}\left[\left|\psi(D_I, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D_I', \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right))\right|^r\right] \quad \text{and}$$
$$\sigma_{\mathsf{train}}^{(r,q)} = \mathbb{E}\left[\left|\psi(D_i, \hat{\eta}\left(D_{\tilde{\mathsf{s}}}\right)) - \psi(D_i, \hat{\eta}(\tilde{D}_{\tilde{\mathsf{s}}}^{(\mathsf{q})}))\right|^r\right]$$

as the rth-order validation and (r, q)th-order training sample stabilities, respectively. Similarly, we refer to the quantity

$$\sigma_{\mathsf{max}}^{(r)} = \max\left\{\sigma_{\mathsf{valid}}^{(r)}, \sigma_{\mathsf{train}}^{(r,1)}\right\} \tag{A.4.1}$$

as the rth-order full sample stability.

In other words, we define separate stabilities associated with resampling data in the subsets s and $\tilde{\mathsf{s}}$. We refer to these subsets as the "validation" and "training" samples, respectively. The training sample stability $\sigma^{(r,q)}_{\mathsf{train}}$ is identical to the sample stability defined in Section 1.4.

With this in place, we give a large deviations bound for the statistic $a(\mathsf{R}_{g,k}, D)$ about its conditional mean. The primary difficulty is accommodating the dependence in the summands of $a(\mathsf{R}_{g,k}, D)$ across elements of the same cross-split. We tackle this problem through the method of exchangeable pairs (Stein, 1986). In particular, we construct a suitable exchangeable pair with a coupling argument due to Chatterjee (2005) and apply a method for deriving concentration inequalities with exchangeable pairs, also due to Chatterjee (2005, 2007). Some preliminary results that facilitate the application of this approach are given in Appendix A.4.4. Proofs for results stated in this appendix are then given in Appendix A.4.5.

Theorem A.4.1. *Suppose that Assumptions 1.4.1 and B.2.3 hold, that the data D are independently and identically distributed, and that the statistic $a(\mathsf{R}_{1,k}, D)$ has a non-zero variance conditional on D almost surely. If the fourth-order split stability $\zeta^{(4)}$ is finite, then, for each $\varepsilon > 0$, the inequality*

$$P\left\{ |a(\mathsf{R}_{g,k}, D) - \overline{a}\,(D)\,| \leq \sqrt{\frac{2^4(2-\varphi k-\varphi)^2}{\delta}\frac{\Gamma_{k,\varphi,b}\log(\varepsilon^{-1})}{g}} \mid D \right\} \geq 1 - \frac{\varepsilon}{2} \quad \text{(A.4.2)}$$

holds with probability greater than $1-\delta$ as D varies, where

$$\Gamma_{k,\varphi,b} = \left(\frac{1-k\varphi}{1-\varphi}\right) 4\sigma^{(2)}_{\mathsf{max}} + \left(\frac{k\varphi-\varphi}{1-\varphi}\right)\sigma^{(2,b-1)}_{\mathsf{train}} \quad \text{(A.4.3)}$$

and $\varphi = b/n$.

Remark A.4.1. The quantity $\Gamma_{k,\varphi,b}$ interpolates between $\sigma^{(2)}_{\mathsf{max}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ as k varies between its bounds 1 and $1/\varphi$. In the case of independent splitting, i.e., $k=1$, the rate of concentration for the statistic $a(\mathsf{R}_{g,k}, D)$ is driven by the full sample stability $\sigma^{(2)}_{\mathsf{max}}$. On the other hand, in the case of cross-splitting, concentration depends only on the training sample stability $\sigma^{(2,b-1)}_{\mathsf{train}}$. In a previous draft of this paper, we give estimates of $\sigma^{(2)}_{\mathsf{valid}}$ and $\sigma^{(2,b-1)}_{\mathsf{train}}$ in applications to treatment effect estimation and cross-validated risk estimation. We show that the training stability is dramatically smaller than the validation stability and decreases rapidly as b decreases.[8] To see this, observe that we should next expect the validation sample-stability $\sigma^{(r)}_{\mathsf{valid}}$ to change with k. That is, residual randomness is smaller for full

[8]See Figure 3 of the preprint posted at `https://arxiv.org/abs/2311.14204v2`.

cross-splitting, i.e., $k = n/b$, than for independent splitting. ■

Next, we derive bounds for the centered conditional moments of $a\left(\mathsf{R}_{g,k}, D\right)$ through an argument closely related to the proof of Theorem A.4.1. Bounds of this form are known as Burkholder-Davis-Gundy inequalities (Burkholder, 1973).

Theorem A.4.2. *Suppose that Assumptions 1.4.1 and B.2.3 hold and that the data D are independent and identically distributed. If $r = 2^{c-1}$ for some positive integer c and the $4r$th-order split stability $\zeta^{(4r)}$ is finite, then, for each $\delta >$, the inequality*

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r} \mid D\right] \leq (2r-1)^r \left(\frac{2^4(2-\varphi k-\varphi)^2}{g}\right)^r \Gamma_{k,\varphi,b}^{(r)}\,, \tag{A.4.4}$$

holds with probability greater than $1-\delta$ as D varies, where

$$\Gamma_{k,\varphi,b}^{(r)} = \left(\frac{1-k\varphi}{1-\varphi}\right) 2^{2r} \sigma_{\mathsf{max}}^{(2r,1)} + \left(\frac{k\varphi-\varphi}{1-\varphi}\right) \sigma_{\mathsf{valid}}^{(2r,b-1)} \tag{A.4.5}$$

and $\varphi = b/n$.

Remark A.4.2. By setting $r = 1$, the inequality (A.4.4) gives a variance bound. In this case, the right-hand side of the inequality (A.4.4) is equal to negative one times the inverse of the right-hand side of the inequality (A.4.2). In this sense, the concentration inequality Theorem A.4.1 can be thought of as an exponential Efron-Stein inequality adapted to the dependence inherent in our problem (Boucheron et al., 2003). For values of r greater than 1, the unconditional quantity $\Gamma_{k,\varphi,b}^{(r)}$ again interpolates between the $2r$th order full sample stability $\sigma_{\mathsf{max}}^{(2r,1)}$ and the $(2r, b-1)$th order training sample stability $\sigma_{\mathsf{valid}}^{(2r,b-1)}$ as k increases from 1 to $1/\varphi$. ■

Remark A.4.3. An analogous, unconditional, result will follow from the same argument. In particular, under the same conditions, we have that

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r}\right] \leq (2r-1)^r \left(\frac{2^4(2-\varphi k-\varphi)^2}{g}\right)^r \Gamma_{k,\varphi,b}^{(r)}\,. \tag{A.4.6}$$

We opt to state the conditional version of the result thought, as, arguably, the conditional behavior of the statistic is of primary interest. ■

In turn, we bound the distance between the conditional distribution of $a\left(\mathsf{R}_{g,k}, D\right)$ and the standard normal distribution. The result follows by combining the bounds derived in Theorem A.4.2 with a standard Berry-Esseen inequality (Shevtsova, 2011).

Theorem A.4.3. *Suppose that Assumptions 1.4.1 and B.2.3 hold, that the data D are independently and identically distributed, and that the statistic $a(\mathsf{R}_{1,k}, D)$ has a non-zero variance conditional on D almost surely. If the eighth-order split stability $\zeta^{(8)}$ is finite, then for all g, the inequality*

$$\begin{aligned}\sup_{z\in\mathbb{R}}\left|P\left\{\frac{a\left(\mathsf{R}_{g,k},D\right)-\overline{a}\left(D\right)}{\sqrt{v_{g,k}\left(D\right)}}\leq z\mid D\right\}-\Phi(z)\right|\\ \lesssim\frac{2^{6}3^{2}}{\delta}\frac{\left(2-\varphi k-\varphi\right)^{3}}{g^{1/2}}\left(\frac{(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{v_{1,k}(D)}\right)^{3/2}\end{aligned}\tag{A.4.7}$$

is satisfied with probability greater than $1-\delta$ as D varies, where $\Phi(\cdot)$ is the standard normal c.d.f.

Remark A.4.4. To the best of our knowledge, the only Stein's method central limit theorem in the Kolmogorov distance, applicable to our setting, is given in Zhang (2022), which generalizes the argument of Shao and Zhang (2019). In Appendix A.4.6, we show that this central limit theorem implies a bound that does not reduce as either g or k increase. ■

A.4.2 Reproducibility

We now characterize accuracy of the nominal reproducibility error of Algorithm 1.

Theorem A.4.4. *Suppose that the conditional variance $\mathrm{Var}(a(\mathsf{r},D)\mid D)$ is strictly positive, almost surely, where r denotes a random collection drawn uniformly from $\mathcal{R}_{n,k,b}$. Suppose that the collections $\mathsf{R}_{\hat{g},k}$ and $\mathsf{R}'_{\hat{g}',k}$ are independently obtained using Algorithm 1. If Assumptions 1.4.1 and B.2.3 hold, the data D are independent and identically distributed, and the eighth-order split stability $\zeta^{(8)}$ is finite, then, for all sufficiently small ξ, the inequality*

$$\begin{aligned}P\left\{|a(\mathsf{R}_{\hat{g},k},D)-a(\mathsf{R}'_{\hat{g}',k},D)|\leq\xi\mid D\right\}-(1-\beta)\,|\\ \lesssim\left(\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^{4}\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2}(v_{1,k}\left(D\right))^{5/2}}\right)^{1/2}\\ \cdot\log^{3/4}\left(\frac{z_{1-\beta/2}}{\xi}\frac{\delta(v_{1,k}(D))^{2}}{(2-\varphi k-\varphi)^{3}(\Gamma^{(2)}_{k,\varphi,b})^{3/4}}\right),\end{aligned}\tag{A.4.8}$$

holds with probability greater than $1-\delta$ as D varies.

Remark A.4.5. Recall the definition of the oracle stopping time $g^\star$ given in (1.3.7). The bound (A.4.17) is obtained by decomposing the reproducibility error into two terms. The first term involves the error in a normal approximation to the difference $a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$. As $g^\star$ is deterministic conditional on D, a bound on this term follows from an argument very similar to the proof of Theorem A.4.3. The second term involves the differences $a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}_{g^\star,k}, D)$ and $a(\mathsf{R}'_{\hat{g}',k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$. The key step in bounding these quantities involves deriving a high probability bound for the difference $\hat{g} - g^\star$. This follows by combining a concentration inequality analogous to Theorem A.4.1 for the estimator $\hat{v}_{g,k}(D)$ with a maximal inequality due to Steiger (1970). ■

A.4.3 Specialization to Cross-Split, Sample-Stable Statistics

In this section, we specialize the results stated in Appendices A.4.1 and A.4.2 to cross-fit, sample-stable statistics. That is, we focus attention to the setting where $n = k \cdot b$ and impose Assumption 1.4.3. By doing this, we prove each of the results stated in Section 1.4.

We begin by stating the following corollary of Theorem A.4.1, which is referenced in Footnote 22.

Corollary A.4.1. *Suppose that Assumptions 1.4.1, 1.4.3, and B.2.3 hold, that the data D are independently and identically distributed, and that the statistic $a(\mathsf{R}_{1,k}, D)$ has a non-zero variance conditional on D almost surely. If the fourth-order split stability $\zeta^{(4)}$ is finite, then, for each $\varepsilon > 0$, the inequality*

$$P\left\{|a(\mathsf{R}_{g,k}, D) - \overline{a}\,(D)\,| \leq \sqrt{\frac{b-1}{n^2}\frac{1}{g}\frac{\log(\varepsilon^{-1})}{\delta}} \mid D\right\} \geq 1 - \varepsilon \tag{A.4.9}$$

holds with probability greater than $1 - \delta$ as D varies.

Proof. Corollary A.4.1 follows immediately from Theorem A.4.1 by restricting attention to the case that $n = k \cdot b$ and applying Assumption 1.4.3. In particular, observe that if $n = k \cdot b$, then $\Gamma_{k,\varphi,b} = \sigma^{(2,b-1)}_{\mathsf{train}}$ and

$$(2 - \varphi k - \varphi)^2 = (1 - \varphi)^2 = \left(\frac{n-b}{b}\right)^2 . \tag{A.4.10}$$

Applying Assumption 1.4.3 then gives

$$\Gamma_{k,\varphi,b} = \sigma^{(2,b-1)}_{\mathsf{train}} \lesssim \left(\frac{\sqrt{b-1}}{n-b}\right)^2 . \tag{A.4.11}$$

Plugging (A.4.10) and (A.4.11) into Theorem A.4.1 gives the desired bound. ■

Next, we show that Theorem A.4.2 implies the following generalization of Theorem 1.4.1 to higher-order moments. Theorem 1.4.1 is a strictly weaker result, as it only holds for the second moment.

Corollary A.4.2. *Suppose that Assumptions 1.4.1, 1.4.3, and B.2.3 hold, the data D are independently and identically distributed, and $n = k \cdot b$. If the 4rth-order split stability $\zeta^{(4)}$ is finite, then, for each $\delta > 0$, the inequality*

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r} \mid D\right] \lesssim \left(\frac{b-1}{n^2}\frac{1}{g}\right)^{r/2} \tag{A.4.12}$$

holds with probability greater than $1 - \delta$ as D varies.

Proof. If $n = k \cdot b$, then $\Gamma^{(r)}_{k,\varphi,b} = \sigma^{(2r,b-1)}_{\mathsf{valid}}$. Applying Assumption 1.4.3 then gives

$$\Gamma^{(r)}_{k,\varphi,b} \lesssim \left(\frac{\sqrt{b-1}}{n-b}\right)^{2r} . \tag{A.4.13}$$

Plugging (A.4.10) and (A.4.13) into (A.4.4) gives

$$\mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)\right)^{2r}\right] \lesssim \left(\frac{1}{nkg}\right)^{r} , \tag{A.4.14}$$

where we omit constants the depend only on r. ■

In turn, we specialize the non-asymptotic central limit theorem Theorem A.4.3 cross-fit, sample-stable statistics. This result suggests that, although the aggregate statistic $a(\mathsf{R}_{g,k}, D)$ is more concentrated at larger values of k, the quality of a normal approximation to its conditional distribution does not necessarily improve. This result is referenced in Footnote 27.

Corollary A.4.3. *Suppose that Assumptions 1.4.1, 1.4.3, and B.2.3 hold, the data D are independently and identically distributed, $n = k \cdot b$, and the conditional variance $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ is strictly positive almost surely, where r denotes a random collection drawn uniformly from $\mathcal{R}_{n,k,b}$. If the 8th-order split stability $\zeta^{(8)}$ is finite, then, for all g, the inequality*

$$\sup_{z \in \mathbb{R}} \left| P\left\{\frac{a\left(\mathsf{R}_{g,k}, D\right) - \overline{a}\left(D\right)}{\sqrt{v_{g,k}\left(D\right)}} \leq z \mid D\right\} - \Phi(z)\right| \lesssim \frac{1}{\delta}\left(\frac{1}{n}\frac{1}{k}\frac{b-1}{v_{1,k}(D)}\right)^{3/2}\sqrt{\frac{1}{g}} \tag{A.4.15}$$

is satisfied with probability greater than $1-\delta$ *as* D *varies, where* $\Phi(\cdot)$ *is the standard normal c.d.f.*

Proof. As before, Corollary A.4.3 follows by restricting attention to the case that $n = k \cdot b$ and imposing Assumption 1.4.3. In particular, Corollary A.4.3 is obtained by plugging the bounds (A.4.10) and (A.4.13) into (A.4.7). ∎

Remark A.4.6. In this case, the quantities g and k enter differently. In particular, observe that, by applying the variance bound (1.4.9) for the case $g = 1$, to lower bound the rate (A.4.15), we get

$$\left(\frac{1}{n}\frac{1}{k}\frac{1}{v_{1,k}(D)}\right)^{3/2}\sqrt{\frac{1}{g}} \gtrsim \sqrt{\frac{1}{g}}\,. \tag{A.4.16}$$

Written differently, Corollary A.4.3 suggests that the quality of a normal approximation to the conditional distribution of the aggregate statistic $a\left(\mathsf{R}_{g,k}, D\right)$ does not increase with k—it only depends on g. ∎

Finally, we show that Theorem 1.4.2 is a corollary of Theorem A.4.4 by restricting attention to the case that $n = k \cdot b$ and applying Assumption 1.4.3. We restate Theorem 1.4.2 below as a corollary for ease of reference.

Corollary A.4.4. *Suppose that the collections* $\mathsf{R}_{\hat{g},k}$ *and* $\mathsf{R}'_{\hat{g}',k}$ *are independently obtained using Algorithm 1. If Assumptions 1.4.1, 1.4.3, and B.2.3 hold, the conditional variance* $v_{1,k}(D) = \mathrm{Var}(a(\mathsf{r}, D) \mid D)$ *is strictly positive, almost surely, the data* D *are independent and identically distributed, and the eighth-order split stability* $\zeta^{(8)}$ *is finite, then for all sufficiently small* ξ*, the inequality*

$$\begin{aligned}\left|P\left\{\left|a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)\right| \geq \xi \mid D\right\} - \beta\right| \\ \lesssim \frac{1}{\delta^{3/4}}\frac{1}{kn}\left(\frac{1}{v_{1,k}(D)}\right)^{5/4}\left(\frac{\xi}{z_{1-\beta/2}}\right)^{1/2},\end{aligned} \tag{A.4.17}$$

holds with probability greater than $1-\delta$ *as* D *varies, where in writing (A.4.17), we have omitted a multiplicative term that converges to zero logarithmically as* ξ *decreases to zero.*

Proof. Observe that if $n = k \cdot b$, then the right-hand-side of the inequality (A.4.17) can be written

$$\left(\frac{\xi}{z_{1-\beta/2}}\frac{\sigma^{(2,b-1)}_{\mathsf{train}}(\sigma^{(4,b-1)}_{\mathsf{train}})^{1/2}}{\delta^{3/2}(v_{1,k}\left(D\right))^{5/2}}\left(\frac{n-b}{n}\right)^4\right)^{1/2}$$

$$\cdot \log^{3/4}\left(\frac{z_{1-\beta/2}}{\xi}\left(\frac{n}{n-b}\right)^3 \frac{\delta(v_{1,k}(D))^2}{(\sigma_{\mathsf{train}}^{(4,b-1)})^{3/4}}\right) . \tag{A.4.18}$$

By applying Assumption 1.4.3, the non-logarithmic factor reduces to

$$\left(\frac{\xi}{z_{1-\beta/2}}\frac{1}{\delta^{3/2}(v_{1,k}(D))^{5/2}}\left(\frac{\sqrt{b-1}}{n}\right)^4\right)^{1/2} \leq \left(\frac{1}{\delta^{3/2}}\frac{\xi}{z_{1-\beta/2}}\left(\frac{1}{kn}\right)^2\left(\frac{1}{v_{1,k}(D)}\right)^{5/2}\right)^{1/2} \tag{A.4.19}$$

as required. ■

A.4.4 Constructing a Stein Representer

Suppose that we are interested in studying the statistic $f(X)$, where X is random variable valued on the separable metric space $\mathcal{X}$. The method of exchangeable pairs has two ingredients. First, we need to construct a random variable X' such that (X, X') is an exchangeable pair. Second, we need to construct an antisymmetric function $F(X, X')$ such that

$$f(X) = \mathbb{E}[F(X, X') \mid X] \tag{A.4.20}$$

almost surely. We will refer to the function $F(X, X')$ as a "Stein representer" for $f(X)$.

In many applications, the Stein representer (A.4.20) induced by a suitable exchangeable pair (X, X') can be derived in closed form. See e.g Chen et al. (2011) and Ross (2011). A closed form derivation is more challenging in our setting, where the collection $\mathsf{R}_{g,k}$ takes the place of the random variable X. To address this issue, we apply a method due to Chatterjee (2005) for constructing Stein representers through a pair of coupled Markov chains induced by an appropriately chosen exchangeable pair.

Chatterjee's construction is founded on a observation, due to Stein (1986), that an exchangeable pair (X, X') induces a reversible Markov kernel K through

$$Kg(x) = \mathbb{E}[g(X') \mid X = x] ,$$

where g is any function satisfying $\mathbb{E}[|g(X)|] < \infty$. Suppose that $\{X_m\}_{m\geq 0}$ and $\{X'_m\}_{m\geq 0}$ are two Markov chains constructed with the kernel induced by (X, X') and coupled in such a way that the marginal distributions of X_m and X'_m depend only on the initial conditions X_0 and X'_0, respectively. Chatterjee makes the following observation. If there exists a constant

C such that

$$\sum_{m=0}^{\infty} \left|\mathbb{E}\left[f(X_m) - f(X'_m) \mid X_0 = x, X'_0 = y\right]\right| \leq C \,, \tag{A.4.21}$$

for each x and y in $\mathcal{X}$, then the function

$$F(x,y) = \sum_{m=0}^{\infty} \mathbb{E}\left[f(X_m) - f(X'_m) \mid X_0 = x, X'_0 = y\right]$$

is a Stein representer for $f\left(X\right)$. See Paulin et al. (2013, 2016) for applications of this idea to the derivation of matrix concentration inequalities. To apply this construction to our setting, we have two tasks. First, we need to specify a suitable exchangeable pair. Second, we need to construct a pair of coupled Markov chains induced by this pair that satisfy the finiteness condition (A.4.21). Throughout, for a vector $x = (x_i)_{i=1}^n$ we let $(x_{-\ell}, y)$ denote the vector formed by replacing the ℓth component of x with y.

Our construction is premised on the observation that the random collection of splits $\mathsf{R}_{g,k}$ can be generated by a random collection of permutations. To see this, let $\mathcal{P}_n$ denote the set of permutations of the set $[n]$, treating each $\pi \in \mathcal{P}_n$ as a bijection from $[n]$ to $[n]$. Observe that each permutation $\pi \in \mathcal{P}_n$ can be associated with an element of $\mathcal{R}_{n,k,b}$, denoted by $\mathsf{r}_k\left(\pi\right) = \left(\mathsf{s}_1\left(\pi\right), \ldots, \mathsf{s}_k\left(\pi\right)\right)$, through

$$\mathsf{s}_i\left(\pi\right) = \left\{\pi\left(k\cdot(i-1)+1\right), \ldots, \pi\left(k\cdot(i-1)+b\right)\right\}.$$

If $\boldsymbol{\pi} = (\pi_i)_{i=1}^g$denotes a collection of permutations drawn independently and uniformly at random from $\mathcal{P}_n$, then the collection $\mathsf{R}_{g,k}\left(\boldsymbol{\pi}\right) = \left(\mathsf{r}_k\left(\pi_i\right)\right)_{i=1}^g$ is equidistributed with the collection $\mathsf{R}_{g,k}$ defined in Section 1.2.

Now, we construct an exchangeable pair $(\boldsymbol{\pi}, \boldsymbol{\pi}')$, keeping in mind that our aim is to verify a condition of the form (A.4.21). For any permutation $\pi \in \mathcal{P}_n$ and indices $i, j \in [n]$, define the updated permutation

$$\hat{\pi}\left(i,j\right)\left(x\right) = \begin{cases} j, & x = i, \\ \pi\left(i\right), & x = \pi^{-1}\left(j\right), \\ \pi\left(x\right), & \text{otherwise.} \end{cases}$$

In other words, $\hat{\pi}\left(i,j\right)$ is identical to π, except that i maps to j and $\pi^{-1}\left(j\right)$ maps to $\pi\left(i\right)$. Let L be distributed uniformly on $[g]$ and let I and J be independently and uniformly distributed on $[n]$. Define the modified collection

$$\boldsymbol{\pi}' = (\pi_{-L}, \hat{\pi}_L(I, J)).$$

and observe that $(\boldsymbol{\pi}, \boldsymbol{\pi}')$ is an exchangeable pair.

With the exchangeable pair $(\boldsymbol{\pi}, \boldsymbol{\pi}')$ in place, we choose a coupled pair of Markov chains that it induces. For each $m \geq 1$, let L_m be distributed uniformly on $[g]$ and let I_m and J_m be distributed uniformly on $[n]$. Construct $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)$ from the pair $(\boldsymbol{\pi}_0, \boldsymbol{\pi}'_0) = (\boldsymbol{\pi}, \boldsymbol{\pi}')$ by setting

$$\boldsymbol{\pi}_m = (\pi_{m-1,-L_m}, \hat{\pi}_{m-1,L_m}(I_m, J_m)) \quad \text{and} \quad \boldsymbol{\pi}'_m = (\pi'_{m-1,-L_m}, \hat{\pi}'_{m-1,L_m}(I_m, J_m)) \tag{A.4.22}$$

recursively. In other words, the Markov chain $\boldsymbol{\pi}'_0$ is initialized by choosing one permutation in $\boldsymbol{\pi}_0$ and swapping one pair of indices. In the mth iteration of the Markov chain $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)_{m\geq 0}$, the permutations π_{m-1,L_m} and π'_{m-1,L_m} are selected and updated so that I_m now maps to J_m. As the iterations proceed, I_m will continue to map to the same index in the L_mth permutation in both collections. That is, in each iteration that a new permutation L and index I are selected, the two collections become more similar. Eventually, every L and I will have been selected, the two collections $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ will be the same, and so $a(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D) - a(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D) = 0$. This will imply the finiteness (A.4.21). This argument is formalized in the proof of the following Lemma.

Lemma A.4.1. *The function*

$$A(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D) = \sum_{m=0}^{\infty} \mathbb{E}\left[\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right)\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right],$$

is finite, antisymmetric, and satisfies the equality

$$\mathbb{E}\left[A(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D) \mid \boldsymbol{\pi}, D\right] = a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \tag{A.4.23}$$

almost surely.

A.4.5 Proofs

A.4.5.1 Proofs for Non-Sequential Results

To obtain the concentration inequality Theorem A.4.1 and Burkholder-Davis-Gundy inequality Theorem A.4.2, we apply the following result due to Chatterjee (2005, 2007), which we have augmented to be applicable under a set of assumptions considered in Paulin et al. (2016).

Lemma A.4.2. *Let $\mathcal{X}$ be a separable metric space and suppose that (X, X') is an exchangeable pair of $\mathcal{X}$-valued random variables. Let $f : \mathcal{X} \to \mathbb{R}$ be a square integrable function and let $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Stein representer for f. For each positive integer (r), define the quantities*

$$U_f^{(r)}(X) = \frac{1}{2}\mathbb{E}\left[\left(f(X) - f(X')\right)^{2r} \mid X\right] \quad \text{and} \quad U_F^{(r)}(X) = \frac{1}{2}\mathbb{E}\left[F(X, X')^{2r} \mid X\right]. \tag{A.4.24}$$

If there exist nonnegative constants u and s such that

$$U_f^{(1)}(X) \le s^{-1}u \quad \text{and} \quad U_F^{(1)}(X) \le su, \tag{A.4.25}$$

then the concentration inequality

$$P\{|f(X)| \ge \delta\} \le 2\exp\left(-t^2/2u\right) \tag{A.4.26}$$

holds for all $t \ge 0$. Moreover, the moment inequality

$$\mathbb{E}\left[f(X)^{2r}\right] \le (2r-1)^r \left(s\mathbb{E}\left[U_F^{(r)}(X)\right] + s^{-1}\mathbb{E}\left[U_f^{(r)}(X)\right]\right) \tag{A.4.27}$$

holds for all positive integers r for any positive constant s.

Throughout, to ease notation, we use the short hand

$$Z = (\mathsf{R}_{g,k}(\boldsymbol{\pi}), D) \quad \text{and} \quad Z' = (\mathsf{R}_{g,k}(\boldsymbol{\pi}'), D)\,.$$

The Markov chains $(Z_m)_{m\ge 0}$ and $(Z'_m)_{m\ge 0}$ are defined accordingly and we simplify $A(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D)$ to $A(Z, Z')$. To apply Lemma A.4.2, we are required to develop bounds for the objects

$$U_a^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[\left(a(Z) - a(Z')\right)^{2r} \mid Z\right] \quad \text{and} \quad U_A^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[A(Z, Z')^{2r} \mid Z\right].$$

Our approach is based on the following Lemma, which combines a generalization of an idea due to Lemma 10.4 of Paulin et al. (2016) with a Markov type bound.

Lemma A.4.3. *Let $r = 2^c$ for some positive integer c. Let $f : \mathcal{Z} \to \mathbb{R}$ be a function such that there exists a square integrable random variable W with*

$$\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f\left(Z_m\right) - f\left(Z_m'\right) \mid Z_0 = Z, Z_0' = Z'\right]\right)^r \leq W \tag{A.4.28}$$

almost surely. If the inequality

$$\mathbb{E}\left[\mathbb{E}\left[f(Z_m) - f(Z_m') \mid Z_0 = Z, Z_0' = Z'\right]^r \mid \boldsymbol{\pi}\right] \leq h_m^r \tag{A.4.29}$$

holds for each $m \geq 0$ and each collection $\boldsymbol{\pi}$, where $(h_m)_{m\geq 0}$ is a deterministic sequence of nonnegative numbers, then the inequalities

$$\frac{1}{2}\mathbb{E}\left[\left(f\left(Z\right) - f\left(Z'\right)\right)^r \mid Z\right] \leq \frac{h_0^r}{\delta} \quad \textit{and} \tag{A.4.30}$$

$$\frac{1}{2}\mathbb{E}\left[\left(\sum_{m=0}^{\infty} \mathbb{E}\left[f\left(Z_m\right) - f\left(Z_m'\right) \mid Z_0 = Z, Z_0' = Z'\right]\right)^r \mid Z\right] \leq \frac{1}{\delta}\left(\sum_{m=0}^{\infty} h_m\right)^r \tag{A.4.31}$$

both hold with probability greater than $1 - \delta$ as D varies

To apply this Lemma, we begin by noting that the inequality

$$\begin{aligned} &\left(\sum_{m=0}^{\infty} \mathbb{E}\left[a\left(Z_m\right) - a\left(Z_m'\right) \mid Z_0 = Z, Z_0' = Z'\right]\right)^2 \\ &\lesssim g^2 n^4 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T\left(\mathsf{s}, U, D\right) - T\left(\mathsf{s}', U, D\right)\right)^2 \end{aligned} \tag{A.4.32}$$

follows from Lemma A.5.1, stated in Appendix A.5.1. Moreover, the right hand side of (A.4.32) is square integrable, as the fourth order split-stability $\zeta^{(4)}$ is finite by assumption. Deterministic bounds of the form (A.4.29) are obtained through the following Lemma.

Lemma A.4.4. *Under Assumptions 1.4.1 and B.2.3, for all integers $m \geq 0$ and $r \geq 1$, the inequality*

$$\begin{aligned} &\mathbb{E}\left[\mathbb{E}\left[a(Z_m) - a(Z_m') \mid Z_0 = Z, Z_0' = Z'\right]^{2r} \mid \boldsymbol{\pi}\right] \\ &\leq 2^{4r}\left(1 - \frac{2}{gn^2}\right)^{2mr}\left(\frac{2n - bk - b}{gn^2}\right)^{2r} \Gamma_{k,\varphi,b}^{(r)} \end{aligned}$$

holds almost surely, where $\Gamma_{k,\varphi,b}^{(r)}$ is defined in the statement of Theorem A.4.2.

Combining Lemmas A.4.3 and A.4.4, we have that

$$U_A^{(r)}(Z) \leq \frac{1}{\delta}\left(\frac{gn^2}{2}\right)^r\left(\frac{2^4\left(2n-bk-b\right)^2}{gn^2}\right)^r\Gamma_{k,\varphi,b}^{(r)}$$

and

$$U_a^{(r)}(Z) \leq \frac{1}{\delta}\left(\frac{2}{gn^2}\right)^r\left(\frac{2^4\left(2n-bk-b\right)^2}{gn^2}\right)^r\Gamma_{k,\varphi,b}^{(r)}$$

with probability $1-\delta$. Corollary A.4.1 is obtained by setting $r=1$ and applying (A.4.26) of Lemma A.4.2 with $s=gn^2/2$ and

$$u=\frac{1}{\delta}\frac{2^4\left(2n-bk-b\right)^2}{gn^2}\Gamma_{n,k,b}^{(1)}.$$

Similarly, Theorem A.4.2 is obtained by applying (A.4.27) of Lemma A.4.2 with $s=(gn^2/2)^r$.

The normal approximation error bound Theorem A.4.3 follows from Theorem A.4.2. To see this, consider the centered statistic

$$a\left(Z\right)-\overline{a}\left(D\right)=\frac{1}{g}\sum_{\ell=1}^{g}\overline{a}\left(\mathsf{r}_\ell,D\right),\quad\text{where}\quad\overline{a}\left(\mathsf{r}_\ell,D\right)=\frac{1}{k}\sum_{i=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right) \tag{A.4.33}$$

for each ℓ in $[g]$. Conditional on the data D, the statistics $\overline{a}\left(\mathsf{r}_\ell,D\right)$ are independent, identically distributed, and mean zero. Moreover, we have that

$$\mathbb{E}\left[\left|\overline{a}\left(\mathsf{r}_\ell,D\right)\right|^3\right]\leq 3^2 2^6\left(2-\varphi k-\varphi\right)^3\left(\Gamma_{k,\varphi,b}^{(2)}\right)^{3/4} \tag{A.4.34}$$

by Hölder's inequality and Theorem A.4.2. Hence, we find that

$$\frac{\mathbb{E}\left[\left|\overline{a}\left(\mathsf{r}_\ell,D\right)\right|^3\mid D\right]}{g^{1/2}\left(v_{1,k}(D)\right)^{3/2}}\leq\frac{3^2 2^6}{\delta}\frac{\left(2-\varphi k-\varphi\right)^3}{\left(v_{1,k}(D)\right)^{3/2}}\frac{\left(\Gamma_{k,\varphi,b}^{(2)}\right)^{3/4}}{g^{1/2}}$$

holds with probability greater than $1-\delta$, by combining (A.4.34) with the Markov inequality. The proof then follows by the Berry-Esseen inequality. See e.g., Corollary 1 of Shevtsova (2011). ■

A.4.5.2 Proofs for Sequential Results

The first step for verifying Theorem A.4.4 is deriving a large deviation bound for the variance estimator $\hat{v}\left(\mathsf{R}_{g,k}, D\right)$ defined in (1.3.4). This is obtained in the following Lemma, which follows from an argument very similar to the proof of Theorem A.4.1.

Lemma A.4.5. *Suppose that Assumptions 1.4.1 and B.2.3 hold and that the data D are independent and identically distributed. If the eighth-order split-stability $\zeta^{(8)}$ is finite, then the conditional concentration inequality*

$$\log \frac{1}{4} P\left\{\left|\frac{\hat{v}(\mathsf{R}_{g,k}, D)}{v_{g,k}\left(D\right)} - 1\right| \geq t \mid D\right\} \lesssim -\frac{\delta\left(v_{1,k}(D)\right)^2}{(2-\varphi k-\varphi)^4}\frac{gt^2}{\Gamma^{(2)}_{k,\varphi,b}}$$

holds for all $t > 0$ with probability greater than $1-\delta$ as D varies.

We focus our analysis on the error

$$\left|P\left\{a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D) \leq \xi \mid D\right\} - (1-\beta/2)\right|. \tag{A.4.35}$$

An analogous argument will yield the same bound for the lower tail. We begin by bounding (A.4.35) with quantities that will be easier to handle in isolation. Define the objects

$$U\left(\mathsf{R}_{g,k}, D\right) = \frac{1}{g^\star}\left(\sum_{i=1}^{g} \overline{a}\left(\mathsf{r}_{i,k}, D\right) - \sum_{i=1}^{g^\star} \overline{a}\left(\mathsf{r}_{i,k}, D\right)\right) \quad \text{and} \tag{A.4.36}$$

$$Q\left(\mathsf{R}_{g,k}, D\right) = \left(1 - \frac{g}{g^\star}\right)\left(a(\mathsf{R}_{g,k}, D) - \overline{a}\left(D\right)\right)\ . \tag{A.4.37}$$

The following Lemma bounds the error (A.4.35) in terms of the error in the normal approximation to the quantity $a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)$ and generic high probability bounds on (A.4.36) and (A.4.37).

Lemma A.4.6. *Define the events*

$$\mathcal{U}_{k,\lambda}\left(D\right) = \left\{\left|U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)\right| \leq \lambda\sqrt{2v_{g^\star,k}\left(D\right)}\right\} \quad \textit{and}$$

$$\mathcal{Q}_{k,\lambda}(D) = \left\{\left|Q(\mathsf{R}_{g,k}, D) - Q(\mathsf{R}'_{\hat{g}',k}, D)\right| \leq \lambda\sqrt{2v_{g^\star,k}\left(D\right)}\right\}.$$

The quantity (A.4.35) is bounded above by

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{a(\mathsf{R}_{g^\star,k},D)-a(\mathsf{R}'_{g^\star,k},D)}{\sqrt{2v_{g^\star,k}\left(D\right)}}\leq z\mid D\right\}-\Phi(z)\right|$$
$$+\,2\lambda+\left(1-P\left\{\mathcal{U}_{k,\lambda}\left(D\right)\cap\mathcal{Q}_{k,\lambda}(D)\mid D\right\}\right)\;, \tag{A.4.38}$$

where $\Phi(\cdot)$ is the standard normal c.d.f.

Thus, it remains to give suitable bounds for the objects in (A.4.38). We state these bounds in terms of the quantities

$$\rho_{k,\varphi,b}(\xi,\beta\mid D)=\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^3(\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta\left(v_{1,k}\left(D\right)\right)^2}\quad\text{and} \tag{A.4.39}$$

$$\lambda_{k,\varphi,b}(\xi,\beta\mid D)=\left(\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^4\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2}(v_{1,k}\left(D\right))^{5/2}}\right)^{1/2}\;, \tag{A.4.40}$$

respectively. Each part of following Lemma follows from an application of Lemma A.4.6.

Lemma A.4.7. *Suppose that Assumptions 1.4.1 and B.2.3 hold, that the data D are independent and identically distributed, that the conditional variance $v_{1,k}(D)=\operatorname{Var}(a(\mathsf{r},D)\mid D)$ is strictly positive almost surely, and that the eighth-order split stability $\zeta^{(8)}$ is finite.*

(i) *The inequality*

$$\left|P\left\{\frac{a(\mathsf{R}_{g^\star,k},D)-a(\mathsf{R}'_{g^\star,k},D)}{\sqrt{2v_{g^\star,k}\left(D\right)}}\leq z\mid D\right\}-\Phi(z)\right|\lesssim\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)$$

is satisfied with probability greater than $1-\delta$ as D varies.

(ii) *The conditional concentration inequality*

$$P\left\{\frac{|U(\mathsf{R}_{\hat{g},k},D)-U(\mathsf{R}'_{\hat{g}',k},D)|}{\sqrt{2v_{g^\star,k}\left(D\right)}}\geq\lambda_{k,\varphi,b}\left(\xi,\beta\mid D\right)\log^{3/4}\left(\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)^{-1}\right)\mid D\right\}$$
$$\lesssim\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)$$

holds with probability greater than $1-\delta$ as D varies.

(iii) *For all sufficiently small ξ, the conditional concentration inequality*

$$P\left\{\frac{|Q(\mathsf{R}_{\hat{g},k},D)-Q(\mathsf{R}'_{\hat{g}',k},D)|}{\sqrt{2v_{g^\star,k}\left(D\right)}}\geq\lambda_{k,\varphi,b}\left(\xi,\beta\mid D\right)\right\}\lesssim\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)$$

holds with probability greater than $1-\delta$ *as* D *varies.*

Now, observe that $\rho_{k,\varphi,b}(\xi,\beta \mid D)$ is smaller than $\lambda_{k,\varphi,b}(\xi,\beta \mid D)$ for all sufficiently small ξ, almost surely. Consequently, by combining Lemma A.4.6 and Lemma A.4.7, we find that

$$\begin{aligned}
&\left|P\left\{|a(\mathsf{R}_{\hat{g},k},D)-a(\mathsf{R}'_{\hat{g}',k},D)|\leq\xi\mid D\right\}-(1-\beta)\right| \\
&\lesssim\left(\frac{\xi}{z_{1-\beta/2}}\frac{(2-\varphi k-\varphi)^4\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2}(v_{1,k}(D))^{5/2}}\right)^{1/2} \\
&\qquad\cdot\log^{3/4}\left(\frac{z_{1-\beta/2}}{\xi}\frac{\delta(v_{1,k}(D))^2}{(2-\varphi k-\varphi)^3(\Gamma^{(2)}_{k,\varphi,b})^{3/4}}\right)
\end{aligned}\tag{A.4.41}$$

with probability greater than $1-\delta$, as required. ■

A.4.6 Comparison with Zhang (2022)

We state a Berry-Esseen bound for $a(\mathsf{R}_{g,k},D)$ through an application of a result due to Zhang (2022). In contrast to the bound stated in Theorem A.4.3, the bound obtained here does not shrink as g increases. On the other hand, the bound obtained below is unconditional. It is straightforward to modify our argument to give an analogous high-probability conditional bound.

Theorem A.4.5. *Let* W *denote a standard normal random variable. Suppose that Assumptions 1.4.1 and B.2.3 hold and that the data* D *are independent and identically distributed. If the eighth-order split stability* $\zeta^{(8)}$ *is finite, then the Berry-Esseen inequality*

$$d_K\left(\frac{a(\mathsf{R}_{g,k},D)-\overline{a}(D)}{\sqrt{\mathbb{E}[v_{g,k}(D)]}},W\right)\leq\frac{4(2-\varphi k-\varphi)^2(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\mathbb{E}[v_{1,k}(D)]}$$

is satisfied.

A.4.6.1 Proof of Theorem A.4.5

We apply the following central limit theorem, due to Zhang (2022). This result generalizes Theorem 2.1 of Shao and Zhang (2019) to accommodate general Stein representers.

Theorem A.4.6 (Theorem 4.1, Zhang, 2022)**.** *Let* $\mathcal{X}$ *be a separable metric space and suppose that* (X,X') *is an exchangeable pair of* $\mathcal{X}$*-valued random variables. Suppose* *that* $f : \mathcal{X} \to \mathbb{R}$ *and* $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *are square-integrable functions such that* F *is antisymmetric and*

$$\mathbb{E}\left[F\left(X, X'\right) \mid X\right] = f\left(X\right)$$

almost surely. Assume that $\mathrm{Var}\left(f\left(X\right)\right)$ *is finite and non-zero and that* $\mathbb{E}\left[f\left(X\right)\right] = 0$*. Define the objects*

$$\overline{f}\left(X\right) = f\left(X\right)/\sqrt{\mathrm{Var}\left(f\left(X\right)\right)}. \tag{A.4.42}$$

$$G\left(X\right) = \frac{1}{2}\mathbb{E}\left[\left(f\left(X\right) - f\left(X'\right)\right)F\left(X, X'\right) \mid X\right] \quad \textit{and} \tag{A.4.43}$$

$$\overline{G}\left(X\right) = \frac{1}{2}\mathbb{E}\left[\left(f\left(X\right) - f\left(X'\right)\right)\left|F\left(X, X'\right)\right| \mid X\right]. \tag{A.4.44}$$

Let W *denote a standard normal random variable. The bound*

$$d_K\left(\overline{f}\left(X\right), W\right) \leq \frac{\mathbb{E}\left[\left|G\left(X\right) - \mathbb{E}\left[G\left(X\right)\right]\right|\right] + \mathbb{E}\left[\left|\overline{G}\left(X\right)\right|\right]}{\mathrm{Var}\left(f\left(X\right)\right)} \tag{A.4.45}$$

is satisfied.

To this end, define the objects

$$B\left(Z\right) = \frac{1}{2}\mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)A\left(Z, Z'\right) \mid Z\right] \quad \text{and}$$
$$\overline{B}\left(Z\right) = \frac{1}{2}\mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)\left|A\left(Z, Z'\right)\right| \mid Z\right].$$

It will suffice to bound the quantity

$$\frac{\mathbb{E}\left[\left|B\left(Z\right) - \mathbb{E}\left[B\left(Z\right)\right]\right|\right] + \mathbb{E}\left[\left|\overline{B}\left(Z\right)\right|\right]}{\mathrm{Var}\left(a\left(Z\right) - \overline{a}\left(Z\right)\right)}. \tag{A.4.46}$$

Observe that

$$\mathbb{E}\left[\left|B\left(X\right) - \mathbb{E}\left[B\left(X\right)\right]\right|\right] \leq \sqrt{\mathrm{Var}\left(B\left(Z\right)\right)} \quad \text{and}$$
$$\mathbb{E}\left[\left|\overline{B}\left(Z\right)\right|\right] \leq \sqrt{\mathrm{Var}\left(\overline{B}\left(Z\right)\right)}$$

by the Cauchy-Schwarz inequality and the fact that $\mathbb{E}\left[\overline{B}\left(X\right)\right] = 0$ by exchangeability. Consequently, as

$$\mathrm{Var}\left(B\left(Z\right)\right) \leq \mathbb{E}\left[B\left(Z\right)^2\right] = \mathbb{E}\left[\left(a\left(Z\right) - a\left(Z'\right)\right)^2 A\left(Z, Z'\right)^2\right] = \mathrm{Var}\left(\overline{B}\left(Z\right)\right)$$

it will suffice to bound

$$2\mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2}A\left(Z,Z'\right)^{2}\right]. \tag{A.4.47}$$

By Young's inequality, we have that

$$\mathbb{E}\left[\left(a\left(Z\right)-a\left(Z'\right)\right)^{2}A\left(Z,Z'\right)^{2}\right]\leq\frac{1}{2}\left(s^{-1}\mathbb{E}\left[U_{a}^{(2)}(Z)\right]+s\mathbb{E}\left[U_{A}^{(2)}(Z)\right]\right),$$

where $U_a^{(2)}(Z)$ and $U_A^{(2)}(Z)$ are defined in **??** A.4.5.1.

Observe that the bound

$$\left(\sum_{m=0}^{\infty}\mathbb{E}\left[a\left(Z_{m}\right)-a\left(Z_{m}'\right)\mid Z_{0}=Z,Z_{0}'=Z'\right]\right)^{4}$$
$$\leq\left(2gn^{2}\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T\left(\mathsf{s},D\right)-T\left(\mathsf{s}',D\right)\right)\right)^{4} \tag{A.4.48}$$

holds by Lemma A.5.1 and that the right-hand side of (A.4.48) is square-integrable as the eighth-order split stability $\zeta^{(8)}$ is finite. Thus, by combining Lemma A.5.3 and Lemma A.4.4, we have that

$$\overline{U}_{A}^{(2)}\leq\left(\frac{gn^{2}}{2}\right)^{2}\left(\frac{4\left(2n-bk-b\right)^{2}}{gn^{2}}\right)^{2}\Gamma_{k,\varphi,b}^{(2)}$$

and

$$\overline{U}_{a}^{(2)}\leq\left(\frac{2}{gn^{2}}\right)^{2}\left(\frac{4\left(2n-bk-b\right)^{2}}{gn^{2}}\right)^{2}\Gamma_{k,\varphi,b}^{(2)}.$$

Hence, by taking $s=\left(gn^{2}/2\right)^{2}$, we find that (A.4.47) is bounded above by

$$\frac{4}{gn^{2}}\left(\left(2n-bk-b\right)^{2}(\Gamma_{k,\varphi,b}^{(2)})^{1/2}\right)$$

Now, we have that

$$\mathrm{Var}\left(a\left(Z\right)-\overline{a}\left(D\right)\right)=\mathbb{E}\left[v_{g,k}\left(D\right)\right]$$

by the law of total variance and the fact that $\mathbb{E}\left[a\left(Z\right)-\overline{a}\left(D\right)\mid D\right]=0$. Consequently we can decompose

$$\mathbb{E}\left[v_{g,k}\left(D\right)\right]=\frac{\mathbb{E}[\phi_{n,b}(D)]+(k-1)\,\mathbb{E}[\gamma_{n,b}(D)]}{kg}. \tag{A.4.49}$$

Hence, we have that (A.4.46) is bounded above by

$$\frac{4\left(2-\varphi k-\varphi\right)^{2}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\mathbb{E}\left[v_{1,k}(D)\right]},$$

as required. ■

A.5. Proofs for Lemmas Stated in Appendix A.4

A.5.1 Proof of Lemma A.4.1

We use the following Lemma in several places.

Lemma A.5.1. *Let ψ and ψ' be two sets, each containing g elements of $\mathcal{P}_n$. The inequality*

$$\begin{aligned}&\sum_{m=0}^{\infty}\left|\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m)\right)-a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m)\right)\mid\boldsymbol{\pi}_0=\boldsymbol{\psi},\boldsymbol{\pi}'_0=\boldsymbol{\psi}',D\right]\right|\\&\leq 2gn^2\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T\left(\mathsf{s},D\right)-T\left(\mathsf{s}',D\right)\right)\end{aligned}$$

holds almost surely.

Observe that the quantity

$$\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T\left(\mathsf{s},D\right)-T\left(\mathsf{s}',D\right)\right)$$

is finite almost surely. Thus, the convergence of the series defining $A\left(\boldsymbol{\pi},\boldsymbol{\pi}'\mid D\right)$ follows from Lemma A.5.1. Define the operator

$$\begin{aligned}K&:\mathcal{F}\to\mathcal{F}\\f\left(\cdot\right)&\mapsto\mathbb{E}\left[f\left(\boldsymbol{\pi}'\right)\mid\boldsymbol{\pi}=\cdot\right],\end{aligned}$$

where $\mathcal{F}$ is the set of all measurable functions supported on the domain of $\boldsymbol{\pi}$. Observe that

$$\begin{aligned}&\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m),D\right)-a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m),D\right)\mid\boldsymbol{\pi}_0=\boldsymbol{\pi},\boldsymbol{\pi}'_0=\boldsymbol{\pi}',D\right]\\&=\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m),D\right)-\overline{a}(D)\mid\boldsymbol{\pi}_0=\boldsymbol{\pi},D\right]-\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m),D\right)-\overline{a}(D)\mid\boldsymbol{\pi}'_0=\boldsymbol{\pi}',D\right]\\&=K^m\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi})\right)-\overline{a}(D)\right)-K^{m+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}')\right)-\overline{a}(D)\right).\end{aligned}$$

Thus, for any m', we have that

$$\begin{aligned}
&\sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] \\
&= a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) - K^{m'+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right)\right). \qquad \text{(A.5.1)}
\end{aligned}$$

By Lemma A.5.1, the partial sums (A.5.1) converge almost everywhere and so the sequence

$$\left(K^{m+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right)\right)\right)_{m\geq 0} \qquad \text{(A.5.2)}$$

also converges almost everywhere. Lemma A.5.1 also implies that the limit of (A.5.2) depends only on D, as

$$K^m\left(a\left(\left(\mathsf{R}_{g,k}\left(\boldsymbol{\psi}\right), D\right)\right) - \overline{a}\left(D\right)\right) - K^m a\left(\left(\mathsf{R}_{g,k}\left(\boldsymbol{\psi}'\right), D\right) - \overline{a}\left(D\right)\right) \to 0$$

for any $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$ each containing g elements of $\mathcal{P}_n$. Therefore, we have that

$$\begin{aligned}
&\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid D\right] \\
&= \mathbb{E}\left[\lim_{n\to\infty}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) - K^{m+1} a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right)\right) \mid D\right] \\
&= \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \mid D\right] - b\left(D\right),
\end{aligned}$$

for some quantity

$$b\left(D\right) = \lim_{m\to\infty} K^m\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right)\right)$$

that depends only on D. Observe that

$$\begin{aligned}
&\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid D\right] \\
&= \mathbb{E}\left[\lim_{m'\to\infty}\sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] \mid D\right] \\
&= \lim_{m'\to\infty}\sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid D\right] \qquad \text{(Dominated Conv.)} \\
&= 0, \qquad \text{(Exchangeability)}
\end{aligned}$$

where the applicability of the Dominated Convergence Theorem follows from Lemma A.5.1.

Thus, as

$$\mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right) \mid D\right] = 0,$$

we can conclude that $b\left(D\right) = 0$ almost surely. Hence, we find that

$$\begin{aligned}
&\mathbb{E}\left[A\left(\boldsymbol{\pi}, \boldsymbol{\pi}' \mid D\right) \mid \boldsymbol{\pi}, D\right] \\
&= \mathbb{E}\left[\lim_{m'\to\infty} \sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] \mid \boldsymbol{\pi}, D\right] \\
&= \lim_{m'\to\infty} \sum_{m=0}^{m'} \mathbb{E}\left[a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D\right) - a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D\right) \mid \boldsymbol{\pi}_0 = \boldsymbol{\pi}, \boldsymbol{\pi}'_0 = \boldsymbol{\pi}', D\right] \\
&\hspace{30em} \text{(Dominated Conv.)} \\
&- \lim_{m'\to\infty} K^{m'+1}\left(a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right)\right) \\
&= a\left(\mathsf{R}_{g,k}(\boldsymbol{\pi}), D\right) - \overline{a}\left(D\right),
\end{aligned}$$

completing the proof. ■

A.5.2 Proof of Lemma A.4.2

First, observe that

$$U_F\left(X\right) \geq \frac{1}{2}\left(\mathbb{E}\left[F\left(X, X'\right) \mid X\right]\right)^2 = \frac{1}{2} f\left(X\right)^2,$$

where the inequality follows from Jensen's inequality and the definition of the Stein representer F. By (A.4.30), we have that

$$su \geq \frac{1}{2} f\left(X\right)^2.$$

Hence, the random variable $f\left(X\right)$ is bounded almost surely.

Now, suppose $h : \mathcal{X} \to \mathbb{R}$ is any measurable function such that $\mathbb{E}\left[h\left(X\right) F\left(X, X'\right)\right] < \infty$. Then, $\mathbb{E}\left[h\left(X\right) f\left(X\right)\right] = \mathbb{E}\left[h\left(X\right) F\left(X, X'\right)\right]$. Using the exchangeability of X and X' and the fact that F is antisymmetric, we have that

$$\mathbb{E}\left[h\left(X\right) F\left(X, X'\right)\right] = \mathbb{E}\left[h\left(X'\right) F\left(X', X\right)\right] = -\mathbb{E}\left[h\left(X'\right) F\left(X, X'\right)\right]$$

and that therefore

$$\mathbb{E}\left[h\left(X\right)f\left(X\right)\right]=\frac{1}{2}\mathbb{E}\left[\left(h\left(X\right)-h\left(X'\right)\right)F\left(X,X'\right)\right]. \tag{A.5.3}$$

Let

$$m\left(\theta\right)=\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\right]$$

denote the moment generating function of $f\left(X\right)$. As $f\left(X\right)$ is bounded almost surely, we can exchange differentiation and expectation in the differentiation of $m\left(\theta\right)$. Thus, we obtain

$$\begin{aligned} m'\left(\theta\right)&=\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)f\left(X\right)\right].\\ &=\frac{1}{2}\mathbb{E}\left[\left(\exp\left(\theta f\left(X\right)\right)-\exp\left(\theta f\left(X'\right)\right)\right)F\left(X,X'\right)\right], \end{aligned} \tag{A.5.4}$$

where the second inequality follows from (A.5.3). To bound $m'\left(\theta\right)$ we apply the following exponential mean-value inequality, stated in a more general form in Paulin et al. (2016).

Lemma A.5.2. *For all constants x, y, and c in $\mathbb{R}$ and $s>0$, it holds that*

$$\left|\left(e^{x}-e^{y}\right)c\right|\leq\frac{1}{4}\left(s\left(x-y\right)^{2}+s^{-1}c^{2}\right)\left(e^{x}+e^{y}\right).$$

In particular, by (A.5.4) and Lemma A.5.2, we obtain the bound

$$\begin{aligned} \left|m'\left(\theta\right)\right|&\leq\frac{1}{2}\mathbb{E}\left[\left|\left(\exp\left(\theta f\left(X\right)\right)-\exp\left(\theta f\left(X'\right)\right)\right)F\left(X,X'\right)\right|\right]\\ &\leq\frac{1}{8}\inf_{t>0}\mathbb{E}\left[\left(t\left(\theta f\left(X\right)-\theta f\left(X'\right)\right)^{2}+t^{-1}F\left(X,X'\right)^{2}\right)\left(\exp\left(\theta f\left(X\right)\right)+\exp\left(\theta f\left(X'\right)\right)\right)\right]\\ &=\frac{\left|\theta\right|}{4}\inf_{t>0}\mathbb{E}\left[\left(t\left(f\left(X\right)-f\left(X'\right)\right)^{2}+t^{-1}F\left(X,X'\right)^{2}\right)\exp\left(\theta f\left(X\right)\right)\right]\\ &=\frac{\left|\theta\right|}{2}\inf_{t>0}\mathbb{E}\Bigg[\left(\frac{t}{2}\mathbb{E}\left[\left(f\left(X\right)-f\left(X'\right)\right)^{2}\mid X\right]+\frac{1}{2t}\mathbb{E}\left[F\left(X,X'\right)^{2}\mid X\right]\right)\\ &\qquad\qquad\cdot\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\mid X\right]\Bigg]\\ &=\frac{\left|\theta\right|}{2}\inf_{t>0}\mathbb{E}\left[\left(tU_{f}\left(X\right)+t^{-1}U_{F}\left(X\right)\right)\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\mid X\right]\right]\\ &\leq\frac{\left|\theta\right|}{2}\mathbb{E}\left[\left(sU_{f}\left(X\right)+s^{-1}U_{F}\left(X\right)\right)\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\mid X\right]\right]\\ &\leq\left|\theta\right|v\mathbb{E}\left[\exp\left(\theta f\left(X\right)\right)\right] \end{aligned}$$

for all $\theta \in \mathbb{R}$. Thus, we have that

$$m'(\theta) \leq u\theta m(\theta)$$

for all $\theta > 0$. As $m(\cdot)$ is a convex function and $m'(0) = 0$, $m'(\theta)$ always has the same sign as θ, we find that

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log m(\theta) \leq u\theta.$$

As a consequence, and by $m(0) = 1$, we have that

$$\log m(\theta) \leq \int_0^\theta ut\mathrm{d}t \leq \frac{u\theta^2}{2}.$$

By the Chernoff bound (see e.g., Equation 2.5, Wainwright, 2019), we have

$$\log P\{f(X) \geq \delta\} \leq \inf_{\theta \geq 0} (\log m(\theta) - \theta\delta) \leq \inf_{\theta \geq 0} \left(\frac{u\theta^2}{2} - \theta\delta\right)$$

Solving this minimization with $\theta = \delta/u$, we find

$$P\{f(X) \geq t\} \leq \exp\left(\frac{-\delta^2}{2u}\right),$$

as required. The analogous lower tail bound follows from an identical argument, which completes the proof of (A.4.26). To prove (A.4.27) we apply the following result stated in Chatterjee (2007).

Theorem A.5.1 (Theorem 1.5, (iii), Chatterjee, 2007)**.** *Reintroduce the notation and assumptions from the statement of Theorem A.4.2. Define*

$$\Delta(X) = \frac{1}{2}\mathbb{E}\left[|F(X, X')(f(X) - f(X'))| \mid X\right].$$

The Burkholder-Davis-Gundy inequality

$$\mathbb{E}\left[f(X)^{2r}\right] \leq (2r-1)^r \mathbb{E}\left[\Delta(X)^r\right]$$

holds for any positive integer r.

The inequality

$$\mathbb{E}\left[\Delta(X)^r\right] = \mathbb{E}\left[\mathbb{E}\left[|(f(X) - f(X'))F(X, X')| \mid X\right]^r\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\mathbb{E}\left[\left|\left(f(X)-f(X')\right)F(X,X')\right|^r \mid X\right]\right] && \text{(Jensen)}\\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\left(s^{-1}\left(f(X)-f(X')\right)^{2r}\right)\left(sF(X,X')^{2r}\right)\right)^{1/2} \mid X\right]\right]\\
&\leq \mathbb{E}\left[\mathbb{E}\left[s^{-1}\left(f(X)-f(X')\right)^{2r}+sF(X,X')^{2r} \mid X\right]\right] && \text{(Young)}\\
&= s^{-1}\mathbb{E}\left[U_f^{(r)}(X)\right]+s\mathbb{E}\left[U_F^{(r)}(X)\right]
\end{aligned}
$$

then completes the proof. ■

A.5.3 Proof of Lemma A.4.3

We apply the following Lemma.

Lemma A.5.3. *Let $(X_m)_{m\geq 0}$ be a sequence of real-valued random variables. Suppose that the inequality*

$$\mathbb{E}\left[X_m^{2^c}\right] \leq h_m^{2^c} \tag{A.5.5}$$

holds for each $m \geq 0$ and positive integer c, where $(h_m)_{m\geq 0}$ is a deterministic sequence of nonnegative real numbers. If there exists a square integrable random variable W such that

$$\left(\sum_{m=0}^{\infty} X_m\right)^{2^c} \leq W$$

almost surely, then the inequality

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m\right)^{2^c}\right] \leq \left(\sum_{m=0}^{\infty} h_m\right)^{2^c}$$

holds almost surely.

Define the objects

$$U_f^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[\left(f(Z)-f(Z')\right)^{2r} \mid Z\right] \quad \text{and}$$

$$U_F^{(r)}(Z) = \frac{1}{2}\mathbb{E}\left[\left(\sum_{m=0}^{\infty}\mathbb{E}\left[f(Z_m)-f(Z'_m) \mid Z_0=Z, Z'_0=Z'\right]\right)^{2r} \mid Z\right].$$

By the (A.4.28), we have that

$$\mathbb{E}\left[\sum_{m=0}^{\infty} f(Z_m) - f(Z'_m) \mid Z_0 = Z, Z'_0 = Z'\right]^{2^c} \leq W.$$

Thus, (A.4.29) guarantees that the conditions of Lemma A.5.3 are satisfied, and we have that

$$\mathbb{E}\left[U_F^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{1}{2}\left(\sum_{i=0}^{\infty} h_i\right)^{2^c} \quad \text{and} \quad \mathbb{E}\left[U_f^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{1}{2} h_0^{2^c}.$$

By Markov's inequality, we obtain

$$P\left\{U_F^{(2^{c-1})}(Z) \geq \frac{2}{\delta}\mathbb{E}\left[U_F^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \quad \text{or} \quad U_f^{(2^{c-1})}(Z) \geq \frac{2}{\delta}\mathbb{E}\left[U_f^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \mid \boldsymbol{\pi}\right\} \leq \delta.$$

Hence, by Lemma A.5.3 and DeMorgan's law, the probability that both

$$U_F^{(2^{c-1})}(Z) \leq \frac{2}{\delta}\mathbb{E}\left[U_F^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{1}{\delta}\left(\sum_{m=0}^{\infty} h_m\right)^{2^c} \quad \text{and}$$
$$U_f^{(2^{c-1})}(Z) \leq \frac{2}{\delta}\mathbb{E}\left[U_f^{(2^{c-1})}(Z) \mid \boldsymbol{\pi}\right] \leq \frac{h_0^{2^c}}{\delta}$$

hold is greater than $1 - \delta$. ■

A.5.4 Proof of Lemma A.4.4

Recall the definition of the collections $(\boldsymbol{\pi}_m, \boldsymbol{\pi}'_m)_{m \geq 0}$ given in (A.4.22). Fix $Z_0 = Z$ and $Z'_0 = Z'$ throughout. For any i in $[n]$, let $\mathsf{s}_{m,\ell}(i)$ denote the element of the collection

$$\mathsf{r}(\pi_{m,\ell}) = (\mathsf{s}_1(\pi_{m,\ell}), \ldots, \mathsf{s}_g(\pi_{m,\ell}))$$

that contains the index i and set $\mathsf{s}_{m,\ell}(i)$ equal to $\varnothing$ if no element of $\mathsf{r}(\pi_{m,\ell})$ contains i.

We begin by defining three events that will determine the structure of our argument. By construction, the collections $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ are either identical or differ in exactly two indices in their Lth element. Let $\mathcal{E}_m$ denote the event that $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ differ. On the event $\mathcal{E}_m$, let $i_{1,m}$ and $i_{2,m}$ denote the two indices in which the Lth elements of $\boldsymbol{\pi}_m$ and $\boldsymbol{\pi}'_m$ differ. Define the random variables B_m and C_m such that, conditional on $\mathcal{E}_m$, each is uniformly distributed on $\{i_{1,m}, i_{2,m}\}$ such that $B_m \neq C_m$. On the complement of $\mathcal{E}_m$, set these indices uniformly at random. Let $\mathcal{F}_m$ denote the event that

$$\mathsf{s}_{m,L}\left(B_m\right) \neq \mathsf{s}_{m,L}\left(C_m\right),$$

i.e., the event that the indices that differ are not in the same element of the collection $\mathsf{r}(\pi_{m,\ell})$. Finally, let $\mathcal{G}_m$ denote the event that

$$\mathsf{s}_{m,L}\left(B_m\right) \neq \varnothing \quad \text{and} \quad \mathsf{s}_{m,L}\left(C_m\right) \neq \varnothing,$$

i.e., the event that the indices that differ are both in the collection $\mathsf{r}(\pi_{m,\ell})$.

By Assumption 1.4.1, the event $\mathcal{H}_m = \mathcal{E}_m \cap \mathcal{F}_m$ is a necessary condition for $a\left(Z_m\right) - a\left(Z'_m\right) \neq 0$. Thus, we can compute

$$\begin{aligned}
P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\} &= P\left\{\mathcal{F}_m \mid \mathcal{E}_m, \mathcal{E}_0\right\} P\left\{\mathcal{E}_m \mid \mathcal{E}_0\right\} \\
&= \Big(P\left\{\mathsf{s}_{m,L}\left(B_m\right) \neq \mathsf{s}_{m,L}\left(C_m\right) \mid \mathsf{s}_{m,L}\left(B_m\right) \neq \varnothing, \mathcal{E}_m\right\} P\left\{\mathsf{s}_{m,L}\left(B_m\right) \neq \varnothing \mid \mathcal{E}_m\right\} \\
&\quad + P\left\{\mathsf{s}_{m,L}\left(B_m\right) \neq \mathsf{s}_{m,L}\left(C_m\right) \mid \mathsf{s}_{m,L}\left(B_m\right) = \varnothing, \mathcal{E}_m\right\} P\left\{\mathsf{s}_{m,L}\left(B_m\right) = \varnothing \mid \mathcal{E}_m\right\}\Big) \\
&\quad \cdot P\left\{\mathcal{E}_m \mid \mathcal{E}_0\right\} \\
&= \left(\frac{n-b}{n-1}\frac{kb}{n} + \frac{kb}{n-1}\frac{n-kb}{b}\right)\left(1 - \frac{2}{gn^2}\right)^m \\
&= \left(\frac{kb\left(2n-kb-b\right)}{n(n-1)}\right)\left(1 - \frac{2}{gn^2}\right)^m
\end{aligned}$$

for all $m \geq 0$ by the law of total probability. Moreover, we have that

$$P\left\{\mathcal{H}_m \mid Z, Z'\right\} \leq P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\}$$

almost surely. Consequently, we find that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[a\left(Z_m\right) - a\left(Z'_m\right) \mid Z, Z'\right]^{2r} \mid \boldsymbol{\pi}\right] \\
&= \mathbb{E}\left[\left(P\left\{\mathcal{H}_m \mid Z, Z'\right\} \mathbb{E}\left[a\left(Z_m\right) - a\left(Z'_m\right) \mid Z, Z', \mathcal{H}_m\right]\right)^{2r} \mid \boldsymbol{\pi}\right] \\
&\leq \left(P\left\{\mathcal{H}_m \mid \mathcal{E}_0\right\}\right)^{2r} \mathbb{E}\left[\mathbb{E}\left[\left(a\left(Z_m\right) - a\left(Z'_m\right)\right)^{2r} \mid Z, Z', \mathcal{H}_m\right] \mid \boldsymbol{\pi}\right] && \text{(Jensen)} \\
&= \left(1 - \frac{2}{gn^2}\right)^{2mr}\left(\frac{kb\left(2n-kb-b\right)}{n(n-1)}\right)^{2r} \mathbb{E}\left[\mathbb{E}\left[\left(a\left(Z_m\right) - a\left(Z'_m\right)\right)^{2r} \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \mid \boldsymbol{\pi}\right] \\
&\leq \left(1 - \frac{2}{gn^2}\right)^{2mr}\left(\frac{2kb\left(2n-kb-b\right)}{n^2}\right)^{2r} \mathbb{E}\left[\left(a\left(Z_m\right) - a\left(Z'_m\right)\right)^{2r} \mid \mathcal{H}_m\right], && \text{(A.5.6)}
\end{aligned}$$

where the final inequality follows from the fact that $\boldsymbol{\pi}$ is uniformly distributed independently of D and the elementary inequality $n/(n-1) \leq 2$. Thus, it remains to bound the expectation in (A.5.6).

To ease notation, we now drop the dependence on m and L. Observe that

$$P\{\mathcal{G} \mid \mathcal{H}\} = P\{\mathsf{s}(B) \neq \varnothing \mid \mathsf{s}(C) \neq \varnothing\} = \frac{kb-b}{n-b}$$

and that therefore

$$\begin{aligned}\mathbb{E}\left[(a(Z)-a(Z'))^{2r} \mid \mathcal{H}, \boldsymbol{\pi}\right] = &\left(\frac{kb-b}{n-b}\right)\mathbb{E}\left[(a(Z)-a(Z'))^{2r} \mid \mathcal{G}\right] \\ &+ \left(\frac{n-kb}{n-b}\right)\mathbb{E}\left[(a(Z)-a(Z'))^{2r} \mid \mathcal{H}\setminus\mathcal{G}\right]. \end{aligned} \tag{A.5.7}$$

Define the sets

$$\hat{\mathsf{s}}_i = \mathsf{s}(i) \setminus i \quad \text{and} \quad \bar{\mathsf{s}} = \tilde{\mathsf{s}}(B) \cap \tilde{\mathsf{s}}(C).$$

With an abuse of notation, we let

$$\begin{aligned}\psi(D_i, \hat{\eta}(D_j, D_{\hat{\mathsf{s}}_k})) &= \psi(D_i, \hat{\eta}(D_j \cap D_{\hat{\mathsf{s}}_k} \cap D_{\bar{\mathsf{s}}})) \quad \text{and} \\ h(D_i, D_j, D_{\hat{\mathsf{s}}_k}) &= \psi(D_i, \hat{\eta}(D_j, D_{\hat{\mathsf{s}}_k})) - \psi(D_j, \hat{\eta}(D_i, D_{\hat{\mathsf{s}}_k})).\end{aligned}$$

Observe that

$$\begin{aligned}&\mathbb{E}\left[(a(Z)-a(Z'))^{2r} \mid \mathcal{H}\setminus\mathcal{G}\right] \\ &= \mathbb{E}\left[(a(Z)-a(Z'))^{2r} \mid \mathcal{H}\setminus\mathcal{G}\right] = \left(\frac{1}{gkb}\right)^{2r}\mathbb{E}\left[h(D_B, D_C, D_{\tilde{s}_C})^{2r}\right]\ . \end{aligned} \tag{A.5.8}$$

On the other hand, we can decompose

$$\begin{aligned}&\mathbb{E}\left[(a(Z)-a(Z'))^{2} \mid \mathcal{G}\right] \\ &= \left(\frac{1}{gkb}\right)^{2r}\mathbb{E}\left[(h(D_B, D_C, D_{\hat{\mathsf{s}}_C}) + h(D_C, D_B, D_{\hat{\mathsf{s}}_B}))^{2r}\right] \\ &= \left(\frac{1}{gkb}\right)^{2r}\mathbb{E}\left[(h(D_B, D_C, D_{\hat{\mathsf{s}}_C}) - h(D_B, D_B, D_{\hat{\mathsf{s}}_B}))^{2r}\right]\ . \end{aligned} \tag{A.5.9}$$

Consequently, by (A.5.7), (A.5.8), and (A.5.9), it suffices to express suitable bounds for the expectations

$$\begin{aligned}
&\mathbb{E}\left[(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat\eta(D_B,D_{\hat{\mathsf{s}}_C})))^{2r}\right] \quad \text{and}\\
&\mathbb{E}\Bigg[\Bigg((\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat\eta(D_B,D_{\hat{\mathsf{s}}_C}))) \\
&\qquad -(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_B}))-\psi(D_C,\hat\eta(D_B,D_{\hat{\mathsf{s}}_B})))\Bigg)^{2r}\Bigg]
\end{aligned} \tag{A.5.10}$$

respectively. To this end, recall that $\tilde D_i$ are independent copies of D_i for each i. Observe that

$$\begin{aligned}
&\mathbb{E}\left[(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat\eta(D_B,D_{\hat{\mathsf{s}}_C})))^{2r}\right]\\
&=\mathbb{E}\left[\left(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))+\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat\eta(D_B,D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right]\\
&=\sum_{q=0}^{2r}\binom{2r}{q}\mathbb{E}\Bigg[\left(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))\right)^{2r-q}\\
&\qquad\left(\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat\eta(D_B,D_{\hat{\mathsf{s}}_C}))\right)^{q}\Bigg] \qquad \text{(Binomial Theorem)}\\
&\le 2^{2r}\mathbb{E}\left[\left(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right]
\end{aligned} \tag{A.5.11}$$

where the inequality follows from the fact that B and C are exchangeable and the Hölder inequality. Similarly, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))\right)^{2r-q}\right]\\
&=\mathbb{E}\left[\left(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))+\psi(D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))-\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))\right)^{2r}\right]\\
&=\sum_{q=0}^{2r}\binom{2r}{q}\mathbb{E}\Bigg[\left(\psi(D_B,\hat\eta(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))\right)^{2r-q}\\
&\qquad\left(\psi(D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))-\psi(\tilde D_B,\hat\eta(\tilde D_C,D_{\hat{\mathsf{s}}_C}))\right)^{q}\Bigg]\\
&\qquad\qquad \text{(Binomial Theorem)}\\
&\le\sum_{q=0}^{2r}\binom{2r}{q}\left(\sigma^{(2r)}_{\mathsf{valid}}\right)^{\frac{2r-q}{2r}}\left(\sigma^{(2r,1)}_{\mathsf{train}}\right)^{\frac{q}{2r}} \qquad \text{(Hölder)}\\
&\le 2^{2r}\sigma^{(2r)}_{\mathsf{max}},
\end{aligned} \tag{A.5.12}$$

where the last inequality follows by the definitions of $\sigma_{\mathsf{valid}}^{(2r)}$ and $\sigma_{\mathsf{train}}^{(2r,1)}$. Thus, we have that

$$\mathbb{E}\left[(\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_C})))^{2r}\right]\leq 2^{4r}\sigma_{\mathsf{max}}^{(2r)} \tag{A.5.13}$$

by (A.5.11) and (A.5.12).

Next, we consider the double difference term (A.5.10). In this case, we have that

$$\begin{aligned}
&\mathbb{E}\Bigg[\bigg((\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_C})))\\
&\qquad\qquad -(\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_B}))-\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_B})))\bigg)^{2r}\Bigg]\\
&=\mathbb{E}\Bigg[\bigg((\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_B})))\\
&\qquad\qquad -(\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_B})))\bigg)^{2r}\Bigg]\\
&=\sum_{q=0}^{2r}\binom{2r}{q}\mathbb{E}\Bigg[\bigg((\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_B})))\\
&\qquad\qquad\qquad \cdot(\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_C}))-\psi(D_C,\hat{\eta}(D_B,D_{\hat{\mathsf{s}}_B})))\bigg)^{2r}\Bigg]\\
&\hspace{30em}\text{(Binomial Theorem)}\\
&\leq 2^{2r}\mathbb{E}\left[(\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_C}))-\psi(D_B,\hat{\eta}(D_C,D_{\hat{\mathsf{s}}_B})))^{2r}\right]=2^{2r}\sigma_{\mathsf{train}}^{(2r,b-1)}
\end{aligned} \tag{A.5.14}$$

where the final inequality follows from the fact that B and C are exchangeable and the Hölder inequality. Putting the pieces together, we have that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[a\left(Z_m\right)-a\left(Z_m'\right)\mid Z,Z'\right]^{2r}\mid\boldsymbol{\pi}\right]\\
&\leq 2^{4r}\left(1-\frac{2}{gn^2}\right)^{2mr}\left(\frac{2n-kb-b}{gn^2}\right)^{2r}\left(\left(\frac{n-kb}{n-b}\right)2^{2r}\sigma_{\mathsf{max}}^{(2r)}+\left(\frac{kb-b}{n-b}\right)\sigma_{\mathsf{train}}^{(2r,b-1)}\right)
\end{aligned}$$

as required. ■

A.5.5 Proof of Lemma A.4.5

Throughout, we us the short hand $\hat{v}_{g,k}\left(Z\right)$ to denote $\hat{v}(\mathsf{R}_{g,k},D)$. We begin by decomposing the estimator $\hat{v}_{g,k}\left(Z\right)$ into two parts that will each be easier to handle when considered in isolation. To this end, define the statistics

$$\tilde{v}_{g,k}(Z) = \frac{1}{g^2k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},Y)-\overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'},D)-\overline{a}(D)\right) \quad \text{and}$$
$$\check{v}_{g,k}(Z) = \left(a(Z)-\overline{a}(D)\right)^2.$$

Observe that both $\tilde{v}_{g,k}(Z)$ and $\check{v}_{g,k}(Z)$ are unbiased for $v(Z)$. Moreover, we can write

$$\begin{aligned}
\hat{v}_{g,k}(Z) &= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-a(Z)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a(Z)\right)\\
&= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a(Z)\right)\\
&\quad+\frac{1}{g(g-1)k}\sum_{\ell=1}^{g}\sum_{i=1}^{g}\left(\overline{a}(D)-a(Z)\right)\left(T(\mathsf{s}_{\ell,i},D)-a(Z)\right)\\
&= \frac{1}{g(g-1)k^2}\sum_{\ell=1}^{g}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a(Z)\right)\\
&= \frac{g}{g-1}\tilde{v}(Z)+\frac{1}{g(g-1)k}\sum_{\ell=1}^{g}\sum_{i=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}(D)\right)\left(\overline{a}(D)-a(Z)\right)\\
&= \frac{g}{g-1}\tilde{v}_{g,k}(Z)-\frac{1}{g-1}\check{v}_{g,k}(Z).
\end{aligned} \tag{A.5.15}$$

Hence, it will suffice to characterize the exponential rates of conditional concentration for the statistics $\tilde{v}_{g,k}(Z)$ and $\check{v}_{g,k}(Z)$. These are established through the application of the following Lemma, which follows from an argument very similar to the proof of Theorem A.4.1.

Lemma A.5.4. *Suppose that Assumptions 1.4.1 and B.2.3 hold and that the data D are independently and identically distributed. If the eighth-order split-stability $\zeta^{(8)}$ is finite, then:*

(i) *The conditional concentration inequality*

$$\log\frac{1}{2}P\left\{\left|\tilde{v}(Z)-v(D)\right|\geq t \mid D\right\} \lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3t^2}{\Gamma^{(2)}_{k,\varphi,b}}, \tag{A.5.16}$$

holds for all $t>0$ with probability greater than $1-\delta$.

(ii) *The conditional concentration inequality*

$$\log\frac{1}{2}P\left\{\left|\breve{v}\left(Z\right)-v\left(D\right)\right|\geq t\mid D\right\}\lesssim-\frac{\delta}{\left(2-\varphi k-\varphi\right)^{4}}\frac{g^{2}t^{2}}{\Gamma_{k,\varphi,b}^{(2)}}\,, \tag{A.5.17}$$

holds for all $t>0$ *with probability greater than* $1-\delta$.

Putting the pieces together, by (A.5.15), we have that

$$\begin{aligned}
&P\left\{\left|\hat{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right|\leq t\mid D\right\}\\
&=P\left\{\left|\frac{g}{g-1}\left(\tilde{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right)-\frac{1}{g-1}\left(\breve{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right)\right|\leq t\mid D\right\}\\
&\geq P\left\{\frac{g}{g-1}\left|\tilde{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right|+\frac{1}{g-1}\left|\breve{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right|\leq t\mid D\right\}\\
&\geq P\left\{\frac{g}{g-1}\left|\tilde{v}_{g,k}\left(Z\right)-v\left(D\right)\right|\leq\frac{\sqrt{g}}{1+\sqrt{g}}t,\frac{1}{g-1}\left|\breve{v}\left(Z\right)-v\left(D\right)\right|\leq\frac{1}{1+\sqrt{g}}t\mid D\right\}\\
&\geq P\left\{\left|\tilde{v}_{g,k}\left(Z\right)-v\left(D\right)\right|\leq\frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}}t\mid D\right\}+P\left\{\left|\breve{v}_{g,k}\left(Z\right)-v\left(D\right)\right|\leq\frac{g-1}{1+\sqrt{g}}t\mid D\right\}-1\\
&\geq 1-4\exp\left(-\frac{\delta}{C(2-\varphi k-\varphi)^{4}}\frac{g^{3}t^{2}}{\Gamma_{k,\varphi,b}^{(2)}}\right)
\end{aligned}$$

with probability greater than $1-\delta$ for some universal constant C, where the last inequality follows by Lemma A.5.4 and the facts that $4g\geq\left(1+\sqrt{g}\right)^{2}$ for all $g\geq 1$ and $(1/4)\,g^{2}\leq\left(g-1\right)^{2}$ for all $g\geq 2$. ■

A.5.6 Proof of Lemma A.4.6

Define the event

$$\mathcal{W}_{\lambda}\left(D\right)=\mathcal{U}_{k,\lambda}\left(D\right)\cap\mathcal{Q}_{k,\lambda}(D),$$

the quantity

$$W\left(D\right)=\left(U(\mathsf{R}_{\hat{g},k},D)-U(\mathsf{R}'_{\hat{g}',k},D\right)+\left(Q(\mathsf{R}_{\hat{g},k},D)-Q(\mathsf{R}'_{\hat{g}',k},D)\right).$$

By the decomposition (A.3.6), we have that

$$\left|P\left\{a(\mathsf{R}_{\hat{g},k},D)-a(\mathsf{R}'_{\hat{g}',k},D)\leq\xi\mid D\right\}-\left(1-\beta/2\right)\right|$$

$$
\begin{aligned}
&= \left| P\left\{ \frac{a(\mathsf{R}_{\hat{g},k}, D) - a(\mathsf{R}'_{\hat{g}',k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq \frac{\xi}{\sqrt{2v_{g^\star,k}(D)}} \mid D \right\} - (1-\beta/2) \right| \\
&= \left| P\left\{ \frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq \frac{\xi}{\sqrt{2v_{g^\star,k}(D)}} - \frac{W(D)}{\sqrt{2v_{g^\star,k}(D)}} \mid D \right\} - (1-\beta/2) \right| \\
&\leq \sup_{q\in[-2\lambda,2\lambda]} \left| P\left\{ \frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq z_{1-\beta/2} + q \right\} - (1-\beta/2) \right| \\
&+ (1 - P\{\mathcal{W}_\lambda(D) \mid D\}) \\
&\leq \sup_{z\in\mathbb{R}} \left| P\left\{ \frac{a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D)}{\sqrt{2v_{g^\star,k}(D)}} \leq z \mid D \right\} - \Phi(z) \right| + 2\lambda + (1 - P\{\mathcal{W}_\lambda(D) \mid D\}),
\end{aligned}
$$

as required. ■

A.5.7 Proof of Lemma A.4.7

A.5.7.1 Part (i)

Observe that we can write

$$
a(\mathsf{R}_{g^\star,k}, D) - a(\mathsf{R}'_{g^\star,k}, D) = \frac{1}{g^\star} \sum_{\ell=1}^{g} \left(\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D) \right),
$$

where $\overline{a}(\cdot, D)$ is defined in (B.2.4). We have that

$$
\begin{aligned}
\mathbb{E}\left[|\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D)|^3 \right] &\leq \left(\mathbb{E}\left[(\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D))^4 \right] \right)^{3/4} && \text{(Hölder)} \\
&\leq 2^6 \left(\mathbb{E}\left[(\overline{a}(\mathsf{r}_\ell, D))^4 \right] \right)^{3/4} \\
&\leq 2^6 3^2 (2 - \varphi k - \varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4},
\end{aligned}
$$

where the second inequality follows from Hölder's inequality, the binomial theorem, and the fact that r_ℓ and r'_ℓ are exchangeable. The final inequality follows from Theorem A.4.2. Consequently, we find that

$$
\begin{aligned}
\frac{\mathbb{E}\left[|\overline{a}(\mathsf{r}_\ell, D) - \overline{a}(\mathsf{r}'_\ell, D)|^3 \mid D \right]}{(g^\star)^{1/2} (2v_{1,k}(D))^{3/2}} &\lesssim \frac{(2 - \varphi k - \varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta (g^\star)^{1/2} (v_{1,k}(D))^{3/2}} && \text{(Markov)} \\
&\lesssim \frac{\xi}{z_{1-\beta/2}} \frac{(2 - \varphi k - \varphi)^3 (\Gamma^{(2)}_{k,\varphi,b})^{3/4}}{\delta (v_{1,k}(D))^2},
\end{aligned}
$$

holds with probability $1-\delta$. The Lemma then follows by the Berry-Esseen inequality. ■

A.5.7.2 Part (ii)

Our bound is based on the following two conditional concentration inequalities. Both arguments are based on a Chernoff-type maximal inequality, due to Steiger (1970).

Lemma A.5.5. *Suppose that Assumptions 1.4.1 and B.2.3 hold, that the data D are independent and identically distributed, and that the eighth-order split ζ^8 stability is finite.*

(i) *For all $t > 0$ and $c > 0$, the condition concentration inequality*

$$\log \frac{1}{2} P\left\{|U(\mathsf{R}_{\hat{g},k}, D)| \geq t\sqrt{v_{g^\star,k}(D)}, |\hat{g}/g^\star - 1| \leq c \mid D\right\} \tag{A.5.18}$$

$$\leq -\frac{\delta v_{1,k}(D)}{2^4(2-\varphi k-\varphi)^2 \Gamma^{(1)}_{k,\varphi,b}} \frac{t^2}{c} \tag{A.5.19}$$

holds with probability greater than $1-\delta$ as D varies.

(ii) *If $g^\star c \geq 2$ and $1/2 > c > 0$, then the conditional concentration inequality*

$$\log \frac{1}{8} P\{|\hat{g} - g^\star| > cg^\star \mid D\} \lesssim -\frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z^2_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \tag{A.5.20}$$

holds with probability greater than $1-\delta$ as D varies.

Observe that

$$\begin{aligned}
&P\left\{|U(\mathsf{R}_{\hat{g},k}, D) - U(\mathsf{R}'_{\hat{g}',k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)}\right\} \\
&\lesssim P\left\{|U(\mathsf{R}_{\hat{g},k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)}\right\} \\
&\leq P\left\{|U(\mathsf{R}_{\hat{g},k}, D)| \geq \lambda\sqrt{2v_{g^\star,k}(D)}, |\hat{g}/g^\star - 1| \leq c \mid D\right\} + P\{|\hat{g}/g^\star - 1| \geq c \mid D\} \\
&\lesssim \exp\left(-\frac{\delta v_{1,k(D)}}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{\lambda^2}{c}\right) + \exp\left(-C\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z^2_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2}\right)
\end{aligned} \tag{A.5.21}$$

for some universal constant C. Hence, it remains to choose c and λ such that the quantity (A.5.21) is less than $\rho_{k,\varphi,b}(\xi, \beta \mid D)$. First, we choose c such that

$$\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z^2_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}} \frac{c^2}{\xi^2} \lesssim \log\left(\rho_{k,\varphi,b}(\xi, \beta \mid D)^{-1}\right)$$

Choosing c by

$$c = \xi \left(\frac{(2-\varphi k-\varphi)^2\,(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{z_{1-\beta/2}\delta^{1/2}(v_{1,k}(D))^{3/2}} \right) \log^{1/2}\left(\rho_{k,\varphi,b}\left(\xi,\beta \mid D\right)^{-1}\right) \tag{A.5.22}$$

suffices. Next, we choose λ such that

$$\frac{\delta v_{1,k(D)}}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}}\frac{\lambda^2}{c} \lesssim \log\left(\rho_{k,\varphi,b}\left(\xi,\beta \mid D\right)^{-1}\right).$$

We can rewrite this condition by plugging in our choice of c through

$$\frac{\lambda^2}{\xi}\frac{z_{1-\beta/2}\delta^{3/2}(v_{1,k}(D))^{5/2}}{(2-\varphi k-\varphi)^4\,\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}} \lesssim \log^{3/2}\left(\rho_{k,\varphi,b}\left(\xi,\beta \mid D\right)^{-1}\right)$$

Choosing λ by

$$\lambda = \left(\frac{\xi}{z_{1-\beta/2}} \frac{(2-\varphi k-\varphi)^4\Gamma^{(1)}_{k,\varphi,b}(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{\delta^{3/2}(v_{1,k}\left(D\right))^{5/2}} \right)^{1/2} \tilde{\lambda}_{k,\varphi,b}\left(D\right), \quad \text{where}$$
$$\tilde{\lambda}_{k,\varphi,b}\left(D\right) = \log^{3/4}\left(\rho_{k,\varphi,b}\left(\xi,\beta \mid D\right)^{-1}\right)$$

will then suffice, as required. ■

A.5.7.3 Part(iii)

Observe that

$$\begin{aligned}
&P\left\{ \left| \left(1-\frac{\hat{g}}{g^\star}\right) \left(a(\mathsf{R}_{\hat{g},k},D) - \overline{a}\left(D\right)\right) \right| \geq \lambda\sqrt{2v_{g^\star,k}\left(D\right)}, |g^\star-\hat{g}| \leq cg^\star \mid D \right\} \\
&= P\left\{ \left| \left(a(\mathsf{R}_{\hat{g},k},D) - \overline{a}\left(D\right)\right) \right| \geq \frac{\lambda}{c}\sqrt{2v_{g^\star,k}\left(D\right)}, |g^\star-\hat{g}| \leq cg^\star \mid D \right\} \\
&\leq P\left\{ \max_{|g^\star-g|\leq cg^\star} \left| \left(a(\mathsf{R}_{g,k},D) - \overline{a}\left(D\right)\right) \right| \geq \frac{\lambda}{c}\sqrt{2v_{g^\star,k}\left(D\right)} \mid D \right\} \\
&\leq P\left\{ \max_{|g^\star-g|\leq cg^\star} \left| \left(a(\mathsf{R}_{g,k},D) - \overline{a}\left(D\right)\right) \right| \geq \frac{\lambda}{c}\left(\frac{\xi}{z_{1-\beta/2}}\right) \mid D \right\} \\
&\leq \sum_{|g^\star-g|\leq cg^\star} P\left\{ \left| \left(a(\mathsf{R}_{g,k},D) - \overline{a}\left(D\right)\right) \right| \geq \frac{\lambda}{c}\left(\frac{\xi}{z_{1-\beta/2}}\right) \right\}.
\end{aligned} \tag{A.5.23}$$

By Theorem A.4.1, we have that

$$P\left\{\left|\left(a(\mathsf{R}_{g,k},D)-\overline{a}\left(D\right)\right)\right|\geq\frac{\lambda}{c}\left(\frac{\xi}{z_{1-\beta/2}}\right)\right\}$$
$$\leq 2\exp\left(-\frac{g\delta}{2^4(2-\varphi k-\varphi)^2\Gamma_{k,\varphi,b}(1)}\frac{\lambda^2}{c^2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2\right)$$

and so is (A.5.23) bounded from above by

$$4cg^\star\exp\left(-\frac{g^\star(1-c)\delta}{4(2-\varphi k-\varphi)^2\Gamma_{k,\varphi,b}(1)}\frac{\lambda^2}{c^2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2\right)$$
$$\lesssim v_{1,k}\left(D\right)c\left(\frac{z_{1-\beta/2}}{\xi}\right)^2\exp\left(-\frac{\delta v_{1,k}(D)}{2^5(2-\varphi k-\varphi)^2\Gamma_{k,\varphi,b}(1)}\frac{\lambda^2}{c^2}\right)$$

if $1-c\geq 1/2$ by the definition of $g^\star$. Thus, we have that

$$P\left\{|Q(\mathsf{R}_{\hat{g},k},D)-Q(\mathsf{R}'_{\hat{g}',k},D)|\geq\lambda\sqrt{2v_{g^\star,k}\left(D\right)}\mid D\right\}$$
$$\lesssim P\left\{\left|\left(1-\frac{\hat{g}}{g^\star}\right)\left(a(\mathsf{R}_{\hat{g},k},D)-\overline{a}\left(D\right)\right)\right|\geq\lambda\sqrt{2v_{g^\star,k}\left(D\right)},|g^\star-\hat{g}|\leq cg^\star\mid D\right\}$$
$$+P\left\{|g^\star-\hat{g}|\geq cg^\star\mid D\right\}$$
$$\lesssim v_{1,k}\left(D\right)c\left(\frac{z_{1-\beta/2}}{\xi}\right)^2\exp\left(-\frac{\delta v_{1,k}(D)}{2^5(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}}\frac{\lambda^2}{c^2}\right) \tag{A.5.24}$$
$$+\exp\left(-\frac{1}{C}\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4}\frac{z_{1-\beta/2}}{\Gamma^{(2)}_{k,\varphi,b}}\frac{c^2}{\xi^2}\right) \tag{A.5.25}$$

for some universal constant C by Lemma A.5.5, Part (ii). If c is chosen to satisfy (A.5.22), the term (A.5.25) will be bounded above by $\rho_{\varphi,k}\left(\xi,\beta\mid D\right)$ for all sufficiently small ξ. By plugging this value of c and $\lambda=\lambda_{\varphi,k}\left(\xi,\beta\mid D\right)$ into (A.5.24), we obtain the term

$$\left(\frac{z^2_{1-\beta/2}}{\xi}\right)\left(\frac{(2-\varphi k-\varphi)^2\,(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}{z_{1-\beta/2}\delta^{1/2}(v_{1,k}(D))^{1/2}}\right)\log^{1/2}\left(\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)^{-1}\right)$$
$$\exp\left(-\frac{1}{\xi}\frac{\delta^{1/2}(v_{1,k}(D))^{3/2}}{(2-\varphi k-\varphi)^2(\Gamma^{(2)}_{k,\varphi,b})^{1/2}}\tilde{\lambda}^U_{k,\varphi,b}\left(D\right)\log\left(\rho_{k,\varphi,b}\left(\xi,\beta\mid D\right)^{-1}\right)\right) \tag{A.5.26}$$

as required. Observe that the term (A.5.26) is bounded above by $\rho_{\varphi,k}\left(\xi,\beta\mid D\right)$, for all sufficiently small ξ, completing the proof. ■

A.5.8 Proofs for Supporting Lemmas

A.5.8.1 Proof of Lemma A.5.1

Let $T_i(\pi_{m,\ell}) = T(\mathsf{s}_i(\pi_{\ell,m}), D)$ and define $T_i(\pi'_{m,\ell})$ analogously. Define

$$R_m(\ell) = \left\{ i \in [n] : \pi_{m,\ell}(i) \neq \pi'_{m,\ell}(i) \right\}$$

and let $\overline{R}_m(\ell) = |R_m(\ell)|$ denote the number of shared indices in the permutations in $\pi_{m,\ell}$ and $\pi'_{m,\ell}$. Observe that

$$\frac{1}{k} \sum_{i=1}^{k} \left(T_i(\pi_{m,\ell}) - T_i\left(\pi'_{m,\ell}\right) \right) = 0 \tag{A.5.27}$$

almost surely for all $m' \geq m$ if and only if $\overline{R}_m(\ell) = n$. Let $N(\ell)$ denote the values of the smallest index m with $\overline{R}_m(\ell) = n$. Observe that

$$\begin{aligned}
&\sum_{m=0}^{\infty} \left| \mathbb{E}\left[a(\mathsf{R}_{g,k}(\boldsymbol{\pi}_m), D) - a(\mathsf{R}_{g,k}(\boldsymbol{\pi}'_m), D) \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D \right] \right| \qquad \text{(A.5.28)} \\
&\leq \sum_{m=0}^{\infty} \frac{1}{g} \frac{1}{k} \sum_{\ell=1}^{g} \sum_{i=1}^{k} \mathbb{E}\left[\left| \left(T_i(\pi_{m,\ell}) - T_i\left(\pi'_{m,\ell}\right) \right) \right| \mid \boldsymbol{\pi}_0 = \boldsymbol{\psi}, \boldsymbol{\pi}'_0 = \boldsymbol{\psi}', D \right] \\
&\leq \frac{1}{g} \frac{1}{k} \sum_{\ell=1}^{g} \sum_{i=1}^{k} \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left| T(\mathsf{s}, D) - T(\mathsf{s}', D) \right| \mathbb{E}\left[N(\ell) \right]
\end{aligned}$$

Hence, it suffices to bound the quantity $\mathbb{E}[N(\ell)]$.

To that end, let $N_r(\ell)$ be the value of the smallest index m with $\overline{R}_m(\ell) \geq r$. We proceed analogously to standard analysis of the coupon collector's problem (see e.g., Section 2.2 of Levin and Peres, 2017). We can evaluate

$$\begin{aligned}
&P\left\{ \overline{R}_m(\ell) \geq r + 1 \mid \overline{R}_{m-1}(\ell) = r \right\} \\
&= \frac{1}{g} P\left\{ \pi_\ell(I_m) \neq \pi'_\ell(I_m), \pi_\ell^{-1}(J_m) \neq \pi_\ell'^{-1}(J_m) \right\} = \frac{(n-r)^2}{gn^2}
\end{aligned}$$

and

$$P\left\{ \overline{R}_m(\ell) < r \mid \overline{R}_{m-1}(\ell) = r \right\} = 0,$$

and thereby obtain the bound

$$\mathbb{E}\left[N\left(\ell\right)\right] = \sum_{r=1}^{n} \mathbb{E}\left[N_r\left(\ell\right) - N_{r-1}\left(\ell\right)\right] \leq \sum_{r=1}^{n} \frac{gn^2}{\left(n-r+1\right)^2} = gn^2 \sum_{r=1}^{n} \frac{1}{r^2} \leq 2gn^2. \tag{A.5.29}$$

Hence, we find that (A.5.28) is upper bounded by

$$2gn^2 \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} T\left(\mathsf{s}, D\right) - T\left(\mathsf{s}', D\right),$$

as required. ■

A.5.8.2 Proof of Lemma A.5.2

Observe that

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{tx} e^{(1-t)y} = \left(x-y\right) e^{tx} e^{(1-t)y}.$$

Thus, we find that

$$\begin{aligned}
c\left(e^x - e^y\right) &= c \int_0^1 \left(\frac{\mathrm{d}}{\mathrm{d}t} e^{tx} e^{(1-t)y}\right) \mathrm{d}t && \text{(Fundamental Theorem of Calculus)} \\
&= c\left(x-y\right) \int_0^1 e^{tx} e^{(1-t)y} \mathrm{d}t \\
&\leq c\left(x-y\right) \int_0^1 \left(te^x + \left(1-t\right) e^y\right) && \text{(Convexity)} \\
&= \frac{c}{2}\left(x-y\right)\left(e^x + e^y\right) \\
&= \frac{1}{2}\left(\left(s^{-1}c^2\left(e^x + e^y\right)\right)\left(s\left(x-y\right)^2\left(e^x + e^y\right)\right)\right)^{1/2} \\
&\leq \frac{1}{4}\left(\left(s^{-1}c^2\right) + s\left(x-y\right)^2\right)\left(e^x + e^y\right), && \text{(AM-GM)}
\end{aligned}$$

as required. ■

A.5.8.3 Proof of Lemma A.5.3

To begin, consider any collection of real numbers $x_1, \ldots, x_{2^c}$, for some positive integer c. For any integer $s > 0$, we have

$$(x_1 \cdots x_{2^c})^{2s} \leq \frac{1}{4}\left(\left(x_{i_1} \cdots x_{i_{2^{c-1}}}\right)^{2s} + \left(x_{i_{2^{c-1}+1}} \cdots x_{2^c}\right)^{2s}\right)^2 \qquad \text{(Young's Inequality)}$$
$$\leq \frac{1}{4}\left(\left(x_{i_1} \cdots x_{i_{2^{c-1}}}\right)^{2(s+1)} + 2\,(x_1 \cdots x_{2^c})^{2s} + \left(x_{i_{2^{c-1}+1}} \cdots x_{2^c}\right)^{2(s+1)}\right)$$

and so

$$(x_1 \cdots x_{2^c})^{2s} \leq \frac{1}{2}\left(x_{i_1} \cdots x_{i_{2^{c-1}}}\right)^{2(s+1)} + \frac{1}{2}\left(x_{i_{2^{c-1}+1}} \cdots x_{2^c}\right)^{2(s+1)}. \qquad \text{(A.5.30)}$$

Consequently, we have that

$$x_1 \cdots x_{2^c} \leq \frac{1}{2^c}\sum_{i=1}^{2^c} x_i^{2c} \qquad \text{(A.5.31)}$$

through 2^r applications of (A.5.30).

Now, to prove the Lemma, we may assume without loss that $h_i > 0$ for all i by continuity. Observe that

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m\right)^{2^c}\right] = \sum_{i_1=0}^{\infty} \cdots \sum_{i_{2^c}=0}^{\infty} \mathbb{E}\left[X_{i_1} \cdots X_{i_{2^c}}\right] \qquad \text{(A.5.32)}$$

by dominated convergence. By writing

$$X_1 \cdots X_{2^c} = \prod_{j=1}^{2^c}\left(\frac{\prod_{k\neq j} h_k}{h_j^{2^{c-1}}}\right)^{1/2^c} X_j$$

we find that

$$\mathbb{E}\left[X_1 \cdots X_{2^c}\right] \leq \frac{1}{2^c}\sum_{j=1}^{2^c} \frac{\prod_{k\neq j} h_k}{h_j^{2^{c-1}}} h_j^{2^c} = \prod_{j=1}^{2^c} h_j$$

by (A.5.31). Hence, by (A.5.32), we have that

$$\mathbb{E}\left[\left(\sum_{m=0}^{\infty} X_m\right)^{2^c}\right] \leq \sum_{i_1=0}^{\infty} \cdots \sum_{i_{2^c}=0}^{\infty} \prod_{j=1}^{2^c} h_{i_j} = \left(\sum_{m=0}^{\infty} h_j\right)^{2^c},$$

as required. ■

A.5.8.4 Proof of Lemma A.5.4

We begin by constructing Stein representers $\tilde{V}\left(Z, Z'\right)$ and $\check{V}\left(Z, Z'\right)$ for the statistic $\tilde{v}_{g,k}\left(Z\right)$ and $\check{v}_{g,k}\left(Z\right)$, respectively. We will use the same exchangeable pair $\left(Z, Z'\right)_{m\geq 0}$ and Markov chain $\left(Z_m, Z'_m\right)_{m\geq 1}$ defined in Section A.4.4. The subsequent result follows from an argument similar to Lemma A.4.1, again based on an idea expressed by Lemma 4.1 of Chatterjee (2005).

Lemma A.5.6.

(i) *Let ψ and ψ' be two sets, each containing g elements of $\mathcal{P}_n$. The inequalities*

$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\left(\tilde{v}_{g,k}\left(Z_m\right)-\tilde{v}_{g,k}\left(Z'_m\right)\right)\mid Z_0=Z, Z'_0=Z'\right]\right|\leq n^2\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2\quad\text{and}$$
$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\left(\check{v}_{g,k}\left(Z_m\right)-\check{v}_{g,k}\left(Z'_m\right)\right)\mid Z_0=Z, Z'_0=Z'\right]\right|\leq 2gn^2\max_{\mathsf{s}\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2$$

hold almost surely.

(ii) *The functions*

$$\tilde{V}\left(Z,Z'\right)=\sum_{m=0}^{\infty}\mathbb{E}\left[\left(\tilde{v}_{g,k}\left(Z_m\right)-\tilde{v}_{g,k}\left(Z'_m\right)\right)\mid Z_0=Z, Z'_0=Z'\right]\quad\text{and}$$
$$\check{V}\left(Z,Z'\right)=\sum_{m=0}^{\infty}\mathbb{E}\left[\left(\check{v}_{g,k}\left(Z_m\right)-\check{v}_{g,k}\left(Z'_m\right)\right)\mid Z_0=Z, Z'_0=Z'\right]$$

are finite and satisfy the equalities

$$\mathbb{E}\left[\tilde{V}\left(Z,Z'\right)\mid Z\right]=\tilde{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\quad\text{and}$$
$$\mathbb{E}\left[\check{V}\left(Z,Z'\right)\mid Z\right]=\check{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)$$

almost surely.

Next, we apply the concentration inequality stated in Theorem A.4.2, due to Chatterjee (2005, 2007). That is, to establish part (i) of the Lemma, it will suffice to characterize constants s and u that satisfy

$$U_{\tilde{v}}\left(Z\right)\leq s^{-1}u\quad\text{and}\quad U_{\tilde{V}}\left(Z\right)\leq su,$$

with probability $1-\delta$, where

$$U_{\tilde{v}}(Z)=\frac{1}{2}\mathbb{E}\left[\left(\tilde{v}(Z)-\tilde{v}(Z')\right)\mid Z\right] \quad \text{and} \quad U_{\tilde{V}}(Z)=\frac{1}{2}\mathbb{E}\left[\tilde{V}(Z,Z')^2\mid Z\right].$$

The same statement holds for part (ii) of the Lemma, for the objects $U_{\check{v}}(Z)$ and $U_{\check{V}}(Z)$ defined analogously.

We obtain such a characterization through the application of Lemma A.4.3. To this end, observe that, by Lemma A.5.6, part (i), the bound

$$\left(\sum_{m=0}^{\infty}\mathbb{E}\left[\left(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\right)\mid Z_0=Z, Z'_0=Z'\right]\right)^2 \tag{A.5.33}$$

$$\leq n^4\max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^4 \tag{A.5.34}$$

holds almost surely. The right hand side of (A.5.34) is square integrable, as the eighth-order split-stability $\zeta^{(8)}$ is finite almost surely. An analogous statement holds for the statistic $\check{v}_{g,k}(Z)$. Thus, deterministic bounds of the form (A.4.29) can be obtained through the application of the following Lemma.

Lemma A.5.7. *Suppose that Assumptions 1.4.1 and B.2.3 hold and that the data D are independent and identically distributed. If the eighth-order split stability $\zeta^{(8)}$ is finite, then*
(i) *The inequality*

$$\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid Z,Z'\right]^2\mid\boldsymbol{\pi}\right]\lesssim\left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{(2-\varphi k-\varphi)^4}{n^2g^4}\right)\Gamma^{(2)}_{k,\varphi,b},$$

holds almost surely for all integers $m\geq 0$, and
(ii) *The inequality*

$$\mathbb{E}\left[\mathbb{E}\left[\check{v}_{g,k}(Z_m)-\check{v}_{g,k}(Z'_m)\mid Z,Z'\right]^2\mid\boldsymbol{\pi}\right]\lesssim\left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{(2-\varphi k-\varphi)^4}{n^2g^3}\right)\Gamma^{(2)}_{k,\varphi,b},$$

holds almost surely for all integers $m\geq 0$.

Thus, by Lemma A.4.3, part (i), the inequalities

$$U_{\tilde{V}}(Z)\lesssim\frac{1}{\delta}\left(\frac{gn^2}{2}\right)^2\left(\frac{(2-\varphi k-\varphi)^4}{n^2g^4}\right)\Gamma^{(2)}_{k,\varphi,b}$$

$$\lesssim \left(\frac{gn^2}{2}\right)\left(\frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^3}\right)\Gamma^{(2)}_{k,\varphi,b}$$

and

$$\begin{aligned} U_{\tilde{v}}(Z) &\lesssim \frac{1}{\delta}\left(\frac{(2-\varphi k-\varphi)^4}{n^2 g^4}\right)\Gamma^{(2)}_{k,\varphi,b} \\ &\lesssim \left(\frac{2}{gn^2}\right)\left(\frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^3}\right)\Gamma^{(2)}_{k,\varphi,b} \end{aligned}$$

both hold with probability greater than $1-\delta$. Hence, we obtain the bound (A.5.16) by applying Theorem A.4.2 and choosing $s = gn^2/2$ and

$$u = \frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^3}\Gamma^{(2)}_{k,\varphi,b}.$$

Analogous inequalities for the objects $U_{\check{v}}(Z)$ and $U_{\check{V}}(Z)$ hold by Lemma A.4.3, part (ii), where in that case

$$u = \frac{1}{\delta}\frac{(2-\varphi k-\varphi)^4}{g^2}n^4\Gamma^{(2)}_{k,\varphi,b}.$$

which similarly implies the bound (A.5.17) by Theorem A.4.2. ■

A.5.8.5 Proof of Lemma A.5.6

Reinstate the notation of the proof of Lemma A.5.1. We begin by noting that

$$\begin{aligned} \max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}} \left(T(\mathsf{s},D)-T(\mathsf{s}',D)\right)^2 &= \max_{\mathsf{s},\mathsf{s}'\in\mathsf{S}_{n,b}} \left(T(\mathsf{s},D)-\overline{a}(D)\right)^2 + \left(T(\mathsf{s}',D)-\overline{a}(D)\right)^2 \\ &\qquad + 2\left(T(\mathsf{s},D)-\overline{a}(D)\right)\left(T(\mathsf{s}',D)-\overline{a}(D)\right) \\ &\geq 4\max_{\mathsf{s}\in\mathsf{S}_{n,b}}\left(T(\mathsf{s},D)-\overline{a}(D)\right)^2. \end{aligned}$$

Observe that

$$\sum_{i,i'=1}^{k}\left(T_i(\pi_{m,\ell})-\overline{a}(D)\right)\left(T_i(\pi_{m,\ell})-\overline{a}(D)\right) - \sum_{i,i'=1}^{k}\left(T_i\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right)\left(T_{i'}\left(\pi'_{m,\ell}\right)-\overline{a}(D)\right) = 0$$

for all for all $m' \geq m$ if and only if $\overline{R}_m(\ell) = n$. Thus, we have that

$$\sum_{m=0}^{\infty}\left|\mathbb{E}\left[\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m)\mid Z_0=(\mathsf{r}(\boldsymbol{\psi}),D), Z'_0=(\mathsf{r}(\boldsymbol{\psi}'),D)\right]\right|$$

$$\leq \sum_{m=0}^{\infty} \frac{1}{g^2k^2} \sum_{\ell=1}^{g} \sum_{i,i'=1}^{k} \mathbb{E}\Bigg[\Bigg| \left(T_i\left(\pi_{m,\ell}\right) - \overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi_{m,\ell}\right) - \overline{a}\left(D\right)\right)$$
$$- \left(T_i\left(\pi'_{m,\ell}\right) - \overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi'_{m,\ell}\right) - \overline{a}\left(D\right)\right)\Bigg| \mid Z_0 = \left(\mathsf{r}\left(\boldsymbol{\psi}\right), D\right), Z'_0 = \left(\mathsf{r}\left(\boldsymbol{\psi}'\right), D\right)\Bigg]$$
$$\leq \frac{2}{g^2k^2} \sum_{\ell=1}^{g} \sum_{i,i'=1}^{k} \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left| \left(T\left(\mathsf{s}, D\right) - \overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}', D\right) - \overline{a}\left(D\right)\right)\right| \mathbb{E}\left[N\left(\ell\right)\right].$$
$$\leq \frac{2}{g} \max_{\mathsf{s} \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - \overline{a}\left(D\right)\right)^2 \mathbb{E}\left[N\left(\ell\right)\right].$$
$$\leq \frac{1}{2g} \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T(\mathsf{s}', D)\right)^2 \mathbb{E}\left[N\left(\ell\right)\right].$$

Similarly, we have that

$$\sum_{m=0}^{\infty} \left| \mathbb{E}\left[\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m) \mid Z_0 = \left(\mathsf{r}\left(\boldsymbol{\psi}\right), D\right), Z'_0 = \left(\mathsf{r}\left(\boldsymbol{\psi}'\right), D\right)\right] \right|$$
$$\leq \sum_{m=0}^{\infty} \frac{1}{g^2k^2} \sum_{\ell,\ell'=1}^{g} \sum_{i,i'=1}^{k} \mathbb{E}\Bigg[\Bigg| \left(T_i\left(\pi_{m,\ell}\right) - \overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi_{m,\ell'}\right) - \overline{a}\left(D\right)\right)$$
$$- \left(T_i\left(\pi'_{m,\ell}\right) - \overline{a}\left(D\right)\right)\left(T_{i'}\left(\pi'_{m,\ell'}\right) - \overline{a}\left(D\right)\right)\Bigg| \mid Z_0 = \left(\mathsf{r}\left(\boldsymbol{\psi}\right), D\right), Z'_0 = \left(\mathsf{r}\left(\boldsymbol{\psi}'\right), D\right)\Bigg]$$
$$\leq \frac{2}{g^2k^2} \sum_{\ell,\ell'=1}^{g} \sum_{i,i'=1}^{k} \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left| \left(T\left(\mathsf{s}, D\right) - \overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}', D\right) - \overline{a}\left(D\right)\right)\right| \mathbb{E}\left[\max\left\{N\left(\ell\right), N\left(\ell'\right)\right\}\right]$$
$$\leq \frac{2}{g^2k^2} \sum_{\ell,\ell'=1}^{g} \sum_{i,i'=1}^{k} \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left| \left(T\left(\mathsf{s}, D\right) - \overline{a}\left(D\right)\right)\left(T\left(\mathsf{s}', D\right) - \overline{a}\left(D\right)\right)\right| \mathbb{E}\left[N\left(\ell\right) + N\left(\ell'\right)\right]$$
$$\leq \frac{1}{2} \max_{\mathsf{s},\mathsf{s}' \in \mathsf{S}_{n,b}} \left(T(\mathsf{s}, D) - T(\mathsf{s}', D)\right)^2 \left|\mathbb{E}\left[N\left(\ell\right) + N\left(\ell'\right)\right]\right.$$

Thus, part (i) of the Lemma follows by the bound (A.5.29). Part (ii) of the Lemma then follows by an argument analogous to the proof of Lemma A.4.1. ■

A.5.8.6 Proof of Lemma A.5.7

We reinstate the notation introduced in the proof of Lemma A.4.4 from Section A.5.4. First, observe that

$$\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_m) - \tilde{v}_{g,k}(Z'_m) \mid Z, Z'\right]^2 \mid \boldsymbol{\pi}\right]$$
$$= \mathbb{E}\left[\left(P\left\{\mathcal{H}_m\right\} \mathbb{E}\left[\tilde{v}_{g,k}(Z_m) - \tilde{v}_{g,k}(Z'_m) \mid \mathcal{H}_m, Z, Z'\right]\right)^2 \mid \boldsymbol{\pi}\right]$$

$$\leq \left(1-\frac{2}{gn^2}\right)^{2m}\left(\frac{2kb\,(2n-bk-b)}{n^2}\right)^2 \mathbb{E}\left[(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m))^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \quad \text{(A.5.35)}$$

as before. Recall the notation

$$\overline{a}(\mathsf{r}_\ell, D) = \frac{1}{k}\sum_{i=1}^{k}(T(\mathsf{s}_{\ell,i}, D) - \overline{a}(D))$$

and observe that

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m))^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \\
&= \frac{1}{g^4}\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D)^2 - \overline{a}(\mathsf{r}(\pi'_{m,L}), D)^2\right)^2 \mid \mathcal{H}_m\right] \\
&= \frac{1}{g^4}\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D) + \overline{a}(\mathsf{r}(\pi'_{m,L}), D)\right)^2\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D) - \overline{a}(\mathsf{r}(\pi'_{m,L}), D)\right)^2 \mid \mathcal{H}_m\right] \\
&\leq \frac{1}{g^4}\left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D) + \overline{a}(\mathsf{r}(\pi'_{m,L}), D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} && \text{(Cauchy-Schwarz)} \\
&\qquad \cdot \left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D) - \overline{a}(\mathsf{r}(\pi'_{m,L}), D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \\
&\leq \frac{2^4}{g^4}\left(\mathbb{E}\left[(\overline{a}(\mathsf{r}(\pi_{m,L}), D) - \overline{a}(D))^4\right]\right)^{1/2} && \text{(Hölder)} \\
&\qquad \cdot \left(\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D) - \overline{a}(\mathsf{r}(\pi'_{m,L}), D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2}\,. && \text{(A.5.36)}
\end{aligned}$$

Observe that

$$\mathbb{E}\left[(\overline{a}(\mathsf{r}(\pi_{m,L}), D) - \overline{a}(D))^4\right] \leq 3^2\left(2^4(2-\varphi k-\varphi)^2\right)^2\Gamma^{(2)}_{k,\varphi,b}$$

by Theorem A.4.2, as Assumption 1.4.1 is maintained and the eighth-order sample-split stability $\zeta^{(8)}$ is finite. In turn, we have that

$$\mathbb{E}\left[\left(\overline{a}(\mathsf{r}(\pi_{m,L}), D) - \overline{a}(\mathsf{r}(\pi'_{m,L}), D)\right)^4 \mid \mathcal{H}_m\right] \leq \frac{4}{k^4 b^4}\Gamma^{(2)}_{k,\varphi,b} \quad \text{(A.5.37)}$$

by (A.5.7), (A.5.13), and (A.5.14). Hence, we have that

$$\mathbb{E}\left[(\tilde{v}_{g,k}(Z_m)-\tilde{v}_{g,k}(Z'_m))^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \leq \frac{3\cdot 2^9}{k^2 b^2 g^4}(2-\varphi k-\varphi)^2\Gamma^{(2)}_{k,\varphi,b}\,. \quad \text{(A.5.38)}$$

Combining (A.5.35), (A.5.37), (A.5.38), we find that

$$\mathbb{E}\left[\mathbb{E}\left[\tilde{v}_{g,k}(Z_m) - \tilde{v}_{g,k}(Z'_m) \mid Z, Z'\right]^2 \mid \boldsymbol{\pi}\right] \leq \left(1 - \frac{2}{gn^2}\right)^{2m} \left(\frac{(3 \cdot 2^{11})\,(2 - \varphi k - \varphi)^4}{n^2 g^4}\right) \Gamma^{(2)}_{k,\varphi,b}\,,$$

which completes the proof of the first part of the Lemma.

Second, following the same argument, we again have that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m) \mid Z, Z'\right]^2 \mid \boldsymbol{\pi}\right] \\
&\leq \left(1 - \frac{2}{gn^2}\right)^{2m} \left(\frac{2kb\,(2n - bk - b)}{n^2}\right)^2 \mathbb{E}\left[\left(\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \qquad \text{(A.5.39)}
\end{aligned}$$

In this case, we can compute

$$\begin{aligned}
&\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \\
&= \mathbb{E}\left[\left(\left(a(Z_m) - \overline{a}(D)\right)^2 - \left(a(Z'_m) - \overline{a}(D)\right)^2\right)^2 \mid \mathcal{H}_m\right] \\
&= \mathbb{E}\left[\left(a(Z_m) - \overline{a}(D) + a(Z'_m) - \overline{a}(D)\right)^2 \left(a(Z_m) - a(Z'_m)\right)^2 \mid \mathcal{H}_m\right] \\
&\leq \left(\mathbb{E}\left[\left(a(Z_m) - \overline{a}(D) + a(Z'_m) - \overline{a}(D)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \qquad \text{(Cauchy-Schwarz)} \\
&\quad \cdot \left(\mathbb{E}\left[\left(a(Z_m) - a(Z'_m)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2} \\
&\leq 2^4 \left(\mathbb{E}\left[\left(a(Z_m) - \overline{a}(D)\right)^4\right]\right)^{1/2} \left(\mathbb{E}\left[\left(a(Z_m) - a(Z'_m)\right)^4 \mid \mathcal{H}_m\right]\right)^{1/2}\,, \qquad \text{(A.5.40)}
\end{aligned}$$

where the last inequality follows from Hölder's inequality. Again we have that

$$\mathbb{E}\left[\left(a(Z_m) - a(Z'_m)\right)^4\right] \leq 3^2 \left(\frac{2^4 \varphi^2 (2 - \varphi k - \varphi)^2}{g}\right)^2 \Gamma^{(2)}_{k,\varphi,b} \qquad \text{(A.5.41)}$$

by Theorem A.4.2, as Assumption 1.4.1 is maintained and the eighth-order sample-split stability $\zeta^{(8)}$ is finite. Similarly. we have that

$$\mathbb{E}\left[\left(a(Z_m) - \overline{a}(D)\right)^4 \mid \mathcal{H}_m\right] \leq \frac{4}{g^4 k^4 b^4} \Gamma^{(2)}_{k,\varphi,b} \qquad \text{(A.5.42)}$$

by (A.5.7), (A.5.13), and (A.5.14). Combining (A.5.39), (A.5.41), (A.5.42), we find that

$$\mathbb{E}\left[\left(\check{v}_{g,k}(Z_m) - \check{v}_{g,k}(Z'_m)\right)^2 \mid \mathcal{H}_m, \boldsymbol{\pi}\right] \leq \left(1 - \frac{2}{gn^2}\right)^{2m} \left(\frac{(3 \cdot 2^9)\,\varphi^4\,(2 - \varphi k - \varphi)^4}{n^2 g^3}\right) \Gamma^{(2)}_{k,\varphi,b}\,,$$

which completes the proof of the second part of the Lemma. ■

A.5.8.7 Proof of Lemma A.5.5, Part (i)

Observe that

$$\begin{aligned}
&P\left\{|U(\mathsf{R}_{\hat{g},k},D)| \geq t\sqrt{v_{g^\star,k}(D)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\} \\
&= P\left\{\sum_{i=1}^{g}\overline{a}(\mathsf{r}_{i,k},D) - \sum_{i=1}^{g^\star}\overline{a}(\mathsf{r}_{i,k},D) \geq tg^\star\sqrt{v_{g^\star,k}(D)}, |\hat{g}-g^\star| \leq cg^\star \mid D\right\} \\
&\leq P\left\{\max_{g^\star(1-c)\leq g\leq g^\star} |\sum_{i=1}^{g}\overline{a}(\mathsf{r}_{i,k},D) - \sum_{i=1}^{g^\star}\overline{a}(\mathsf{r}_{i,k},D)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\} \\
&+ P\left\{\max_{g^\star\leq g\leq g^\star(1+c)} |\sum_{i=1}^{g}\overline{a}(\mathsf{r}_{i,k},D) - \sum_{i=1}^{g^\star}\overline{a}(\mathsf{r}_{i,k},D)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\} \\
&= P\left\{\max_{g^\star(1-c)\leq g\leq g^\star} |\sum_{i=1}^{g^\star-g}\overline{a}(\mathsf{r}_{i,k},D)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\} \\
&+ P\left\{\max_{g^\star\leq g\leq g^\star(1+c)} |\sum_{i=1}^{g-g^\star}\overline{a}(\mathsf{r}_{i,k},D)| \geq tg^\star\sqrt{v_{g^\star,k}(D)} \mid D\right\}. \qquad \text{(A.5.43)}
\end{aligned}$$

To bound the two probabilities (A.5.43), we apply the following Chernoff-type variation to the Kolmogorov maximal inequality, due to Steiger (1970).

Theorem A.5.2 (Steiger, 1970)**.** *Let S_i, $i = 1, 2, \ldots$, be a real-valued martingale sequence. If the moment generating function*

$$m_n(\theta) = \mathbb{E}[\exp(\theta S_n)]$$

is finite for all positive θ, then the inequality

$$\log\ P\left\{\max_{1\leq n'\leq n} S_{n'} > t\right\} \leq \inf_{\theta>0}(\log m_n(\theta) - \theta t),$$

holds.

Conditional on D, the random variables

$$\overline{a}(\mathsf{r}_{i,k},D), \quad i = 1, 2, \ldots,$$

are mean-zero, independent, and identically distributed. Thus, the partial sums

$$S_m = \sum_{i=1}^{m} \overline{a}\left(\mathsf{r}_{i,k}, D\right)$$

are a martingale sequence. Moreover, though inspection of the proof of Lemma A.4.2, we find that Theorem Corollary A.4.1 implies that

$$\begin{aligned}
&\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\theta S_{\lfloor cg^\star \rfloor}\right) \mid D\right] - \theta\tau\right) \\
&= \inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\frac{\theta}{\lfloor cg^\star \rfloor} S_{\lfloor cg^\star \rfloor}\right) \mid D\right] - \theta \frac{\tau}{\lfloor cg^\star \rfloor}\right) \\
&\le -\frac{\lfloor cg^\star \rfloor}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{\delta\tau^2}{\left(\lfloor cg^\star \rfloor\right)^2} \\
&\le -\frac{\delta}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{\tau^2}{cg^\star}
\end{aligned}$$

with probability greater than $1-\delta$, as Assumption 1.4.1 holds and the fourth-order split stability $\zeta^{(4)}$ is finite. Thus, by setting

$$\tau = tg^\star\sqrt{v_{g^\star,k}\left(D\right)} = \sqrt{g^\star} t \left(v_{1,k}(D)\right)^{1/2}$$

we find that

$$\begin{aligned}
&\log \frac{1}{2} P\left\{\left|U(\mathsf{R}_{\hat{g},k}, D)\right| \ge t\sqrt{v_{g^\star,k}\left(D\right)}, \left|\hat{g}-g^\star\right| \le cg^\star \mid D\right\} \\
&\le -\frac{\delta v_{1,k}(D)}{2^4(2-\varphi k-\varphi)^2\Gamma^{(1)}_{k,\varphi,b}} \frac{t^2}{c}
\end{aligned}$$

with probability greater than $1-\delta$, by Theorem A.5.2 and the inequality (A.5.43).

A.5.8.8 Proof of Lemma A.5.5, Part (ii)

Observe that

$$P\left\{\left|\hat{g}-g^\star\right| > cg^\star \mid D\right\} = P\left\{g^\star\left(1+c\right) < \hat{g} \mid D\right\} + P\left\{\hat{g} < g^\star\left(1-c\right) \mid D\right\}. \quad \text{(A.5.44)}$$

We begin by handling the first term. We have that

$$P\left\{g^\star\left(1+c\right) < \hat{g} \mid D\right\}$$

$$\leq P\left\{\hat{v}_{g^\star(1+c),k}\left(Z\right)-v_{g^\star(1+c),k}\left(D\right)+v_{g^\star(1+c),k}\left(D\right)>\frac{1}{2}\left(\frac{\xi}{z_{1-\beta/2}}\right)^2\mid D\right\}$$
$$\leq P\left\{\hat{v}_{g^\star(1+c),k}\left(Z\right)-v_{g^\star(1+c),k}\left(D\right)>\frac{v_{1,k}(D)}{g^\star}-\frac{v_{1,k}(D)}{g^\star\left(1+c\right)}\mid D\right\}$$
$$=P\left\{\hat{v}_{g^\star(1+c),k}\left(Z\right)-v_{g^\star(1+c),k}\left(D\right)>\frac{v_{1,k}(D)}{g^\star}\frac{c}{1+c}\mid D\right\},$$

where the first inequality follows the definition of $g^\star$. Thus, as Assumption 1.4.1 holds and the eighth-order split stability $\zeta^{(8)}$ is finite, we have that

$$\log\frac{1}{4}P\left\{g^\star\left(1+c\right)<\hat{g}\mid D\right\}\lesssim-\frac{\delta(v_{1,k}(D))^2}{(2-\varphi k-\varphi)^4}\frac{g^\star(1+c)c^2}{\Gamma^{(2)}_{k,\varphi,b}}$$
$$\lesssim-\frac{\delta(v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4}\frac{z^2_{1-\beta/2}c^2}{\xi^2\Gamma^{(2)}_{k,\varphi,b}}$$

with probability greater than $1-\delta$, by Lemma A.4.5, the definition of $g^\star$ and the assumption that $0<c<1/2$.

Next, we bound the second term in (A.5.44). Define the partial sums

$$A_g=\sum_{\ell=1}^{g}\left(\frac{1}{k^2}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\left(T(\mathsf{s}_{\ell,i'},D)-\overline{a}\left(D\right)\right)-v_{1,k}(D)\right)\quad\text{and}$$
$$B_g=\sum_{\ell=1}^{g}\left(\frac{1}{k^2}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\left(T(\mathsf{s}_{\ell,i'},D)-\overline{a}\left(D\right)\right)\right.$$
$$\left.+2\sum_{\ell'=1}^{\ell-1}\left(T(\mathsf{s}_{\ell,i},D)-\overline{a}\left(D\right)\right)\left(T(\mathsf{s}_{\ell',i'},D)-\overline{a}\left(D\right)\right)-v_{1,k}(D)\right).$$

Observe that the equality

$$\hat{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)=\frac{g}{g-1}\left(\tilde{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right)-\frac{1}{g-1}\left(\check{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)\right),$$

from the proof of Lemma A.4.5, implies that

$$\hat{v}_{g,k}\left(Z\right)-v_{g,k}\left(D\right)=\frac{1}{g\left(g-1\right)}\sum_{\ell=1}^{g}\frac{1}{k^2}\sum_{i,i'=1}^{k}\left(T(\mathsf{s}_{\ell,i},D)-a\left(Z\right)\right)\left(T(\mathsf{s}_{\ell,i'},D)-a\left(Z\right)\right)-\frac{v_{1,k}(D)}{g}$$
$$=\frac{1}{g-1}\frac{1}{g}A_g-\frac{1}{g-1}\frac{1}{g^2}B_g$$

for all positive g. Thus, we can write

$$\begin{aligned}
&P\left\{\hat{g} < g^{\star}\left(1-c\right) \mid D\right\} \\
&= P\left\{\min_{2\leq g'\leq g^{\star}(1-c)} \hat{v}_{g',k}\left(Z\right) \leq \left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\} \\
&= P\left\{\min_{2\leq g'\leq g^{\star}(1-c)} \frac{\left(g'A_{g'} - B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2} + \frac{1}{g'}v_{1,k}(D) \leq \left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\},
\end{aligned}$$

where the first equality follows from the definition of $\hat{g}$. Now, in the event that

$$\min_{2\leq g'\leq g^{\star}(1-c)} \frac{\left(g'A_{g'} - B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2} + \frac{1}{g'}v_{1,k}(D) \leq \left(\frac{\xi}{z_{1-\beta/2}}\right)^2, \tag{A.5.45}$$

it must also be the case that

$$\min_{2\leq g'\leq g^{\star}(1-c)} \left(g'-1\right)\left(g'\right)^2 \left(\frac{\left(g'A_{g'} - B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2} + \frac{1}{g'}v_{1,k}(D)\right) \leq (\hat{g}-1)(\hat{g})^2 \left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \tag{A.5.46}$$

as $\hat{g} \leq g^{\star}\left(1-c\right)$ necessarily. But then, similarly, (A.5.46) implies that

$$\min_{2\leq g'\leq g^{\star}(1-c)} \left(g'A_{g'} - B_{g'}\right) + \frac{(\hat{g}-1)(\hat{g})^2}{g'}v_{1,k}(D) \leq (\hat{g}-1)(\hat{g})^2 \left(\frac{\xi}{z_{1-\beta/2}}\right)^2,$$

and in turn

$$\begin{aligned}
&\min_{2\leq g'\leq g^{\star}(1-c)} \frac{\left(g'A_{g'} - B_{g'}\right)}{\left(g^{\star}\left(1-c\right)-1\right)\left(g^{\star}\left(1-c\right)\right)^2} \\
&\leq \frac{(\hat{g}-1)(\hat{g})^2}{\left(g^{\star}\left(1-c\right)-1\right)\left(g^{\star}\left(1-c\right)\right)^2}\left(\frac{1}{g^{\star}-1} - \frac{1}{g^{\star}\left(1-c\right)}\right)v_{1,k}(D) \\
&= \frac{(\hat{g}-1)(\hat{g})^2}{\left(g^{\star}\left(1-c\right)-1\right)\left(g^{\star}\left(1-c\right)\right)^2}\left(\frac{1-g^{\star}c}{g^{\star}\left(g^{\star}-1\right)\left(1-c\right)}\right)v_{1,k}(D),
\end{aligned}$$

are then also true. Finally, again as (A.5.45) is equivalent to $\hat{g} < g^{\star}\left(1-c\right)$, we have that

$$\begin{aligned}
&P\left\{\min_{2\leq g'\leq g^{\star}(1-c)} \frac{\left(g'A_{g'} - B_{g'}\right)}{\left(g'-1\right)\left(g'\right)^2} + \frac{1}{g'}v_{1,k}(D) \leq \left(\frac{\xi}{z_{1-\beta/2}}\right)^2 \mid D\right\} \\
&\leq P\left\{\min_{2\leq g'\leq g^{\star}(1-c)} \frac{g'A_{g'} - B_{g'}}{\left(g^{\star}\left(1-c\right)-1\right)\left(g^{\star}\left(1-c\right)\right)^2} \leq \left(\frac{1-g^{\star}c}{g^{\star}\left(g^{\star}-1\right)\left(1-c\right)}\right)v_{1,k}(D) \mid D\right\}.
\end{aligned}$$

Now, observe that we can write

$$
\begin{aligned}
&P\left\{\min_{2\le g'\le g^\star(1-c)} \frac{g' A_{g'} - B_{g'}}{\left(g^\star(1-c)-1\right)\left(g^\star(1-c)\right)^2} \le \left(\frac{1-g^\star c}{g^\star\left(g^\star-1\right)(1-c)}\right) v_{1,k}(D) \mid D\right\}\\
&= P\left\{\max_{2\le g'\le g^\star(1-c)} \frac{B_{g'} - g' A_{g'}}{\left(g^\star(1-c)-1\right)\left(g^\star(1-c)\right)^2} \ge \left(\frac{g^\star c-1}{g^\star\left(g^\star-1\right)(1-c)}\right) v_{1,k}(D) \mid D\right\} .
\end{aligned}
$$

We bound this term by combining the argument used to establish Lemma A.4.5 with an application of Theorem A.5.2. To this end, observe that

$$
\begin{aligned}
&P\left\{\max_{2\le g'\le g} \frac{B_{g'} - g' A_{g'}}{(g-1)\,g^2} \le t \mid D\right\}\\
&\ge P\left\{\max_{2\le g'\le g} \frac{-1}{g-1}\frac{g'}{g^2} A_{g'} \le \frac{1}{1+\sqrt{g}} t \mid D\right\} + P\left\{\max_{2\le g'\le g} \frac{1}{g-1}\frac{1}{g^2} B_{g'} \le \frac{\sqrt{g}}{1+\sqrt{g}} t \mid D\right\} - 1\\
&\ge P\left\{\max_{2\le g'\le g} \frac{g}{g-1}\frac{-1}{g^2} A_{g'} \le \frac{\sqrt{g}}{1+\sqrt{g}} t \mid D\right\} + P\left\{\max_{2\le g'\le g} \frac{1}{g-1}\frac{1}{g^2} B_{g'} \le \frac{1}{1+\sqrt{g}} t \mid D\right\} - 1\\
&\ge P\left\{\max_{2\le g'\le g} -A_{g'} \le g^2 \frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}} t \mid D\right\} + P\left\{\max_{2\le g'\le g} |B_{g'}| \le g^2 \frac{g-1}{1+\sqrt{g}} t \mid D\right\} - 1
\end{aligned}
\tag{A.5.47}
$$

for any $t > 0$ and any positive integer D. Observe that the partial sums A_g and B_g are both martingale sequences. As Assumption 1.4.1 holds and the eighth-order split stability $\zeta^{(8)}$ is finite, though inspection of the proof of Lemma A.4.2, Lemma A.5.4 implies that implies that

$$
\begin{aligned}
&\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\theta A_g\right) \mid D\right] - \theta g^2 \frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}} t\right)\\
&= \inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\frac{\theta}{g^2} A_g\right) \mid D\right] - \theta \frac{1}{\sqrt{g}}\frac{g-1}{1+\sqrt{g}} t\right)\\
&\lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3 t^2}{\Gamma^{(2)}_{k,\varphi,b}}
\end{aligned}
\tag{A.5.48}
$$

and

$$
\begin{aligned}
&\inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\theta B_g\right) \mid D\right] - \theta g^2 \frac{g-1}{1+\sqrt{g}} t\right)\\
&= \inf_{\theta>0}\left(\log \mathbb{E}\left[\exp\left(\frac{\theta}{g^2} B_g\right) \mid D\right] - \theta \frac{g-1}{1+\sqrt{g}} t\right)\\
&\lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4}\frac{g^3 t^2}{\Gamma^{(2)}_{k,\varphi,b}}
\end{aligned}
\tag{A.5.49}
$$

each with probability greater than $1-\delta$, where we have used the facts that $4g \leq \left(1+\sqrt{g}\right)^2$ for all $g \geq 1$ and $(1/4)\, g^2 \geq (g-1)^2$ for all $g \geq 2$. Hence, by Theorem A.5.2, and plugging (A.5.48) and (A.5.49) into (A.5.47), we find that

$$\log \frac{1}{4} P \left\{ \max_{2 \leq g' \leq g} \frac{B_{g'} - g' A_{g'}}{(g'-1)\,(g')^2} \geq t \mid D \right\} \lesssim -\frac{\delta}{(2-\varphi k-\varphi)^4} \frac{g^3 t^2}{\Gamma^{(2)}_{k,\varphi,b}}$$

with probability greater than $1-\delta$. Consequently, by the definition of $g^\star$, we find that

$$\begin{aligned}
&\log \frac{1}{4} P \left\{ g^\star - \hat{g} \geq c \mid D \right\} \\
&= \log \frac{1}{4} P \left\{ \max_{2 \leq g' \leq g^\star(1-c)} \frac{B_{g'} - g' A_{g'}}{(g^\star(1-c)-1)\,(g^\star(1-c))^2} \geq \left(\frac{g^\star c - 1}{g^\star (g^\star-1)(1-c)} \right) v_{1,k}(D) \mid D \right\} \\
&\lesssim -\frac{\delta (v_{1,k}(D))^2}{(2-\varphi k-\varphi)^4} \frac{g^\star (1-c)(g^\star c-1)^2}{(g^\star-1)^2 \Gamma^{(2)}_{k,\varphi,b}} \\
&\lesssim -\frac{\delta (v_{1,k}(D))^2}{(2-\varphi k-\varphi)^4} \frac{g^\star c^2}{\Gamma^{(2)}_{k,\varphi,b}} \\
&\lesssim -\frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z^2_{1-\beta/2} c^2}{\xi^2 \Gamma^{(2)}_{k,\varphi,b}}
\end{aligned}$$

where in the second to last inequality we have used the facts that $\frac{1}{2}x \leq x-1$ for $x \geq 2$ and $g^\star c \geq 2$. Thus, putting the pieces together, we find that

$$\log \frac{1}{8} P \left\{ |\hat{g} - g^\star| > c g^\star \mid D \right\} \lesssim -\frac{\delta (v_{1,k}(D))^3}{(2-\varphi k-\varphi)^4} \frac{z^2_{1-\beta/2} c^2}{\xi^2 \Gamma^{(2)}_{k,\varphi,b}}$$

with probability greater than $1-\delta$, as required. ■

Appendix B

Supplemental Appendices for Chapter 2

with Vasilis Syrgkanis

B.1. Proofs for Results Supporting Theorem 2.1.1

B.1.1 Proof of Lemma 2.2.3

B.1.1.1 Part (i)

The result follows from an iterative application of the following symmetrization inequality.

Lemma B.1.1 (Vershynin (2018), Exercise 6.4.5)**.** *Let $Z_1, \ldots, Z_n$ be independent, mean-zero, real-valued random variables. If $V_1, \ldots, V_n$ denote an independent sequence of i.i.d. Rademacher random variables, then the inequality*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} Z_i\right|^q\right]^{1/q} \leq 2\mathbb{E}\left[\left|\sum_{i=1}^{n} V_i Z_i\right|^q\right]^{1/q} \tag{B.1.1}$$

holds for each $q \geq 1$.

In particular, observe that we can write

$$\sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)}) = \sum_{i_1=1}^{\lfloor n/m \rfloor} Z_{i_1}^{(1)} \tag{B.1.2}$$

where the summands

$$Z_i^{(1)} = \sum_{i_2=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} g(D_i^{(1)}, D_{i_2}^{(2)}, \ldots, D_{i_m}^{(m)}) \,. \tag{B.1.3}$$

are independently distributed conditional on the replications $\mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}$. Moreover, the complete degeneracy condition (2.2.21) implies that

$$\begin{aligned} &\mathbb{E}[Z_i^{(1)} \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}] \\ &= \sum_{i_2=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} \mathbb{E}[g(D_i^{(1)}, D_{i_2}^{(2)}, \ldots, D_{i_m}^{(m)}) \mid D_{i_2}^{(2)}, \ldots, D_{i_m}^{(m)}] = 0 \,. \end{aligned} \tag{B.1.4}$$

Thus, by applying Lemma B.1.1 conditional on $\mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}$ and taking expectations, we obtain

$$\begin{aligned} &\mathbb{E}\left[\left|\frac{1}{\lfloor n/m \rfloor^m} \sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)})\right|^q\right]^{1/q} \\ &\leq \frac{2}{\lfloor n/m \rfloor^m} \mathbb{E}\left[\left|\sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} V_{i_1}^{(1)} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)})\right|^q\right]^{1/q} \,. \end{aligned} \tag{B.1.5}$$

Hence, by iterating the same argument $m-1$ more times, first conditionally on all blocks except $\mathbf{D}_n^{(2)}$, then on all blocks except $\mathbf{D}_n^{(3)}$, and so on, we can conclude that

$$\begin{aligned} &\mathbb{E}\left[\left|\frac{1}{\lfloor n/m \rfloor^m} \sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)})\right|^q\right]^{1/q} \\ &\leq \frac{2^m}{\lfloor n/m \rfloor^m} \mathbb{E}\left[\left|\sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} V_{i_1}^{(1)} \cdots V_{i_m}^{(m)} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)})\right|^q\right]^{1/q} \,, \end{aligned} \tag{B.1.6}$$

as required. ■

B.1.1.2 Part (ii)

The result can be obtained as a specialization of the following standard moment inequality for homogeneous Rademacher chaos, often referred to as the Bonami inequality.

Lemma B.1.2 (Theorem 3.2.2, De la Pena and Giné (1999)). *Fix a collection of real-valued quantities* $\{z_{\mathsf{s}} : \mathsf{s} \in \mathcal{S}_{n,b}\}$ *and let* $V_1, \ldots, V_n$ *denote an independent collection of Rademacher random variables. Consider the homogeneous Rademacher chaos of order* b*, given by*

$$Z_b = \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} V_{\mathsf{s}} z_{\mathsf{s}} \;, \tag{B.1.7}$$

where $V_{\mathsf{s}} = \Pi_{i \in \mathsf{s}} V_i$ *for each subset* s *in* $[n]$*. The moment inequality*

$$\mathbb{E}\left[|Z_b|^q\right] \leq q^{bq/2} (\Delta_b)^{q/2} \;, \quad \textit{where} \quad \Delta_b = \sum_{s \in \mathcal{S}_{n,b}} (z_{\mathsf{s}})^2 \;, \tag{B.1.8}$$

holds for every $q > 2$*.*

To apply Lemma B.1.2, we require some further notation. In particular, let $\overline{n} = \lfloor n/m \rfloor \cdot m$, define the injection

$$h : [m] \times [\lfloor n/m \rfloor] \to [\overline{n}] \quad \text{by} \quad h(\ell, i) = (\ell - 1)\lfloor n/m \rfloor + i \;, \tag{B.1.9}$$

and enumerate the sets

$$\mathsf{s}(i_1, \ldots, i_m) = \{h(1, i_1), \ldots, h(m, i_m)\} \in \mathcal{S}_{\overline{n},m} \tag{B.1.10}$$

for each $(i_1, \ldots, i_m)$ in $[\lfloor n/m \rfloor]^m$. Observe that the mapping $(i_1, \ldots, i_m) \mapsto \mathsf{s}(i_1, \ldots, i_m)$ is a bijection. Thus, we can define the array

$$z_{\mathsf{s}} = \begin{cases} g_{i_1,\ldots,i_m} \;, & \text{if } \mathsf{s} = \mathsf{s}(i_1, \ldots, i_m) \text{ for some } (i_1, \ldots, i_m) \in [\lfloor n/m \rfloor]^m \;, \\ 0 \;, & \text{otherwise} \;, \end{cases} \tag{B.1.11}$$

and the sequence of i.i.d. Rademacher variables

$$V_{h(\ell,i)} = V_i^{(\ell)} \;, \tag{B.1.12}$$

for each $\ell \in [m]$ and $i \in [\lfloor n/m \rfloor]$, such that

$$\sum_{i_1=1}^{\lfloor n/m \rfloor} \cdots \sum_{i_m=1}^{\lfloor n/m \rfloor} V_{i_1}^{(1)} \cdots V_{i_m}^{(m)} g_{(i_1,\ldots,i_m)} = \sum_{\mathsf{s} \in \mathcal{S}_{\overline{n},m}} V_{\mathsf{s}} z_{\mathsf{s}} \tag{B.1.13}$$

almost surely. Therefore, Lemma B.1.2 implies that

$$\mathbb{E}\left[\left|\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}V_{i_1}^{(1)}\cdots V_{i_m}^{(m)}g_{(i_1,\ldots,i_m)}\right|^q\right]^{1/q}$$
$$=\mathbb{E}\left[\left|\sum_{\mathsf{s}\in\mathcal{S}_{\overline{n},m}}V_{\mathsf{s}}z_{\mathsf{s}}\right|^q\right]^{1/q}$$
$$\leq q^{m/2}\left(\sum_{\mathsf{s}\in\mathcal{S}_{\overline{n},m}}z_{\mathsf{s}}^2\right)^{1/2}\leq q^{m/2}\left(\sum_{i_1=1}^{\lfloor n/m\rfloor}\cdots\sum_{i_m=1}^{\lfloor n/m\rfloor}g_{(i_1,\ldots,i_m)}^2\right)^{1/2} \tag{B.1.14}$$

for each $q>2$, as required. ■

B.1.2 Proof of Lemma 2.2.4

B.1.2.1 Part (i)

It will suffice to establish the following more general statement. Fix integers $m\geq 1$ and $k\geq 1$ and let $g:\mathcal{D}^m\to\mathbb{R}_+$ be a symmetric, non-negative kernel. Define the quantity

$$A_{m,k}(g)=\frac{1}{k^m}\sum_{i_1=1}^{k}\cdots\sum_{i_m=1}^{k}g(D_{i_1}^{(1)},\ldots,D_{i_m}^{(m)})\,. \tag{B.1.15}$$

The inequality

$$\mathbb{E}\left[|A_{m,k}(g)|^q\right]^{1/q}\leq C^m\sum_{t=0}^{m}\left(\frac{q}{k}\right)^t\Gamma_{t,q}(g)$$

holds for each $q\geq 1\vee\log(k)$.

We proceed by induction on m. To verify the base case, associated with $m=1$, it suffices to show that

$$\mathbb{E}\left[|A_{1,k}(g)|^q\right]^{1/q}=\mathbb{E}\left[\left|\frac{1}{k}\sum_{i_1=1}^{k}g(D_{i_1}^{(1)})\right|^q\right]^{1/q}\lesssim\Gamma_{0,q}(g)+\frac{q}{k}\Gamma_{1,q}(g)$$
$$=\mathbb{E}[g(D_{i_1}^{(1)})]+\frac{q}{k}\mathbb{E}\left[|g(D_{i_1}^{(1)})|^q\right]^{1/q} \tag{B.1.16}$$

holds for each $q\geq\log(k)$. This bound follows immediately from the following specialization of Rosenthal's inequality for non-negative random variables.

Lemma B.1.3. *Let $Z_1, \ldots, Z_k$ be i.i.d. non-negative random variables. Then, the inequality*

$$\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k} Z_i\right|^q\right]^{1/q} \lesssim \mathbb{E}[Z_1] + \frac{q}{k}\mathbb{E}\left[|Z_1|^q\right]^{1/q} \tag{B.1.17}$$

holds for each $q \geq 1 \vee \log(k)$.

Proof. Rosenthal's inequality for non-negative random variables, stated as Theorem 15.10 in Boucheron et al. (2013), implies that

$$\mathbb{E}\left[\left|\sum_{i=1}^{k} Z_i\right|^q\right]^{1/q} \lesssim k\mathbb{E}[Z_1] + q\mathbb{E}\left[\max_{i\in[k]} |Z_i|^q\right]^{1/q} . \tag{B.1.18}$$

As $q \geq \log(k)$, we have that $k^{1/q} \leq e$, and therefore

$$\mathbb{E}\left[\max_{i\in[k]} |Z_i|^q\right]^{1/q} \leq k^{1/q}\mathbb{E}\left[|Z_i|^q\right]^{1/q} \leq e\mathbb{E}\left[|Z_i|^q\right]^{1/q} . \tag{B.1.19}$$

The result follows by plugging (B.1.19) into (B.1.18) and dividing by k. ■

Thus, it suffices to verify the inductive step. That is, assume that the bound (B.1.16) has been established for kernels of order $m-1$. Observe that we can write

$$A_{m,k}(g) = \frac{1}{k^m}\sum_{i_1=1}^{k}\cdots\sum_{i_m=1}^{k} g(D_{i_1}^{(1)}, \ldots, D_{i_m}^{(m)}) = \frac{1}{k}\sum_{i_1=1}^{k} Z_{i_1}^{(1)} \tag{B.1.20}$$

where

$$Z_{i_1}^{(1)} = \frac{1}{k^{m-1}}\sum_{i_2=1}^{k}\cdots\sum_{i_m=1}^{k} g(D_{i_1}^{(1)}, D_{i_2}^{(2)}, \ldots, D_{i_m}^{(m)}) . \tag{B.1.21}$$

Conditional on the replicates $\mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}$, the variables $Z_1^{(1)}, \ldots, Z_k^{(1)}$ are i.i.d. and non-negative. Therefore, Lemma B.1.3 implies that

$$\begin{aligned} &\mathbb{E}\left[|A_{m,k}(g)|^q \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}\right]^{1/q} \\ &\lesssim \mathbb{E}[Z_1^{(1)} \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}] + \frac{q}{k}\mathbb{E}\left[|Z_1^{(1)}|^q \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}\right]^{1/q} . \end{aligned} \tag{B.1.22}$$

Hence, Minkowski's inequality implies that

$$\mathbb{E}\left[|A_{m,k}(g)|^q\right]^{1/q} \lesssim \mathbb{E}\left[\mathbb{E}[Z_1^{(1)} \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}]^q\right]^{1/q} + \frac{q}{k}\mathbb{E}\left[|Z_1^{(1)}|^q\right]^{1/q} \tag{B.1.23}$$

as $q \geq 1$. Now, by defining the $m-1$ order kernels

$$g_0(d_2, \ldots, d_m) = \mathbb{E}[g(D_1, d_2, \ldots, d_m)] \quad \text{and} \quad g_d(d_2, \ldots, d_m) = g(d, d_2, \ldots, d_m) \tag{B.1.24}$$

we can identify

$$\mathbb{E}[Z_1^{(1)} \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}] = A_{m-1,k}(g_0) \quad \text{and} \quad Z_1^{(1)} = A_{m-1,k}(g_{D_{i_1}^{(1)}}) \;, \tag{B.1.25}$$

respectively. Hence, the inductive hypothesis implies that

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}[Z_1^{(1)} \mid \mathbf{D}_n^{(2)}, \ldots, \mathbf{D}_n^{(m)}]^q\right]^{1/q} &\leq C^{m-1} \sum_{t=0}^{m-1} \left(\frac{q}{k}\right)^t \Gamma_{t,q}(g_0) \quad \text{and} \\ \mathbb{E}\left[|Z_1^{(1)}|^q\right]^{1/q} &\leq C^{m-1} \sum_{t=0}^{m-1} \left(\frac{q}{k}\right)^t \mathbb{E}\left[\Gamma_{t,q}(g_{D_{i_1}^{(1)}})^q\right]^{1/q} \;. \end{aligned} \tag{B.1.26}$$

By recognizing

$$\begin{aligned} \Gamma_{t,q}(g_0) &= \mathbb{E}\left[\mathbb{E}[g(D_1, \ldots, D_m) \mid D_2, \ldots, D_{t+1}]^q\right]^{1/q} = \Gamma_{t,q}(g) \quad \text{and} \\ \mathbb{E}\left[\Gamma_{t,q}(g_{D_1^{(1)}})^q\right]^{1/q} &= \mathbb{E}\left[\mathbb{E}[g(D_1, \ldots, D_m) \mid D_1, \ldots, D_{t+1}]^q\right]^{1/q} = \Gamma_{t+1,q}(g) \end{aligned} \tag{B.1.27}$$

for each $0 \leq t \leq m-1$ and plugging the bounds (B.1.26) into (B.1.23), we find that

$$\begin{aligned} \mathbb{E}\left[|A_{m,k}(g)|^q\right]^{1/q} &\leq C^m \sum_{t=0}^{m-1} \left(\frac{q}{k}\right)^t \Gamma_{t,q}(g) + C^m \frac{q}{k} \sum_{t=0}^{m-1} \left(\frac{q}{k}\right)^t \Gamma_{t+1,q}(g) \\ &\leq C^m \sum_{t=0}^{m} \left(\frac{q}{k}\right)^t \Gamma_{t,q}(g) \;, \end{aligned} \tag{B.1.28}$$

as required. ■

B.1.2.2 Part (ii)

We begin by stating two facts that will be applied at several points throughout the ensuing argument. First, we record the following property of sub-exponential random variables.

Lemma B.1.4 (Vershynin (2018), Proposition 2.7.1)**.**
Let Z be a real-valued random variable. Then the moment inequality

$$\mathbb{E}[|Z|^q]^{1/q} \leq Cq\|Z\|_{\psi_1} \tag{B.1.29}$$

holds for each $q \geq 1$.

Second, we note that if Z' is an independent copy of Z, then the triangle inequality implies that

$$\|Z - Z'\|_{\psi_1} \leq \|Z\|_{\psi_1} + \|Z'\|_{\psi_1} = 2\|Z\|_{\psi_1} \ . \tag{B.1.30}$$

Thus, by Jensen's inequality, we obtain the bound

$$\|f^{(t)}(D_{[t]})\|_{\psi_1} \leq \|\nabla_1 \cdots \nabla_t f(D_{[b]})\| \leq 2^t \phi \tag{B.1.31}$$

for each integer $0 \leq t \leq b$.

Now, our interest is in the quantity

$$\Gamma_{t,q}((f^{(m)})^2) = \mathbb{E}\left[\mathbb{E}[f^{(m)}(D_{[m]})^2 \mid D_1, \ldots, D_t]^q\right]^{1/q} \tag{B.1.32}$$

for each integer $0 \leq t \leq m$. We first consider the case $0 \leq t \leq m-1$. To this end, for each deterministic collection $(d_1, \ldots, d_t)$ in $\mathcal{D}^t$, define the symmetric, order $b - t$ kernel

$$\begin{aligned} &g_{d_1,\ldots,d_t}(d_{t+1}, \ldots, d_b) \\ &= (-1)^t \mathbb{E}\left[\nabla_1 \cdots \nabla_t f(D_{[b]}) \mid D_1 = d_1, \ldots, D_t = d_t, D_{t+1} = d_{t+1}, \ldots, D_b = d_b\right] \end{aligned} \tag{B.1.33}$$

and let $g^{(1)}_{d_1,\ldots,d_t}, \ldots, g^{(b-t)}_{d_1,\ldots,d_t}$ denote its Hoeffding projections, as defined in Part (i) of Lemma C.4.1. By Part (i) of Lemma C.4.1, these projections are uncorrelated, and so the decomposition

$$\mathbb{E}\left[g_{d_1,\ldots,d_t}(D_{t+1}, \ldots, D_b)^2\right] = \mathbb{E}[g_{d_1,\ldots,d_t}(D_{t+1}, \ldots, D_{t+\ell})]^2$$

$$+\sum_{\ell=1}^{b-t}\binom{b-t}{\ell}\mathbb{E}\left[\left(g^{(\ell)}_{d_1,\dots,d_t}(D_{t+1},\dots,D_{t+\ell})\right)^2\right] \quad \text{(B.1.34)}$$

holds. Now, observe that

$$\begin{aligned}
&g^{(m-t)}_{d_1,\dots,d_t}(d_{t+1},\dots,d_m) && \text{(B.1.35)}\\
&= (-1)^{m-t}\mathbb{E}\left[\nabla_{t+1}\cdots\nabla_m g_{d_1,\dots,d_t}(D_{t+1},\dots,D_b) \mid D_{t+1}=d_{t+1},\dots,D_m=d_m\right]\\
&= (-1)^m\mathbb{E}\left[\nabla_1\cdots\nabla_m f(D_{[b]}) \mid D_1=d_1,\dots,D_m=d_m\right]\\
&= f^{(m)}(d_1,\dots,d_m)\ , && \text{(B.1.36)}
\end{aligned}$$

where the first and last equalities follow by definition and the second equality follows from the fact that the stochastic difference operators act on distinct coordinates, and thereby commute. As a consequence, the decomposition (B.1.34) implies that

$$\begin{aligned}
&\binom{b-t}{m-t}\mathbb{E}\left[f^{(m)}(D_1,\dots,D_t,D_{t+1},\dots,D_m)^2 \mid D_1,\dots,D_t\right]\\
&\le \mathbb{E}\left[g_{D_1,\dots,D_t}(D_{t+1},\dots,D_b)^2 \mid D_1,\dots,D_t\right] && \text{(B.1.37)}
\end{aligned}$$

holds almost surely. Therefore, we obtain the bound

$$\Gamma_{t,q}((f^{(m)})^2) \le \binom{b-t}{m-t}^{-1}\mathbb{E}\left[\mathbb{E}\left[g_{D_1,\dots,D_t}(D_{t+1},\dots,D_b)^2 \mid D_1,\dots,D_t\right]^q\right]^{1/q}\ . \quad \text{(B.1.38)}$$

Now, Jensen's inequality for conditional expectation and Lemma B.1.4 imply that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[g_{D_1,\dots,D_t}(D_{t+1},\dots,D_b)^2 \mid D_1,\dots,D_t\right]^q\right]^{1/q}\\
&\le \mathbb{E}\left[|g_{D_1,\dots,D_t}(D_{t+1},\dots,D_b)|^{2q}\right]^{1/q}\\
&\le (Cq)^2\|g_{D_1,\dots,D_t}(D_{t+1},\dots,D_b)\|^2_{\psi_1} \le (Cq)^2\phi^2 4^t\ , && \text{(B.1.39)}
\end{aligned}$$

where the last step follows from (B.1.31). Hence, as

$$\binom{b-t}{m-t}^{-1} \le \binom{b}{m}^{-1}\left(\frac{eb}{m}\right)^t\ , \quad \text{(B.1.40)}$$

we can conclude that

$$\Gamma_{t,q}\big((f^{(m)})^2\big) \leq \phi^2 (Cq)^2 C^t \binom{b}{m}^{-1} \left(\frac{b}{m}\right)^t \tag{B.1.41}$$

for each $0 \leq t \leq m-1$.

It remains to consider the case $t = m$. In this case, we have that

$$\Gamma_{m,q}\big((f^{(m)})^2\big) = \mathbb{E}[|f^{(m)}(D_{[m]})|^{2q}]^{1/q} \leq \phi^2 (Cq)^2 4^m \,. \tag{B.1.42}$$

by Lemma B.1.4 and (B.1.31). The bound $\binom{b}{m} \leq (eb/m)^m$ then implies that

$$4^m \leq C^m \binom{b}{m}^{-1} \left(\frac{b}{m}\right)^m \,, \tag{B.1.43}$$

completing the proof. ■

B.1.3 Proof of Corollary 2.1.1

The result follows by considering each of the terms in the decomposition

$$\begin{aligned} &U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] \\ &= \frac{b}{n}\sum_{i=1}^n u^{(1)}(D_i) + \left(U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n}\sum_{i=1}^n u^{(1)}(D_i)\right) \,. \end{aligned} \tag{B.1.44}$$

The first quantity can be handled through the application of the following standard Bernstein-type bound, stated, e.g., in Song et al. (2019).

Lemma B.1.5 (Lemma A.2, Song et al. (2019))**.**
Let $Z_1, \ldots, Z_n$ be independent, centered, random vectors in $\mathbb{R}^d$. Define the quantity

$$\sigma^2 = \max_{j \in [d]} \sum_{i=1}^n \mathbb{E}\left[Z_{i,j}^2\right] \tag{B.1.45}$$

and assume that $\|Z_{ij}\|_{\psi_1} \leq \phi$ for all $i \in [n]$ and $j \in [d]$. The inequality

$$P\left\{\Big\|\sum_{i=1}^n Z_i\Big\|_\infty \geq C\left(\sigma \log^{1/2}(dg) + \phi \log(nd)\left(\log(nd) + \log(g)\right)\right)\right\} \lesssim \frac{1}{g}$$

holds for any constant $g > 0$.

In particular, observe that

$$\max_{j\in[d]}\sum_{i=1}^{n}\frac{b^2}{n^2}\mathbb{E}[u^{(1)}(D_i)^2]=\frac{b^2}{n}\overline{\sigma}_b^2 \quad \text{and} \quad \|\frac{b}{n}u^{(1)}(D_i)\|_{\psi_1}\leq\frac{b}{n}\phi \tag{B.1.46}$$

Thus, Lemma B.1.5 implies that

$$\left\|\frac{b}{n}\sum_{i=1}^{n}u^{(1)}(D_i)\right\|_{\infty}\lesssim\frac{b\overline{\sigma}_b}{n^{1/2}}\log^{1/2}(nd)+\phi\frac{b}{n}\log^2(nd) \tag{B.1.47}$$

with probability greater than $1-C/nd$. In turn, Theorem 2.1.1 implies that

$$\left\|U_{n,b}(u)-\mathbb{E}[u(D_{[b]})]-\frac{b}{n}\sum_{i=1}^{n}u^{(1)}(D_i)\right\|_{\infty}\lesssim\phi\frac{b}{n}\log^2(nd) \tag{B.1.48}$$

with probability greater than $1-C/nd$. Hence, we can conclude that

$$\left\|U_{n,b}(u)-\mathbb{E}[u(D_{[b]})]\right\|_{\infty}\lesssim\sqrt{\frac{b^2\overline{\sigma}_b^2\log(nd)}{n}}+\phi\frac{b}{n}\log^2(nd) \tag{B.1.49}$$

with probability greater than $1-C/nd$, by a union bound. ■

B.1.4 Proof of Corollary 2.1.2

We give the details of the proof of the upper bound encoded in (2.1.9). The lower bound will follow from an analogous argument. Fix a rectangle $\mathsf{R}=[a_\ell,a_u]$ in $\mathcal{R}$, where a_ℓ and a_u are vectors in $\mathbb{R}^d$ with $a_\ell\leq a_u$, interpreted component-wise. Define the enlarged rectangle $\mathsf{R}_t=[a_\ell-t\mathbf{1}_d,a_u+t\mathbf{1}_d]$ for each $t>0$ as well as the normalized Hájek projection

$$\hat{u}^{(1)}(D_i)=\Sigma_b^{-1/2}u^{(1)}(D_i)\ . \tag{B.1.50}$$

Observe that the decomposition

$$\begin{aligned}\sqrt{\frac{n}{b^2}}\Sigma_b^{-1/2}(U_{n,b}(u)-\mathbb{E}[u(D_{[b]})])&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{u}^{(1)}(D_i)\\&+\sqrt{\frac{n}{b^2}}\Sigma_b^{-1/2}\left(U_{n,b}(u)-\mathbb{E}[u(D_{[b]})]-\frac{b}{n}\sum_{i=1}^{n}u^{(1)}(D_i)\right)\end{aligned} \tag{B.1.51}$$

implies the upper bound

$$P\left\{\sqrt{\frac{n}{b^2}}\Sigma_b^{-1/2}(U_{n,b}(u)-\mathbb{E}[u(D_{[b]})])\in \mathsf{R}\right\}-P\left\{\Sigma_b^{-1/2}Z\in \mathsf{R}\right\} \tag{B.1.52}$$

$$\leq \left|P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{u}^{(1)}(D_i)\in \mathsf{R}_t\right\}-P\left\{\Sigma_b^{-1/2}Z\in \mathsf{R}_t\right\}\right| \tag{B.1.53}$$

$$+\left|P\left\{\Sigma_b^{-1/2}Z\in \mathsf{R}_t\right\}-P\left\{\Sigma_b^{-1/2}Z\in \mathsf{R}\right\}\right| \tag{B.1.54}$$

$$+P\left\{\|\sqrt{\frac{n}{b^2}}\Sigma_b^{-1/2}\left(U_{n,b}(u)-\mathbb{E}[u(D_{[b]})]-\frac{b}{n}\sum_{i=1}^n u^{(1)}(D_i)\right)\|_\infty > t\right\}. \tag{B.1.55}$$

We bound the three terms appearing in (B.1.52) in succession.

We begin by bounding the normal approximation term (B.1.53) through the application of the following quantitative central limit theorem, stated as Theorem 2.1 of Chernozhukov et al. (2022).

Lemma B.1.6 (Theorem 2.1, Chernozhukov et al., 2022)**.** *Let $X_1,\dots,X_n$ be a collection of independent, centered, random vectors in $\mathbb{R}^d$ and let Z be a centered Gaussian random vector with covariance matrix*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_iX_i^\top\right]. \tag{B.1.56}$$

If there exist absolute constants c, C_1, and φ such that the bounds

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_{i,j}^2\right]\geq c\,,\quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_{i,j}^4\right]\leq C_1\varphi^2\,,\quad \text{and}\quad \|X_{i,j}\|_{\psi_1}\leq\varphi \tag{B.1.57}$$

hold, then the inequality

$$\sup_{\mathsf{R}\in\mathcal{R}}\left|P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\in \mathsf{R}\right\}-P\left\{Z\in \mathsf{R}\right\}\right|\leq C_2\left(\frac{\varphi^2\log^5(dn)}{n}\right)^{1/4} \tag{B.1.58}$$

holds for some constant C_2 that depends only on c and C_1.

In particular, observe that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(\hat{u}^{(1)}(D_i)\right)^2\right]=\frac{\mathbb{E}\left[\left(u^{(1)}(D_i)\right)^2\right]}{\sigma_{b,j}^2}=1 \tag{B.1.59}$$

for each coordinate j. Moreover, it holds that

$$\|\hat{u}^{(1)}(x^{(j)}, D_i)\|_{\psi_1} \leq (\phi/\underline{\sigma}_b) \tag{B.1.60}$$

for each coordinate j by the boundedness condition (2.1.8). Similarly, it holds that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{u}^{(1)}(x^{(j)}, D_i)\right)^4\right] \leq (\phi/\underline{\sigma}_b)^2 \tag{B.1.61}$$

for each coordinate j, again by boundedness. Consequently, as

$$\mathrm{Var}(\hat{u}^{(1)}(D_i)) = \Sigma_b^{-1/2}\,\mathrm{Var}\,(Z)\,\Sigma_b^{-1/2}, \tag{B.1.62}$$

Lemma B.1.6 implies that the bound

$$\left|P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{u}^{(1)}(D_i) \in \mathsf{R}_t\right\} - P\left\{\Sigma_b^{-1/2} Z \in \mathsf{R}_t\right\}\right| \lesssim \left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n}\right)^{1/4} \tag{B.1.63}$$

holds.

Next, to bound the difference in the Gaussian probabilities (B.1.54), we apply the following anti-concentration inequality, stated in Chernozhukov et al. (2017) and often referred to as Nazarov's inequality.

Lemma B.1.7 (Theorem 1, Chernozhukov et al., 2017)**.** *Let $Z = (Z_j)_{j=1}^d$ be a centered Gaussian random vector in $\mathbb{R}^d$ such that $\mathbb{E}[Z_j^2] \geq c$ for all j in $[d]$ and some constant c. For every $z \in \mathbb{R}^d$ and $t > 0$, the inequality*

$$P\{Z \leq z + t\} - P\{Z \leq z\} \lesssim \frac{t}{c}\sqrt{\log d}$$

holds.

In particular, Lemma B.1.7 gives

$$\left|P\left\{\Sigma_b^{-1/2} Z \in \mathsf{R}_t\right\} - P\left\{\Sigma_b^{-1/2} Z \in \mathsf{R}\right\}\right| \lesssim t\sqrt{\log(d)} \tag{B.1.64}$$

for all $t > 0$. By choosing

$$t = C\frac{1}{\sqrt{\log(dn)}}\left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n}\right)^{1/4} \tag{B.1.65}$$

we find that the bound

$$\left| P\left\{ \Sigma_b^{-1/2} Z \in \mathsf{R}_t \right\} - P\left\{ \Sigma_b^{-1/2} Z \in \mathsf{R} \right\} \right| \lesssim \left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n} \right)^{1/4} \tag{B.1.66}$$

holds.

Finally, we bound the term (B.1.55). Observe that it is without loss of generality to assume that

$$\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n} < 1 \tag{B.1.67}$$

as otherwise the bound (2.1.9) holds vacuously. Now, observe that Theorem 2.1.1 implies that

$$\left\| U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n} \sum_{i=1}^{n} u^{(1)}(D_i) \right\|_\infty \lesssim \phi \frac{b}{n} \log^2(dn) \tag{B.1.68}$$

with probability at least $1 - C/nd$. Consequently, it holds that

$$\begin{aligned} &\left\| \sqrt{\frac{n}{b^2}} \Sigma_b^{-1/2} \left(U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n} \sum_{i=1}^{n} u^{(1)}(D_i) \right) \right\|_\infty \\ &\lesssim \phi \frac{1}{\sqrt{n}\underline{\sigma}_b} \log^2(dn) = \frac{1}{\sqrt{\log(dn)}} \left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n} \right)^{1/2} \end{aligned} \tag{B.1.69}$$

with probability at least $1 - C/nd$, where the final inequality follows from (B.1.67).

Hence, by plugging the bounds (B.1.63), (B.1.66), and (B.1.69) into (B.1.52), we find that

$$\begin{aligned} &P\left\{ \sqrt{\frac{n}{b^2}} \Sigma_b^{-1/2} (U_{n,b}(u) - \mathbb{E}[u(D_{[b]})]) \in \mathsf{R} \right\} - P\left\{ \Sigma_b^{-1/2} Z \in \mathsf{R} \right\} \\ &\lesssim \left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n} \right)^{1/4} + \frac{1}{nd} \lesssim \left(\frac{\phi^2 \log^5(dn)}{\underline{\sigma}_b^2 n} \right)^{1/4}, \end{aligned} \tag{B.1.70}$$

as required. ■

B.1.5 Proof of Remark 2.2.1

Assume that the components of each observation $D_i = (D_{i,j})_{j=1}^d$ are independent and have a standard Gaussian distribution. Consider the kernel

$$u_j(d_1, \ldots, d_b) = \frac{\phi}{16b} \sum_{1 \le r < s \le b} d_{r,j} d_{s,j} \tag{B.1.71}$$

where $d_k = (d_{k,j})_{j=1}^d$ in $\mathbb{R}^d$. We claim that the bound

$$\|u_j(D_{[b]})\|_{\psi_1} \le \phi \tag{B.1.72}$$

and the lower bound

$$P\left\{ \left\| U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n} \sum_{i=1}^n u^{(1)}(D_i) \right\|_\infty \ge c \frac{\phi b \log(nd)}{n} \right\} \ge \frac{c}{nd} \tag{B.1.73}$$

hold for this choice.

We begin by verifying the bound (B.1.72). Consider the quantity

$$Q_j = \frac{1}{b} \sum_{1 \le r < s \le b} D_{r,j} D_{s,j} \; . \tag{B.1.74}$$

Let $A_b \in \mathbb{R}^{b \times b}$ be the symmetric matrix with zeros on the diagonal and off-diagonal entries equal to $1/(2b)$. The eigenvalues of A_b are given by

$$\lambda_1 = \frac{b-1}{2b} \quad \text{and} \quad \lambda_2 = \cdots = \lambda_b = -\frac{1}{2b} \tag{B.1.75}$$

where we note that

$$\sum_{\ell=1}^b |\lambda_\ell| = \frac{b-1}{b} \le 1 \quad \text{and} \quad \sum_{\ell=1}^b \lambda_\ell = 0 \; . \tag{B.1.76}$$

Observe that we can write

$$Q_j = D_j^\top A_b D_j \; . \tag{B.1.77}$$

Thus, by the orthogonal invariance of the standard Gaussian law, there exist independent standard Gaussian replicates $G_{1,j}, \ldots, G_{b,j}$ such that

$$Q_j = \sum_{\ell=1}^{b} \lambda_\ell G_\ell^2 = \sum_{\ell=1}^{b} \lambda_\ell (G_\ell^2 - 1) \,, \tag{B.1.78}$$

where the second equality follows from (B.1.76). Consequently, as each replicate G_j satisfies the bound

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\frac{|G_j^2 - 1|}{16}\right)\right] \leq \exp(1/16)\mathbb{E}[\exp(G_j^2/16)] \\ &\leq \exp(1/16)(1-1/8)^{-1/2} < 2 \,, \end{aligned} \tag{B.1.79}$$

we can evaluate

$$\begin{aligned} \mathbb{E}\left[\exp\left(\frac{|Q_j|}{16}\right)\right] &\leq \mathbb{E}\left[\exp\left(\frac{\sum_{\ell=1}^{b} |\lambda_\ell||G_\ell^2 - 1|}{16}\right)\right] \\ &= \prod_{\ell=1}^{b} \mathbb{E}\left[\exp\left(\frac{|G_\ell^2 - 1|}{16}\right)\right]^{|\lambda_\ell|} \leq 2^{\sum_{\ell=1}^{b} |\lambda_\ell|} \leq 2 \,, \end{aligned} \tag{B.1.80}$$

where the final equality follows from (B.1.76). Hence, the bound

$$\|Q_j\|_{\psi_1} \leq 16 \quad \text{and thereby} \quad \|u_j(D_{[b]})\|_{\psi_1} \leq \phi \tag{B.1.81}$$

hold by the definition of the ψ_1-Orlicz norm.

Next, we verify the bound (B.1.73). First, observe that

$$\mathbb{E}[u_j(D_{[b]})] = 0 \tag{B.1.82}$$

and that

$$\mathbb{E}[u_j(D_{[b]}) \mid D_1] = \frac{\phi}{16b}\left(\sum_{\ell=2}^{b} D_{1,j}\mathbb{E}[D_{\ell,j}] + \sum_{2 \leq r < s \leq b} \mathbb{E}[D_{r,j}D_{s,j}]\right) = 0 \,. \tag{B.1.83}$$

Thus, the Hájek projection is given by

$$u^{(1)}(D_1) = \mathbb{E}[u(D_{[b]}) \mid D_1] - \mathbb{E}[u(D_{[b]})] = 0 \tag{B.1.84}$$

almost surely. As a consequence, it holds that

$$\left\| U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i) \right\|_{\infty} = \left\| U_{n,b} \right\|_{\infty} \tag{B.1.85}$$

almost surely.

Now, let $U_{n,b,j}(u)$ denote the jth component of $U_{n,b}(u)$. Observe that, as each unordered pair $\{i,k\}$ appears in exactly $\binom{n-2}{b-2}$ subsets s in $\mathcal{S}_{n,b}$, we can evaluate

$$\begin{aligned} U_{n,b,j}(u) = \binom{n}{b}^{-1} \sum_{\mathsf{s}\in\mathcal{S}_{n,b}} u_j(D_{\mathsf{s}}) &= \frac{\phi}{16b}\binom{n}{b}^{-1}\binom{n-2}{b-2}\sum_{1\le i<k\le n} D_{i,j}D_{k,j} \\ &= \frac{\phi(b-1)}{16n(n-1)}\sum_{1\le i<k\le n} D_{i,j}D_{k,j} \ . \end{aligned} \tag{B.1.86}$$

By taking an orthonormal change of coordinates in $\mathbb{R}^n$ and identifying

$$Z_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{i,j} \sim \mathsf{N}(0,1) \tag{B.1.87}$$

we can see that there exists $W_j \sim \chi^2_{n-1}$ independent of Z_j such that

$$\sum_{i=1}^{n} D_{i,j}^2 = Z_j^2 + W_j \ . \tag{B.1.88}$$

Thus, we can write

$$\begin{aligned} U_{n,b,j}(u) = \frac{\phi(b-1)}{16n(n-1)}\sum_{1\le i<k\le n} D_{i,j}D_{k,j} &= \frac{\phi(b-1)}{32n(n-1)}\left(\left(\sum_{i=1}^{n} D_{i,j}\right)^2 - \sum_{i=1}^{n} D_{i,j}^2\right) \\ &= \frac{\phi(b-1)}{32n(n-1)}\left(nZ_j^2 - (Z_j^2 + W_j)\right) \\ &= \frac{\phi(b-1)}{32n(n-1)}\left((n-1)Z_j^2 - W_j\right) \ , \end{aligned} \tag{B.1.89}$$

almost surely. Thus, the representations (B.1.85) and (B.1.89) implies that, on the event

$$\mathcal{A} = \left\{ Z_1^2 \ge \log(end) + 2, W_1 \le 2(n-1) \right\} \ , \tag{B.1.90}$$

it holds that

$$\left\|U_{n,b}(u) - \mathbb{E}[u(D_{[b]})] - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(D_i)\right\|_\infty = \left\|U_{n,b}(u)\right\|_\infty \geq \phi\frac{b-1}{32n}\log(end)\ . \quad \text{(B.1.91)}$$

Now, observe that

$$P\{Z_j \geq x\} \geq cx^{-1}\exp(-x^2/2) \quad \text{and} \quad P\{W_j \leq 2(n-1)\} \geq 1/2\ , \qquad \text{(B.1.92)}$$

where the first inequality is a standard lower bound for Gaussian random variables and W_j follows from Markov's inequality. Hence, it holds that

$$\begin{aligned} P\{\mathcal{A}\} &\geq \frac{1}{2}P\{Z_1^2 \geq \log(end)+2\} \\ &\gtrsim (\log(end)+2)^{-1/2}(end)^{-1/2} \gtrsim (nd)^{-1}\ . \end{aligned} \qquad \text{(B.1.93)}$$

The claim is verified by combining (B.1.91) and (B.1.93). ■

B.2. An Abstract Bound on Coverage Error

In this Appendix, we give an abstract bound on the accuracy of the nominal coverage probability for the confidence region introduced in Definition 2.3.1. We make no use of the kernel structure expressed in (2.3.2). The proof is given in Appendix B.2.1. We require several mild smoothness restrictions on the moment function $M(\cdot;\theta,g)$. In contrast to the set of assumptions specified in Section 2.3.3, we do not require moment linearity. Instead, we impose the following generalization of Part (iii) of Assumption 2.3.3.

Assumption B.2.1 (Moment Restrictions)**.**
The moment function $M(x;\theta_0,g_0)$ is twice continuously differentiable in its second argument. Let

$$M^{(1)}(x;\theta,g) = \frac{\partial}{\partial\theta'}M(x;\theta',g)|_{\theta'=\theta} \quad \textit{and} \quad H(x;\theta,g) = \frac{\partial^2}{\partial^2\theta'}M(x;\theta',g)|_{\theta'=\theta} \quad \text{(B.2.1)}$$

denote the Jacobian and Hessian of $M(\cdot;\theta,g)$ in θ, respectively. The Jacobian $M^{(1)}(x;\theta,g)$ is uniformly Lipschitz in its second argument and bounded from below in the sense that

$$\sup_{P\in\mathbf{P}}\sup_{j\in[d]}\left|M^{(1)}(x^{(j)};\theta,g) - M^{(1)}(x^{(j)};\theta,g_0)\right| \lesssim \|g-g_0\|_{2,\infty} \quad \textit{and} \qquad \text{(B.2.2)}$$

$$\inf_{P\in\mathbf{P}}\inf_{j\in[d]}\left|M^{(1)}(x^{(j)};\theta,g)\right|\geq c \tag{B.2.3}$$

for each g and θ and some positive constant c. The Hessian $H(x;\theta,g)$ is uniformly bounded as x, θ, and g vary over their respective domains.

Moreover, we impose an analogous restriction on the centered empirical moment

$$\overline{M}_n(x;\theta,g)=M_n(x;\theta,g,\mathbf{D}_n)-\mathbb{E}\left[M_n(x;\theta,g,\mathbf{D}_n)\right]\ , \tag{B.2.4}$$

where we have made the dependence on $\mathbf{D}_n$ implicit to ease notation.

Assumption B.2.2 (Empirical Smoothness)**.** *The centered empirical moment* (B.2.4) *is twice continuously differentiable in its second argument. Let*

$$\overline{H}_n(x;\theta,g)=\frac{\partial^2}{\partial^2\theta'}\overline{M}_n(x;\theta',g_0)|_{\theta'=\theta} \tag{B.2.5}$$

denote the Hessian of $\overline{M}_n(\cdot;\theta,g)$ in θ, respectively. The Hessian $\overline{H}_n(x;\theta,g)$ is uniformly bounded almost surely as x, θ, and g vary over their respective domains.

Next, we impose a set of high-level restrictions on the structure of the empirical conditional moment (2.3.2). At times we refer to the normalized statistic

$$W_n(x)=-(M^{(1)}(x;g_0))^{-1}\overline{M}_n(x;\theta_0,g_0)\ . \tag{B.2.6}$$

First, we impose a condition that ensures that (B.2.6) is approximately linear. Fix a generic sequence $\overline{q}_{n,d}$.

Assumption B.2.3 (Approximate Linearity)**.**
There exists a function $\overline{u}(\cdot,\cdot)$, a constant $\varphi\geq 1$, and a real-valued sequence $\delta_{n,u}$ such that

$$\mathbb{E}\left[\overline{u}(x^{(j)},D_i)\right]=0\ ,\quad \|\overline{u}(x^{(j)},D_i)\|_{\psi_1}\leq\varphi\ , \tag{B.2.7}$$

and

$$\mathbb{E}\left[\overline{u}^4(x^{(j)},D_i)\right]\leq \mathrm{Var}(\overline{u}(x^{(j)},D_i))\varphi^2 \tag{B.2.8}$$

hold for all j in $[d]$ and P in $\mathbf{P}$. Define the moment $\lambda_j^2=\mathrm{Var}(\overline{u}(x^{(j)},D_i))$ and set $\underline{\lambda}^2=$

$\min_{j\in[d]} \lambda_j^2$. *The higher-order moment inequality*

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\left(W_n(x^{(j)}) - \frac{1}{n}\sum_{i=1}^{n}\overline{u}(x^{(j)}, D_i)\right)\right|^q\right]^{1/q} \leq q^2\delta_{n,u} \;. \tag{B.2.9}$$

holds for all j in $[d]$, $2 \vee 2\log(nd) \leq q \leq \overline{q}_{n,d}$, and uniformly over P in $\mathbf{P}$,

Remark B.2.1. Compare Assumption B.2.3 with the bound

$$\mathbb{E}\left[|U_{n,b}(u_j) - \mathbb{E}[u_j(D_{[b]})] - \frac{b}{n}\sum_{i=1}^{n} u_j^{(1)}(D_i)|^{2\log(nd)}\right]^{1/(2\log(dn))} \lesssim \phi\frac{b}{n}\log^2(dn) \tag{B.2.10}$$

obtained from Lemma 2.2.2. The bound (B.2.10) implies Theorem 2.1.1 by Markov's inequality. ■

Second, we impose several restrictions relating to the estimators $\hat{\theta}_n(\boldsymbol{x}^{(d)})$ and $\hat{g}_n$. A closely related collection of conditions is stated in Chernozhukov et al. (2018).

Assumption B.2.4 (Bias, Consistency, and Stochastic Equicontinuity)**.** *Recall the definition of the object $\underline{\lambda}^2$ introduced in Assumption B.2.3. Let*

$$\overline{M}_n^{(1)}(x;\theta,g) = \frac{\partial}{\partial\theta'}\overline{M}_n(x;\theta',g)|_{\theta'=\theta} \tag{B.2.11}$$

denote the Jacobian of the centered empirical moment $\overline{M}_n(x;\theta,g)$.
***(i)** Define the quantity*

$$\mathsf{Bias}_n(x;\theta,g) = M(x;\theta,g) - \mathbb{E}\left[M_n(x;\theta,g,\boldsymbol{D}_n)\right] \;. \tag{B.2.12}$$

There exists a sequence $\delta_{n,B}$ such that

$$\sup_{P\in\mathbf{P}}\sup_{g\in\mathcal{G}}\sqrt{\frac{n}{\underline{\lambda}^2}}\|\mathsf{Bias}_n(\boldsymbol{x}^{(d)};\theta(\boldsymbol{x}^{(d)}),g)\|_\infty \lesssim (1+\|\theta(\boldsymbol{x}^{(d)})\|_\infty)\delta_{n,B} \tag{B.2.13}$$

uniformly over any vector $\theta(\boldsymbol{x}^{(d)}) = \{\theta(x^{(j)})\}_{j=1}^d$.
***(ii)** There exist sequences $\delta_{n,m}$, $\delta_{n,g}$, and $\delta_{n,\theta}$ such that, uniformly over P in $\mathbf{P}$,*

$$\left(\frac{n}{\underline{\lambda}^2}\right)^{1/4}\mathbb{E}\left[\left|\overline{M}_n^{(1)}(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q} \leq q\delta_{n,m} \;, \tag{B.2.14}$$

$$\left(\frac{n}{\underline{\lambda}^2}\right)^{1/4}\mathbb{E}\left[\|\hat{g}_n - g_0\|_{2,\infty}^q\right]^{1/q} \leq q\delta_{n,g} \;, \quad \textit{and} \tag{B.2.15}$$

$$\left(\frac{n}{\underline{\lambda}^2}\right)^{1/4} \mathbb{E}\left[\left|\hat{\theta}_n(x^{(j)}) - \theta_0(x^{(j)})\right|^q\right]^{1/q} \le q\delta_{n,\theta} \tag{B.2.16}$$

holds for all j in $[d]$ and $2 \vee 2\log(nd) \le q \le \overline{q}_{n,d}$.

(iii) *There exist sequences $\delta_{n,S}$ and $\delta_{n,J}$ such that, uniformly over P in $\mathbf{P}$,*

$$\sqrt{\frac{n}{\underline{\lambda}^2}}\mathbb{E}\left[\left|\overline{M}_n(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n) - \overline{M}_n(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q} \le q\delta_{n,S} \tag{B.2.17}$$

and

$$\sqrt{\frac{n}{\underline{\lambda}^2}}\mathbb{E}\left[\left|\overline{M}_n^{(1)}(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n) - \overline{M}_n^{(1)}(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q} \le q\delta_{n,J}\ , \tag{B.2.18}$$

holds for all j in $[d]$ and $2 \vee 2\log(nd) \le q \le \overline{q}_{n,d}$.

The following theorem gives a non-asymptotic bound on the error in the nominal coverage probability of the confidence regions introduced in Definition 2.3.1.

Theorem B.2.1. *Collect the error sequences*

$$\delta_n = \delta_{n,g}^2 + \delta_{n,\theta}^2 + \delta_{n,m}^2 + \delta_{n,B} + \delta_{n,S} + \underline{\lambda}^{1/2}n^{-1/4}\delta_{n,\theta}\left(\delta_{n,B} + \delta_{n,J}\right) + \delta_{n,u}$$

and assume that $\delta_{\varepsilon n} \le C_\varepsilon \delta_n$ for any $0 < \varepsilon < 1$. Suppose that the Neyman orthogonal moment function $M(x;\theta_0,g_0)$ satisfies Assumption B.2.1 and Part (i) of *Assumption 2.3.3 and that the centered empirical moment function $\overline{M}_n(x;\theta_0,g_0)$ satisfies Assumption B.2.2. If Assumption B.2.3* and *Assumption B.2.4 hold, then the confidence region defined in Definition 2.3.1 satisfies*

$$\sup_{P\in\mathbf{P}}\left|P\left\{\theta_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)})\right\} - (1-\alpha)\right| \lesssim \left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n\log^3(dn) + \frac{1}{nd}\ . \tag{B.2.19}$$

B.2.1 Proof of Theorem B.2.1

Throughout, without loss, we will assume that

$$\left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n\log^3(dn) + \frac{1}{nd} \le c\ , \tag{B.2.20}$$

as otherwise the desired bound (B.2.19) is vacuous.

It will suffice to show that

$$\sup_{z\in\mathbb{R}} \left| P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n(\boldsymbol{x}^{(d)})\|_\infty < z\right\} - P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n^*(\boldsymbol{x}^{(d)})\|_\infty < z \mid \mathbf{D}_n\right\} \right|$$
$$\lesssim \left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^3(dn) + \frac{1}{nd} \tag{B.2.21}$$

with probability greater than $1 - C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn) + (nd)^{-1})$. To see this, let $\mathsf{cv}(\gamma)$ denote the $1-\gamma$ quantile of $\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n(\boldsymbol{x}^{(d)})\|_\infty$ for each γ in $(0,1)$ and fix

$$\beta_{n,d} = \left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^3(dn) + \frac{1}{nd}\ . \tag{B.2.22}$$

Observe that (B.2.21) implies that

$$P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n^*(\boldsymbol{x}^{(d)})\|_\infty < \mathsf{cv}(\alpha - \beta_{n,d}) \mid \mathbf{D}_n\right\}$$
$$\geq P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n(\boldsymbol{x}^{(d)})\|_\infty < \mathsf{cv}(\alpha - \beta_{n,d})\right\} - \beta_{n,d} \geq 1-\alpha \tag{B.2.23}$$

and

$$P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n^*(\boldsymbol{x}^{(d)})\|_\infty < \mathsf{cv}(\alpha + \beta_{n,d}) \mid \mathbf{D}_n\right\}$$
$$\leq P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n(\boldsymbol{x}^{(d)})\|_\infty < \mathsf{cv}(\alpha + \beta_{n,d})\right\} + \beta_{n,d} \leq 1-\alpha \tag{B.2.24}$$

each with probability greater than $1 - C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn) + (nd)^{-1})$. Thus, recalling that $\hat{\mathsf{cv}}(\alpha)$ denotes the $1-\alpha$ quantile of $\sqrt{n}\|\hat{\Lambda}_n^{-1/2}R_n^*(\boldsymbol{x}^{(d)})\|_\infty$ conditioned on the data $\mathbf{D}_n$, (B.2.23) and (B.2.24) imply that

$$P\left\{\mathsf{cv}(\alpha+\beta_{n,d}) < \hat{\mathsf{cv}}(\alpha) < \mathsf{cv}(\alpha - \beta_{n,d})\right\}$$
$$\gtrsim 1 - \left(\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}\right)^{1/2} - \frac{1}{nd} \geq 1 - \beta_{n,d}\ , \tag{B.2.25}$$

where the second inequality follows from the fact that the normalization (B.2.20) implies that

$$\left(\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}\right)^{1/2} \lesssim \left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4}\ . \tag{B.2.26}$$

Consequently, as we can write

$$P\left\{\boldsymbol{\theta}_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)}, \mathbf{D}_n)\right\} = P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2} R_n(\boldsymbol{x}^{(d)})\|_\infty \leq \hat{\mathsf{cv}}(\alpha)\right\} \tag{B.2.27}$$

by definition, the inequality (B.2.25) implies that

$$\begin{aligned} &P\left\{\boldsymbol{\theta}_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)}, \mathbf{D}_n)\right\} \\ &\leq P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2} R_n(\boldsymbol{x}^{(d)})\|_\infty \leq \mathsf{cv}(\alpha - \beta_{n,d})\right\} + \beta_{n,d} \lesssim 1 - \alpha + \beta_{n,d} \end{aligned} \tag{B.2.28}$$

and

$$\begin{aligned} &P\left\{\boldsymbol{\theta}_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)}, \mathbf{D}_n)\right\} \\ &\geq P\left\{\sqrt{n}\|\hat{\Lambda}_n^{-1/2} R_n(\boldsymbol{x}^{(d)})\|_\infty \leq \mathsf{cv}(\alpha + \beta_{n,d})\right\} - \beta_{n,d} \gtrsim 1 - \alpha - \beta_{n,d} \end{aligned} \tag{B.2.29}$$

respectively, as required.

Hence, the remainder of the proof is devoted to establishing the bound (B.2.21). To do this, we require some additional notation. Let $\mathcal{R}$ denote the set of hyper-rectangles in $\mathbb{R}^d$. Let Z denote a centered Gaussian random vector with covariance matrix $\mathrm{Var}(\overline{u}(\boldsymbol{x}^{(d)}, D_i))$. Let Λ be the diagonal matrix with components $\lambda_j^2 = \mathrm{Var}(\overline{u}(x^{(j)}, D_i))$. To verify (B.2.21) it will suffice to establish that

$$\begin{aligned} &\sup_{\mathsf{R}\in\mathcal{R}} \left| P\left\{\sqrt{n}\hat{\Lambda}_n^{-1/2} R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R}\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}\right\} \right| \\ &\qquad \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^3(dn) + \frac{1}{nd} \end{aligned} \tag{B.2.30}$$

and that

$$\begin{aligned} &\sup_{\mathsf{R}\in\mathcal{R}} \left| P\left\{\sqrt{n}\hat{\Lambda}_n^{-1/2} R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R} \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}\right\} \right| \\ &\qquad \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^3(dn) \end{aligned} \tag{B.2.31}$$

with probability greater than $1 - Cn^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn) - (nd)^{-1}$. That is, the probability bound (B.2.21) follows from (B.2.30) and (B.2.31) by considering hyper-rectangles of the form $\mathsf{R} = [-\infty\mathbf{1}_d, z\mathbf{1}_d]$ and applying the triangle inequality.

We provide the details for the proofs of the upper bounds encoded in the absolute inequalities (B.2.30) and (B.2.31), respectively. Analogous lower bounds will follow from the same argument. In particular, fix a rectangle $\mathsf{R} = [a_\ell, a_u]$ in $\mathcal{R}$, where a_ℓ and a_u are vectors in $\mathbb{R}^d$ with $a_\ell \le a_u$, interpreted component-wise, and define the enlarged rectangle $\mathsf{R}_t = [a_\ell - t\mathbf{1}_d, a_u + t\mathbf{1}_d]$ for each $t > 0$. We obtain upper bounds

$$\begin{aligned}
&P\left\{\sqrt{n}\hat{\Lambda}_n^{-1/2}R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R}\right\} - P\left\{\Lambda^{-1/2}Z \in \mathsf{R}\right\} \\
&\quad \le \left|P\left\{\sqrt{n}\Lambda^{-1/2}R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R}_t\right\} - P\left\{\Lambda^{-1/2}Z \in \mathsf{R}_t\right\}\right| && \text{(B.2.32)}\\
&\qquad + \left|P\left\{\Lambda^{-1/2}Z \in \mathsf{R}_t\right\} - P\left\{\Lambda^{-1/2}Z \in \mathsf{R}\right\}\right| && \text{(B.2.33)}\\
&\qquad + P\left\{\sqrt{n}\|(\Lambda^{-1/2} - \hat{\Lambda}_n^{-1/2})R_n(\boldsymbol{x}^{(d)})\|_\infty \ge t\right\}, && \text{(B.2.34)}
\end{aligned}$$

and similarly

$$\begin{aligned}
&P\left\{\sqrt{n}\hat{\Lambda}_n^{-1/2}R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R} \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2}Z \in \mathsf{R}\right\} \\
&\quad \le \left|P\left\{\sqrt{n}\Lambda^{-1/2}R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R}_t \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2}Z \in \mathsf{R}_t\right\}\right| && \text{(B.2.35)}\\
&\qquad + \left|P\left\{\Lambda^{-1/2}Z \in \mathsf{R}_t\right\} - P\left\{\Lambda^{-1/2}Z \in \mathsf{R}\right\}\right| && \text{(B.2.36)}\\
&\qquad + P\left\{\sqrt{n}\|(\Lambda^{-1/2} - \hat{\Lambda}_n^{-1/2})R_n^*(\boldsymbol{x}^{(d)})\|_\infty \ge t \mid \mathbf{D}_n\right\}, && \text{(B.2.37)}
\end{aligned}$$

respectively. It thereby remains to obtain suitable bounds for each term, (B.2.32) through (B.2.37).

To bound the Gaussian approximation errors (B.2.32) and (B.2.35), we apply the following Theorem, which establishes a generic quantitative central limit for the statistic R_n in addition to a generic quantitative conditional central limit theorem for the half-sample bootstrap.

Theorem B.2.2. *Suppose that the moment function $M(x;\theta_0,g_0)$ satisfies Assumption B.2.1 and Part (i) of Assumption 2.3.3 and that Assumptions B.2.3 and B.2.4 hold.*

(i) *The inequality*

$$\begin{aligned}
&\sup_{R\in\mathcal{R}}\sup_{P\in\mathbf{P}}\left|P\left\{\sqrt{n}R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R}\right\} - P\left\{Z \in \mathsf{R}\right\}\right| \\
&\lesssim \frac{\varphi^{1/2}}{\underline{\lambda}^{1/2}}\left(\frac{\log^5(dn)}{n}\right)^{1/4} + \delta_n \log^{5/2}(dn) + \frac{1}{nd} && \text{(B.2.38)}
\end{aligned}$$

holds.

(ii) *If the bootstrap root is constructed with the Half-Sample bootstrap, then the inequality*

$$\begin{aligned}
&\sup_{R\in\mathcal{R}}\sup_{P\in\mathbf{P}}\left|P\left\{\sqrt{n}R_n^*(\boldsymbol{x}^{(d)})\in\mathsf{R}\mid\mathbf{D}_n\right\}-P\left\{Z\in\mathsf{R}\right\}\right| \\
&\lesssim\frac{\varphi^{1/2}}{\underline{\lambda}^{1/2}}\left(\frac{\log^5(dn)}{n}\right)^{1/4}+\delta_n\log^{5/2}(dn) \qquad \text{(B.2.39)}
\end{aligned}$$

holds with probability greater than $1-C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn)+(nd)^{-1})$.

In particular, Theorem B.2.2 implies that

$$\begin{aligned}
&\left|P\left\{\sqrt{n}\Lambda^{-1/2}R_n(\boldsymbol{x}^{(d)})\in\mathsf{R}_t\right\}-P\left\{\Lambda^{-1/2}Z\in\mathsf{R}_t\right\}\right| \\
&=\left|P\left\{\sqrt{n}R_n(\boldsymbol{x}^{(d)})\in\Lambda^{1/2}\mathsf{R}_t\right\}-P\left\{Z\in\Lambda^{1/2}\mathsf{R}_t\right\}\right| \\
&\lesssim\frac{\varphi^{1/2}}{\underline{\lambda}^{1/2}}\left(\frac{\log^5(dn)}{n}\right)^{1/4}+\delta_n\log^{5/2}(dn)+\frac{1}{nd} \qquad \text{(B.2.40)}
\end{aligned}$$

and analogously

$$\begin{aligned}
&\left|P\left\{\sqrt{n}\Lambda^{-1/2}R_n^*(\boldsymbol{x}^{(d)})\in\mathsf{R}_t\mid\mathbf{D}_n\right\}-P\left\{\Lambda^{-1/2}Z\in\mathsf{R}_t\right\}\right| \\
&=\left|P\left\{\sqrt{n}R_n^*(\boldsymbol{x}^{(d)})\in\Lambda^{1/2}\mathsf{R}_t\mid\mathbf{D}_n\right\}-P\left\{Z\in\Lambda^{1/2}\mathsf{R}_t\right\}\right| \\
&\lesssim\frac{\varphi^{1/2}}{\underline{\lambda}^{1/2}}\left(\frac{\log^5(dn)}{n}\right)^{1/4}+\delta_n\log^{5/2}(dn) \qquad \text{(B.2.41)}
\end{aligned}$$

with probability greater than $1-C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn)+(nd)^{-1})$, as $\Lambda^{1/2}\mathsf{R}_t$ is a hyper-rectangle.

To bound the differences in the Gaussian probabilities (B.2.33) and (B.2.36), we apply Lemma B.1.7. In particular, we have that

$$\left|P\left\{\Lambda^{-1/2}Z\in\mathsf{R}_t\right\}-P\left\{\Lambda^{-1/2}Z\in\mathsf{R}\right\}\right|\lesssim t\sqrt{\log d} \qquad \text{(B.2.42)}$$

for all $t>0$.

Finally, we bound the terms (B.2.34) and (B.2.37) resulting from variance estimation. Observe that

$$\begin{aligned}
\|(\Lambda^{-1/2}-\hat{\Lambda}_n^{-1/2})R_n(\boldsymbol{x}^{(d)})\|_\infty &=\|(I-\hat{\Lambda}_n^{-1/2}\Lambda^{1/2})\Lambda^{-1/2}R_n(\boldsymbol{x}^{(d)})\|_\infty \\
&\le\sup_{j\in[d]}|\lambda_j/\hat{\lambda}_{n,j}-1|\cdot\|\Lambda^{-1/2}R_n(\boldsymbol{x}^{(d)})\|_\infty \qquad \text{(B.2.43)}
\end{aligned}$$

and analogously

$$\|(\Lambda^{-1/2} - \hat{\Lambda}_n^{-1/2}) R_n^*(\boldsymbol{x}^{(d)})\|_\infty \leq \sup_{j\in[d]} |\lambda_j/\hat{\lambda}_{n,j} - 1| \cdot \|\Lambda^{-1/2} R_n^*(\boldsymbol{x}^{(d)})\|_\infty \,. \tag{B.2.44}$$

We give probability bounds for the terms on the right-hand sides of (B.2.43) and (B.2.44). Observe that the Borell-TIS inequality (e.g., Theorem 2.1.1 of Adler and Taylor, 2009) implies that

$$P\left\{\|\Lambda^{-1/2} Z\|_\infty \geq C\sqrt{\log dn}\right\} \leq n^{-1} \,.$$

Thus, Theorem B.2.2 implies that

$$P\left\{\sqrt{n}\|\Lambda^{-1/2} R_n(\boldsymbol{x}^{(d)})\|_\infty \geq C\sqrt{\log dn}\right\}$$
$$\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^{5/2}(dn) + \frac{1}{nd} \tag{B.2.45}$$

and that

$$P\left\{\sqrt{n}\|\Lambda^{-1/2} R_n^*(\boldsymbol{x}^{(d)})\|_\infty \geq C\sqrt{\log dn} \mid \mathbf{D}_n\right\}$$
$$\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^{5/2}(dn) \tag{B.2.46}$$

with probability greater than $1 - C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn) + (nd)^{-1})$. In turn, to bound the discrepancy between the bootstrap variance estimate $\hat{\lambda}_{n,j}$ and λ_j, we apply the following Lemma.

Lemma B.2.1. *Suppose that the conditions of Theorem B.2.1 hold. If the bootstrap root is constructed with the Half-Sample bootstrap, then*

$$\sup_{j\in[d]}\left|\frac{\hat{\lambda}_{n,j}}{\lambda_j} - 1\right| \lesssim \frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}} + \frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}$$
$$+ \log^2(dn)\delta_n + \log^4(dn)\delta_n^2 \tag{B.2.47}$$

with probability greater than $1 - C/nd$.

To apply Lemma B.2.1, observe that

$$\left|\frac{\lambda_j}{\hat{\lambda}_{n,j}} - 1\right| = \frac{|\lambda_j/\hat{\lambda}_{n,j} - 1|}{\hat{\lambda}_{n,j}/\lambda_j} \,. \tag{B.2.48}$$

Thus, the event that

$$\sup_{j\in[d]}\left|\frac{\hat{\lambda}_{n,j}}{\lambda_j}-1\right|\le\frac{1}{2}\quad\text{implies}\quad\frac{\hat{\lambda}_{n,j}}{\lambda_j}\ge\frac{1}{2}\,. \tag{B.2.49}$$

Hence, on this event, we find that

$$\sup_{j\in[d]}\left|\frac{\lambda_j}{\hat{\lambda}_{n,j}}-1\right|\le 2\sup_{j\in[d]}\left|\frac{\hat{\lambda}_{n,j}}{\lambda_j}-1\right|\,. \tag{B.2.50}$$

Observe now that the normalization (B.2.20) implies that

$$\begin{aligned}&\frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}}+\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}+\log^2(dn)\delta_n+\log^4(dn)\delta_n^2\\&\lesssim\left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4}+\delta_n\log^3(dn)<c\end{aligned} \tag{B.2.51}$$

and so the condition (B.2.49) is satisfied by the event characterized in Lemma B.2.1. Thus, by combining the bounds (B.2.43), (B.2.45), and (B.2.50) with Lemma B.2.1, we find that

$$\begin{aligned}&P\Bigg\{\sqrt{n}\|(\Lambda^{-1/2}-\hat{\Lambda}_n^{-1/2})R_n(\boldsymbol{x}^{(d)})\|_\infty\\&\qquad\gtrsim\left(\frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}}+\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}+\log^2(dn)\delta_n+\log^4(dn)\delta_n^2\right)\sqrt{\log dn}\Bigg\}\\&\lesssim\left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4}+\delta_n\log^{5/2}(dn)+\frac{1}{nd}\,.\end{aligned} \tag{B.2.52}$$

Similarly, by combining the bounds (B.2.44), (B.2.46), and (B.2.50) with Lemma (B.2.1), we find that

$$\begin{aligned}&P\Bigg\{\sqrt{n}\|(\Lambda^{-1/2}-\hat{\Lambda}_n^{-1/2})R_n^*(\boldsymbol{x}^{(d)})\|_\infty\\&\qquad\gtrsim\left(\frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}}+\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}+\log^2(dn)\delta_n+\log^4(dn)\delta_n^2\right)\sqrt{\log dn}\mid\mathbf{D}_n\Bigg\}\\&\lesssim\left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4}+\delta_n\log^{5/2}(dn)\end{aligned} \tag{B.2.53}$$

with probability greater than $1-C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn)+(nd)^{-1})$.

To put the pieces together, recall the upper bounds (B.2.32) and (B.2.35). By setting

$$t = \left(\frac{\varphi}{\underline{\lambda}} \sqrt{\frac{\log(dn)}{n}} + \frac{\varphi^2 \log^3(dn)}{\underline{\lambda}^2 n} + \log^2(dn)\delta_n + \log^4(dn)\delta_n^2 \right) \sqrt{\log dn}\,, \tag{B.2.54}$$

the bounds (B.2.40), (B.2.42), and (B.2.52) imply that

$$\begin{aligned}
&P\left\{\sqrt{n}\hat{\Lambda}_n^{-1/2} R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R}\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}\right\} \\
&\quad\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4} + \delta_n \log^{5/2}(dn) \\
&\qquad + \left(\frac{\varphi}{\underline{\lambda}} \sqrt{\frac{\log(dn)}{n}} + \frac{\varphi^2 \log^3(dn)}{\underline{\lambda}^2 n} + \log^2(dn)\delta_n + \log^4(dn)\delta_n^2 \right) \log(dn) + \frac{1}{nd} \\
&\quad\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4} + \delta_n \log^3(dn) + \frac{1}{nd}\,.
\end{aligned} \tag{B.2.55}$$

Indeed, the second inequality follows from the normalization (B.2.20). In particular,

$$\frac{\varphi}{\underline{\lambda}} \sqrt{\frac{\log(dn)}{n}} \log(dn) + \frac{\varphi^2 \log^3(dn)}{\underline{\lambda}^2 n} \log(dn) \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4}, \tag{B.2.56}$$

while

$$\log^3(dn)\delta_n + \log^5(dn)\delta_n^2 \lesssim \delta_n \log^3(dn)\,. \tag{B.2.57}$$

Similarly, the bounds (B.2.41), (B.2.42), and (B.2.53) imply that

$$\begin{aligned}
&P\left\{\sqrt{n}\hat{\Lambda}_n^{-1/2} R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R} \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}\right\} \\
&\quad\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4} + \delta_n \log^{5/2}(dn) \\
&\qquad + \left(\frac{\varphi}{\underline{\lambda}} \sqrt{\frac{\log(dn)}{n}} + \frac{\varphi^2 \log^3(dn)}{\underline{\lambda}^2 n} + \log^2(dn)\delta_n + \log^4(dn)\delta_n^2 \right) \log(dn) \\
&\quad\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4} + \delta_n \log^3(dn) + \frac{1}{nd}\,,
\end{aligned} \tag{B.2.58}$$

with probability greater than $1 - C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn) + (nd)^{-1})$. Hence, the bounds (B.2.55) and (B.2.58) verify the bounds encoded in (B.2.30) and (B.2.31), as required.

B.2.2 Proof of Theorem B.2.2

B.2.2.1 Part (i)

Throughout, we take $\theta_0(x) = 0$ for all x without loss of generality. We begin by giving a high-probability bound on the difference

$$\|R_n(\boldsymbol{x}^{(d)}) - W_n(\boldsymbol{x}^{(d)})\|_\infty \tag{B.2.59}$$

which will be needed at a later point in the proof. To this end, let x be any component of the vector $\boldsymbol{x}^{(d)}$. By a Taylor expansion about $\theta_0(x)$, we have that

$$\begin{aligned} M(x;\hat{\theta}_n(x),\hat{g}_n) - M(x;\theta_0(x),\hat{g}_n) &= (\hat{\theta}_n(x)-\theta_0(x))M^{(1)}(x;\theta_0(x),\hat{g}_n) \\ &\quad + (\hat{\theta}_n(x)-\theta_0(x))^2 H(x;\tilde{\theta}_0(x),\hat{g}_n) \end{aligned} \tag{B.2.60}$$

for some $\tilde{\theta}_0(x)$ between $\hat{\theta}_n(x)$ and $\theta_0(x)$. Moreover, we can write

$$\begin{aligned} &(\hat{\theta}_n(x)-\theta_0(x))M^{(1)}(x;\theta_0(x),\hat{g}_n) \\ &= (\hat{\theta}_n(x)-\theta_0(x))M^{(1)}(x;\theta_0(x),g_0) \\ &+ (\hat{\theta}_n(x)-\theta_0(x))\left(M^{(1)}(x;\theta_0(x),\hat{g}_n) - M^{(1)}(x;\theta_0(x),g_0)\right) \end{aligned} \tag{B.2.61}$$

and

$$\begin{aligned} M(x;\hat{\theta}_n(x),\hat{g}_n) - M(x;\theta_0(x),\hat{g}_n) &= (M(x;\hat{\theta}_n(x),\hat{g}_n) - M_n(x;\hat{\theta}_n(x),\hat{g}_n,\mathbf{D}_n)) \\ &\quad + (M(x;\theta_0(x),g_0) - M(x;\theta_0(x),\hat{g}_n)) \\ &= \left(M(x;\hat{\theta}_n(x),\hat{g}_n) - \mathbb{E}\left[M_n(x;\hat{\theta}_n(x),\hat{g}_n,\mathbf{D}_n)\right]\right) \\ &\quad + \left(\mathbb{E}\left[M_n(x;\hat{\theta}_n(x),\hat{g}_n,\mathbf{D}_n)\right] - M_n(x;\hat{\theta}_n(x),\hat{g}_n,\mathbf{D}_n)\right) \\ &\quad + (M(x;\theta_0(x),g_0) - M(x;\theta_0(x),\hat{g}_n))\ . \end{aligned} \tag{B.2.62}$$

Thus, by the identity

$$\begin{aligned} \overline{M}_n(x;\hat{\theta}_n(x),\hat{g}_n) &= \overline{M}_n(x;\hat{\theta}_n(x),\hat{g}_n) - \overline{M}_n(x;\theta_0(x),\hat{g}_n) \\ &\quad + \overline{M}_n(x;\theta_0(x),\hat{g}_n) - \overline{M}_n(x;\theta_0(x),g_0) + \overline{M}_n(x;\theta_0(x),g_0)\ , \end{aligned} \tag{B.2.63}$$

the equalities (B.2.60), (B.2.61), and (B.2.62) imply that

$$
\begin{aligned}
& M^{(1)}(x;\theta_0(x),g_0)(\hat{\theta}_n(x)-\theta_0(x)) \\
& = M^{(1)}(x;\theta_0(x),\hat{g}_n)(\hat{\theta}_n(x)-\theta_0(x)) \\
& - \left(M^{(1)}(x;\theta_0(x),\hat{g}_n)-M^{(1)}(x;\theta_0,g_0)\right)(\hat{\theta}_n(x)-\theta_0(x)) \\
& = -\overline{M}_n(x;\theta_0(x),g_0) && \text{(B.2.64)} \\
& \quad + \mathsf{Bias}(x;\hat{\theta}_n(x),\hat{g}_n) + \mathsf{Nuis}(x;\theta_0(x),\hat{g}_n) && \text{(B.2.65)} \\
& \quad + \mathsf{Stoch}^{(1)}(x;\hat{\theta}_n(x),\hat{g}_n) + \mathsf{Stoch}^{(2)}(x;\hat{g}_n) && \text{(B.2.66)} \\
& \quad - (\hat{\theta}_n(x)-\theta_0(x))^2 H(x;\tilde{\theta}_0(x),\hat{g}_n) && \text{(B.2.67)} \\
& \quad - (\hat{\theta}_n(x)-\theta_0(x))\left(M^{(1)}(x;\theta_0(x),\hat{g}_n)-M^{(1)}(x;\theta_0(x),g_0)\right) && \text{(B.2.68)}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathsf{Bias}(x;\hat{\theta}_n(x),\hat{g}_n) &= M(x;\hat{\theta}_n(x),\hat{g}_n) - \mathbb{E}\left[M_n(x;\hat{\theta}_n(x),\hat{g}_n)\right]\ , && \text{(B.2.69)} \\
\mathsf{Nuis}(x;\theta_0(x),\hat{g}_n) &= M(x;\theta_0(x),g_0) - M(x;\theta_0(x),\hat{g}_n)\ , && \text{(B.2.70)} \\
\mathsf{Stoch}^{(1)}(x;\hat{\theta}_n(x),\hat{g}_n) &= \overline{M}_n(x;\hat{\theta}_n(x),\hat{g}_n) - \overline{M}_n(x;\theta_0(x),\hat{g}_n)\ , \quad \text{and} && \text{(B.2.71)} \\
\mathsf{Stoch}^{(2)}(x;\hat{g}_n) &= \overline{M}_n(x;\theta_0(x),g_0) - \overline{M}_n(x;\theta_0(x),\hat{g}_n)\ , && \text{(B.2.72)}
\end{aligned}
$$

respectively.

We now give moment bounds for the terms (B.2.65), (B.2.66), (B.2.67), and (B.2.68). Fix any $q \geq 2 \vee 2\log(nd)$. We repeatedly use the fact that, for any random vector $Z(\boldsymbol{x}^{(d)}) = (Z(x^{(j)}))_{j=1}^d$,

$$
\mathbb{E}\left[\|Z(\boldsymbol{x}^{(d)})\|_\infty^q\right]^{1/q} \leq d^{1/q}\max_{j\in[d]}\mathbb{E}\left[|Z(x^{(j)})|^q\right]^{1/q} \lesssim \max_{j\in[d]}\mathbb{E}\left[|Z(x^{(j)})|^q\right]^{1/q}\ , \tag{B.2.73}
$$

where the final inequality follows from the lower bound on q. To handle (B.2.65), observe that Parts (i) and (ii) of Assumption B.2.4 imply that

$$
\begin{aligned}
& \mathbb{E}\left[\left|\sqrt{\frac{n}{\lambda^2}}\left\|\mathsf{Bias}_n(\boldsymbol{x}^{(d)};\hat{\theta}_n(\boldsymbol{x}^{(d)}),\hat{g}_n)\right\|_\infty\right|^q\right]^{1/q} \\
& \lesssim \delta_{n,B}\left(1+\mathbb{E}\left[\|\hat{\theta}_n(\boldsymbol{x}^{(d)})\|_\infty^q\right]^{1/q}\right) \\
& \lesssim \delta_{n,B}\left(1+q\frac{\lambda^{1/2}}{n^{1/4}}\delta_{n,\theta}\right) \lesssim q^2\left(\delta_{n,B}+\frac{\lambda^{1/2}}{n^{1/4}}\delta_{n,\theta}\delta_{n,B}\right)\ . && \text{(B.2.74)}
\end{aligned}
$$

Moreover, a Taylor expansion, Neyman orthogonality, second-order smoothness, i.e., Assumption B.2.1, and Assumption B.2.4, Part (ii), give that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\left\|\mathsf{Nuis}(\boldsymbol{x}^{(d)};\theta_0(\boldsymbol{x}^{(d)}),\hat{g}_n)\right\|_\infty\right|^q\right]^{1/q} \\
&\lesssim \mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\|\hat{g}_n-g_0\|_{2,\infty}^2\right|^q\right]^{1/q} = \mathbb{E}\left[\left|\left(\frac{n}{\underline{\lambda}^2}\right)^{1/4}\|\hat{g}_n-g_0\|_{2,\infty}\right|^{2q}\right]^{1/q} \lesssim q^2\delta_{n,g}^2 \,.
\end{aligned} \tag{B.2.75}$$

Next, we handle the term (B.2.66). By Assumption B.2.2, a Taylor expansion gives

$$\begin{aligned}
\mathsf{Stoch}^{(1)}(x;\hat{\theta}_n(x),\hat{g}_n) &= (\hat{\theta}_n(x)-\theta_0(x))\overline{M}_n^{(1)}(x;\theta_0(x),\hat{g}_n) \\
&\quad + (\hat{\theta}_n(x)-\theta_0(x))^2\overline{H}_n(x;\tilde{\theta}_0(x),\hat{g}_n)
\end{aligned} \tag{B.2.76}$$

for some, potentially different, $\tilde{\theta}_0(x)$ between $\hat{\theta}_n(x)$ and $\theta_0(x)$. To bound this term, observe that

$$\begin{aligned}
|\overline{M}_n^{(1)}(x;\theta_0(x),\hat{g}_n)| &\le |\overline{M}_n^{(1)}(x;\theta_0(x),g_0)| \\
&\quad + |\overline{M}_n^{(1)}(x;\theta_0(x),\hat{g}_n)-\overline{M}_n^{(1)}(x;\theta_0(x),g_0)| \,.
\end{aligned} \tag{B.2.77}$$

Hence, Assumption B.2.2 and Assumption B.2.4, Parts (ii) and (iii), together with Holder's inequality, imply that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\left\|\mathsf{Stoch}^{(1)}(\boldsymbol{x}^{(d)};\hat{\theta}_n(\boldsymbol{x}^{(d)}),\hat{g}_n)\right\|_\infty\right|^q\right]^{1/q} \\
&\lesssim q^2\delta_{n,\theta}\delta_{n,m} + q^2\frac{\underline{\lambda}^{1/2}}{n^{1/4}}\delta_{n,\theta}\delta_{n,J} + q^2\delta_{n,\theta}^2 \,.
\end{aligned} \tag{B.2.78}$$

Moreover, Assumption B.2.4, Part (iii), gives that

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\left\|\mathsf{Stoch}^{(2)}(\boldsymbol{x}^{(d)};\hat{g}_n)\right\|_\infty\right|^q\right]^{1/q} \lesssim q\delta_{n,S} \lesssim q^2\delta_{n,S} \,. \tag{B.2.79}$$

Finally, we handle the terms (B.2.67) and (B.2.68). Assumption B.2.4, Part (ii), and Assumption B.2.1 imply that

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\left\|(\hat{\theta}_n(\boldsymbol{x}^{(d)})-\theta_0(\boldsymbol{x}^{(d)}))^2H(\boldsymbol{x}^{(d)};\tilde{\theta}_0(\boldsymbol{x}^{(d)}),\hat{g}_n)\right\|_\infty\right|^q\right]^{1/q} \lesssim q^2\delta_{n,\theta}^2 \,, \tag{B.2.80}$$

and, by the Lipschitz property of $M^{(1)}(\cdot;\theta,g)$ in g,

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\right\|(\hat{\theta}_n(\boldsymbol{x}^{(d)})-\theta_0(\boldsymbol{x}^{(d)})) \\ \left(M^{(1)}(\boldsymbol{x}^{(d)};\theta_0(\boldsymbol{x}^{(d)}),\hat{g}_n)-M^{(1)}(\boldsymbol{x}^{(d)};\theta_0(\boldsymbol{x}^{(d)}),g_0)\right)\Big\|_\infty\Big|^q\right]^{1/q} \\ \lesssim q^2\delta_{n,\theta}\delta_{n,g}\ . \tag{B.2.81}$$

Together, the decomposition (B.2.64) and the lower-boundedness of the Jacobian $M^{(1)}(\cdot;\theta,g)$ imply that

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}\left\|R_n(\boldsymbol{x}^{(d)})-W_n(\boldsymbol{x}^{(d)})\right\|_\infty\right|^q\right]^{1/q} \tag{B.2.82}$$
$$\lesssim q^2\left(\delta_{n,g}^2+\delta_{n,\theta}^2+\delta_{n,B}+\delta_{n,S}+\delta_{n,\theta}\left(\delta_{n,m}+\delta_{n,g}+\underline{\lambda}^{1/2}n^{-1/4}\left(\delta_{n,B}+\delta_{n,J}\right)\right)\right)$$
$$\lesssim q^2\left(\delta_{n,g}^2+\delta_{n,\theta}^2+\delta_{n,m}^2+\delta_{n,B}+\delta_{n,S}+\underline{\lambda}^{1/2}n^{-1/4}\delta_{n,\theta}\left(\delta_{n,B}+\delta_{n,J}\right)\right),$$

where the final inequality uses $ab\le a^2+b^2$, applied to the products $\delta_{n,\theta}\delta_{n,m}$ and $\delta_{n,\theta}\delta_{n,g}$.

With this in place, we turn to the proof of the normal approximation to $R_n(\boldsymbol{x}^{(d)})$ on hyper-rectangles. Fix a rectangle $\mathsf{R}=[a_\ell,a_u]$ in $\mathcal{R}$. For the sake of exposition, we give the details of the proof of the upper bound

$$P\left\{\sqrt{n}R_n(\boldsymbol{x}^{(d)})\in\mathsf{R}\right\}-P\left\{\Lambda^{-1/2}Z\in\mathsf{R}\right\} \\ \lesssim\left(\frac{\varphi^2\log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4}+\delta_n\log^{5/2}(dn)+\frac{1}{nd}\ . \tag{B.2.83}$$

The matching lower bound will follow from a very similar argument. Consider the decomposition

$$R_n(x)=\left(W_n(x)-\frac{1}{n}\sum_{i=1}^n\overline{u}(x,D_i)\right)+\frac{1}{n}\sum_{i=1}^n\overline{u}(x,D_i)+\Delta_n(x),\quad\text{where} \tag{B.2.84}$$
$$\Delta_n(x)=R_n(x)-W_n(x)\ , \tag{B.2.85}$$

and we recall that the function $\overline{u}(\cdot,\cdot)$ is defined in Assumption B.2.3. Define the normalized functions

$$\hat{u}(\boldsymbol{x}^{(d)},D_i)=\Lambda^{-1/2}\overline{u}(\boldsymbol{x}^{(d)},D_i)\quad\text{and}\quad\hat{W}_n(\boldsymbol{x}^{(d)})=\Lambda^{-1/2}W_n(\boldsymbol{x}^{(d)})$$

and the analogously normalized rectangle $\tilde{\mathsf{R}} = [\Lambda^{-1/2}a_\ell, \Lambda^{-1/2}a_u]$ and the enlarged rectangle $\tilde{\mathsf{R}}_t = [\Lambda^{-1/2}a_\ell - \mathbf{1}_d t, \Lambda^{-1/2}a_u + \mathbf{1}_d t]$. Observe that the decomposition (B.2.84) yields the upper bound

$$
\begin{aligned}
&P\left\{\sqrt{n}R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R}\right\} - P\left\{Z \in \mathsf{R}\right\} \\
&= P\left\{\sqrt{n}\Lambda^{-1/2}R_n(\boldsymbol{x}^{(d)}) \in \tilde{\mathsf{R}}\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}\right\} \\
&\quad\leq \left|P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{u}(\boldsymbol{x}^{(d)}, D_i) \in \tilde{\mathsf{R}}_t\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}_t\right\}\right| && \text{(B.2.86)}\\
&\qquad + \left|P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}_t\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}\right\}\right| && \text{(B.2.87)}\\
&\qquad + P\left\{\sqrt{n}\|\hat{W}_n(\boldsymbol{x}^{(d)}) - \frac{1}{n}\sum_{i=1}^{n}\hat{u}(\boldsymbol{x}^{(d)}, D_i)\|_\infty \geq \frac{1}{2}t\right\} && \text{(B.2.88)}\\
&\qquad + P\left\{\sqrt{n}\|\Lambda^{-1/2}\Delta_n(\boldsymbol{x}^{(d)})\|_\infty \geq \frac{1}{2}t\right\} && \text{(B.2.89)}
\end{aligned}
$$

for each $t > 0$. We proceed by providing appropriate bounds for the terms (B.2.86) through (B.2.89).

We begin by bounding the normal approximation term through an application of Lemma B.1.6. In particular, observe that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\hat{u}^2(x^{(j)}, D_i)\right] = 1 \tag{B.2.90}$$

by definition and that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\hat{u}^4(x^{(j)}, D_i)\right] \leq (\varphi/\underline{\lambda})^2 \tag{B.2.91}$$

by Assumption B.2.3. Similarly we have that

$$\|\hat{u}(x^{(j)}, D_i)\|_{\psi_1} \leq (\varphi/\underline{\lambda}) \tag{B.2.92}$$

by Assumption B.2.3. Consequently, as

$$\operatorname{Var}(\hat{u}(\boldsymbol{x}^{(d)}, D_i)) = \Lambda^{-1/2}\operatorname{Var}(Z)\,\Lambda^{-1/2}, \tag{B.2.93}$$

by definition, Lemma B.1.6 implies that the bound

$$\left| P\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{u}(\mathbf{x}^{(d)}, D_i) \in \tilde{\mathsf{R}}_t \right\} - P\left\{ \Lambda^{-1/2} Z \in \tilde{\mathsf{R}}_t \right\} \right| \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4} . \quad \text{(B.2.94)}$$

holds.

In turn, to bound the terms (B.2.87), we have that

$$\left| P\left\{ \Lambda^{-1/2} Z \in \tilde{\mathsf{R}}_t \right\} - P\left\{ \Lambda^{-1/2} Z \in \tilde{\mathsf{R}} \right\} \right| \lesssim t\sqrt{\log d}\,. \quad \text{(B.2.95)}$$

by Lemma B.1.7. To bound the terms (B.2.88), taking $q = 2\log(nd)$, Assumption B.2.3 implies that

$$\sqrt{n} \| \hat{W}_n(\boldsymbol{x}^{(d)}) - \frac{1}{n} \sum_{i=1}^{n} \hat{u}(\boldsymbol{x}^{(d)}, D_i) \|_\infty \lesssim \delta_{n,u} \log^2(dn) \quad \text{(B.2.96)}$$

with probability greater than $1 - C/nd$, by applying Markov's inequality and a union bound. Likewise, the bound (B.2.82) implies that

$$\begin{aligned} &\sqrt{n} \| \Lambda^{-1/2} \Delta_n(\boldsymbol{x}^{(d)}) \|_\infty \\ &\lesssim (\delta_{n,g}^2 + \delta_{n,\theta}^2 + \delta_{n,m}^2 + \delta_{n,B} + \delta_{n,S} + \underline{\lambda}^{1/2} n^{-1/4} \delta_{n,\theta}(\delta_{n,B} + \delta_{n,J})) \log^2(dn) \end{aligned} \quad \text{(B.2.97)}$$

with probability greater than $1 - C/nd$, by applying Markov's inequality and a union bound.

Thus, by choosing $t = C\delta_n \log^2(dn)$, the sum of term (B.2.88) and term (B.2.89) is upper bounded by C/nd. Hence, by plugging this choice of t into (B.2.95), we can conclude that

$$\begin{aligned} &\left| P\left\{ \sqrt{n} R_n(\boldsymbol{x}^{(d)}) \in \mathsf{R} \right\} - P\left\{ \Lambda^{-1/2} Z \in \mathsf{R} \right\} \right| \\ &\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n} \right)^{1/4} + \delta_n \log^{5/2}(dn) + \frac{1}{nd} , \end{aligned} \quad \text{(B.2.98)}$$

as required. ∎

B.2.2.2 Part (ii)

The result follows from an argument whose structure is similar to the proof of Theorem B.2.2, Part (i). Again, we take $\theta_0(x) = 0$ for all x, without loss of generality. We are interested in studying the discrepancy

$$R_n^*(x) = \hat{\theta}_{\mathsf{h}}(x) - \hat{\theta}_n(x) = (\hat{\theta}_{\mathsf{h}}(x) - \theta_0(x)) - R_n(x).$$

In terms of the decomposition (B.2.84), we can write

$$R_n(x) = \left(W_n(x) - \frac{1}{n}\sum_{i=1}^{n} \overline{u}(x, D_i) \right) + \frac{1}{n}\sum_{i=1}^{n} \overline{u}(x, D_i) + \Delta_n(x) \ , \tag{B.2.99}$$

where $\Delta_n(x)$ is defined in (B.2.85). On the other hand, as $\hat{\theta}_{\mathsf{h}}(x)$ is constructed with a random half-sample h of the data $\mathbf{D}_n$, we have that

$$\hat{\theta}_{\mathsf{h}}(x) - \theta_0(x) = \left(\frac{2}{n}\sum_{i\in\mathsf{h}} \overline{u}(x, D_i) - W_{\mathsf{h}}(x) \right) + \frac{2}{n}\sum_{i\in\mathsf{h}} \overline{u}(x, D_i) + \Delta_{\mathsf{h}}(x) \ , \tag{B.2.100}$$

where $U_{\mathsf{h}}(x)$ and $\Delta_{\mathsf{h}}(x)$ are constructed with the half-sample h.

The proof of Theorem B.2.2, Part (i), worked by giving a high probability bound for the first and third term in (B.2.99) and showing that the second term satisfies a central limit theorem. Here, as we are interested in giving a bound conditioned on the data $\mathbf{D}_n$, we show that the difference between the second terms in (B.2.99) and (B.2.100) satisfies a central limit theorem on the event that the first and third terms in (B.2.99) and (B.2.100) satisfy a specified bound, which we show holds with high probability. To this end, let $\hat{W}_{\mathsf{h}}(\boldsymbol{x}^{(d)})$ is defined analogously to $\hat{W}_n(\boldsymbol{x}^{(d)})$ and let $\mathcal{F}_n(t)$ and $\mathcal{F}_{\mathsf{h}}(t)$ denote the events that

$$\sqrt{\frac{n}{\underline{\lambda}^2}}\|\Delta_n(\boldsymbol{x}^{(d)})\|_\infty \le t/4 \quad \text{and} \quad \sqrt{\frac{n}{\underline{\lambda}^2}}\|\Delta_{\mathsf{h}}(\boldsymbol{x}^{(d)})\|_\infty \le t/4 \ , \tag{B.2.101}$$

respectively. Similarly, let $\mathcal{H}_n(t)$ and $\mathcal{H}_{\mathsf{h}}(t)$ denote the events that

$$\begin{aligned} &\sqrt{n}\|\frac{1}{n}\sum_{i=1}^{n} \hat{u}(\boldsymbol{x}^{(d)}, D_i) - \hat{W}_n(\boldsymbol{x}^{(d)})\|_\infty \le t/4 \quad \text{and} \\ &\sqrt{n}\|\frac{2}{n}\sum_{i\in\mathsf{h}} \hat{u}(\boldsymbol{x}^{(d)}, D_i) - \hat{W}_{\mathsf{h}}(\boldsymbol{x}^{(d)})\|_\infty \le t/4 \end{aligned} \tag{B.2.102}$$

respectively. Define the event $\mathcal{E}_n(t) = \mathcal{F}_n(t) \cap \mathcal{F}_{\mathsf{h}}(t) \cap \mathcal{H}_n(t) \cap \mathcal{H}_{\mathsf{h}}(t)$.

Fix a hyper-rectangle R in $\mathcal{R}$. As before, we prove only the requisite upper bound. Recall the definitions of the transformed rectangle $\tilde{\mathsf{R}}$ and the enlarged transformed rectangle

$\tilde{\mathsf{R}}_t$. On the event $\mathcal{E}_n(t)$, we have

$$\begin{aligned}
&P\left\{\sqrt{n}R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R} \mid \mathbf{D}_n\right\} - P\left\{Z \in \mathsf{R}\right\} \\
&= P\left\{\sqrt{n}\Lambda^{-1/2}R_n^*(\boldsymbol{x}^{(d)}) \in \tilde{\mathsf{R}} \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}\right\} \\
&\leq P\left\{\frac{2}{\sqrt{n}}\sum_{i\in\mathsf{h}} \hat{u}(\boldsymbol{x}^{(d)}, D_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{u}(\boldsymbol{x}^{(d)}, D_i) \in \tilde{\mathsf{R}}_t \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}_t\right\} \\
&+ |P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}_t\right\}| \qquad \text{(B.2.103)}
\end{aligned}$$

for each $t > 0$. As the data D_h are drawn independently and identically with distribution P in $\mathbf{P}$ and we have assumed that $\delta_{n/2} \lesssim \delta_n$, by setting $t = C\delta_n \log^2(dn)$, bounds (B.2.97) and (B.2.96) imply that the event $\mathcal{E}_n(t)$ occurs with probability greater than $1 - C/nd$. Thus, Lemma B.1.7 implies that

$$\begin{aligned}
&P\left\{\sqrt{n}R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R} \mid \mathbf{D}_n\right\} - P\left\{Z \in \mathsf{R}\right\} \qquad \text{(B.2.104)} \\
&\leq P\left\{\frac{2}{\sqrt{n}}\sum_{i\in\mathsf{h}} \hat{u}(\boldsymbol{x}^{(d)}, D_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{u}(\boldsymbol{x}^{(d)}, D_i) \in \tilde{\mathsf{R}}_t \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2}Z \in \tilde{\mathsf{R}}_t\right\} \\
&\qquad \text{(B.2.105)} \\
&+ \delta_n \log^{5/2}(dn)
\end{aligned}$$

with probability greater than $1 - C/nd$. Hence, it suffices to bound the term (B.2.105).

To this end, we apply a coupling argument introduced in Yadlowsky et al. (2023), which is similar to a Poissonization technique studied in Præstgaard and Wellner (1993) (see also Section 3.6.2 of Van Der Vaart and Wellner (1996)). In particular, let V_i be a random variable taking the value 1 when i is an element of the subset h and taking the value -1 otherwise. Observe that

$$\frac{2}{n}\sum_{i\in\mathsf{h}} \hat{u}(\boldsymbol{x}^{(d)}, D_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{u}(\boldsymbol{x}^{(d)}, D_i) = \frac{1}{n}\sum_{i=1}^{n} V_i\hat{u}(\boldsymbol{x}^{(d)}, D_i)\,. \qquad \text{(B.2.106)}$$

Let $\tilde{V}_1, \ldots, \tilde{V}_n$ denote a collection of random variables valued on $\{-1, 1\}$. We define their joint distribution as follows. Let Q_n denote a random variable with distribution $\mathsf{Bin}\,(n, 1/2)$. If $Q_n \geq n/2$, then choose $Q_n - n/2$ indices i in $[n]$ with $V_i = -1$ and set $\tilde{V}_i = 1$. If $Q_n < n/2$, then choose $n/2 - Q_n$ indices with $V_i = 1$ and set $\tilde{V}_i = -1$. Set $\tilde{V}_i = V_i$ for all other units. Observe that the collection $\tilde{V}_i$ are independent and identically distributed

Rademacher random variables. With this in place, we obtain the decomposition

$$\frac{1}{n}\sum_{i=1}^{n} V_i \hat{u}(\boldsymbol{x}^{(d)}, D_i) = \frac{1}{n}\sum_{i=1}^{n} \tilde{V}_i \hat{u}(\boldsymbol{x}^{(d)}, D_i) + \frac{1}{n}\sum_{i=1}^{n}\left(V_i - \tilde{V}_i\right)\hat{u}(\boldsymbol{x}^{(d)}, D_i)\ . \tag{B.2.107}$$

Let $\mathcal{V}(t')$ denote the event that

$$\sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(V_i - \tilde{V}_i\right)\hat{u}(\boldsymbol{x}^{(d)}, D_i)\right\|_\infty < t'\ . \tag{B.2.108}$$

Fix any rectangle $\mathsf{R}' = [a'_\ell, a'_u]$ and define the enlarged rectangle $\mathsf{R}'_{t'} = [a'_\ell - t'\mathbf{1}_d, a'_u + t'\mathbf{1}_d]$. On the event $\mathcal{V}(t')$, the decomposition (B.2.107) implies that

$$P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \hat{u}(\boldsymbol{x}^{(d)}, D_i) \in \mathsf{R}' \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}'\right\}$$

$$\leq P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{V}_i \hat{u}(\boldsymbol{x}^{(d)}, D_i) \in \mathsf{R}'_{t'} \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}'_{t'}\right\} \tag{B.2.109}$$

$$+\left|P\left\{\Lambda^{-1/2} Z \in \mathsf{R}'_{t'}\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}'\right\}\right| \tag{B.2.110}$$

for all $t > 0$. Lemma B.1.7 implies that (B.2.110) is less than $t'\sqrt{\log(d)}$. To handle the term (B.2.109), we apply the following quantitative central limit theorem, stated as Lemma 4.6 in Chernozhukov et al. (2022).

Lemma B.2.2 (Lemma 4.6, Chernozhukov et al., 2022)**.** *Consider the setting and assumptions of Lemma B.1.6. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ collect the observed data and let $\tilde{V}_1, \ldots, \tilde{V}_n$ be a collection of independent Rademacher random variables. We have that*

$$\sup_{\mathsf{R}\in\mathcal{R}}\left|P\left\{n^{-1/2}\sum_{i=1}^{n} X_i \in \mathsf{R}\right\} - P\left\{n^{-1/2}\sum_{i=1}^{n}\tilde{V}_i X_i \in \mathsf{R} \mid \boldsymbol{X}_n\right\}\right| \tag{B.2.111}$$

$$\leq C_3\left(\frac{\varphi^2\log^5(dn)}{n}\right)^{1/4}. \tag{B.2.112}$$

with probability greater than

$$1 - C\frac{\varphi\log^{3/2}(dn)}{n^{1/2}} \tag{B.2.113}$$

for some constant C_3 that depends only on the constants C_2 and c defined in the statement of Lemma B.1.6.

Thus, on the event $\mathcal{V}(t')$, Lemma B.1.6 and Lemma B.2.2 imply that

$$P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \hat{u}(\boldsymbol{x}^{(d)}; D_i) \in \mathsf{R}' \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2} Z \in \mathsf{R}'\right\} \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + t'\sqrt{\log(d)}\,,$$

with probability greater than $1 - Cn^{-1/2}\underline{\lambda}^{-1}\varphi \log^{3/2}(dn)$.

Hence, it suffices to give a high probability bound on $\mathcal{V}(t')$. To this end, observe that

$$G_n = \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(V_i - \tilde{V}_i\right)\hat{u}(\boldsymbol{x}^{(d)}; D_i)\right\|_\infty \tag{B.2.114}$$

is equidistributed with

$$\left\|\frac{2}{\sqrt{n}}\sum_{i=1}^{|Q_n - n/2|}\hat{u}(\boldsymbol{x}^{(d)}; D_i)\right\|_\infty. \tag{B.2.115}$$

Consider the decomposition

$$\begin{aligned} P\left\{G_n \geq t'\right\} &\leq P\left\{G_n \geq t', |Q_n - n/2| \leq \delta\frac{n}{2}\right\} + P\left\{|Q_n - n/2| \geq \delta\frac{n}{2}\right\} \\ &\leq P\left\{\max_{1\leq k\leq \delta\frac{n}{2}}\left\|\frac{2}{\sqrt{n}}\sum_{i=1}^{k}\hat{u}(\boldsymbol{x}^{(d)}; D_i)\right\|_\infty \geq t'\right\} + P\left\{|Q_n - n/2| \geq \delta\frac{n}{2}\right\}, \end{aligned} \tag{B.2.116}$$

for some $\delta > 0$ to be chosen. Observe that

$$P\left\{|Q_n - n/2| \geq \frac{\delta n}{2}\right\} \leq 2\exp\left(-\frac{\delta^2 n}{6}\right) \tag{B.2.117}$$

by the multiplicative Chernoff bound. To bound the first term in (B.2.116), we combine Lemma B.1.5 with the following Lévy type inequality for independent random vectors, due to Montgomery-Smith (1993).

Lemma B.2.3 (Theorem 1.1.5, De la Pena and Giné (1999))**.** *Let $Z_1, \ldots, Z_n$ be independent random vectors in $\mathbb{R}^d$. There exist universal constants C_1 and C_2 such that*

$$P\left\{\max_{1\leq k\leq n}\|\sum_{i=1}^{k} Z_i\|_\infty > t\right\} \leq C_1 P\left\{\|\sum_{i=1}^{n} Z_i\|_\infty > \frac{t}{C_2}\right\} \tag{B.2.118}$$

for all $t > 0$.

In particular, as

$$\|\frac{2}{\sqrt{n}}\hat{u}(x^{(j)};D_i)\|_{\psi_1} \le \frac{2}{\sqrt{n}}\frac{\varphi}{\underline{\lambda}} ,$$

and

$$\max_{j\in[d]}\sum_{k=1}^{\delta n/2}\mathbb{E}\left[\left(\frac{2}{\sqrt{n}}\hat{u}(x^{(j)};D_i)\right)^2\right] = 2\delta , \tag{B.2.119}$$

Lemma B.1.5 and Lemma B.2.3 imply that

$$P\left\{\max_{1\le k\le \delta\frac{n}{2}}\left\|\frac{2}{\sqrt{n}}\sum_{i=1}^{k}\hat{u}\left(\boldsymbol{x}^{(d)};D_i\right)\right\|_\infty \ge C\left(\delta\log^{1/2}(dn)+\frac{2}{\sqrt{n}}\frac{\varphi}{\underline{\lambda}}\log^2(dn)\right)\right\} \lesssim \frac{1}{n} . \tag{B.2.120}$$

Now, the choice

$$\delta = C\sqrt{\frac{\log n}{n}} \tag{B.2.121}$$

gives

$$P\left\{|Q_n - n/2| \ge \frac{\delta n}{2}\right\} \lesssim \frac{1}{n} \tag{B.2.122}$$

by (B.2.117). Plugging this choice into (B.2.120) yields

$$P\left\{\max_{1\le k\le \delta\frac{n}{2}}\left\|\frac{2}{\sqrt{n}}\sum_{i=1}^{k}\hat{u}\left(\boldsymbol{x}^{(d)};D_i\right)\right\|_\infty \ge C\frac{1}{\sqrt{n}}\frac{\varphi}{\underline{\lambda}}\log^2(dn)\right\} \lesssim \frac{1}{n} . \tag{B.2.123}$$

Hence, we find that the inequality

$$\begin{aligned} P\Bigg\{\bigg\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(V_i-\tilde{V}_i\right)\hat{u}(\boldsymbol{x}^{(d)},D_i)\bigg\|_\infty \\ \ge C\left(\frac{\varphi^2\log^4(dn)}{\underline{\lambda}^2 n}\right)^{1/2}\Bigg\} \lesssim \frac{1}{n} \end{aligned} \tag{B.2.124}$$

holds for all n sufficiently large.

Thus, by setting

$$t' = C\left(\frac{\varphi^2\log^4(dn)}{\underline{\lambda}^2 n}\right)^{1/2} ,$$

we find that

$$P\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\hat{u}(\boldsymbol{x}^{(d)},D_i)\in\mathsf{R}' \mid \mathbf{D}_n\right\} - P\left\{\Lambda^{-1/2}Z\in\mathsf{R}'\right\}$$

$$\lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \sqrt{\log(d)} \left(\frac{\varphi^2 \log^4(dn)}{\underline{\lambda}^2 n}\right)^{1/2} \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4}, \tag{B.2.125}$$

with probability greater than $1 - Cn^{-1/2}\underline{\lambda}^{-1}\varphi \log^{3/2}(dn)$. Putting the pieces together, the inequalities (B.2.104) and (B.2.125) imply that

$$P\left\{\sqrt{n} R_n^*(\boldsymbol{x}^{(d)}) \in \mathsf{R} \mid \mathbf{D}_n\right\} - P\left\{Z \in \mathsf{R}\right\} \lesssim \left(\frac{\varphi^2 \log^5(dn)}{\underline{\lambda}^2 n}\right)^{1/4} + \delta_n \log^{5/2}(dn)$$

with probability greater than $1 - C(n^{-1/2}\varphi\underline{\lambda}^{-1}\log^{3/2}(dn) + (nd)^{-1})$, as required. ■

B.2.3 Proof of Lemma B.2.1

Again, we take $\theta_0(x) = 0$ for all x, without loss of generality. Recall from the proof of Theorem B.2.2, Part (ii), that V_i is a random variable taking the value 1 when i is an element of the subset h, and taking the value -1 otherwise, and that

$$\frac{2}{n}\sum_{i\in\mathsf{h}} \overline{u}(x, D_i) - \frac{1}{n}\sum_{i=1}^{n} \overline{u}(x, D_i) = \frac{1}{n}\sum_{i=1}^{n} V_i \overline{u}(x, D_i) . \tag{B.2.126}$$

Define the objects

$$T_n(x) = \frac{1}{n}\sum_{i=1}^{n} \overline{u}(x, D_i) - W_n(x) \quad \text{and} \quad T_{\mathsf{h}}(x) = \frac{2}{n}\sum_{i\in\mathsf{h}} \overline{u}(x, D_i) - W_{\mathsf{h}}(x) . \tag{B.2.127}$$

and recall the objects $\Delta_{\mathsf{h}}(x^{(j)})$ and $\Delta_n(x^{(j)})$ introduced in (B.2.85) and (B.2.100), respectively.

Observe that we can write

$$\begin{aligned} \hat{\lambda}_{n,j}^2 &= n\mathbb{E}_V\left[\left(R_n^*(x^{(j)})\right)^2\right] \qquad \text{(B.2.128)} \\ &= n\mathbb{E}_V\left[\left(\frac{1}{n}\sum_{i=1}^{n} V_i\overline{u}(x^{(j)}, D_i) - T_{\mathsf{h}}(x^{(j)}) + T_n(x^{(j)}) + \Delta_{\mathsf{h}}(x^{(j)}) - \Delta_n(x^{(j)})\right)^2\right] , \end{aligned}$$

where the notation $\mathbb{E}_V[\cdot]$ denotes that the expectation is evaluated only over the random variables $V_1, \dots, V_n$. Thus, by defining the quantities

$$\overline{\lambda}_{n,j}^2 = n\mathbb{E}_V\left[\left(\frac{1}{n}\sum_{i=1}^n V_i\overline{u}(x^{(j)}, D_i)\right)^2\right] \quad \text{and} \tag{B.2.129}$$
$$\Gamma_{n,j} = n\mathbb{E}_V\left[\left(T_n(x^{(j)}) - T_{\mathsf{h}}(x^{(j)}) + \Delta_{\mathsf{h}}(x^{(j)}) - \Delta_n(x^{(j)})\right)^2\right]$$

we obtain the bound

$$|\hat{\lambda}_{n,j}^2 - \lambda_j^2| \le |\overline{\lambda}_{n,j}^2 - \lambda_j^2| + 2\overline{\lambda}_{n,j}\Gamma_{n,j}^{1/2} + \Gamma_{n,j} \tag{B.2.130}$$

by expanding the square and applying the Cauchy-Schwarz inequality.

Now, we obtain a large deviation bound for $\Gamma_{n,j}$. Observe that Assumption B.2.3 and (B.2.82) imply that

$$\mathbb{E}\left[\left(\frac{n}{\underline{\lambda}^2}T_n(x^{(j)})^2\right)^{q/2}\right] \lesssim q^{2q}\delta_n^q \quad \text{and} \quad \mathbb{E}\left[\left(\frac{n}{\underline{\lambda}^2}\Delta_n(x^{(j)})^2\right)^{q/2}\right] \lesssim q^{2q}\delta_n^q\,. \tag{B.2.131}$$

respectively. Likewise, Jensen's inequality implies that

$$\mathbb{E}\left[\left(\mathbb{E}_V\left[\frac{n}{\underline{\lambda}^2}T_{\mathsf{h}}(x^{(j)})^2\right]\right)^{q/2}\right] \le \mathbb{E}\left[\mathbb{E}_V\left[\left|\sqrt{\frac{n}{\underline{\lambda}^2}}T_{\mathsf{h}}(x^{(j)})\right|^q\right]\right] \lesssim q^{2q}\delta_n^q \tag{B.2.132}$$

and analogously

$$\mathbb{E}\left[\left(\mathbb{E}_V\left[\frac{n}{\underline{\lambda}^2}\Delta_{\mathsf{h}}(x^{(j)})^2\right]\right)^{q/2}\right] \lesssim q^{2q}\delta_n^q\,. \tag{B.2.133}$$

Thus, by taking $q = 2 \vee 2\log(nd)$ and applying Markov and a union bound, we find that

$$\sup_{j\in[d]} \frac{n}{\underline{\lambda}^2}\Big(\mathbb{E}_V\left[T_{\mathsf{h}}(x^{(j)})^2\right] + \mathbb{E}_V\left[\Delta_{\mathsf{h}}(x^{(j)})^2\right] + T_n(x^{(j)})^2 + \Delta_n(x^{(j)})^2\Big) \lesssim \log^4(dn)\delta_n^2 \tag{B.2.134}$$

with probability at least $1 - C/nd$. Consequently, we find that

$$\sup_{j\in[d]} \Gamma_{n,j} \lesssim \underline{\lambda}^2\log^4(dn)\delta_n^2 \tag{B.2.135}$$

with probability at least $1 - C/nd$.

Next, we derive a large deviation bound for the difference $|\overline{\lambda}^2_{n,j} - \lambda_j^2|$. Observe that we can evaluate

$$\begin{aligned}\mathbb{E}\left[V_i V_{i'}\right] &= \frac{1}{2}\mathbb{E}\left[V_i \mid V_{i'} = 1\right] - \frac{1}{2}\mathbb{E}\left[V_{i'} \mid V_i = -1\right] \\ &= \frac{1}{2}\left(\frac{n/2-1}{n-1} - \frac{n/2}{n-1}\right) - \frac{1}{2}\left(\frac{n/2}{n-1} - \frac{n/2-1}{n-1}\right) = -\frac{1}{n-1}\end{aligned} \tag{B.2.136}$$

and thereby

$$\begin{aligned}\overline{\lambda}^2_{n,j} &= \frac{1}{n}\sum_{i=1}^{n} \overline{u}^2(x^{(j)}, D_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{i' \neq i} \mathbb{E}\left[V_i V_{i'}\right] \overline{u}(x^{(j)}, D_i)\overline{u}(x^{(j)}, D_{i'}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \overline{u}^2(x^{(j)}, D_i) - \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{i' \neq i} \overline{u}(x^{(j)}, D_i)\overline{u}(x^{(j)}, D_{i'}) \,.\end{aligned} \tag{B.2.137}$$

We now bound the two terms appearing in (B.2.137). First, we apply the following Rosenthal type inequality.

Lemma B.2.4 (Theorem 9, Boucheron et al., 2005)**.** *Let $X_1, \ldots, X_b$ be independent centered random variables. For any integer $q \geq 2$, we have that*

$$\|\frac{1}{b}\sum_{i=1}^{b} X_i\|_q \lesssim \sqrt{\frac{\mathrm{Var}(X_i) q}{b}} + \frac{q}{b}\|\max_{i\in[b]}|X_i|\|_q.$$

In particular, the variables

$$\overline{u}^2(x^{(j)}, D_i) - \lambda_j^2 \tag{B.2.138}$$

are independent and mean-zero. Moreover, Assumption B.2.3 implies that

$$\mathbb{E}\left[\left(\overline{u}^2(x^{(j)}, D_i) - \lambda_j^2\right)^2\right] \leq \mathbb{E}\left[\overline{u}^4(x^{(j)}, D_i)\right] \leq \lambda_j^2 \varphi^2 \,. \tag{B.2.139}$$

In turn, the bound $\|\overline{u}(x^{(j)}, D_i)\|_{\psi_1} \leq \varphi$ gives

$$\left\|\max_{i\in[n]}\left|\overline{u}^2(x^{(j)}, D_i) - \lambda_j^2\right|\right\|_q \lesssim q^2\varphi^2 \tag{B.2.140}$$

for every $q \geq 2 \vee 2\log(nd)$. Therefore, Lemma B.2.4 implies that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(\overline{u}^2(x^{(j)}, D_i) - \lambda_j^2\right)\right|^q\right]^{1/q} \lesssim \lambda_j \varphi \sqrt{\frac{q}{n}} + \frac{\varphi^2 q^3}{n}\,. \tag{B.2.141}$$

By taking $q = 2 \vee 2\log(nd)$, applying Markov and a union bound gives

$$\sup_{j\in[d]} \frac{1}{\lambda_j^2}\left|\frac{1}{n}\sum_{i=1}^{n}\overline{u}^2(x^{(j)}, D_i) - \lambda_j^2\right| \lesssim \frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}} + \frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n} \tag{B.2.142}$$

with probability greater than $1 - C/nd$.

To handle the second term, we apply the following sub-exponential formulation of the Hanson and Wright (1971) concentration inequality for quadratic forms, due to Götze et al. (2021).

Lemma B.2.5 (Proposition 1.1, Götze et al. (2021))**.** *Let $X_1, \ldots, X_n$ be independent, centered, random variables satisfying $\mathbb{E}[X_i^2] = \sigma_i^2$ and $\|X_i\|_{\psi_1} \leq \varphi$. If $A = (a_{i,i'})$ is any symmetric $n \times n$ matrix, then the inequality*

$$\begin{aligned} P&\left\{\left|\sum_{i=1}^{n}\sum_{i'=1}^{n} a_{i,i'}X_iX_{i'} - \sum_{i=1}^{n}\sigma_i^2 a_{i,i}\right| \geq t\right\} \\ &\leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{\varphi^4\|A\|_F^2}, \frac{t^{1/2}}{\varphi\|A\|_{\mathsf{op}}^{1/2}}\right)\right) \end{aligned} \tag{B.2.143}$$

holds for any $t \geq 0$, where $\|\cdot\|_F$ and $\|\cdot\|_{\mathsf{op}}$ denote the Frobenius and ℓ_2 operator norms, respectively.

In particular, let A denote the $n \times n$ matrix with zeroes on the diagonal and $(n(n-1))^{-1}$ in every other entry. In this case

$$\|A\|_F^2 = \frac{1}{n}\frac{1}{n-1} \quad \text{and} \quad \|A\|_{\mathsf{op}}^{1/2} = \frac{1}{\sqrt{n}} \tag{B.2.144}$$

and so Lemma B.2.5 implies that

$$\sup_{j\in[d]} \frac{1}{\lambda_j^2}\left|\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{i'\neq i}\overline{u}(x^{(j)}, D_i)\overline{u}(x^{(j)}, D_{i'})\right| \lesssim \frac{\varphi^2}{\underline{\lambda}^2 n}\log^2(dn) \tag{B.2.145}$$

with probability greater than $1 - C/nd$.

Hence, by combining (B.2.137), (B.2.142), and (B.2.145), we find that

$$\sup_{j\in[d]}\left|\frac{\overline{\lambda}_{n,j}^2}{\lambda_j^2}-1\right|\lesssim\frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}}+\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n} \tag{B.2.146}$$

with probability greater than $1-C/nd$. Thus, by plugging (B.2.135) and (B.2.146) into the bound (B.2.130), we obtain

$$\begin{aligned}\sup_{j\in[d]}\left|\frac{\hat{\lambda}_{n,j}^2}{\lambda_j^2}-1\right| &\le \sup_{j\in[d]}\left|\frac{\overline{\lambda}_{n,j}^2}{\lambda_j^2}-1\right|+2\sup_{j\in[d]}\frac{\overline{\lambda}_{n,j}}{\lambda_j}\sup_{j\in[d]}\frac{\Gamma_{n,j}^{1/2}}{\lambda_j}+\sup_{j\in[d]}\frac{\Gamma_{n,j}}{\lambda_j^2}\\ &\lesssim\frac{\varphi}{\underline{\lambda}}\sqrt{\frac{\log(dn)}{n}}+\frac{\varphi^2\log^3(dn)}{\underline{\lambda}^2 n}+\log^2(dn)\delta_n+\log^4(dn)\delta_n^2 \end{aligned} \tag{B.2.147}$$

with probability greater than $1-C/nd$, as required. ■

B.3. Proof of Theorem 2.3.1

B.3.1 Proof of Theorem 2.3.1

The result follows from an application of Theorem B.2.1. We begin by verifying the requisite assumptions. Theorem B.2.1 requires that the moment function $M(\cdot;\theta_0,g_0)$ is Neyman Orthogonal and satisfies Assumption B.2.1. Neyman Orthogonality is also imposed in the conditions of Theorem 2.3.1, and so is satisfied. Assumption B.2.1 requires that the moment function $M(x;\theta_0,g_0)$ is twice continuously differentiable and has a Jacobian function

$$M^{(1)}(x;g)=\frac{\partial}{\partial\theta'}M(x;\theta,g)\mid_{\theta'=\theta} \tag{B.3.1}$$

that satisfies the bounds (B.2.2) and (B.2.3). Twice continuous differentiability follows immediately from Assumption 2.3.2, which states that the moment function $m(\cdot;\theta,g)$ satisfies the linear representation

$$m(d;\theta,g)=m^{(1)}(d;g)\cdot\theta+m^{(2)}(d;g)\ , \tag{B.3.2}$$

with bounded functions $m^{(1)}(\cdot;g)$ and $m^{(2)}(\cdot;g)$. In particular, the representation (B.3.2) implies that the Jacobian can be written

$$M^{(1)}(x;\theta,g)=\mathbb{E}[m^{(1)}(D_i;g)\mid X_i=x] \tag{B.3.3}$$

and that the Hessian can be written

$$H(x;\theta,g) = \frac{\partial^2}{\partial^2\theta'} M(x;\theta,g_0)\mid_{\theta'=\theta} = 0\ . \tag{B.3.4}$$

Here, our ability to exchange integration and differentiation follows from boundedness of $m(\cdot;\theta,g)$. Consequently, the bounds (B.2.2) and (B.2.3) follow from the bounds (2.3.26) and (2.3.27) stated as Part (iii) of Assumption 2.3.3.

Theorem B.2.1 also requires that the centered empirical moment $\overline{M}_n(x;\theta,g)$ satisfy Assumption B.2.2. Assumption B.2.2 imposes the restriction that $\overline{M}_n(x;\theta,g)$ is twice continuously differentiable and that its Hessian

$$\overline{H}_n(x;\theta,g) = \frac{\partial^2}{\partial^2\theta'} \overline{M}_n(x;\theta',g_0)|_{\theta'=\theta} \tag{B.3.5}$$

is uniformly bounded almost surely as x, θ, and g vary over their respective domains. Recall that, in our case, we have that

$$\overline{M}_n(x;\theta,g) = \sum_{i=1}^{n} \left(K(x,X_i)m(D_i;\theta,g) - \mathbb{E}\left[K(x,X_i)m(D_i;\theta,g)\right]\right)\ . \tag{B.3.6}$$

Twice continuous differentiability again follows from Assumption 2.3.2, i.e., the linear representation (B.3.2). In turn, we have that $\overline{H}_n(x;\theta,g) = 0$ almost surely, verifying the uniform boundedness condition. Moreover, for future reference, the empirical Jacobian is given by

$$\overline{M}_n^{(1)}(x;g) = \sum_{i=1}^{n} \Big(K(x,X_i)m^{(1)}(D_i;g) - \mathbb{E}\left[K(x,X_i)m^{(1)}(D_i;g)\right]\Big) \tag{B.3.7}$$

again by Assumption 2.3.2.

We now quantify the generic sequence δ_n^u defined in Assumption B.2.3. Recall the normalized statistic $W_n(x)$ defined in (B.2.6). By the definition (2.3.13), this quantity can be written

$$\begin{aligned} W_n(x) &= \frac{1}{r}\sum_{q=1}^{r} u(x;D_{\mathsf{s}_q},\xi_{\mathsf{s}_q},\theta_0,g_0)\ , \quad \text{where} \\ u(x;D_{\mathsf{s}},\xi_{\mathsf{s}},\theta,g) &= -M^{(1)}(x;g_0)^{-1}\sum_{i\in\mathsf{s}} \Big(\kappa(x,X_i,D_{\mathsf{s}},\xi_{\mathsf{s}})m(D_i;\theta,g) \\ &\qquad - \mathbb{E}\big[\kappa(x,X_i,D_{\mathsf{s}},\xi_{\mathsf{s}})m(D_i;\theta,g)\big]\Big)\ . \end{aligned}$$

Define the de-randomized kernel function and Hájek projection

$$\tilde{u}(x; D) = \mathbb{E}\left[u(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, \theta_0(x), g_0) \mid D_{\mathsf{s}} = D\right] \quad \text{and} \tag{B.3.8}$$

$$\tilde{u}^{(1)}(x; D) = \mathbb{E}\left[u(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, \theta_0(x), g_0) \mid i \in \mathsf{s}, D_i = D\right] , \tag{B.3.9}$$

respectively, in addition to the quantities

$$\sigma_{b,j}^2 = \mathrm{Var}(\tilde{u}^{(1)}(x^{(j)}, D_i)) \quad \text{and} \quad \underline{\sigma}_b^2 = \min_{j \in [d]} \sigma_{b,j}^2 . \tag{B.3.10}$$

The result below follows from an application of Theorem 2.1.1.

Lemma B.3.1 (Asymptotic Linearity)**.** *Suppose that the de-randomized kernel function* (B.3.8) *is bounded by $\phi \geq 1$ almost surely and is invariant to permutations of its second argument. Suppose that b and r are chosen to satisfy $n \leq \sqrt{r}b$. If the inequality*

$$\frac{(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi b \log(dn)}{n} \leq C \tag{B.3.11}$$

holds, then

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{b^2 \underline{\sigma}_b^2}}\left(W_n(x^{(j)}) - \frac{b}{n}\sum_{i=1}^n \tilde{u}^{(1)}(x^{(j)}; D_i)\right)\right|^q\right]^{1/q} \lesssim q^2 \left(\frac{(1 + \|\theta_0(\boldsymbol{x}^{(d)})\varphi^2\|_\infty)^2}{\underline{\sigma}_b^2 n}\right)^{1/2} , \tag{B.3.12}$$

for each $j \in [d]$ and each $2 \vee 2\log(nd) \leq q \leq cn/b$.

In particular, observe that Assumption 2.3.2 implies that the de-randomized kernel function (B.3.8) is bounded by ϕ, up to a constant that depends only on $\mathbf{P}$. Permutation invariance again follows by assumption. We may assume that the first condition in (B.3.11) holds, as otherwise the desired bound is vacuously true. Consequently, Assumption B.2.3 holds with the choice

$$\delta_{n,u} = C\left(\frac{b\phi^2}{n}\right)^{1/2} \tag{B.3.13}$$

by incrementality and the fact that $\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty$ is bounded as P varies over $\mathbf{P}$.

Next, we quantify the sequences introduced in Assumption B.2.4. The lemma below follows from arguments similar to those used in Wager and Athey (2018) and Oprescu et al. (2019).

Lemma B.3.2 (Bias, Consistency, and Stochastic Equicontinuity)**.**
Suppose that the conditions of Theorem 2.3.1 are satisfied.

(i) *The bound*

$$\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\|\mathsf{Bias}_n(\boldsymbol{x}^{(d)};\theta(\boldsymbol{x}^{(d)}),\hat{g}_n)\|_\infty \lesssim (1+\|\theta(\boldsymbol{x}^{(d)})\|_\infty)\delta_{n,B}\ , \quad \text{where} \qquad \text{(B.3.14)}$$
$$\delta_{n,B} = \left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/2}\varepsilon_b$$

holds.

(ii) *The bound*

$$\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}\mathbb{E}\left[\left|\overline{M}_n^{(1)}(x^{(j)};g_0)\right|^q\right]^{1/q} \lesssim q\delta_{n,m}\ , \quad \text{where} \quad \delta_{n,m} = \left(\frac{\phi^4}{\underline{\sigma}_b^2 n}\right)^{1/4} \qquad \text{(B.3.15)}$$

holds for each $q \geq 2 \vee 2\log(dn)$ *satisfying*

$$\frac{\phi^2 b q^2}{n} \leq c \qquad \text{(B.3.16)}$$

and each $j \in [d]$.

(iii) *The bounds*

$$\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\mathbb{E}\left[\left|\overline{M}_n(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n) - \overline{M}_n(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q} \lesssim q\delta_{n,S}\ , \quad \text{where}$$
$$\delta_{n,S} = \left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/2}\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + \varepsilon_b + (1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\right)\ , \qquad \text{(B.3.17)}$$

and

$$\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\mathbb{E}\left[\left|\overline{M}_n^{(1)}(x^{(j)};\hat{g}_n) - \overline{M}_n^{(1)}(x^{(j)};g_0)\right|^q\right]^{1/q} \lesssim q\delta_{n,J}\ , \quad \text{where}$$
$$\delta_{n,J} = \left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/2}\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + \varepsilon_b + (1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\right) \qquad \text{(B.3.18)}$$

hold for each $q \geq 2 \vee 2\log(dn)$ *satisfying*

$$\frac{\phi^2(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)^2 b q^2}{n} + \varepsilon_b \leq c \qquad \text{(B.3.19)}$$

and each $j \in [d]$.

(iv) *The bound*

$$\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4} \mathbb{E}\left[\left|\hat{\theta}_n(x^{(j)}) - \theta_0(x^{(j)})\right|^q\right]^{1/q} \lesssim q\delta_{n,\theta}\,, \quad \textit{where}$$

$$\delta_{n,\theta} = \left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + (1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\varepsilon_b + (1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\right) \tag{B.3.20}$$

holds for each $q \geq 2 \vee 2\log(dn)$ *satisfying*

$$\varepsilon_b \leq 1\,, \quad q\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} \leq 1\,, \quad \textit{and} \quad \frac{\phi^2(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)^2 bq^2}{n} + \varepsilon_b \leq 1 \tag{B.3.21}$$

and each $j \in [d]$.

Observe that the choice

$$\overline{u}(x, D_i) = b \cdot \tilde{u}^{(1)}(x, D_i)\,,$$

suggested by Lemma B.3.1 implies that

$$b\phi^2 \gtrsim \underline{\lambda}^2 = b^2\underline{\sigma}_b^2 \gtrsim b\,, \tag{B.3.22}$$

by Assumption 2.3.2 and incrementality. We set $\overline{q}_{n,d} = C\log(dn)$. We may assume that there exists some $c \leq 1$ such that

$$\left(\delta_{n,g}^2 + \sqrt{\frac{n}{b}}\varepsilon_b\right)\log^3(dn) \leq c \quad \text{and} \quad \frac{b\phi^4\log^8(dn)}{n} \leq c\,, \tag{B.3.23}$$

as otherwise the bound is vacuously true. Consequently, all of the conditions of Lemma B.3.2 are satisfied for every $q \leq \overline{q}_{n,d}$. Thus, Lemma B.3.2 and Lemma B.3.1 imply that we can set

$$\delta_{n,u} = C\left(\frac{\phi^2}{\underline{\sigma}_b^2 n}\right)^{1/2} \lesssim \left(\frac{b\phi^2}{n}\right)^{1/2}\,, \tag{B.3.24}$$

$$\delta_{n,B} = C\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\varepsilon_b \lesssim \sqrt{\frac{n}{b}}\varepsilon_b\,, \tag{B.3.25}$$

$$\delta_{n,m} = C\left(\frac{\phi^4}{\underline{\sigma}_b^2 n}\right)^{1/4} \lesssim \left(\frac{b\phi^4}{n}\right)^{1/4}\,, \tag{B.3.26}$$

$$\delta_{n,\theta} = C\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + \varepsilon_b + \phi\sqrt{\frac{b}{n}}\right)$$

$$\lesssim \delta_{n,g} + \left(\frac{n}{b}\right)^{1/4} \varepsilon_b + \left(\frac{b}{n}\right)^{1/4} \phi\,, \quad \text{and} \tag{B.3.27}$$

$$\delta_{n,S} = \delta_{n,J} = C\log^{1/2}(dn)\left(\frac{b}{n}\right)^{1/4} \delta_{n,g} + C\sqrt{\frac{n}{b}}\varepsilon_b + C\phi\sqrt{\frac{b}{n}}\log(dn)\,, \tag{B.3.28}$$

respectively, where we have repeatedly used the fact that $\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty$ is uniformly bounded.

With these results in place, we apply Theorem B.2.1. We begin by giving a suitable upper bound for the sequence

$$\delta_n = \delta_{n,g}^2 + \delta_{n,\theta}^2 + \delta_{n,m}^2 + \delta_{n,B} + \delta_{n,S} + \delta_{n,u} + \underline{\lambda}^{1/2} n^{-1/4} \delta_{n,\theta}\left(\delta_{n,B} + \delta_{n,J}\right) \tag{B.3.29}$$

introduced in the statement of Theorem B.2.1. Observe that

$$\frac{\underline{\lambda}^{1/2}}{n^{1/4}}(\delta_{n,B} + \delta_{n,J})\delta_{n,\theta} \lesssim \frac{b^{1/4}}{n^{1/4}}(\delta_{n,B} + \delta_{n,J})\delta_{n,\theta} \lesssim \delta_{n,B} + \delta_{n,J}\,, \tag{B.3.30}$$

as otherwise the desired bound would be vacuous, where the first inequality follows from (B.3.22). Consequently, we find that, if $\delta_{n,J} = \delta_{n,S}$, then

$$\delta_n \lesssim \delta_{n,g}^2 + \delta_{n,\theta}^2 + \delta_{n,m}^2 + \delta_{n,B} + \delta_{n,S} + \delta_{n,u}\,. \tag{B.3.31}$$

Plugging in the choices specified above gives

$$\delta_n \lesssim \delta_{n,g}^2 + \sqrt{\frac{n}{b}}\varepsilon_b + \sqrt{\frac{b}{n}}\phi^2 + \left(\frac{b}{n}\right)^{1/4} \delta_{n,g}\log^{1/2}(dn) + \phi\sqrt{\frac{b}{n}}\log(dn) \tag{B.3.32}$$

$$\lesssim \delta_{n,g}^2 + \sqrt{\frac{n}{b}}\varepsilon_b + \phi^2\sqrt{\frac{b}{n}}\log(dn)\,. \tag{B.3.33}$$

where, to get the second inequality, we apply the elementary bound

$$\left(\frac{b}{n}\right)^{1/4} \delta_{n,g}\log^{1/2}(dn) \lesssim \delta_{n,g}^2 + \sqrt{\frac{b}{n}}\log(dn)$$

and use $\phi \geq 1$. Hence, the inequality (B.3.33), the Orlicz-norm bound

$$\|\overline{u}(x^{(j)}, D)\|_{\psi_1} \leq b \cdot (1 + |\theta_0(x^{(j)})|)\phi \lesssim b\phi\,,$$

and Theorem B.2.1 imply that

$$\begin{aligned}
&\sup_{P\in\mathbf{P}} \left|P\left\{\theta_0(\boldsymbol{x}^{(d)}) \in \hat{\mathcal{C}}(\boldsymbol{x}^{(d)})\right\} - (1-\alpha)\right| \\
&\lesssim \left(\frac{b\phi^2\log^5(dn)}{n}\right)^{1/4} + \left(\delta_{n,g}^2 + \sqrt{\frac{n}{b}}\varepsilon_b + \phi^2\sqrt{\frac{b}{n}}\log(dn)\right)\log^3(dn) + \frac{1}{nd} \\
&\lesssim \left(\frac{b\phi^4\log^8(dn)}{n}\right)^{1/4} + \left(\delta_{n,g}^2 + \sqrt{\frac{n}{b}}\varepsilon_b\right)\log^3(dn)\ , \qquad \text{(B.3.34)}
\end{aligned}$$

as required. ■

B.3.2 Proof of Lemma B.3.1

To ease notation, we define the parameter

$$\phi(\theta_0) = (1 + \|\theta_0(x^{(j)})\|)\phi$$

and drop dependence on $\theta(x)$ and g when writing $u(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, \theta(x), g)$, as these values will be taken to be $\theta_0(x)$ and g_0 throughout. We are interested in studying the discrepancy

$$W_n(x^{(j)}) - \frac{b}{n}\sum_{i=1}^{n}\tilde{u}^{(1)}(x^{(j)}; D_i)\ . \qquad \text{(B.3.35)}$$

Define the quantities

$$\hat{U}_n(x^{(j)}) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} u(x^{(j)}; D_{\mathsf{s}}, \xi_{\mathsf{s}}) \quad \text{and} \quad \overline{U}_n(x^{(j)}) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} \tilde{u}(x^{(j)}; D_{\mathsf{s}}) \qquad \text{(B.3.36)}$$

for some collection of independent random variables $\boldsymbol{\xi} = (\xi_{\mathsf{s}})_{\mathsf{s}\in\mathcal{S}_{n,b}}$ having the same distribution as ξ. The statistics $\hat{U}_n(x^{(j)})$ and $\overline{U}_n(x^{(j)})$ are the complete, randomized and de-randomized, U-statistics associated with the randomized d-dimensional kernel $u(x^{(j)}; \cdot, \cdot)$, respectively. Consider the decomposition

$$\begin{aligned}
&\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(W_n(x^{(j)}) - \frac{b}{n}\sum_{i=1}^{n}\tilde{u}^{(1)}(x^{(j)}; D_i)\right) \\
&\quad = \sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(W_n(x^{(j)}) - \hat{U}_n(x^{(j)})\right) + \sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(\hat{U}_n(x^{(j)}) - \overline{U}_n(x^{(j)})\right) \\
&\qquad + \sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(\overline{U}_n(x^{(j)}) - \frac{b}{n}\sum_{i=1}^{n}\tilde{u}^{(1)}(x^{(j)}; D_i)\right)\ . \qquad \text{(B.3.37)}
\end{aligned}$$

We bound each of the three terms in (B.3.37) in succession.

A higher-order moment bound for the third term in (B.3.37) is obtained by applying the same argument used to verify Lemma B.3.1. In particular, Part (ii) of Lemma C.4.1 and Lemma B.3.1 imply that

$$\mathbb{E}\left[\left|\overline{U}_n(x^{(j)}) - \frac{b}{n}\sum_{i=1}^{n}\tilde{u}^{(1)}(x^{(j)}; D_i)\right|^q\right]^{1/q} \lesssim \phi q^2 \frac{b}{n} \tag{B.3.38}$$

for each $2 \vee 2\log(nd) \leq q \leq cn/b$. Analogous bounds for the first two terms in (B.3.37) are obtained through the application of the following Lemma.

Lemma B.3.3. *Let $\mathbf{D}_n = (D_i)_{i=1}^n$ and $\boldsymbol{\xi} = (\xi_{\mathsf{s}})_{\mathsf{s}\in\mathcal{S}_{n,b}}$ denote two independent collections of independent and identically distributed random variables. Consider the real-valued function $u(D_{\mathsf{s}}, \xi_{\mathsf{s}})$. Assume that the absolute value of each component of $u(D_{\mathsf{s}}, \xi_{\mathsf{s}})$ is bounded by the constant $\phi \geq 1$ almost surely.*

(i) *If the bound*

$$q^{1/2}\frac{\phi b}{n} \leq 1 \tag{B.3.39}$$

holds for some $q \geq 2 \vee 2\log(dn)$, then the bound

$$\mathbb{E}\left[\left|\binom{n}{b}^{-1}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}}\left(u(D_{\mathsf{s}}, \xi_{\mathsf{s}}) - \mathbb{E}\left[u(D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mid \mathbf{D}_{\mathsf{s}}\right]\right)\right|^q\right]^{1/q} \lesssim q^{1/2}\frac{\phi b}{n} \tag{B.3.40}$$

holds for all $b > 2$.

(ii) *If, in addition, the sets $(\mathsf{s}_q)_{q=1}^r$ are drawn independently and identically from $\mathcal{S}_{n,b}$ and $n \leq \sqrt{r}b$, then*

$$\mathbb{E}\left[\left|\frac{1}{r}\sum_{q=1}^{r}\left(u(D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) - \mathbb{E}\left[u(D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) \mid \mathbf{D}_n, \boldsymbol{\xi}\right]\right)\right|^q\right]^{1/q} \lesssim q^{1/2}\frac{b\phi}{n} \tag{B.3.41}$$

holds.

In particular, observe that

$$\hat{U}_n(x^{(j)}) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} u(x^{(j)}; D_{\mathsf{s}}, \xi_{\mathsf{s}}) = \mathbb{E}[u(x^{(j)}; D_{\mathsf{s}_g}, \xi_{\mathsf{s}_g}) \mid \mathbf{D}_n, \boldsymbol{\xi}]$$

and that therefore we can write

$$W_n(x^{(j)})-\hat{U}_n(x^{(j)}) = \frac{1}{r}\sum_{q=1}^{r} Z_q, \quad \text{with} \quad Z_q = u(x^{(j)}; D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q})-\mathbb{E}[u(x^{(j)}; D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) \mid \mathbf{D}_n, \boldsymbol{\xi}] \,.$$

Consequently, by the normalization $q^{1/2}\phi(\theta_0)b/n \leq 1$, Part (ii) of Lemma B.3.3 implies that

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(W_n(x^{(j)}) - \hat{U}_n(x^{(j)})\right)\right|^q\right]^{1/q} \lesssim \sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}q^{1/2}\frac{b\phi(\theta_0)}{n} = \left(\frac{\phi(\theta_0)^2 q}{\underline{\sigma}_b^2 n}\right)^{1/2} . \tag{B.3.42}$$

In turn, we can similarly write

$$\hat{U}_n(x^{(j)})-\overline{U}_n(x^{(j)}) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} Z_{\mathsf{s}} \quad \text{with} \quad Z_{\mathsf{s}} = u(x^{(j)}; D_{\mathsf{s}}, \xi_{\mathsf{s}})-\mathbb{E}\left[u(x^{(j)}; D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mid D_{\mathsf{s}}\right] \,.$$

Thus, again by the normalization $q^{1/2}\phi(\theta_0)b/n \leq 1$, Part (i) of Lemma B.3.3 implies that

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(\hat{U}_n(x^{(j)}) - \overline{U}_n(x^{(j)})\right)\right|^q\right]^{1/q} \lesssim \left(\frac{\phi(\theta_0)^2 q}{\underline{\sigma}_b^2 n}\right)^{1/2} . \tag{B.3.43}$$

Putting the pieces together, the bounds (B.3.38), (B.3.42), and (B.3.43) imply that

$$\mathbb{E}\left[\left|\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}\left(W_n(x^{(j)}) - \frac{b}{n}\sum_{i=1}^{n}\tilde{u}^{(1)}(x^{(j)}; D_i)\right)\right|^q\right]^{1/q} \lesssim q^2\left(\frac{\phi(\theta_0)^2}{\underline{\sigma}_b^2 n}\right)^{1/2} , \tag{B.3.44}$$

for each $j \in [d]$ and each $2 \vee 2\log(nd) \leq q \leq cn/b$, as required. ■

B.3.3 Proof of Lemma B.3.2

Throughout, we use the shorthand $N_b = \binom{n}{b}$.

B.3.3.1 Part (i)

Observe that

$$\begin{aligned}
\mathbb{E}\left[M_n(x;\theta,g)\right] &= \mathbb{E}\left[\sum_{i=1}^{n} K(x, X_i)m(D_i;\theta,g)\right] \\
&= \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}}\sum_{i\in\mathsf{s}}\mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi)m(D_i;\theta,g)\right]
\end{aligned}$$

$$= \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} \sum_{i \in \mathsf{s}} \mathbb{E}\left[\mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi) \mid X_i, D_{\mathsf{s}_{-i}}\right] \mathbb{E}\left[m(D_i; \theta, g) \mid X_i, D_{\mathsf{s}_{-i}}\right]\right]$$

(Honesty)

$$= \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} \sum_{i \in \mathsf{s}} \mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi) \mathbb{E}\left[m(D_i; \theta, g) \mid X_i\right]\right]$$

$$= \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} \sum_{i \in \mathsf{s}} \mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi) M(X_i; \theta, g)\right] .$$

Therefore, the normalization

$$\sum_{i \in \mathsf{s}} \kappa(x, X_i, \mathsf{s}, \xi) = 1 \tag{B.3.45}$$

implies that

$$\mathsf{Bias}_n(x; \theta, g) = \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} \sum_{i \in \mathsf{s}} \mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi) \left(M(X_i; \theta, g) - M(x; \theta, g)\right)\right] .$$

By the boundedness part of Assumption 2.3.2 and Part (iii) of Assumption 2.3.3, we find that

$$\mathsf{Bias}_n(x; \theta, g) \lesssim (1 + |\theta|) \mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi) \|X_i - x\|_\infty\right] \le (1 + |\theta|) \varepsilon_b , \tag{B.3.46}$$

where the final inequality follows from the definition of the shrinkage ε_b and the normalization (B.3.45). ■

B.3.3.2 Part (ii)

Define the function

$$J(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}) = \sum_{i \in \mathsf{s}} \left(\kappa(x, X_i, \mathsf{s}, \xi_{\mathsf{s}}) m^{(1)}(D_i; g_0) - \mathbb{E}\left[\kappa(x, X_i, \mathsf{s}, \xi_{\mathsf{s}}) m^{(1)}(D_i; g_0)\right]\right)$$

and observe that

$$\overline{M}_n^{(1)}(x; g_0) = \frac{1}{r} \sum_{q=1}^{r} J(x; D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) . \tag{B.3.47}$$

Consider the decomposition

$$\overline{M}_n^{(1)}(x; g_0) = \tilde{A}(x) + \hat{A}(x) + \overline{A}(x) , \tag{B.3.48}$$

where

$$\tilde{A}(x) = \frac{1}{r}\sum_{q=1}^{r}\left(J(x; D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) - \mathbb{E}\left[J(x; D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) \mid \mathbf{D}_n, \boldsymbol{\xi}\right]\right) , \tag{B.3.49}$$

$$\hat{A}(x) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}}\left(J(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}) - \mathbb{E}\left[J(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mid D_{\mathsf{s}}\right]\right) , \quad \text{and} \tag{B.3.50}$$

$$\overline{A}(x) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}}\mathbb{E}\left[J(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mid D_{\mathsf{s}}\right] , \tag{B.3.51}$$

respectively. We again apply Lemma B.3.3 to bound (B.3.49) and (B.3.50). In particular, by Assumption 2.3.2, the normalization (B.3.16), and the restriction $n \leq b\sqrt{r}$, Lemma B.3.3 implies that, for each $j \in [d]$,

$$\mathbb{E}\left[\left|\tilde{A}(x^{(j)})\right|^q\right]^{1/q} \lesssim q^{1/2}\frac{b\phi}{n} \quad \text{and} \tag{B.3.52}$$

$$\mathbb{E}\left[\left|\hat{A}(x^{(j)})\right|^q\right]^{1/q} \lesssim q^{1/2}\frac{b\phi}{n} \tag{B.3.53}$$

respectively. In turn, to bound the term (B.3.51), write

$$\overline{J}_j(D_{\mathsf{s}}) = \mathbb{E}\left[J(x^{(j)}; D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mid D_{\mathsf{s}}\right] .$$

Observe that $\overline{J}_j(\cdot)$ is symmetric, mean-zero, and bounded by a constant multiple of ϕ. Moreover,

$$\overline{A}(x^{(j)}) = U_{n,b}(\overline{J}_j) .$$

Let $\overline{J}_j^{(\ell)}$ denote the ℓ-th Hoeffding projection of $\overline{J}_j$. By Part (ii) of Lemma C.4.1, we have

$$\overline{A}(x^{(j)}) = \frac{b}{n}\sum_{i=1}^{n}\overline{J}_j^{(1)}(D_i) + \sum_{\ell=2}^{b}\binom{b}{\ell}U_{n,\ell}(\overline{J}_j^{(\ell)}) . \tag{B.3.54}$$

We bound the two terms in (B.3.54) in succession. By Hoeffding orthogonality,

$$b\,\mathbb{E}\left[\left(\overline{J}_j^{(1)}(D_i)\right)^2\right] \leq \mathbb{E}\left[\overline{J}_j(D_{[b]})^2\right] \lesssim \phi^2 .$$

Moreover, $|\overline{J}_j^{(1)}(D_i)| \lesssim \phi$ almost surely. Therefore, Lemma B.2.4 gives

$$\begin{aligned}\mathbb{E}\left[\left|\frac{b}{n}\sum_{i=1}^{n}\overline{J}_j^{(1)}(D_i)\right|^q\right]^{1/q} &\lesssim b\left(\sqrt{\frac{q}{n}\mathbb{E}\left[\left(\overline{J}_j^{(1)}(D_i)\right)^2\right]}+\frac{q}{n}\left\|\max_{i\in[n]}\left|\overline{J}_j^{(1)}(D_i)\right|\right\|_q\right)\\ &\lesssim \phi\sqrt{\frac{bq}{n}}+\phi\frac{bq}{n}\lesssim q\phi\sqrt{\frac{b}{n}}\,. \end{aligned}\tag{B.3.55}$$

Next, Lemma B.3.1 and Minkowski's inequality imply that

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{\ell=2}^{b}\binom{b}{\ell}U_{n,\ell}(\overline{J}_j^{(\ell)})\right|^q\right]^{1/q} &\le \sum_{\ell=2}^{b}\mathbb{E}\left[\left|\binom{b}{\ell}U_{n,\ell}(\overline{J}_j^{(\ell)})\right|^q\right]^{1/q}\\ &\lesssim \phi q\sum_{\ell=2}^{b}\left(\frac{Cqb}{n}\right)^{\ell/2}\lesssim \phi q^2\frac{b}{n}\,. \end{aligned}\tag{B.3.56}$$

where the final inequality follows from the geometric series formula and the fact that $bq/n \le c$, which is implied by (B.3.16). Since $bq^2/n \le c$ is also implied by (B.3.16), the right-hand side of (B.3.56) is bounded by $q\phi\sqrt{b/n}$. Combining (B.3.54), (B.3.55), and (B.3.56), we obtain

$$\mathbb{E}\left[\left|\overline{A}(x^{(j)})\right|^q\right]^{1/q}\lesssim q\phi\sqrt{\frac{b}{n}}\,. \tag{B.3.57}$$

Thus, the bounds (B.3.52), (B.3.53), and (B.3.57) imply that

$$\mathbb{E}\left[\left|\overline{M}_n^{(1)}(x^{(j)};g_0)\right|^q\right]^{1/q}\lesssim q\phi\sqrt{\frac{b}{n}}\,. \tag{B.3.58}$$

Multiplying both sides of (B.3.58) by

$$\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}$$

gives

$$\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}\mathbb{E}\left[\left|\overline{M}_n^{(1)}(x^{(j)};g_0)\right|^q\right]^{1/q}\lesssim q\phi\frac{1}{\underline{\sigma}_b^{1/2}n^{1/4}}=q\left(\frac{\phi^4}{\underline{\sigma}_b^2 n}\right)^{1/4}\,, \tag{B.3.59}$$

as required. ∎

B.3.3.3 Part (iii)

We give the details of the proof of the stated moment bound on the discrepancy

$$\overline{M}_n(x^{(j)}; \theta_0(x^{(j)}), \hat{g}_n) - \overline{M}_n(x^{(j)}; \theta_0(x^{(j)}), g_0) \tag{B.3.60}$$

only. The argument giving the analogous bound associated with the term $\overline{M}_n^{(1)}(x^{(j)}; \cdot)$ is analogous. Define the functions

$$F(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, g) = \sum_{i \in \mathsf{s}} f(x, X_i; D_{\mathsf{s}}, \xi_{\mathsf{s}}, g) \quad \text{and} \quad \overline{F}(x; D_{\mathsf{s}}, g) = \mathbb{E}_{\xi_{\mathsf{s}}}\left[F(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, g)\right] , \tag{B.3.61}$$

where

$$\begin{aligned} f(x, X_i; D_{\mathsf{s}}, \xi_{\mathsf{s}}, g) &= \kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) \left(m(D_i; \theta_0, g) - m(D_i; \theta_0, g_0)\right) \\ &\quad - \mathbb{E}\left[\kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) \left(m(D_i; \theta_0, g) - m(D_i; \theta_0, g_0)\right)\right] \end{aligned} \tag{B.3.62}$$

and the notation $\mathbb{E}_A[\cdot]$ indicates that we are evaluating the expectation over the randomness in A. Define the quantities

$$\tilde{F}_n(x; g) = \frac{1}{r} \sum_{q=1}^{r} \left(F(x; D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}, g) - \mathbb{E}_{\mathsf{s}}\left[F(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, g) \mid \mathbf{D}_n, \boldsymbol{\xi}\right]\right) , \tag{B.3.63}$$

$$\hat{F}_n(x; g) = \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} \left(F(x; D_{\mathsf{s}}, \xi_{\mathsf{s}}, g) - \overline{F}(x; D_{\mathsf{s}}, g)\right) , \quad \text{and} \tag{B.3.64}$$

$$\overline{F}_n(x; g) = \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} \overline{F}(x; D_{\mathsf{s}}, g) , \tag{B.3.65}$$

and consider the decomposition

$$\overline{M}_n(x; \theta_0(x), g) - \overline{M}_n(x; \theta_0(x), g_0) = \tilde{F}_n(x; g) + \hat{F}_n(x; g) + \overline{F}_n(x; g) . \tag{B.3.66}$$

We again apply Lemma B.3.3 to bound (B.3.63) and (B.3.64). In particular, by Assumption 2.3.2, the normalization (B.3.19), and the restriction $n \leq b\sqrt{r}$, Lemma B.3.3 implies that

$$\mathbb{E}\left[\left|\tilde{F}_n(x^{(j)}; \hat{g}_n)\right|^q\right]^{1/q} \lesssim q^{1/2} \frac{b(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi}{n} \quad \text{and} \tag{B.3.67}$$

$$\mathbb{E}\left[\left|\hat{F}_n(x^{(j)}; \hat{g}_n)\right|^q\right]^{1/q} \lesssim q^{1/2} \frac{b(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi}{n} . \tag{B.3.68}$$

Here, we have used the statistical independence of $\hat{g}_n$ and $\mathbf{D}_n$, together with the fact that the preceding bounds hold uniformly over $g \in \mathcal{G}$.

To bound the term (B.3.65), first observe that

$$\mathbb{E}\left[\overline{F}(x; D_{\mathsf{s}}, g)\right] = 0 \quad \text{and} \quad \left|\overline{F}(x; D_{\mathsf{s}}, g)\right| \lesssim (1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi \tag{B.3.69}$$

by the boundedness part of Assumption 2.3.2. In turn, observe that

$$\begin{aligned}
\mathbb{E}\left[\left(\overline{F}(x; D_{\mathsf{s}}, g)\right)^2\right] &\leq \mathbb{E}\left[\sum_{i\in\mathsf{s}} \kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) \mathbb{E}\left[(m(D_i; \theta_0, g) - m(D_i; \theta_0, g_0))^2 \mid X_i\right]\right] \\
&\lesssim \mathbb{E}\left[\sum_{i\in\mathsf{s}} \kappa(x, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) V(x, g)\right] + \varepsilon_b \\
&\lesssim \|g - g_0\|_{2,\infty}^2 + \varepsilon_b \ ,
\end{aligned} \tag{B.3.70}$$

where the first inequality follows from Honesty and Jensen's inequality, the second inequality follows from Assumption 2.3.3, Part (ii), and the definition of the shrinkage rate ε_b, and the third inequality follows from Assumption 2.3.3, Part (ii), and the normalization that $\sum_{i\in\mathsf{s}} \kappa(x, X_i, D_{\mathsf{s}}, \xi) = 1$ almost surely.

Let $\overline{F}^{(\ell)}(x; \cdot, g)$ denote the ℓ-th Hoeffding projection of the kernel $\overline{F}(x; \cdot, g)$. By Part (ii) of Lemma C.4.1, we have

$$\overline{F}_n(x; g) = \frac{b}{n} \sum_{i=1}^n \overline{F}^{(1)}(x; D_i, g) + \sum_{\ell=2}^b \binom{b}{\ell} U_{n,\ell}\left(\overline{F}^{(\ell)}(x; \cdot, g)\right) \ . \tag{B.3.71}$$

By Hoeffding orthogonality and (B.3.70),

$$b\mathbb{E}\left[\left(\overline{F}^{(1)}(x; D_i, g)\right)^2\right] \leq \mathbb{E}\left[\left(\overline{F}(x; D_{\mathsf{s}}, g)\right)^2\right] \lesssim \|g - g_0\|_{2,\infty}^2 + \varepsilon_b \ . \tag{B.3.72}$$

Moreover,

$$\left|\overline{F}^{(1)}(x; D_i, g)\right| \lesssim (1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi$$

almost surely. Hence, Lemma B.2.4 gives

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{b}{n} \sum_{i=1}^n \overline{F}^{(1)}(x^{(j)}; D_i, g)\right|^q\right]^{1/q} \\
&\qquad \lesssim \sqrt{\frac{bq}{n}} \left(\|g - g_0\|_{2,\infty} + \varepsilon_b^{1/2}\right) + \frac{b}{n}(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q \ .
\end{aligned} \tag{B.3.73}$$

Next, Lemma B.3.1 and Minkowski's inequality imply that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{\ell=2}^{b}\binom{b}{\ell}U_{n,\ell}\left(\overline{F}^{(\ell)}(x^{(j)};\cdot,g)\right)\right|^q\right]^{1/q} \\
&\quad\leq\sum_{\ell=2}^{b}\mathbb{E}\left[\left|\binom{b}{\ell}U_{n,\ell}\left(\overline{F}^{(\ell)}(x^{(j)};\cdot,g)\right)\right|^q\right]^{1/q} \\
&\quad\lesssim(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q\sum_{\ell=2}^{b}\left(\frac{Cqb}{n}\right)^{\ell/2}\lesssim(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2\frac{b}{n}\,, \qquad \text{(B.3.74)}
\end{aligned}$$

where the final inequality follows from the geometric series formula and (B.3.19). Combining (B.3.71), (B.3.73), and (B.3.74), we find that, for each fixed $g\in\mathcal{G}$,

$$\mathbb{E}\left[\left|\overline{F}_n(x^{(j)};g)\right|^q\right]^{1/q}\lesssim\sqrt{\frac{bq}{n}}\left(\|g-g_0\|_{2,\infty}+\varepsilon_b^{1/2}\right)+\frac{b}{n}(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2\,. \qquad \text{(B.3.75)}$$

Consequently, by the elementary bound

$$\sqrt{\frac{bq}{n}}\varepsilon_b^{1/2}\leq\frac{1}{2}\frac{bq}{n}+\frac{1}{2}\varepsilon_b$$

and the fact that $\hat{g}_n$ is statistically independent of $\mathbf{D}_n$, the rate condition on $\hat{g}_n$ gives

$$\begin{aligned}
\mathbb{E}\left[\left|\overline{F}_n(x^{(j)};\hat{g}_n)\right|^q\right]^{1/q}&\lesssim q^{3/2}\sqrt{\frac{b}{n}}\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+\sqrt{\frac{bq}{n}}\varepsilon_b^{1/2} \\
&\qquad+\frac{b}{n}(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2 \\
&\lesssim q^{3/2}\sqrt{\frac{b}{n}}\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+\varepsilon_b+\frac{b}{n}(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2\,.
\end{aligned} \qquad \text{(B.3.76)}$$

Putting the pieces together, the decomposition (B.3.66) and the bounds (B.3.67), (B.3.68), and (B.3.76) imply that

$$\begin{aligned}
&\mathbb{E}\left[\left|\overline{M}_n(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n)-\overline{M}_n(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q} \\
&\quad\lesssim q^{3/2}\sqrt{\frac{b}{n}}\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+\varepsilon_b+\frac{b}{n}(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2\,, \qquad \text{(B.3.77)}
\end{aligned}$$

where we have absorbed the bounds (B.3.67) and (B.3.68) into the final term. By the normalization (B.3.19), we have

$$q^{1/2}\sqrt{\frac{b}{n}} \lesssim 1 \quad \text{and} \quad q(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}} \lesssim 1\,.$$

Therefore,

$$\mathbb{E}\left[\left|\overline{M}_n(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n) - \overline{M}_n(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q}$$
$$\lesssim q\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + \varepsilon_b + (1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\right)\,. \tag{B.3.78}$$

Multiplying both sides by

$$\sqrt{\frac{n}{b^2\underline{\sigma}_b^2}}$$

gives (B.3.17). ■

B.3.3.4 Part (iv)

Recall the decomposition

$$M^{(1)}(x;\theta_0(x),g_0)(\hat{\theta}_n(x)-\theta_0(x)) \tag{B.3.79}$$
$$= -\overline{M}_n(x;\theta_0(x),g_0) \tag{B.3.80}$$
$$+\,\mathsf{Bias}_n(x;\hat{\theta}_n(x),\hat{g}_n) + \mathsf{Nuis}(x;\theta_0(x),\hat{g}_n) \tag{B.3.81}$$
$$+\,\mathsf{Stoch}^{(1)}(x;\hat{\theta}_n(x),\hat{g}_n) + \mathsf{Stoch}^{(2)}(x;\hat{g}_n) \tag{B.3.82}$$
$$-\,(\hat{\theta}_n(x)-\theta_0(x))\left(M^{(1)}(x;\theta_0(x),\hat{g}_n) - M^{(1)}(x;\theta_0(x),g_0)\right) \tag{B.3.83}$$

stated as (B.2.68) in the Proof of Theorem B.2.2, Part (i), where we note that the various terms appearing in (B.3.79) are defined in (B.2.69) through (B.2.72). Here, we have used the fact that the Hessian $H(x;\theta,g)=0$ almost surely, by Assumption 2.3.2. We also use Part (iv) of Assumption 2.3.3, which implies that $\hat{\theta}_n(x)$ and $\theta_0(x)$ are uniformly bounded.

First, observe that, by an argument identical to the argument used to establish Part (ii) of this Lemma, i.e., (B.3.58), we have that

$$\mathbb{E}\left[\left|\overline{M}_n(x^{(j)};\theta_0(x^{(j)}),g_0)\right|^q\right]^{1/q} \lesssim q(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\,. \tag{B.3.84}$$

Moreover, as before, Parts (iii) and (iv) of Assumption 2.3.3 imply that

$$\mathbb{E}\left[\left|(\hat{\theta}_n(x^{(j)}) - \theta_0(x^{(j)}))\left(M^{(1)}(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n) - M^{(1)}(x^{(j)};\theta_0(x^{(j)}),g_0)\right)\right|^q\right]^{1/q} \lesssim \mathbb{E}\left[\|\hat{g}_n - g_0\|_{2,\infty}^q\right]^{1/q} \lesssim q\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} \,. \tag{B.3.85}$$

In turn, observe that Part (i) of this Lemma and Part (iv) of Assumption 2.3.3 give that

$$\left|\mathsf{Bias}_n(x^{(j)};\hat{\theta}_n(x^{(j)}),\hat{g}_n)\right| \lesssim (1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\varepsilon_b \,. \tag{B.3.86}$$

and that Neyman orthogonality gives

$$\mathbb{E}\left[\left|\mathsf{Nuis}(x^{(j)};\theta_0(x^{(j)}),\hat{g}_n)\right|^q\right]^{1/q} \lesssim \mathbb{E}\left[\|\hat{g}_n - g_0\|_{2,\infty}^{2q}\right]^{1/q} \lesssim q^2\sqrt{\frac{b}{n}}\delta_{n,g}^2 \lesssim q\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} \,, \tag{B.3.87}$$

where the final inequality follows from the normalization $q(b/n)^{1/4}\delta_{n,g} \leq 1$.

Moreover, as before, Assumption B.2.2 and a Taylor expansion gives

$$\begin{aligned} |\mathsf{Stoch}^{(1)}(x;\hat{\theta}_n(x),\hat{g}_n)| &= |(\hat{\theta}_n(x) - \theta_0(x))\overline{M}_n^{(1)}(x;\theta_0(x),\hat{g}_n)| \\ &\leq |\hat{\theta}_n(x) - \theta_0(x)||\overline{M}_n^{(1)}(x;\theta_0(x),g_0)| \\ &\quad + |\hat{\theta}_n(x) - \theta_0(x)||\overline{M}_n^{(1)}(x;\theta_0(x),\hat{g}_n) - \overline{M}_n^{(1)}(x;\theta_0(x),g_0)| \,, \end{aligned} \tag{B.3.88}$$

where we have used the fact that the empirical Hessian $\overline{H}_n(x;\tilde{\theta}_0(x),\hat{g}_n) = 0$ almost surely, by Assumption 2.3.2. Consequently, Part (iv) of Assumption 2.3.3, Part (ii) of this Lemma, and Part (iii) of this Lemma imply that

$$\begin{aligned} &\mathbb{E}\left[\left|\mathsf{Stoch}^{(1)}(x^{(j)};\hat{\theta}_n(x^{(j)}),\hat{g}_n)\right|^q\right]^{1/q} \\ &\quad\lesssim q(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}} \\ &\qquad + q^{3/2}\sqrt{\frac{b}{n}}\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + \varepsilon_b + \frac{b}{n}(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2 \\ &\quad\lesssim q(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}} + q\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + q(1 + \|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\varepsilon_b \,. \end{aligned} \tag{B.3.89}$$

Here, the final inequality follows from the normalizations (B.3.21). Similarly, Part (iii) of this Lemma gives

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathsf{Stoch}^{(2)}(x^{(j)};\hat{g}_n)\right|^q\right]^{1/q} \\
&\quad\lesssim q^{3/2}\sqrt{\frac{b}{n}}\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+\varepsilon_b+\frac{b}{n}(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi q^2 \\
&\quad\lesssim q\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+q(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\varepsilon_b+q(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\,. \qquad \text{(B.3.90)}
\end{aligned}$$

Again, the final inequality follows from the normalization (B.3.21).

Consequently, plugging the bounds (B.3.84), (B.3.85), (B.3.86), (B.3.87), (B.3.89), and (B.3.90) into the decomposition (B.3.79), applying Minkowski's inequality, and using the fact that $M^{(1)}(x;\theta_0(x),g_0)$ is bounded from below by Part (iii) of Assumption 2.3.3, we find that

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{\theta}_n(x^{(j)})-\theta_0(x^{(j)})\right|^q\right]^{1/q} \\
&\quad\lesssim q\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+q(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\varepsilon_b+q(1+\|\theta_0(\boldsymbol{x}^{(d)})\|_\infty)\phi\sqrt{\frac{b}{n}}\,. \qquad \text{(B.3.91)}
\end{aligned}$$

Multiplying both sides of (B.3.91) by

$$\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}$$

gives

$$\left(\frac{n}{b^2\underline{\sigma}_b^2}\right)^{1/4}\mathbb{E}\left[\left|\hat{\theta}_n(x^{(j)})-\theta_0(x^{(j)})\right|^q\right]^{1/q}\lesssim q\delta_{n,\theta}\,, \qquad \text{(B.3.92)}$$

as required. ■

B.3.4 Proof of Lemma B.3.3

Throughout, we use the shorthand $N_b=\binom{n}{b}$. We begin by considering the quantity

$$\frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}}Z_{\mathsf{s}},\quad\text{where}\quad Z_{\mathsf{s}}=u(D_{\mathsf{s}},\xi_{\mathsf{s}})-\mathbb{E}\left[u(D_{\mathsf{s}},\xi_{\mathsf{s}})\mid\mathbf{D}_{\mathsf{s}}\right]\,. \qquad \text{(B.3.93)}$$

Conditioned on the data $\mathbf{D}_n$, the observations Z_{s}, $\mathsf{s} \in \mathcal{S}_{n,b}$, are centered and mutually independent. Moreover, each component of Z_{s} is bounded by 2ϕ almost surely. Hence, by Lemma B.2.4, we obtain

$$\mathbb{E}\left[\left|\frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} Z_{\mathsf{s},j}\right|^q \mid \mathbf{D}_n\right]^{1/q} \lesssim \phi\left(\sqrt{\frac{q}{N_b}} + \frac{q}{N_b}\right) . \tag{B.3.94}$$

As $N_b^{-1/2} \lesssim b/n$ for $b > 2$, we have that

$$\phi\left(\sqrt{\frac{q}{N_b}} + \frac{q}{N_b}\right) \lesssim \frac{b\phi q^{1/2}}{n} + \phi q\frac{b^2}{n^2} \lesssim \frac{b\phi q^{1/2}}{n} , \tag{B.3.95}$$

where the last inequality follows from the normalization $\phi b q^{1/2}/n \leq 1$ and the fact that $\phi \geq 1$. Taking expectations over $\mathbf{D}_n$ verifies (B.3.40).

Next, consider the quantity

$$\frac{1}{r}\sum_{q=1}^{r} Z_q, \quad \text{where} \quad Z_q = u(D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) - \mathbb{E}\left[u(D_{\mathsf{s}_q}, \xi_{\mathsf{s}_q}) \mid \mathbf{D}_n, \boldsymbol{\xi}\right] . \tag{B.3.96}$$

Conditioned on $\mathbf{D}_n$ and $\boldsymbol{\xi}$, the observations Z_q, $q \in [r]$, are centered and mutually independent. Moreover, each component of Z_q is bounded by 2ϕ almost surely. Thus, the same argument used above gives

$$\mathbb{E}\left[\left|\frac{1}{r}\sum_{q=1}^{r} Z_q\right|^q \mid \mathbf{D}_n, \boldsymbol{\xi}\right]^{1/q} \lesssim \phi\left(\sqrt{\frac{q}{r}} + \frac{q}{r}\right) . \tag{B.3.97}$$

Now, the restriction $n \leq \sqrt{r}b$ implies that

$$\phi\left(\sqrt{\frac{q}{r}} + \frac{q}{r}\right) \lesssim \frac{b\phi q^{1/2}}{n} + \phi q\frac{b^2}{n^2} \lesssim \frac{b\phi q^{1/2}}{n} , \tag{B.3.98}$$

where the final inequality again follows from the normalization $\phi b q^{1/2}/n \leq 1$ and the fact that $\phi \geq 1$. Taking expectations over $\mathbf{D}_n$ and $\boldsymbol{\xi}$ verifies (B.3.41). ■

B.4. Additional Results and Discussion

B.4.1 Stochastic Equicontinuity without Sample Splitting

In the main text, we assume that the nuisance parameter estimator $\hat{g}_n$ is computed on a separate sample, independent of the data $\mathbf{D}_n$. In this section, we give additional conditions under which the nuisance parameter estimator $\hat{g}_n$ can be computed on the same data used to construct the estimator $\hat{\theta}_n(\boldsymbol{x}^{(d)})$. Our analysis builds on a similar result given in Chen et al. (2022), who give a pointwise analysis of unconditional moment estimators. By contrast, we give a high-dimensional analysis of conditional moment estimators constructed with subsampled kernels.

In the proof of Theorem 2.3.1, the statistical independence of the estimator $\hat{g}_n$ and the data $\mathbf{D}_n$ is only needed in the proof of Lemma B.3.2, Part (iii). Thus, our objective is to prove an analogous result, without imposing this restriction. In particular, consider the functions

$$a(x; D_i, D_{\mathsf{s}}, g) = \kappa(x, X_i; D_{\mathsf{s}})m(D_i, g) \quad \text{and} \quad A(x; g) = \mathbb{E}\left[\kappa(x, X_i; D_{\mathsf{s}})m(D_i, g)\right],$$

where $\kappa(\cdot, \cdot; D_{\mathsf{s}})$ is a kernel function and $m(D_i, g)$ is a moment function. We are interested in the complete, deterministic, U-statistic

$$\binom{n}{b}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} F(\boldsymbol{x}^{(d)}; D_{\mathsf{s}}, \hat{g}_n) \tag{B.4.1}$$

where the kernel function $F(x; D_{\mathsf{s}}, g)$ is given by

$$\begin{aligned} F(x; D_{\mathsf{s}}, g) &= \sum_{i \in \mathsf{s}} f(x; D_i, D_{\mathsf{s}}, g), \\ f(x; D_i, D_{\mathsf{s}}, g) &= a(x; D_i, D_{\mathsf{s}}, g) - a(x; D_i, D_{\mathsf{s}}, g_0) - \left(A(x; g) - \mathbb{E}\left[A(x; g_0)\right]\right), \end{aligned}$$

and $\hat{g}_n$ is some estimator of the nuisance parameter g_0 computed with the data $\mathbf{D}_n$. That is, for the sake of simplicity, we have restricted our attention to complete, deterministic subsampled kernel estimators. Incomplete, random, subsampled kernel estimators can be accommodated through the same methods applied repeatedly throughout the arguments supporting Theorem 2.3.1.

B.4.1.1 Regularity and Stability

To state our additional conditions, we require some additional notation. Let $\mathbf{D}'_n = (D'_i)_{i=1}^n$ denote an independent copy of $\mathbf{D}_n$. Let $\hat{g}_n^{(-\mathsf{s})}$ denote a version of the estimator $\hat{g}_n$ formed with all observations in $\mathbf{D}_n$, except that the observations D_{s} are replaced by the observations D'_{s}. Define the norm

$$\|g - \tilde{g}\|_{2q,\infty} = \max_{k\in[h]} \max_{j\in[d]} \left(\mathbb{E}\left[\left(g^{(k)}(Z_i) - \tilde{g}^{(k)}(Z_i) \right)^{2q} \mid X_i = x^{(j)} \right] \right)^{1/2q}$$

for any nuisance parameters $g = (g^{(k)})_{k=1}^h$ and $\tilde{g} = (\tilde{g}^{(k)})_{k=1}^h$ in the space $\mathcal{G}$ and any positive integer $q > 0$.

First, we require that moment $m(D_i, g)$ satisfies a pair of smoothness conditions.

Assumption B.4.1. *Define the higher-order variogram*

$$V^{(q)}(x; g, g') = \mathbb{E}\left[(m(D_i, g) - m(D_i, g'))^{2q} \mid X_i = x \right]$$

for each g and g' in $\mathcal{G}$. For each integer $q > 0$, the Lipschitz condition

$$|V^{(q)}(x; g, g') - V^{(q)}(x'; g, g')| \lesssim \|x - x'\|_\infty^q \tag{B.4.2}$$

holds for each x and x' in $\mathcal{X}$ and the mean-square continuity condition

$$\sup_{j\in[d]} V^{(q)}(x^{(j)}; g, g') \lesssim \|g - g'\|_{2q,\infty}^{2q} \tag{B.4.3}$$

holds for each g and g' in $\mathcal{G}$.

Assumption B.4.1 is a higher-order analogue to Part (ii) of Assumption 2.3.3.

Second, we impose the following higher-order shrinkage condition.

Assumption B.4.2. *Fix a sequence ε_b. The kernel $\kappa(\cdot, \cdot, D_{\mathsf{s}}, \xi_{\mathsf{s}})$ satisfies the higher-order shrinkage bound*

$$\sup_{P\in\mathbf{P}} \sup_{j\in[d]} \mathbb{E}\left[\max\left\{ \|X_i - x^{(j)}\|_\infty : \kappa(x^{(j)}, X_i, D_{\mathsf{s}}, \xi_{\mathsf{s}}) > 0 \right\}^q \right]^{1/q} \le \varepsilon_b \,. \tag{B.4.4}$$

each integer $q \ge 2 \vee 2\log(nd)$.

Third, we impose the following higher-order moment stability conditions.

Assumption B.4.3. *There exists some increasing sequence $\gamma(q)$, which may depend on the parameters n, b, and d, such that the L_q-norm stability bounds*

$$\|m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})\|_q \lesssim q\gamma(q)\frac{b}{n} \quad \text{and} \tag{B.4.5}$$

$$\|m(D_1'', \hat{g}_n) - m(D_1'', \hat{g}_n^{(-[b])})\|_q \lesssim q\gamma(q)\frac{b}{n} \tag{B.4.6}$$

hold, where D_1'' is an independent copy of D_1.

Assumption B.4.3 quantifies the sensitivity of the moment $m(D_1, \hat{g}_n)$ to re-sampling b elements of the data $\mathbf{D}_n$. Analogous conditions are studied in Abou-Moustafa and Szepesvári (2019) and Chen et al. (2022). We discuss moments and estimators that satisfy Assumption B.4.3 in **??** B.4.1.3.

B.4.1.2 Stochastic Equicontinuity

The following Theorem gives a suitable bound for the U-statistic (B.4.1).

Theorem B.4.1. *Suppose that the kernel $\kappa(x, X_i; D_{\mathsf{s}})$ satisfies Assumption 2.3.1 and Assumption B.4.2, that the moment function $m(D_i, g)$ satisfies Assumption B.4.1, and that the nuisance parameter estimator satisfies the higher-order moment bound*

$$\sup_{P \in \mathbf{P}} \mathbb{E}\left[\|\hat{g}_n - g_0\|_{2q,\infty}^{2q}\right]^{1/(2q)} \lesssim q\left(\frac{b}{n}\right)^{1/4} \delta_{n,g} \tag{B.4.7}$$

for some sequence $\delta_{n,g}$. If Assumption B.4.3 holds, then the bound

$$\mathbb{E}\left[\left|\frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} F(x^{(j)}; D_{\mathsf{s}}, \hat{g}_n)\right|^q\right]^{1/q} \lesssim q^{3/2}\sqrt{\frac{b}{n}}\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g} + \varepsilon_b^{1/2} + b^{1/q}\gamma(2q)\right) \tag{B.4.8}$$

holds uniformly over $P \in \mathbf{P}$, for each $j \in [d]$ and each even integer $q \geq 2 \vee 2\log(nd)$.

Remark B.4.1. Compare the bound (B.4.8) to the rate given in Lemma B.3.2, Part (iii). Taking $q = \overline{q}_{n,d} \asymp \log(dn)$ and writing

$$\gamma_{n,b}^{\star} = b^{1/\overline{q}_{n,d}}\gamma(2\overline{q}_{n,d}) \ ,$$

the bound (B.4.8) gives the same-sample analogue of the stochastic equicontinuity rate

$$\sqrt{\frac{b}{n}}\log^{3/2}(dn)\left(\left(\frac{b}{n}\right)^{1/4}\delta_{n,g}+\varepsilon_b^{1/2}+\gamma^{\star}_{n,b}\right)\,.$$

The first term is comparable to the corresponding term in Lemma B.3.2, Part (iii), up to the additional $\log^{1/2}(dn)$ factor generated by the generalized Efron–Stein inequality. Moreover, the elementary inequality

$$\sqrt{\frac{b\log(dn)}{n}}\varepsilon_b^{1/2}\leq\frac{b\log(dn)}{n}+\varepsilon_b$$

allows the shrinkage term to be compared directly with the ε_b term appearing in Lemma B.3.2, Part (iii), under the same small-order normalizations used in the proof of Theorem 2.3.1. Thus, relative to the sample-split argument, the principal additional contribution is the stability term

$$\sqrt{\frac{b}{n}}\log^{3/2}(dn)\gamma^{\star}_{n,b}\,.$$

Consequently, under the conditions of Theorem B.4.1, if this stability contribution is of smaller order than the stochastic equicontinuity rate in Lemma B.3.2, Part (iii), a result analogous to Theorem 2.3.1 will hold without sample splitting. We discuss conditions under which $\gamma^{\star}_{n,b}$ decreases in **??** B.4.1.3. ■

Remark B.4.2. Theorem B.4.1 follows from an argument similar to the argument developed in Chen et al. (2022). There is one important difference. Chen et al. (2022) apply a "double-centering trick," due to Kumar et al. (2013), to get a variance bound. To obtain a high-dimensional bound, with a logarithmic dependence on the dimension d, we replace this step with an application of a generalized Efron-Stein inequality, due to Boucheron et al. (2005). ■

B.4.1.3 Verifying Theorem B.4.1

In this section, we study conditions under which Assumption B.4.3 is satisfied. First, we impose an additional smoothness condition on the moment function $m(D_i, g)$. Throughout, we write the nuisance parameter estimator $\hat{g}_n$ as $\hat{g}_n = (\hat{g}_{k,n})_{k=1}^{h}$.

Assumption B.4.4. *The inequalities*

$$\|m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})\|_{2q} \lesssim \max_{k\in[h]} \mathbb{E}\left[\sup_{z\in\mathcal{Z}} \left|\hat{g}_{k,n}(z) - \hat{g}_{k,n}^{(-[b])}(z)\right|^{2q}\right]^{1/(2q)} \tag{B.4.9}$$

and

$$\|m(D_i'', \hat{g}_n) - m(D_i'', \hat{g}_n^{(-[b])})\|_{2q} \lesssim \max_{k\in[h]} \mathbb{E}\left[\left|\hat{g}_{k,n}(Z_i'') - \hat{g}_{k,n}^{(-[b])}(Z_i'')\right|^{2q}\right]^{1/(2q)} \tag{B.4.10}$$

hold for any positive integer q.

The bound (B.4.9) stipulates that the higher-order moments of $m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})$ are smaller than the supremum of the higher-order moments of $\hat{g}_{k,n}(z) - \hat{g}_{k,n}^{(-[b])}(z)$ over the space $\mathcal{Z}$. The bound (B.4.10) is weaker, as we only need to control the distance $\hat{g}_{k,n}(z) - \hat{g}_{k,n}^{(-[b])}(z)$ at $z = Z_i'$.

Bounds of the form specified by Assumption B.4.4 are satisfied by many standard choices of moment function. For example, consider the moment function

$$m(D_i, g) = \mu(1, Z_i) - \mu(0, Z_i) + \beta(W_i, Z_i)(Y_i - \mu(W_i, Z_i)) \tag{B.4.11}$$

where the nuisance function g collects the moment parameters $g = (\mu, \beta)$. Observe that (B.4.11) is analogous to the moment function (2.3.4) used as a running example in the main text.

Lemma B.4.1. *If the quantities*

$$|\hat{\beta}_n(W_i, Z_i)| \quad \textit{and} \quad |Y_i - \hat{\mu}_n(W_i, Z_i)| \tag{B.4.12}$$

are bounded almost surely, then it holds that

$$\begin{aligned}
&\|m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})\|_{2p} \\
&\lesssim \max\Bigg\{ \max_{w\in\{0,1\}} \left\{\|\hat{\mu}_n(w, Z_i) - \hat{\mu}_n^{(-[b])}(w, Z_i)\|_{2p}\right\}, \\
&\qquad \|\hat{\mu}_n(W_i, Z_i) - \hat{\mu}_n^{(-[b])}(W_i, Z_i)\|_{2p}, \|\hat{\beta}_n(W_i, Z_i) - \hat{\beta}_n^{(-[b])}(W_i, Z_i)\|_{2p}\Bigg\}, \quad \textit{and}
\end{aligned} \tag{B.4.13}$$

$$\begin{aligned}
&\|m(D_i'', \hat{g}_n) - m(D_i'', \hat{g}_n^{(-[b])})\|_{2p} \\
&\lesssim \max\Bigg\{ \max_{w\in\{0,1\}} \left\{\|\hat{\mu}_n(w, Z_i'') - \hat{\mu}_n^{(-[b])}(w, Z_i'')\|_{2p}\right\},
\end{aligned} \tag{B.4.14}$$

$$\|\hat{\mu}_n(W_i, Z_i'') - \hat{\mu}_n^{(-[b])}(w, Z_i'')\|_{2p}, \|\hat{\beta}_n(W_i, Z_i'') - \hat{\beta}_n^{(-[b])}(W_i, Z_i'')\|_{2p}\Big\},$$

respectively, where D_i'' is an independent copy of D_i and Z_i'' is an independent copy of Z_i.

Now, we consider the higher-order stability of the nuisance parameter estimator $\hat{g}_n(z)$. For the sake of simplicity, we assume that $\hat{g}_n(z)$ is scalar valued and takes the form of a complete, deterministic, U-statistic

$$\hat{g}_n(z) = \binom{n}{b'}^{-1} \sum_{\mathsf{s} \in \mathcal{S}_{n,b'}} u_z(D_{\mathsf{s}}), \tag{B.4.15}$$

where $u_z(D_{\mathsf{s}})$ is some deterministic kernel function of order b'. Again, let $\hat{g}_n^{(-[b])}(z)$ be constructed analogously to $\hat{g}_n(z)$ using the $\mathbf{D}_n$, but replacing $D_{[b]}$ with an independent copy $D'_{[b]}$.

Lemma B.4.2. *Suppose that the kernel function $u_z(D_{\mathsf{s}})$ satisfies the bound $|u_z(D_{\mathsf{s}})| \leq \phi$ almost surely for each z in $\mathcal{Z}$. If the estimator $\hat{g}_n(z)$ is given by (B.4.15), then it holds that*

$$\|\hat{g}_n(z) - \hat{g}_n^{(-[b])}(z)\|_q \lesssim \frac{\sqrt{b'b}}{n} q\phi + \frac{b'}{n} q^2 \phi. \tag{B.4.16}$$

for each z in $\mathcal{Z}$ and $2\log(dn) \leq q \leq cn/b'$

Remark B.4.3. Under Assumption B.4.4, Lemma B.4.2 implies that

$$\|m(D_1'', \hat{g}_n) - m(D_1'', \hat{g}_n^{(-[b])})\|_q \lesssim \frac{\sqrt{b'b}}{n} q\phi + \frac{b'}{n} q^2 \phi. \tag{B.4.17}$$

Thus, relative to Assumption B.4.3, we can take

$$\gamma(q) \lesssim \phi \left(\sqrt{\frac{b'}{b}} + q\frac{b'}{b} \right).$$

Consequently, at the choice $q \asymp \log(dn)$, invoked in Theorem B.4.1, this sequence is small provided

$$\frac{b'}{b} \to 0 \quad \text{and} \quad \log(dn)\frac{b'}{b} \to 0\,.$$

In particular, Lemma B.4.2 suggests that the stability contribution in (B.4.8) is negligible if b' is sufficiently small relative to b. ■

Remark B.4.4. Assumption B.4.3 entails two stability bounds, (B.4.5) and (B.4.6). Under Assumption B.4.4, Lemma B.4.2 quantifies the sequence $\gamma(q)$ introduced on the right-hand side of (B.4.5). By the discussion above, if $b' = o(b)$, this sequence is sufficiently small for a result analogous to Lemma B.3.2, Part (iii) to hold. Handling the bound (B.4.6) requires developing a more refined argument. In particular, under Assumption B.4.4, to bound the right-hand-side of (B.4.5), we must account for the supremum over the space $\mathcal{Z}$ within the expectation. Such a generalization should be reasonably straightforward, potentially through a chaining argument. Roughly speaking, we should expect the supremum to contribute a term like $\log(p)$ to the stability, where p is the dimension of $\mathcal{Z}$. In this case, in settings where p is not high-dimensional, this would not generate any issues. This should be contrasted with analogous suprema taken over the parameter space of a decision tree (which are generated if stochastic equicontinuity is controlled with a union bound). Here the dimension of the parameter space is roughly $2^{\text{Depth of Tree}}$, which may be large enough to make a material difference. However, in order to operationalize this intuition, we would need to place further restrictions sufficient for the smoothness of the Hájek projection of $\hat{g}_n(z)$. We leave this for future work. ■

B.4.2 Cross-Fitting

Theorem 2.3.1 holds under the assumption that the nuisance parameter estimator $\hat{g}_n$ is computed on a data set that is statistically independent of the data used to construct the confidence region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$. This condition can be achieved by splitting the data into two parts, at the cost of reducing statistical precision and introducing randomness independent of the observed data. In Appendix B.4.1, we introduce further restrictions that allow the nuisance parameter estimator $\hat{g}_n$ and the confidence region $\hat{\mathcal{C}}(\boldsymbol{x}^{(d)})$ to be computed using the same data. However, in practice, these conditions are more stringent, and it may be unclear whether they are satisfied for a given nuisance parameter estimator $\hat{g}_n$.

In this section, we detail an alternative procedure for constructing a confidence region similar to the confidence region specified in Definition 2.3.1, whose validity will hold under the same conditions imposed in Theorem 2.3.1. The procedure is based on a cross-fit estimator. To introduce this estimator, we require some additional notation. Let $\mathcal{R}_{n,k}$ denote the set of partitions of $[n]$ into k equally sized and mutually exclusive subsets. That is, for each $\mathsf{r} = (\mathsf{s}_1, \ldots, \mathsf{s}_k)$ in $\mathcal{R}_{n,k}$, the sets $\mathsf{s}_1, \ldots, \mathsf{s}_k$ are mutually exclusive, have union equal to $[n]$, and are each of size n/k. Throughout, we set $q = n/k$ and let $\tilde{\mathsf{s}}$ denote the complement of the set s in $[n]$. Finally, for each subset s of $[n]$, let $\hat{g}_{\mathsf{s}}$ denote a version of the nuisance parameter estimator $\hat{g}_n$ computed with the data D_s.

The cross-fit estimator is computed as follows. For each subset s in $\mathcal{S}_{n,q}$, let the estimator $\hat{\theta}_\mathsf{s}(\boldsymbol{x}^{(d)})$ be constructed by solving the empirical conditional moment equality (2.3.2) using the data D_s and the nuisance parameter estimator $\hat{g}_{\tilde{\mathsf{s}}}$. That is, the nuisance parameter estimator is computed on the sample of data whose indices are not in s. The k-fold cross-fit estimator is given by

$$\hat{\theta}_\mathsf{r}(\boldsymbol{x}^{(d)}) = \frac{1}{k}\sum_{\mathsf{s}\in\mathsf{r}} \hat{\theta}_\mathsf{s}(\boldsymbol{x}^{(d)}) , \tag{B.4.18}$$

where r denotes a random element of $\mathcal{R}_{n,k}$.

It is somewhat unclear how to implement the half-sample bootstrap with the estimator $\hat{\theta}_\mathsf{r}(\boldsymbol{x}^{(d)})$. We propose the following computationally efficient variant. Let $\mathcal{H}(\mathsf{s})$ denote the set of half-samples of the set s, i.e., the set of subsets of s that contain exactly half of its elements. We say that the collection $\mathsf{H} = (\mathsf{h}_\ell)_{\ell=1}^k$ is a half-sample of the partition $\mathsf{r} = (\mathsf{s}_\ell)_{\ell=1}^k$ if each h_ℓ is an element of $\mathcal{H}(\mathsf{s}_\ell)$, i.e., if each h_ℓ is a half-sample of s_ℓ. For each ℓ in $[k]$, let the estimator $\hat{\theta}_{\mathsf{h}_\ell}(\boldsymbol{x}^{(d)})$ be constructed by solving the empirical conditional moment equality (2.3.2) using the data D_{h_ℓ} and the nuisance parameter estimator $\hat{g}_{\tilde{\mathsf{s}}_\ell}$, i.e., using the same nuisance parameter estimates used to construct (B.4.18). The half-sample k-fold cross-split estimator is given by

$$\hat{\theta}_\mathsf{H}(\boldsymbol{x}^{(d)}) = \frac{1}{k}\sum_{\mathsf{h}\in\mathsf{H}} \hat{\theta}_\mathsf{h}(\boldsymbol{x}^{(d)}) . \tag{B.4.19}$$

In this way, once the k-fold cross-fit estimator (B.4.18) has been computed, the nuisance parameter estimates $\hat{g}_{\tilde{\mathsf{s}}_1}, \ldots, \hat{g}_{\tilde{\mathsf{s}}_k}$ do not need to be recomputed to construct the half-sample k-fold cross-split estimator (B.4.19). Simultaneous confidence intervals can then be constructed analogously to the intervals introduced in Definition 2.3.1 by using the half-sample k-fold cross-split bootstrap root

$$R_n^*(\boldsymbol{x}^{(d)}) = \hat{\theta}_\mathsf{H}(\boldsymbol{x}^{(d)}) - \hat{\theta}_\mathsf{r}(\boldsymbol{x}^{(d)}) \tag{B.4.20}$$

in place of the half-sample bootstrap root (2.3.8). In other words, when implementing the half-sample bootstrap based on an estimator constructed with k-fold cross-fitting, one can avoid recomputing nuisance parameters if the half-samples are "stratified" across the k-folds. The error bound given in Theorem 2.3.1 will generalize to the confidence region based on the bootstrap root (B.4.20) through a straightforward argument, so long as k is bounded.

B.4.3 Data

In this appendix, we document our treatment of the Banerjee et al. (2015) data.[1] The data from the graduation program implemented in Pakistan are considered in Chen and Ritzwoller (2023). Here, we consider the data from the graduation program implemented in Ghana, as it has a larger sample size.

The data record measurements of many pre-treatment and post-treatment outcomes for 2,606 individuals in the northern region of Ghana. Baseline survey measurements were made prior to the allocation of treatment. A multifaceted treatment was randomly allocated to 632 of the individuals. Banerjee et al. (2015) consider data from two endline surveys, made two years and three years after the initial asset transfer, respectively. For the purpose of this paper, we consider only data from the baseline survey, records of the treatment allocation, and measurements from the first endline survey. We omit data from 164 attrited individuals, none of whom were assigned to the treatment.

The covariate vector Z_i is composed of measurements of 16 pre-treatment outcomes. Four of these outcomes are associated with consumption: total monthly consumption and total monthly consumption on food, non-food, and durable commodities. Each consumption variable is measured in 2014 US dollars. We transform total monthly consumption to logs, base 10. Three of the variables are associated with assets, each of which is an index constructed from survey data measuring total assets, total productive assets, and total household assets. We transform the total assets measurement to logs, base 10. Five of the outcomes are associated with food security. These consist of four binary variables indicating different aspects of food security, e.g. did a child skip a meal, in addition to an index aggregating these measurements.[2] The final four variables are associated with finance and income: the total amount of outstanding loans, the total amount of savings, income from agriculture, and total income from business. The covariate vector X_i collects the total monthly consumption and assets for each individual. The outcome Y_i measures the total assets two years after the initial asset transfer. Again, we transform these measurements to logs, base 10.

In Figure 2.1 and Figure 2.2, we set the subsample proportion b/n equal to 0.05. In constructing the nuisance parameter estimate, we set $b/n = 0.025$. We use $r = 200$ bootstrap replicates throughout. We use 20,000 trees to construct Figure 2.1 and 2,000 trees in each

[1] The data from Banerjee et al. (2015) were acquired from `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NHIXNT` on September 10, 2021.

[2] There are 2 individuals with missing values for the food security index. We impute these values with the median values of the food security index.

bootstrap replicate to construct Figure 2.2 and throughout the simulation.

B.4.4 Simulation Details

B.4.4.1 Calibration

We calibrate a simulation to the Banerjee et al. (2015) data using a collection of GANs. This approach to simulation design was proposed by Athey et al. (2021). Roughly speaking, a GAN is a pair of neural networks. The objective of the first network, the generator, is to generate data that looks like the Banerjee et al. (2015) data. The objective of the second network, the discriminator, is to discriminate between the real Banerjee et al. (2015) data and the data generated by the generator. These networks compete iteratively until convergence. The idea is that, after convergence, the generator is a good proxy for the true process that generated the Banerjee et al. (2015) data. We use the "WGAN" package associated with Athey et al. (2021).

To calibrate our simulation, we estimate three GANs. The first GAN estimates the distribution of the covariates X_i, i.e., baseline consumption and baseline total assets. The second GAN estimates the distribution of Z_i conditional on X_i. Recall that Z_i collects all baseline covariates, other than the covariates in X_i. The third GAN estimates the distribution of Y_i conditioned on Z_i, X_i, and W_i. To generate an observation D_i, we generate X_i, generate Z_i conditioned on X_i, and generate the potential outcomes $Y_i(1)$ and $Y_i(0)$ from the estimated distributions of Y_i conditioned on Z_i, X_i, and $W_i = 1$ and $W_i = 0$, respectively. The treatment indicator W_i is drawn i.i.d., Bernoulli with the observed probability in the Banerjee et al. (2015) data and we set $Y_i = Y_i(W_i)$. In this way, we know the true treatment effect $Y_i(1) - Y_i(0)$ for each unit in our simulation. We draw 10 million observations D_i with this process. In the simulation, datasets of various sizes are sampled from this collection.

We use a related procedure to determine the true CATE $\theta_0(x)$ at each value x in the query-vector $\mathbf{x}^{(d)}$ (i.e., the centers of each of the rectangles displayed in Figure 2.1). Specifically, for each x in $\mathbf{x}^{(d)}$, we draw 100,000 observations from the distribution of Z_i conditioned on $X_i = x$. We then draw observations $Y_i(1)$ and $Y_i(0)$ for each of these replicates, and compute the average of the true treatment effects $Y_i(1) - Y_i(0)$. Figure B.1 displays these pseudo-true values of the CATE $\theta_0(x)$. Our simulation design captures many of the same features of the data recovered by GRF, but gives a somewhat smoother picture of the CATE.

Figure B.2 displays a scatterplot comparing the moments from the Banerjee et al. (2015) data to those generated by our calibrated simulation. The distributions match quite closely.

Figure B.3 compares a scatter plot of the observed values of baseline consumption and baseline assets in the Banerjee et al. (2015) data with a heat-map of the distribution of these covariates in our simulation. The limits of the horizontal and vertical axes in this Figure match Figures 2.1 and 2.2 displayed in the main text. Some observations fall outside of the limits of this figure. The quartiles of baseline log consumption are 3.33, 3.76, and 4.20. The quartiles of baseline assets are -0.45, -0.71, and 0.03.

B.4.4.2 Additional Results

Figure B.4 and Figure B.5 give versions of Figure 2.1 and Figure 2.2, constructed using the 2-fold cross-split estimator (B.4.18) and the 2-fold cross-split bootstrap root (B.4.20). The estimates and confidence bounds are quantitatively and qualitatively very similar.

Figure B.6 displays results for the simulation presented in Section 3.6.4, analogous to Figure 2.3, for the 2-fold cross-split estimator (B.4.18) and the 2-fold cross-split bootstrap root (B.4.20). The confidence region is very slightly anti-conservative, so long as b/n is decreasing as n increases. The measurements of bias and variance exhibit patterns very similar to the patterns displayed in Figure 2.3.

B.5. Proofs Supporting Appendix B.4

B.5.1 Proof of Theorem B.4.1

Throughout, we let x denote an arbitrary element of $\boldsymbol{x}^{(d)}$ and use the shorthand $N_b = \binom{n}{b}$. We are interested in giving a higher-order moment bound for the quantity

$$\frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} F(x; D_{\mathsf{s}}, \hat{g}_n) = \frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}}\sum_{i\in\mathsf{s}}\Big(a(x; D_i, D_{\mathsf{s}}, \hat{g}_n) - a(x; D_i, D_{\mathsf{s}}, g_0) - \left(A(x;\hat{g}_n) - A(x; g_0)\right)\Big)\,. \tag{B.5.1}$$

We begin by decomposing (B.5.1) into three terms that will be easier to handle in isolation. In particular, by the representation (2.2.19) the quantity (B.5.1) can be re-expressed as

$$\frac{1}{N_b}\sum_{\mathsf{s}\in\mathcal{S}_{n,b}} F(x; D_{\mathsf{s}}, \hat{g}_n) = \frac{1}{n!}\sum_{\pi\in\mathcal{P}_n}\left\lfloor\frac{n}{b}\right\rfloor^{-1}\sum_{\ell=1}^{\lfloor n/b\rfloor} F_1(x; \mathsf{s}_{\pi,\ell}) \tag{B.5.2}$$

$$- \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \left\lfloor \frac{n}{b} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/b \rfloor} F_2(x; \mathsf{s}_{\pi,\ell}) \tag{B.5.3}$$

$$+ \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \left\lfloor \frac{n}{b} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/b \rfloor} F_3(x; \mathsf{s}_{\pi,\ell}) \tag{B.5.4}$$

where

$$\begin{aligned}
F_1(x; \mathsf{s}) &= \sum_{i \in \mathsf{s}} a(x; D_i, D_{\mathsf{s}}, \hat{g}_n) - a(x; D_i, D_{\mathsf{s}}, \hat{g}_n^{(-\mathsf{s})}) \\
F_2(x; \mathsf{s}) &= \sum_{i \in \mathsf{s}} A(x; \hat{g}_n) - A(x; \hat{g}_n^{(-\mathsf{s})}), \quad \text{and} \\
F_3(x; \mathsf{s}) &= \sum_{i \in \mathsf{s}} a(x; D_i, D_{\mathsf{s}}, \hat{g}_n^{(-\mathsf{s})}) - a(x; D_i, D_{\mathsf{s}}, g_0) \\
&\qquad - (A(x; \hat{g}_n^{(-\mathsf{s})}) - A(x; g_0)),
\end{aligned}$$

respectively. We give suitable probability bounds for the terms in (B.5.2) through (B.5.4).

We begin by considering the term (B.5.2). Fix any integer $q \geq 1$. We show that

$$\left\| \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \left\lfloor \frac{n}{b} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/b \rfloor} F_1(x; \mathsf{s}_{\pi,\ell}) \right\|_{2q} \lesssim 2qn^{-1} b\gamma(2q) b^{1/2q} . \tag{B.5.5}$$

To this end, observe that

$$\begin{aligned}
\| \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \left\lfloor \frac{n}{b} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/b \rfloor} F_1(x; \mathsf{s}_{\pi,\ell}) \|_{2q} &\leq \| \sum_{i=1}^{b} \kappa(x, X_i; D_{[b]})(m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})) \|_{2q} \\
&\leq \| \max_{i \in [b]} |m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})|^2 \|_q^{1/2},
\end{aligned} \tag{B.5.6}$$

where the first inequality follows from Jensen's inequality and the second inequality follows from Assumption 2.3.1, Part (ii). Lemma 2.2.2 of Van Der Vaart and Wellner (1996), implies that, if $A_1, \ldots, A_n$ are any collection of real-valued random variables, then

$$\| \max_{i \in [n]} |A_i| \|_q \lesssim n^{1/q} \max_{i \in [n]} \| |A_i| \|_q. \tag{B.5.7}$$

Thus, we have that

$$\| \max_{i \in [b]} |m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])})|^2 \|_q^{1/2} \leq b^{1/(2q)} \| m(D_i, \hat{g}_n) - m(D_i, \hat{g}_n^{(-[b])}) \|_{2q}.$$

$$\lesssim 2qn^{-1}b\gamma(2q)b^{1/2q}$$

where the second inequality follows from Assumption B.4.3. Hence, the condition (B.5.5) holds.

We now turn to the term (B.5.3). In this case, we show that

$$\left\|\frac{1}{n!}\sum_{\pi\in\mathcal{P}_n}\left\lfloor\frac{n}{b}\right\rfloor^{-1}\sum_{\ell=1}^{\lfloor n/b\rfloor}F_2(x;\mathsf{s}_{\pi,\ell})\right\|_{2q}\lesssim 2qn^{-1}b\gamma(2q)b^{1/2q}\,. \tag{B.5.8}$$

To this end, let $\mathbf{D}_n'' = (D_i'')_{i=1}^n$ be another independent copy of $\mathbf{D}_n$. Observe that

$$\begin{aligned}
&\left\|\frac{1}{n!}\sum_{\pi\in\mathcal{P}_n}\left\lfloor\frac{n}{b}\right\rfloor^{-1}\sum_{\ell=1}^{\lfloor n/b\rfloor}F_2(x;\mathsf{s}_{\pi,\ell})\right\|_{2q}\\
&\quad\le\left\|\sum_{i=1}^{b}\left(A(x;\hat{g}_n)-A(x;\hat{g}_n^{(-[b])})\right)\right\|_{2q}\\
&\quad\le\left\|\sum_{i=1}^{b}\kappa(x,X_i'';D_{[b]}'')\left(m(D_i'',\hat{g}_n)-m(D_i'',\hat{g}_n^{(-[b])})\right)\right\|_{2q}\\
&\quad\le\left\|\max_{i\in[b]}\left|m(D_i'',\hat{g}_n)-m(D_i'',\hat{g}_n^{(-[b])})\right|^2\right\|_q^{1/2}\\
&\quad\le b^{1/(2q)}\left\|m(D_i'',\hat{g}_n)-m(D_i'',\hat{g}_n^{(-[b])})\right\|_{2q}\\
&\quad\lesssim 2qn^{-1}b\gamma(2q)b^{1/2q}\,,
\end{aligned} \tag{B.5.9}$$

where the first two inequalities follow from Jensen's inequality, the third inequality from Assumption 2.3.1, Part (ii), the fourth inequality follows from (B.5.7), and the final inequality follows from the independent-copy stability bound in Assumption B.4.3. Hence, the condition (B.5.8) holds.

Finally, we consider the term (B.5.4). The argument here is somewhat more involved. To simplify exposition, let $k=\lfloor n/b\rfloor$ and define the sets $\mathsf{s}_\ell=\{(\ell-1)b+1,\ldots,b\ell\}$ for each ℓ in $1,\ldots,k$. It will suffice to show that

$$\left\|\frac{1}{k}\sum_{\ell=1}^{k}F_3(x;\mathsf{s}_\ell)\right\|_q\lesssim\sqrt{\frac{1}{k}}\left(\left(\frac{1}{k}\right)^{1/4}\delta_{n,g}+\varepsilon_b^{1/2}+b^{1/q}\gamma(q)\right)q^{3/2} \tag{B.5.10}$$

as Jensen's inequality implies that

$$\left\| \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \left\lfloor \frac{n}{b} \right\rfloor^{-1} \sum_{\ell=1}^{\lfloor n/b \rfloor} F_3(x; \mathsf{s}_{\pi,\ell}) \right\|_q \leq \left\| \frac{1}{k} \sum_{\ell=1}^{k} F_3(x; \mathsf{s}_\ell) \right\|_q . \tag{B.5.11}$$

To establish the bound (B.5.10), we apply the following generalized Efron-Stein inequality, due to Boucheron et al. (2005). See also Chapter 15.2 of Boucheron et al. (2013).

Lemma B.5.1 (Theorem 2, Boucheron et al., 2005)**.** *Let $X = (X_i)_{i=1}^n$ be a sequence of independent random variables and let $X' = (X_i')_{i=1}^n$ denote an independent copy of X. Let $X^{(-i)}$ be constructed by taking X and replacing X_i with X_i'. Consider the random variable $f(X)$, where $f(\cdot)$ is any real-valued function. It holds that*

$$\| f(X) - \mathbb{E}\left[f(X) \right] \|_q \lesssim \sqrt{q} \left\| \sum_{i=1}^{n} \left(f(X) - f(X^{(-i)}) \right)^2 \right\|_{q/2}^{1/2}$$

for any $q \geq 2$.

Before applying Lemma B.5.1, observe that

$$\mathbb{E}\left[\frac{1}{k} \sum_{\ell=1}^{k} F_3(x; \mathsf{s}_\ell) \right] = 0 . \tag{B.5.12}$$

Indeed, conditional on $\hat{g}_n^{(-\mathsf{s}_\ell)}$, the block D_{s_ℓ} is independent of $\hat{g}_n^{(-\mathsf{s}_\ell)}$. Thus, by Honesty we find,

$$\mathbb{E}\Bigg[\sum_{i \in \mathsf{s}_\ell} (a(x; D_i, D_{\mathsf{s}_\ell}, \hat{g}_n^{(-\mathsf{s}_\ell)}) - a(x; D_i, D_{\mathsf{s}_\ell}, g_0) - \\ (A(x; \hat{g}_n^{(-\mathsf{s}_\ell)}) - A(x; g_0))) \mid | \hat{g}_n^{(-\mathsf{s}_\ell)} \Bigg] = 0 . \tag{B.5.13}$$

Taking expectations verifies (B.5.12).

To apply Lemma B.5.1, let $\hat{g}_n^{(-\mathsf{s}_\ell, -\mathsf{s}_r)}$ denote a version of the estimator $\hat{g}_n$ formed with all observations in $\mathbf{D}_n$, except that the observations D_{s_ℓ} and D_{s_r} are replaced by the observations D'_{s_ℓ} and D''_{s_r}, respectively. Additionally, let $\mathbf{D}'''_n$ be an additional, independent, copy of $\mathbf{D}_n$ and let $\tilde{g}_n^{(-\mathsf{s}_\ell)}$ denote a version of the estimator $\hat{g}_n$ formed with all observations in $\mathbf{D}_n$, except that the observations D_{s_ℓ} are replaced by the observations D'''_{s_ℓ}. Let the quantity $F_3^{(-r)}(x; \mathsf{s}_\ell)$

be given by

$$F_3^{(-r)}(x;\mathsf{s}_\ell)=\sum_{i\in\mathsf{s}_\ell}\Bigg(a(x;D_i,D_{\mathsf{s}_\ell},\hat{g}_n^{(-\mathsf{s}_\ell,-\mathsf{s}_r)})-a(x;D_i,D_{\mathsf{s}_\ell},g_0)\\ -\left(A(x;\hat{g}_n^{(-\mathsf{s}_\ell,-\mathsf{s}_r)})-A(x;g_0)\right)\Bigg)$$

if $r\neq\ell$, and by

$$F_3^{(-\ell)}(x;\mathsf{s}_\ell)=\sum_{i\in\mathsf{s}_\ell}\Bigg(a(x;D_i'',D_{\mathsf{s}_\ell}'',\tilde{g}_n^{(-\mathsf{s}_\ell)})-a(x;D_i'',D_{\mathsf{s}_\ell}'',g_0)\\ -\left(A(x;\tilde{g}_n^{(-\mathsf{s}_\ell)})-A(x;g_0)\right)\Bigg)$$

otherwise. Combining (B.5.12) with Lemma B.5.1 implies that

$$\left\|\frac{1}{k}\sum_{\ell=1}^{k}F_3(x;\mathsf{s}_\ell)\right\|_q\\ \leq\sqrt{q}\left\|\sum_{\ell=1}^{k}\left(\frac{1}{k}\left(F_3(x;\mathsf{s}_\ell)-F_3^{(-\ell)}(x;\mathsf{s}_\ell)\right)+\frac{1}{k}\sum_{r\neq\ell}\left(F_3(x;\mathsf{s}_r)-F_3^{(-\ell)}(x;\mathsf{s}_r)\right)\right)^2\right\|_{q/2}^{1/2}. \tag{B.5.14}$$

We obtain a suitable bound for (B.5.14) through the application of the following Lemma.

Lemma B.5.2. *Suppose that the conditions of Theorem B.4.1 hold. Then*

$$\left\|\sum_{\ell=1}^{k}\left(\frac{1}{k}\left(F_3(x^{(j)};\mathsf{s}_\ell)-F_3^{(-\ell)}(x^{(j)};\mathsf{s}_\ell)\right)+\frac{1}{k}\sum_{r\neq\ell}\left(F_3(x^{(j)};\mathsf{s}_r)-F_3^{(-\ell)}(x^{(j)};\mathsf{s}_r)\right)\right)^2\right\|_{q/2}^{1/2}\\ \lesssim\sqrt{\frac{1}{k}}\left(\left(\frac{1}{k}\right)^{1/4}\delta_{n,g}+\varepsilon_b^{1/2}+b^{1/q}\gamma(q)\right)q \tag{B.5.15}$$

for each $j\in[d]$ and each even integer q.

Consequently, the bound (B.5.14) and Lemma B.5.2 imply (B.5.10).

Putting the pieces together, by the decomposition (B.5.2)–(B.5.4), Minkowski's inequality, and the bounds (B.5.5), (B.5.8), and (B.5.10), we find that

$$\begin{aligned}
&\left\| \frac{1}{N_b} \sum_{\mathsf{s} \in \mathcal{S}_{n,b}} F(x; D_{\mathsf{s}}, \hat{g}_n) \right\|_q \\
&\quad \lesssim q \frac{b}{n} \gamma(2q) b^{1/(2q)} + q^{3/2} \sqrt{\frac{1}{k}} \left(\left(\frac{1}{k} \right)^{1/4} \delta_{n,g} + \varepsilon_b^{1/2} + b^{1/q} \gamma(q) \right) \\
&\quad \lesssim q^{3/2} \sqrt{\frac{b}{n}} \left(\left(\frac{b}{n} \right)^{1/4} \delta_{n,g} + \varepsilon_b^{1/2} + b^{1/q} \gamma(2q) \right) , \qquad \text{(B.5.16)}
\end{aligned}$$

where the final inequality follows from $k = \lfloor n/b \rfloor$, the monotonicity of $\gamma(\cdot)$, and the fact that $b \le n$. ■

B.5.2 Proof of Lemma B.5.2

Throughout, we let x denote an arbitrary element of $\boldsymbol{x}^{(d)}$ and use the shorthand $N_b = \binom{n}{b}$. Observe that

$$\left\| \sum_{\ell=1}^{k} \left(\frac{1}{k} \left(F_3(x; \mathsf{s}_\ell) - F_3^{(-\ell)}(x^{(j)}; \mathsf{s}_\ell) \right) + \frac{1}{k} \sum_{r \neq \ell} \left(F_3(x; \mathsf{s}_r) - F_3^{(-\ell)}(x; \mathsf{s}_r) \right) \right)^2 \right\|_{q/2} \qquad \text{(B.5.17)}$$

$$\begin{aligned}
&\le \sum_{\ell=1}^{k} \left\| \left(\frac{1}{k} \left(F_3(x; \mathsf{s}_\ell) - F_3^{(-\ell)}(x^{(j)}; \mathsf{s}_\ell) \right) + \frac{1}{k} \sum_{r \neq \ell} \left(F_3(x; \mathsf{s}_r) - F_3^{(-\ell)}(x; \mathsf{s}_r) \right) \right)^2 \right\|_{q/2} \\
&\lesssim \left\| \sum_{\ell=1}^{k} \frac{1}{k^2} \left(F_3(x; \mathsf{s}_\ell) - F_3^{(-\ell)}(x^{(j)}; \mathsf{s}_\ell) \right)^2 + \frac{1}{k} \sum_{\ell=1}^{k} \sum_{r \neq \ell} \left(F_3(x; \mathsf{s}_r) - F_3^{(-\ell)}(x; \mathsf{s}_r) \right)^2 \right\|_{q/2} \\
&\lesssim \frac{1}{k} \| F_3(x; \mathsf{s}_\ell) \|_q^2 \qquad \text{(B.5.18)} \\
&+ k \| F_3(x; \mathsf{s}_r) - F_3^{(-\ell)}(x; \mathsf{s}_r) \|_q^2, \qquad \text{(B.5.19)}
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from Cauchy-Schwarz and Jensen's inequality, and the final inequality follows from the triangle inequality.

Consider the term (B.5.18). Observe that

$$\mathbb{E} \left[| F_3(x; D_{\mathsf{s}_1}) |^q \right]$$

$$= \mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_1}\left(a(x;D_i,D_{\mathsf{s}_1},\hat{g}_n^{(-\mathsf{s}_1)}) - a(x;D_i,D_{\mathsf{s}_1},g_0) - \left(A(x;\hat{g}_n^{(-\mathsf{s}_1)}) - A(x;g_0)\right)\right)\right|^q\right]$$

$$\lesssim \mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_1}\left(a(x;D_i,D_{\mathsf{s}_1},\hat{g}_n^{(-\mathsf{s}_1)}) - a(x;D_i,D_{\mathsf{s}_1},g_0)\right)\right|^q\right] \tag{B.5.20}$$

$$+ \mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_1}\left(A(x;\hat{g}_n^{(-\mathsf{s}_1)}) - A(x;g_0)\right)\right|^q\right] \tag{B.5.21}$$

by the triangle inequality. We begin by bounding the term (B.5.20). Observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_1}\kappa(x,X_i;D_{\mathsf{s}_1})\left(m(D_i,\hat{g}_n^{(-\mathsf{s}_1)}) - m(D_i,g_0)\right)\right|^q\right] \\
&\leq \mathbb{E}\left[\sum_{i\in\mathsf{s}_1}\kappa(x,X_i;D_{\mathsf{s}_1})\left|m(D_i,\hat{g}_n^{(-\mathsf{s}_1)}) - m(D_i,g_0)\right|^q\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\sum_{i\in\mathsf{s}_1}\kappa(x,X_i;D_{\mathsf{s}_1})\mathbb{E}\left[\left|m(D_i,\hat{g}_n^{(-\mathsf{s}_1)}) - m(D_i,g_0)\right|^q \mid X_i,\hat{g}_n^{(-\mathsf{s}_1)}\right] \mid \hat{g}_n^{(-\mathsf{s}_1)}\right]\right] \\
&\lesssim \mathbb{E}\left[\|\hat{g}_n^{(-\mathsf{s}_1)} - g_0\|_{q,\infty}^q\right] + \varepsilon_b^{q/2},
\end{aligned} \tag{B.5.22}$$

where the first inequality follows from Jensen's inequality, the second inequality follows from Honesty, i.e., Assumption 2.3.1, Part (i), and the final inequality follows from Assumption B.4.1, Assumption B.4.2, and Assumption 2.3.1, Part (ii).

To handle the second term (B.5.21), let $\mathbf{D}_n'' = (D_i'')_{i=1}^n$ be another independent copy of $\mathbf{D}_n$. By Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_1}\left(A(x;\hat{g}_n^{(-\mathsf{s}_1)}) - A(x;g_0)\right)\right|^q\right] \\
&\leq \mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_1}\kappa(x,X_i'';D_{\mathsf{s}_1}'')\left(m(D_i'',\hat{g}_n^{(-\mathsf{s}_1)}) - m(D_i'',g_0)\right)\right|^q\right] \\
&\lesssim \mathbb{E}\left[\|\hat{g}_n^{(-\mathsf{s}_1)} - g_0\|_{q,\infty}^q\right] + \varepsilon_b^{q/2},
\end{aligned} \tag{B.5.23}$$

where the final inequality follows from the same argument used to obtain (B.5.22). Hence, the bounds (B.5.22) and (B.5.23) imply that

$$\|F_3(x;\mathsf{s}_1)\|_q^2 \lesssim \left(q^q\left(\frac{b}{n}\right)^{q/4}\delta_{n,g}^q + \varepsilon_b^{q/2}\right)^{2/q}$$

$$\lesssim q^2 \left(\left(\frac{b}{n}\right)^{1/2} \delta_{n,g}^2 + \varepsilon_b \right), \tag{B.5.24}$$

where we have used the rate condition (B.4.7) and the fact that $\hat{g}_n^{(-\mathsf{s}_1)}$ and $\hat{g}_n$ are identically distributed.

Next, consider the term (B.5.19). For $r \neq \ell$, observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|F_3(x;\mathsf{s}_r) - F_3^{(-\ell)}(x;\mathsf{s}_r)\right|^q\right] \\
&= \mathbb{E}\Bigg[\Bigg|\sum_{i\in\mathsf{s}_r}\left(a(x;D_i,D_{\mathsf{s}_r},\hat{g}_n^{(-\mathsf{s}_r)}) - a(x;D_i,D_{\mathsf{s}_r},\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right) \\
&\qquad - \sum_{i\in\mathsf{s}_r}\left(A(x;\hat{g}_n^{(-\mathsf{s}_r)}) - A(x;\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right)\Bigg|^q\Bigg] \\
&\lesssim \mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_r}\left(a(x;D_i,D_{\mathsf{s}_r},\hat{g}_n^{(-\mathsf{s}_r)}) - a(x;D_i,D_{\mathsf{s}_r},\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right)\right|^q\right] && \text{(B.5.25)} \\
&\quad + \mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_r}\left(A(x;\hat{g}_n^{(-\mathsf{s}_r)}) - A(x;\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right)\right|^q\right] && \text{(B.5.26)}
\end{aligned}$$

by the triangle inequality. To bound the quantity (B.5.25), observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_r}\kappa(x,X_i;D_{\mathsf{s}_r})\left(m(D_i,\hat{g}_n^{(-\mathsf{s}_r)}) - m(D_i,\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right)\right|^q\right] \\
&\leq \mathbb{E}\left[\max_{i\in[b]}\left|m(D_i'',\hat{g}_n) - m(D_i'',\hat{g}_n^{(-[b])})\right|^q\right] \\
&\leq b\mathbb{E}\left[\left|m(D_i'',\hat{g}_n) - m(D_i'',\hat{g}_n^{(-[b])})\right|^q\right] \\
&\lesssim b\left(qn^{-1}b\gamma(q)\right)^q, && \text{(B.5.27)}
\end{aligned}$$

where the first inequality follows from Jensen's inequality, Assumption 2.3.1, Part (ii), and exchangeability, and the final inequality follows from the independent-copy stability bound in Assumption B.4.3.

Similarly, to bound the quantity (B.5.26), let $\mathbf{D}_n''$ be an independent copy of $\mathbf{D}_n$. By Jensen's inequality,

$$\mathbb{E}\left[\left|\sum_{i\in\mathsf{s}_r}\left(A(x;\hat{g}_n^{(-\mathsf{s}_r)}) - A(x;\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right)\right|^q\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left|\sum_{i\in \mathsf{s}_r}\kappa(x,X_i'';D_{\mathsf{s}_r}'')\left(m(D_i'',\hat{g}_n^{(-\mathsf{s}_r)})-m(D_i'',\hat{g}_n^{(-\mathsf{s}_r,-\mathsf{s}_\ell)})\right)\right|^q\right]\\
&\leq \mathbb{E}\left[\max_{i\in[b]}\left|m(D_i'',\hat{g}_n)-m(D_i'',\hat{g}_n^{(-[b])})\right|^q\right]\\
&\lesssim b\left(qn^{-1}b\gamma(q)\right)^q, \qquad \text{(B.5.28)}
\end{aligned}
$$

where the final inequality again follows from the independent-copy stability bound in Assumption B.4.3. Consequently, (B.5.27) and (B.5.28) imply that

$$
\|F_3(x;\mathsf{s}_r)-F_3^{(-\ell)}(x;\mathsf{s}_r)\|_q^2 \lesssim b^{2/q}\left(qn^{-1}b\gamma(q)\right)^2 . \qquad \text{(B.5.29)}
$$

Putting the pieces together, (B.5.17), (B.5.24), and (B.5.29) imply that

$$
\begin{aligned}
&\left\|\sum_{\ell=1}^{k}\left(\frac{1}{k}\left(F_3(x;\mathsf{s}_\ell)-F_3^{(-\ell)}(x;\mathsf{s}_\ell)\right)+\frac{1}{k}\sum_{r\neq\ell}\left(F_3(x;\mathsf{s}_r)-F_3^{(-\ell)}(x;\mathsf{s}_r)\right)\right)^2\right\|_{q/2}^{1/2}\\
&\lesssim \left[\frac{q^2}{k}\left(\left(\frac{b}{n}\right)^{1/2}\delta_{n,g}^2+\varepsilon_b\right)+kb^{2/q}\left(qn^{-1}b\gamma(q)\right)^2\right]^{1/2}\\
&\lesssim \sqrt{\frac{1}{k}}\left(\left(\frac{1}{k}\right)^{1/4}\delta_{n,g}+\varepsilon_b^{1/2}+b^{1/q}\gamma(q)\right)q,
\end{aligned}
$$

where the final inequality uses $k=\lfloor n/b\rfloor$. ■

B.5.3 Proof of Lemma B.4.1

Consider the decomposition

$$
\begin{aligned}
&m(D_i,\hat{g}_n)-m(D_i,\hat{g}_n^{(-[b])})\\
&=\left(\hat{\mu}_n(Z_i,1)-\hat{\mu}_n^{(-[b])}(Z_i,1)\right)-\left(\hat{\mu}_n(Z_i,0)-\hat{\mu}_n^{(-[b])}(Z_i,0)\right)\\
&\quad-\hat{\beta}_n^{(-[b])}(W_i,Z_i)\left(\hat{\mu}_n(Z_i,W_i)-\hat{\mu}_n^{(-[b])}(Z_i,W_i)\right)\\
&\quad+\left(Y_i-\hat{\mu}_n(Z_i,W_i)\right)\left(\hat{\beta}_n(W_i,Z_i)-\hat{\beta}_n^{(-[b])}(W_i,Z_i)\right).
\end{aligned}
$$

We have that

$$
\begin{aligned}
&\|m(D_i,\hat{g}_n)-m(D_i,\hat{g}_n^{(-[b])})\|_{2p}\\
&\leq \|\hat{\mu}_n(Z_i,1)-\hat{\mu}_n^{(-[b])}(Z_i,1)\|_{2p}+\|\hat{\mu}_n(Z_i,0)-\hat{\mu}_n^{(-[b])}(Z_i,0)\|_{2p}
\end{aligned}
$$

$$
\begin{aligned}
&+ \|\hat{\beta}_n^{(-[b])}(W_i, Z_i)(\hat{\mu}_n(Z_i, W_i) - \hat{\mu}_n^{(-[b])}(Z_i, W_i))\|_{2p} \\
&+ \|(Y_i - \hat{\mu}_n(Z_i, W_i))(\hat{\beta}_n(W_i, Z_i) - \hat{\beta}_n^{(-[b])}(W_i, Z_i))\|_{2p} \\
\lesssim\; & \|\hat{\mu}_n(Z_i, 1) - \hat{\mu}_n^{(-[b])}(Z_i, 1)\|_{2p} + \|\hat{\mu}_n(Z_i, 0) - \hat{\mu}_n^{(-[b])}(Z_i, 0)\|_{2p} \\
&+ \|\hat{\mu}_n(Z_i, W_i) - \hat{\mu}_n^{(-[b])}(Z_i, W_i)\|_{2p} + \|\hat{\beta}_n(W_i, Z_i) - \hat{\beta}_n^{(-[b])}(W_i, Z_i)\|_{2p},
\end{aligned}
$$

where the first inequality follows from the triangle inequality and the second inequality follows from the boundedness condition (B.4.12), applied to both the original and resampled nuisance estimators. This verifies the condition (B.4.13). An identical argument verifies the condition (B.4.14). ■

B.5.4 Proof of Lemma B.4.2

To ease notation, we omit dependence on the evaluation point z. Recall the definition of the Hájek projection

$$u^{(1)}(D) = \mathbb{E}\left[u(D_{[b']}) \mid D_1 = D\right] - \mathbb{E}\left[u(D_{[b']})\right],$$

where D is an independent copy of D_1. Let

$$\nu^2 = \mathrm{Var}(u(D_{[b']})) \quad \text{and} \quad \sigma_{b'}^2 = \mathrm{Var}(u^{(1)}(D))$$

denote the kernel variance and Hájek projection variance, respectively. Consider the decomposition

$$
\begin{aligned}
\hat{g}_n(\mathbf{D}_n) - \hat{g}_n(\mathbf{D}_n^{(-[b])}) = &\frac{b'}{n}\sum_{i=1}^{b}\left(u^{(1)}(D_i) - u^{(1)}(D_i')\right) && \text{(B.5.30)} \\
&+ \left(\hat{g}_n(\mathbf{D}_n) - \mathbb{E}[u(D_{[b']})] - \frac{b'}{n}\sum_{i=1}^{n} u^{(1)}(D_i)\right) && \text{(B.5.31)} \\
&- \left(\hat{g}_n(\mathbf{D}_n^{(-[b])}) - \mathbb{E}[u(D_{[b']})] - \frac{b'}{n}\left(\sum_{i=1}^{b} u^{(1)}(D_i') + \sum_{i=b+1}^{n} u^{(1)}(D_i)\right)\right). && \text{(B.5.32)}
\end{aligned}
$$

The result is obtained by giving higher-order moment bounds for each of the terms (B.5.30), (B.5.31), and (B.5.32).

We begin by considering the terms (B.5.31) and (B.5.32). Observe that the decomposition

(2.2.5), Lemma 2.2.2, and the triangle inequality imply that

$$\left\| \hat{g}_n(\mathbf{D}_n) - \mathbb{E}[u(D_{[b']})] - \frac{b'}{n}\sum_{i=1}^{n} u^{(1)}(D_i) \right\|_q \le \phi q \sum_{m=2}^{b'} \left(\frac{Cqb'}{n}\right)^{m/2} \lesssim \phi q^2 \frac{b'}{n} \tag{B.5.33}$$

for $q \le cn/b'$. Likewise,

$$\left\| \hat{g}_n(\mathbf{D}_n^{(-[b])}) - \mathbb{E}[u(D_{[b']})] - \frac{b'}{n}\left(\sum_{i=1}^{b} u^{(1)}(D_i') + \sum_{i=b+1}^{n} u^{(1)}(D_i)\right) \right\|_q$$
$$\le \phi q \sum_{m=2}^{b'} \left(\frac{Cqb'}{n}\right)^{m/2} \lesssim \phi q^2 \frac{b'}{n}\,. \tag{B.5.34}$$

To handle the term (B.5.30), we apply Lemma B.2.4. In particular, Hoeffding orthogonality gives

$$b'\sigma_{b'}^2 \le \mathrm{Var}(u(D_{[b']})) \lesssim \phi^2,$$

and so $\sigma_{b'}^2 \lesssim \phi^2/b'$. Therefore,

$$\left\| \frac{b'}{n}\sum_{i=1}^{b} \left(u^{(1)}(D_i) - u^{(1)}(D_i')\right) \right\|_q$$
$$= \frac{b'b}{n}\left\| \frac{1}{b}\sum_{i=1}^{b} \left(u^{(1)}(D_i) - u^{(1)}(D_i')\right) \right\|_q$$
$$\lesssim \frac{b'b}{n}\left(\sqrt{\frac{\sigma_{b'}^2 q}{b}} + q\frac{\phi}{b}\right) \lesssim \frac{\sqrt{b'b}}{n} q\phi + \frac{b'}{n} q\phi\,. \tag{B.5.35}$$

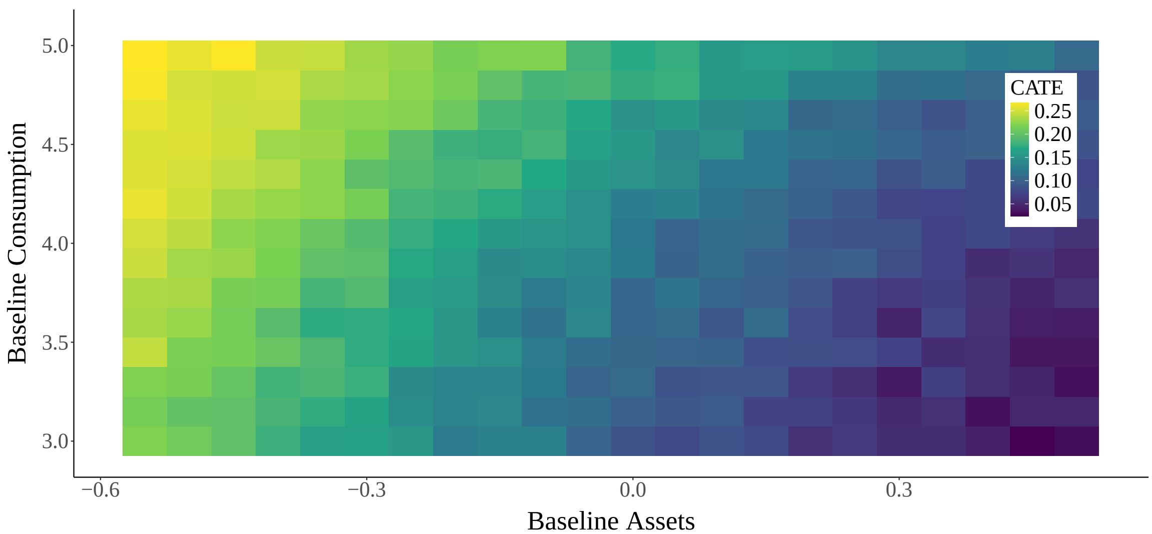
Putting the pieces together, we find that

$$\|\hat{g}_n(\mathbf{D}_n) - \hat{g}_n(\mathbf{D}_n^{(-[b])})\|_q \lesssim \frac{\sqrt{b'b}}{n} q\phi + \frac{b'}{n} q^2\phi$$

by combining the bounds (B.5.33), (B.5.34), and (B.5.35), as required. ■

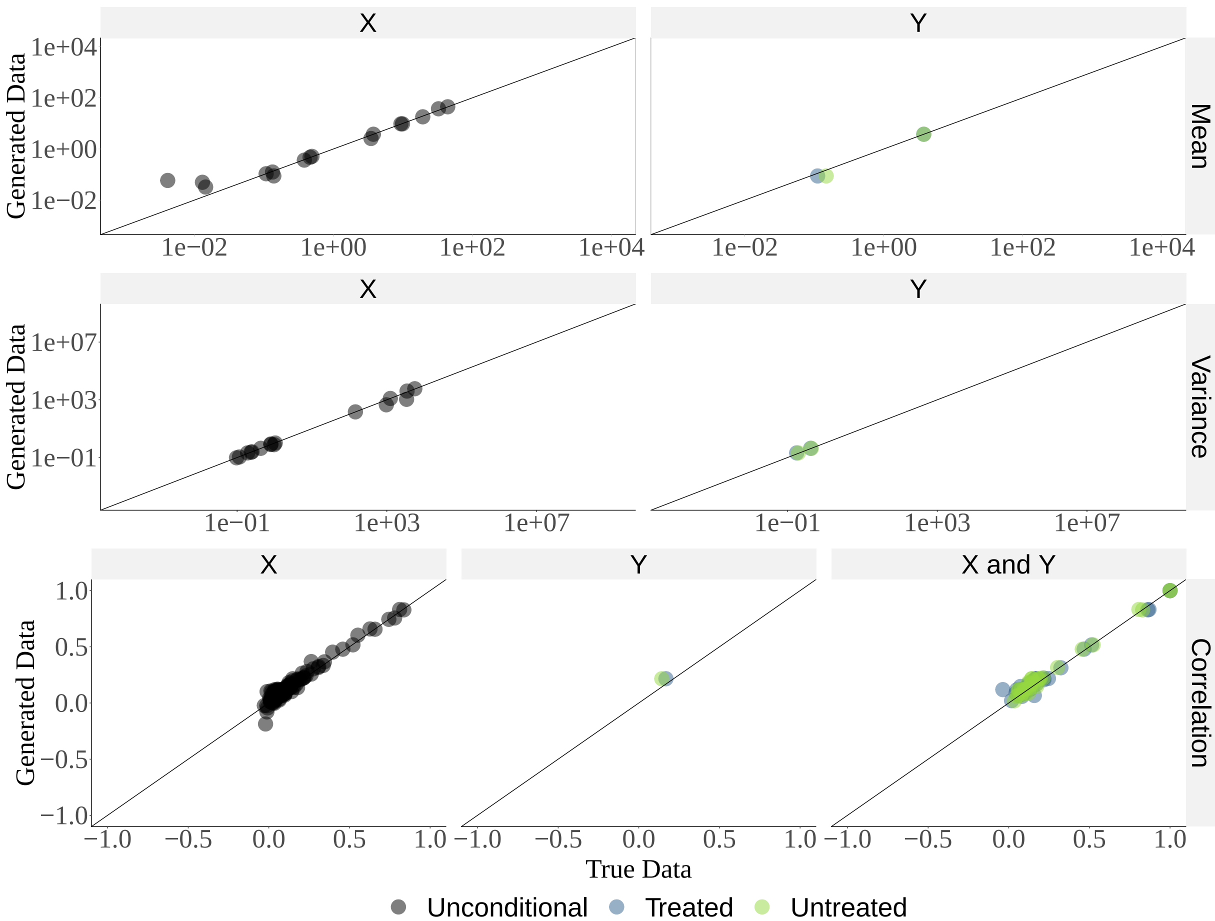
B.6. Supplemental Figures

Figure B.1: Calibrated CATEs



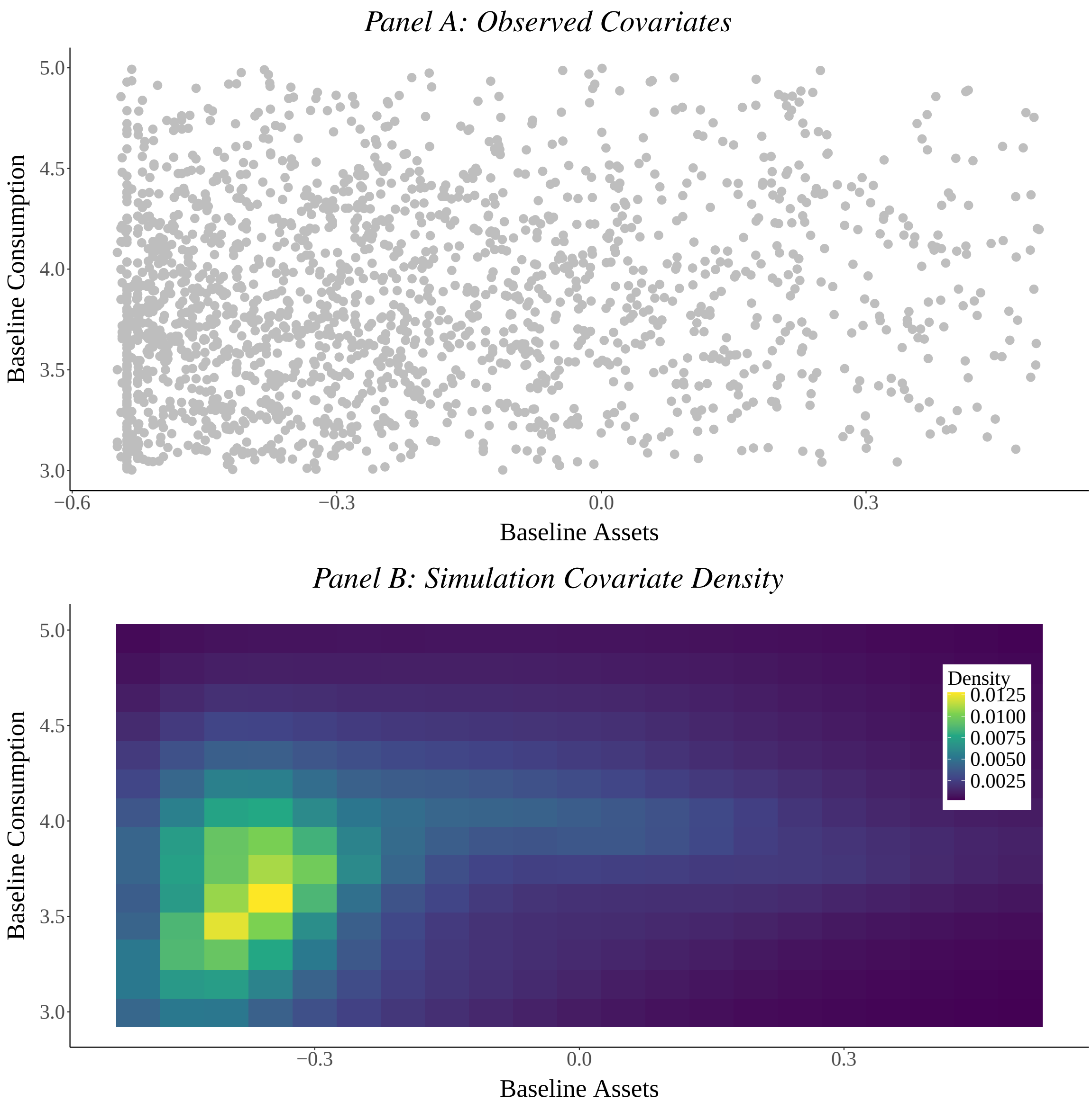
Notes: Figure B.1 displays the “true” value of CATE used in our calibrated simulation.

Figure B.2: Validation



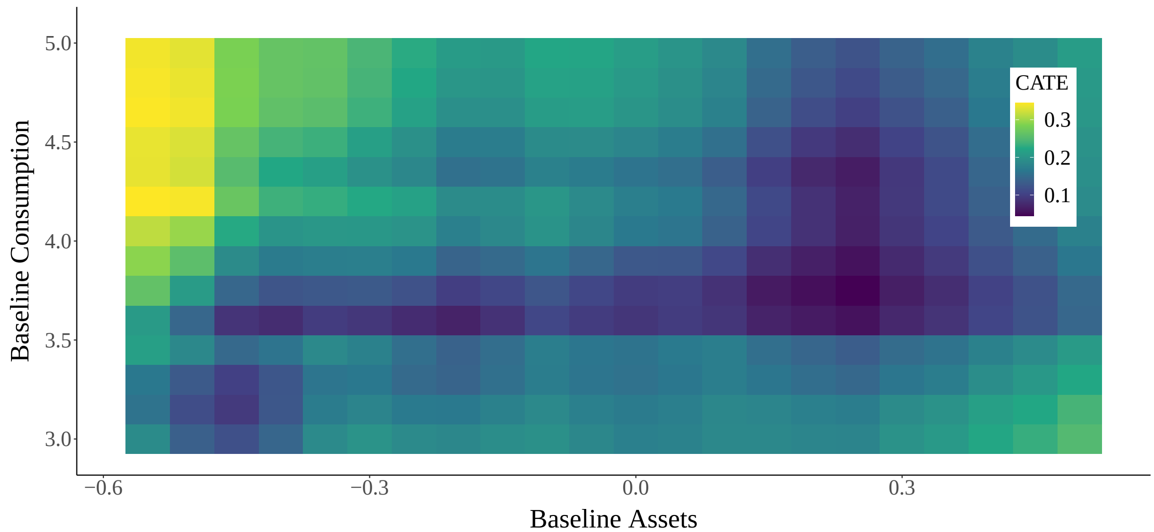
Notes: Figure B.2 displays scatterplots comparing the moments of the data from Banerjee et al. (2015) to the GAN-generated simulation data. Columns differentiate between different types of variables. Rows differentiate between different types of moments. The x-axis of each sub-panel measures the moments of the true data. The y-axis of each sub-panel measures the moments of the generated data. The x and y axes in the first two rows are displayed in log-scale. A forty-five degree line is displayed in all sub-panels. Blue and green dots denote moments conditioned on treatment being set to one and zero, respectively. Black dots denote unconditioned moments.

Figure B.3: Covariate Density



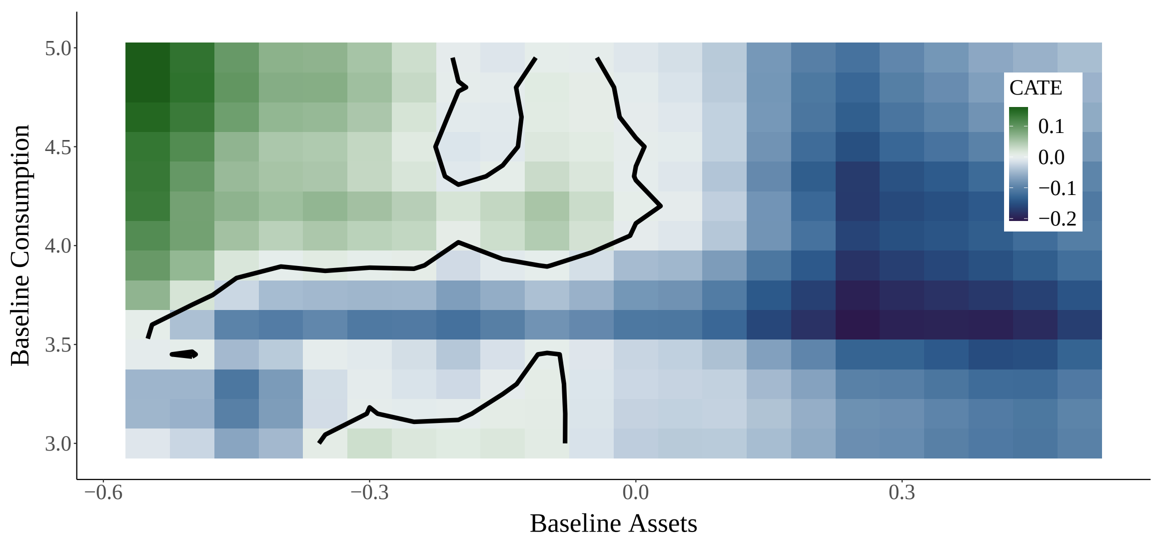
Notes: Panel A of Figure B.3 displays a scatter plot of the observed values of baseline consumption and baseline assets in the Banerjee et al. (2015) data. The horizontal and vertical axes display the baseline monthly consumption, normalized to 2014 dollars on a logarithmic scale base 10, and an index for baseline assets, respectively. Panel B displays a heat-map giving the density of the joint distribution of baseline consumption and baseline assets associated with our calibrated simulation.

Figure B.4: CATE Estimates, 2-Fold Cross-Fitting



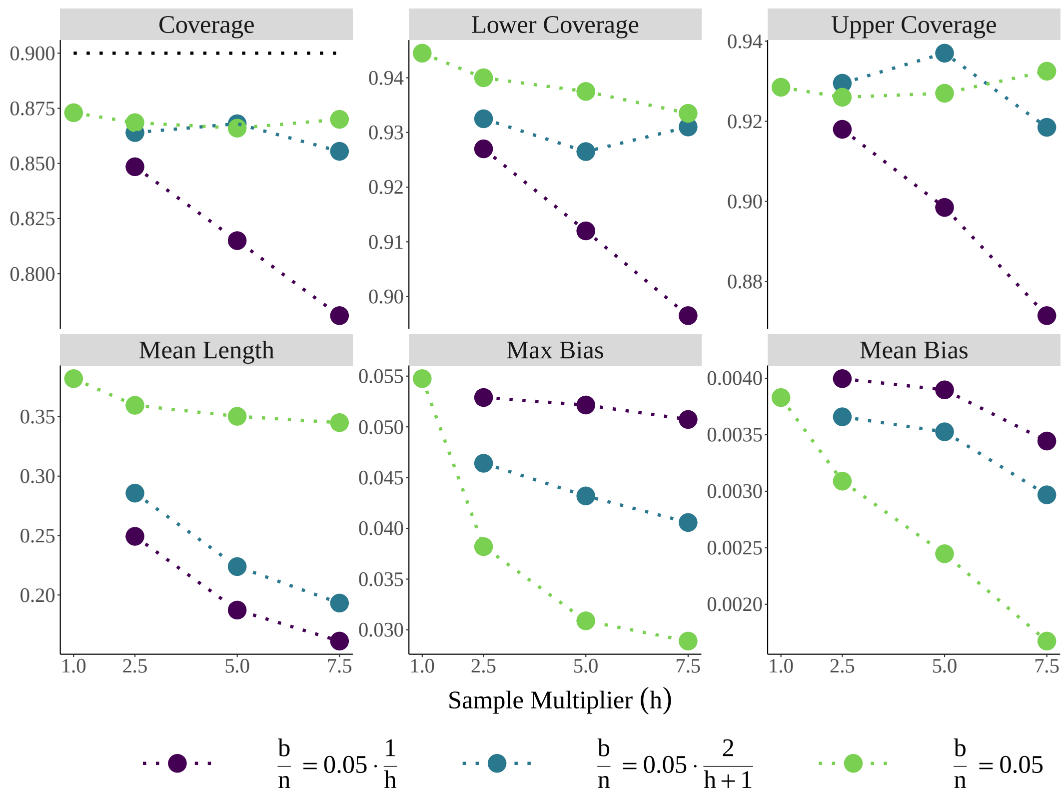
Notes: Figure B.4 displays a heat map giving CATE estimates for the intervention studied in Banerjee et al. (2015) on post-treatment assets, using the 2-fold cross-split estimator (B.4.18). Other features of the display are analogous to Figure 2.1.

Figure B.5: Half-Sample Confidence Region Lower Bound, 2-Fold Cross-Fitting



Notes: Figure B.5 displays a heat map giving half-sample lower confidence bound for the CATE of the intervention studied in Banerjee et al. (2015) on post-treatment total assets. Other features of the display are analogous to Figure 2.2.

Figure B.6: Performance, 2-Fold Cross-Fitting



Notes: Figure B.6 displays several measurements of the performance of the confidence intervals formulated in Definition 2.3.1, constructed using the 2-fold cross-split estimator (B.4.18) and the 2-fold cross-split bootstrap root (B.4.20), in a simulation calibrated to the Banerjee et al. (2015) data. Other features of the display are analogous to Figure 2.3.

Appendix C

Supplemental Appendices for Chapter 3

C.1. Proofs for Decomposition Theorems

In this Appendix, we give the proofs for Theorems 3.2.1 and 3.4.1 Both results hold under two additional regularity conditions, stated in Appendix C.1.1. Proofs for these results, and each of the supporting Lemmas, are given in Appendices C.1.2 and C.1.3, respectively.

C.1.1 Additional Regularity Conditions

Theorem 3.2.1 holds under two, additional regularity conditions. The first condition stipulates that various conditional expectations of the outcome have first and second derivatives that exist and are on the order of a constant.

Assumption C.1.1. *For any unit i and distinct units j and k, the map*

$$(\delta, w) \mapsto \mathbb{E}[Y_i \mid D_{k,j} = \delta, W_j = w, S_j] \tag{C.1.1}$$

is twice continuously differentiable, almost surely. Moreover:
***(i)** For any units i and j, it holds that*

$$\partial_w \, \mathbb{E}[Y_i \mid W_j = w, S_j = s] = O(1) \tag{C.1.2}$$

uniformly over each w and s in their respective domains.
***(ii)** For distinct units i, j, and k, it holds that*

$$\partial^2_{\delta,w} \, \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w, S_j = s] = O(1) \tag{C.1.3}$$

uniformly over each δ, w, and s in their respective domains.

The second condition asks that the conditional expectation of the proximity measure of interest is not much larger than its variance. This restriction is used to ensure that several of the terms in the error in the representation (3.2.7) are negligible. We additionally ask that the conditional variance of treatments is on the order of a constant.

Assumption C.1.2. *It holds that*

$$\mathbb{E}[D_{i,\ell} \mid S_\ell] \lesssim \mathrm{Var}(D_{i,\ell} \mid S_\ell) \quad \textit{and} \quad \mathrm{Var}(W_\ell \mid S_\ell) \asymp 1 \tag{C.1.4}$$

uniformly almost surely.

The regularity conditions needed for Theorem 3.4.1 are direct generalizations of Assumptions 3.2.3, C.1.1, and C.1.2. In particular, each restriction additionally condition on the covariates $H_{i,j}$.

Assumption C.1.3. *For any unit i and distinct units j and k, the map*

$$(\delta, w) \mapsto \mathbb{E}[Y_i \mid D_{k,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j] \tag{C.1.5}$$

is twice continuously differentiable, almost surely. Moreover:
(i) *For any units i and j, it holds that*

$$\partial_w \, \mathbb{E}[Y_i \mid W_j = w, X_j^{\mathsf{out}}, \overline{S}_j] = O(1) \tag{C.1.6}$$

uniformly over each w in its domain.
(ii) *For distinct units i, j, and k, it holds that*

$$\partial^2_{\delta,w} \, \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j] = O(1) \tag{C.1.7}$$

uniformly over each δ and w in their respective domains.
(iii) *For distinct units i, j, and k, it holds that*

$$\partial^2_{\delta,w} \, \mathbb{E}[Y_i \mid D_{k,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j] = o(n^{-1/2}) \tag{C.1.8}$$

uniformly over each δ and w in their respective domains.

Assumption C.1.4. *It holds that*

$$\mathbb{E}[D_{i,\ell} \mid H_{i,\ell}, S_\ell] \lesssim \mathrm{Var}(D_{i,\ell} \mid H_{i,\ell}, S_\ell) \quad \textit{and} \quad \mathrm{Var}(W_\ell \mid \overline{S}_j) \asymp 1 \tag{C.1.9}$$

uniformly almost surely.

C.1.2 Proofs for Theorems 3.2.1 and 3.4.1

Theorems 3.2.1 and 3.4.1 are founded on three general Lemmas. The first Lemma facilitates the evaluation of several expectations that recur throughout the argument.

Lemma C.1.1. *The random variables $(Y_i)_{i=1}^n$ are identically distributed. The array $(M_{i,j})_{i,j\in[n]}$ satisfies*

$$M_{i,j} = M_n(S_i, S_j) \tag{C.1.10}$$

and is supported on an interval whose radius is bounded by a constant. The random variables $(S_i, W_i)_{i=1}^n$ are i.i.d. and satisfy the condition

$$\mathbb{E}[W_i \mid S_i] = 0\ . \tag{C.1.11}$$

Moreover, the conditional variance $\mathrm{Var}(W_i \mid S_i)$ is bounded by a constant almost surely. Define the terms

$$\tilde{\Xi}_i = \sum_{j\neq i}\left(M_{i,j} - \frac{1}{n}\sum_{k\neq j} M_{k,j}\right)W_j - \frac{1}{n}\sum_{k\neq i} M_{k,i}W_i \quad \text{and} \quad \tilde{M}_{i,\ell} = M_{i,\ell} - \mathbb{E}[M_{i,\ell} \mid S_\ell]\ . \tag{C.1.12}$$

It holds that

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[Y_i\hat{\Xi}_i] = \frac{n-2}{n}&\left(\mathbb{E}[Y_i\tilde{M}_{i,\ell}W_\ell] - \mathbb{E}[Y_i\tilde{M}_{q,\ell}W_\ell]\right) \qquad \text{(C.1.13)}\\ &+ \frac{n-2}{n^2}\mathbb{E}[Y_iM_{q,\ell}W_\ell] - \frac{n-1}{n^2}\mathbb{E}[Y_iM_{\ell,i}W_i] + \frac{1}{n^2}\mathbb{E}[Y_iM_{i,\ell}W_\ell]\end{aligned}$$

and

$$\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\hat{\Xi}_i^2] - \frac{n-2}{n}\mathbb{E}\left[\mathrm{Var}(M_{i,\ell} \mid S_\ell)\,\mathrm{Var}(W_\ell \mid S_\ell)\right] = O(n^{-2})\ , \tag{C.1.14}$$

respectively.

The second Lemma gives a decomposition of the expectations of products of three random variables that satisfy particular restrictions on their conditional dependencies. The result relies on an application of a method of argument due to Kolesár and Plagborg-Møller (2024), who build on ideas due to Yitzhaki (1996) and Angrist and Krueger (1999).

Lemma C.1.2. *Fix the random variables Y, W, Z, and U. Assume that the distributions of W and Z, conditioned U, have support contained in the bounded sets $\mathcal{W}$ and $\mathcal{Z}$ almost surely and that the restriction*

$$W \perp\!\!\!\perp Z \mid U \tag{C.1.15}$$

holds. If the random variable Y is real-valued and the first derivatives and second mixed-partial derivative of the conditional expectation $(z, w) \mapsto \mathbb{E}[Y \mid Z = z, W = w, U]$ exist and are uniformly bounded on $\mathcal{Z} \times \mathcal{W}$, then the equalities

$$\mathbb{E}[Y(W - \mathbb{E}[W \mid U])] = \int_{\mathcal{W}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W \geq w\}(W - \mathbb{E}[W \mid U]) \mid U]\partial_w \mathbb{E}[Y \mid W = w, U]] \,\mathrm{d}w \tag{C.1.16}$$

and

$$\begin{aligned} &\mathbb{E}[Y(Z - \mathbb{E}[Z \mid U])(W - \mathbb{E}[W \mid U])] \qquad\qquad (\text{C.1.17})\\ &= \int_{\mathcal{Z}} \int_{\mathcal{W}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W \geq w\}(W - \mathbb{E}[W \mid U]) \mid U]\mathbb{E}[\mathbb{I}\{Z \geq z\}(Z - \mathbb{E}[Z \mid U]) \mid U] \\ &\qquad \partial^2_{z,w}\mathbb{E}[Y \mid Z = z, W = w, U]] \,\mathrm{d}w \,\mathrm{d}z \end{aligned}$$

hold.

The third Lemma gives a helpful, well-known decomposition of particular conditional variances.

Lemma C.1.3. *Fix the random variables Z and U. If Z is a real-valued and supported on $\mathcal{Z}$, then*

$$\mathrm{Cov}(\mathbb{I}\{Z \geq z\}, Z \mid U) \geq 0 \quad \textit{and} \quad \int_{\mathcal{Z}} \mathrm{Cov}(\mathbb{I}\{Z \geq z\}, Z \mid U) \,\mathrm{d}z = \mathrm{Var}(Z \mid U)\,, \tag{C.1.18}$$

almost surely, for each z in $\mathcal{Z}$.

C.1.2.1 Proof of Theorem 3.2.1

Define the risk

$$R_n(\theta) = \mathbb{E}\left[\min_{\alpha \in \mathbb{R}} \left\{\sum_{i=1}^{n} (Y_i - \alpha - \theta \cdot \Delta_i)^2\right\}\right]\,. \tag{C.1.19}$$

By the Frisch-Waugh-Lovell Theorem, we can write

$$R_n(\theta) = \mathbb{E}\left[\sum_{i=1}^{n}\left((Y_i - \tilde{Y}_n) - \theta \cdot \tilde{\Delta}_i\right)^2\right] , \tag{C.1.20}$$

where $\tilde{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and

$$\begin{aligned}
\tilde{\Delta}_i &= \sum_{j \neq i} D_{i,j}W_j - \frac{1}{n}\sum_{k=1}^{n}\sum_{m \neq k} D_{k,m}W_m \\
&= \sum_{j \neq i} D_{i,j}W_j - \sum_{m=1}^{n}\frac{1}{n}\sum_{k \neq m} D_{k,m}W_m \\
&= \sum_{j \neq i}\left(D_{i,j} - \frac{1}{n}\sum_{k \neq j} D_{k,j}\right)W_j - \frac{1}{n}\sum_{k \neq i} D_{k,i}W_i \ .
\end{aligned} \tag{C.1.21}$$

Thus, the parameter $\overline{\theta}_n$ admits the representation

$$\overline{\theta}_n = \frac{\sum_{i=1}^{n}\mathbb{E}[(Y_i - \tilde{Y}_n)\tilde{\Delta}_i]}{\sum_{i=1}^{n}\mathbb{E}[\tilde{\Delta}_i^2]} = \frac{\sum_{i=1}^{n}\mathbb{E}[Y_i\tilde{\Delta}_i]}{\sum_{i=1}^{n}\mathbb{E}[\tilde{\Delta}_i^2]} \tag{C.1.22}$$

by the Projection Theorem and the fact that $\sum_{i=1}^{n}\tilde{\Delta}_i = 0$ by construction.

We begin by simplifying the representation (C.1.22) through an application of Lemma C.1.1. Observe that the conditions of Lemma C.1.1 are implied by Assumptions 3.2.1 and 3.2.2. Hence, we obtain the expansion

$$\begin{aligned}
\overline{\theta}_n = {} & \frac{\mathbb{E}[Y_i\tilde{D}_{i,q}W_q] - \mathbb{E}[Y_i\tilde{D}_{\ell,q}W_q]}{\mathbb{E}[\mathrm{Var}(D_{i,q} \mid S_i)\,\mathrm{Var}(W_i \mid S_i)]} \\
& + \frac{1}{n}\frac{\mathbb{E}[Y_i\tilde{D}_{\ell,q}W_q] - \mathbb{E}[Y_i\tilde{D}_{\ell,i}W_i]}{\mathbb{E}[\mathrm{Var}(D_{i,q} \mid S_q)\,\mathrm{Var}(W_q \mid S_q)]} \\
& + \frac{1}{n}\frac{\mathbb{E}[\mathbb{E}[Y_iW_q \mid S_q]\mathbb{E}[D_{\ell,q} \mid S_q]] - \mathbb{E}[\mathbb{E}[Y_qW_q \mid S_q]\mathbb{E}[D_{\ell,q} \mid S_q]}{\mathbb{E}[\mathrm{Var}(D_{i,q} \mid S_q)\,\mathrm{Var}(W_q \mid S_q)]} \\
& + \frac{1}{n^2}\frac{\mathbb{E}[Y_i\tilde{D}_{i,q}W_q]}{\mathbb{E}[\mathrm{Var}(D_{i,q} \mid S_q)\,\mathrm{Var}(W_q \mid S_q)]} + \frac{1}{n^2}\frac{\mathbb{E}[\mathbb{E}[Y_iW_q \mid S_q]\mathbb{E}[D_{i,q} \mid S_q]]}{\mathbb{E}[\mathrm{Var}(D_{i,q} \mid S_q)\,\mathrm{Var}(W_q \mid S_q)]} + O(n^{-2}) ,
\end{aligned} \tag{C.1.23}$$

where

$$\tilde{D}_{i,\ell} = D_{i,\ell} - \mathbb{E}[D_{i,\ell} \mid S_\ell] \ . \tag{C.1.24}$$

We further re-express the parameter $\overline{\theta}_n$ through several applications of Lemma C.1.2.

Observe that Assumptions 3.2.1 and 3.2.2 imply that

$$W_\ell \perp\!\!\!\perp D_{i,\ell} \mid S_\ell \quad \text{and} \quad \mathbb{E}[W_\ell \mid S_\ell] = 0\ . \tag{C.1.25}$$

Thus, Lemma C.1.2 implies that

$$\begin{aligned}
\mathbb{E}[Y_i \tilde{D}_{i,q} W_q] &= \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W_q \geq w\} W_q \mid S_q] \mathbb{E}[\mathbb{I}\{D_{i,q} \geq \delta\} \tilde{D}_{i,q} \mid S_q] \\
&\qquad\qquad \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, S_q]] \, \mathrm{d}w \, \mathrm{d}\delta\ , \\
\mathbb{E}[Y_i \tilde{D}_{\ell,q} W_q] &= \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W_q \geq w\} W_q \mid S_q] \mathbb{E}[\mathbb{I}\{D_{\ell,q} \geq \delta\} \tilde{D}_{\ell,q} \mid S_q] \\
&\qquad\qquad \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, S_q]] \, \mathrm{d}w \, \mathrm{d}\delta\ , \\
\mathbb{E}[Y_i W_q \mid S_q] &= \int_{\overline{\mathcal{W}}} \mathbb{E}[\mathbb{I}\{W_q \geq w\} W_q \mid S_q] \partial_w \mathbb{E}[Y_i \mid W_q = w, S_q] \, \mathrm{d}w, \quad \text{and} \\
\mathbb{E}[Y_q W_q \mid S_q] &= \int_{\overline{\mathcal{W}}} \mathbb{E}[\mathbb{I}\{W_q \geq w\} W_q \mid S_q] \partial_w \mathbb{E}[Y_q \mid W_q = w, S_q] \, \mathrm{d}w\ ,
\end{aligned} \tag{C.1.26}$$

respectively. Consequently, by plugging each of the representations (C.1.26) into the expression (C.1.23), we find that

$$\begin{aligned}
\overline{\theta}_n = &\int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_q)(\partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, S_q] \\
&\qquad\qquad - \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, S_q])] \, \mathrm{d}w \, \mathrm{d}\delta \\
&+ \frac{1}{n} \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_q)(\partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, S_q] \\
&\qquad\qquad - \partial^2_{\delta,w} \mathbb{E}[Y_q \mid D_{\ell,q} = \delta, W_q = w, S_q])]] \, \mathrm{d}w \, \mathrm{d}\delta \\
&+ \frac{1}{n^2} \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_q) \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, S_q]] \, \mathrm{d}w \, \mathrm{d}\delta \\
&+ \frac{1}{n} \int_{\overline{\mathcal{W}}} \mathbb{E}\left[\lambda'(w \mid S_q) \left(\left(\frac{n+1}{n}\right) \partial_w \mathbb{E}[Y_i \mid W_q = w, S_q] \right.\right. \\
&\qquad\qquad \left.\left. - \partial_w \mathbb{E}[Y_q \mid W_q = w, S_q] \right)\right] \mathrm{d}w + O(n^{-2})\ ,
\end{aligned} \tag{C.1.27}$$

where

$$\begin{aligned}
\lambda(\delta, w \mid S_\ell) &= \frac{\mathrm{Cov}(\mathbb{I}\{D_{i,\ell} \geq \delta\}, D_{i,\ell} \mid S_\ell) \, \mathrm{Cov}(\mathbb{I}\{W_\ell \geq w\}, W_\ell \mid S_\ell)}{\mathbb{E}[\mathrm{Var}(D_{i,\ell} \mid S_\ell) \, \mathrm{Var}(W_\ell \mid S_\ell)]} \quad \text{and} \\
\lambda'(w) &= \frac{\mathbb{E}[D_{i,\ell} \mid S_\ell] \, \mathrm{Cov}(\mathbb{I}\{W_\ell \geq w\}, W_\ell \mid S_q)}{\mathbb{E}[\mathrm{Var}(D_{i,\ell} \mid S_\ell) \, \mathrm{Var}(W_\ell \mid S_\ell)]}\ ,
\end{aligned} \tag{C.1.28}$$

respectively. Now, observe that Lemma C.1.3 implies that

$$\lambda(\delta, w \mid S_\ell) \geq 0 \quad \text{and} \quad \mathbb{E}\left[\int_{\mathcal{D}} \int_{\mathcal{W}} \lambda(\delta, w \mid S_\ell)\, \mathrm{d}w\, \mathrm{d}\delta\right] = 1\,, \tag{C.1.29}$$

respectively. Thus, Assumptions 3.2.3 and C.1.1 imply that

$$\begin{aligned}
&\int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_j) \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, S_q]]\, \mathrm{d}w\, \mathrm{d}\delta = o(n^{-1/2})\,, && \text{(C.1.30)}\\
&\frac{1}{n} \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_q)(\partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, S_q] \\
&\qquad\qquad - \partial^2_{\delta,w} \mathbb{E}[Y_q \mid D_{\ell,q} = \delta, W_q = w, S_q])]]\, \mathrm{d}w\, \mathrm{d}\delta = o(n^{-1/2})\,, \quad \text{and} \\
&\frac{1}{n^2} \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_q) \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, S_q]]\, \mathrm{d}w\, \mathrm{d}\delta = O(n^{-2})\,.
\end{aligned}$$

In turn, Assumption C.1.2, Lemma C.1.3, and Cauchy-Schwarz imply that $\lambda'(w \mid S_\ell) \geq 0$ and

$$\begin{aligned}
\mathbb{E}[\lambda'(w \mid S_\ell)] &\lesssim \frac{\mathbb{E}[\mathbb{E}[D_{i,\ell} \mid S_\ell] \sqrt{\operatorname{Var}(W_\ell \mid S_\ell)}]}{\mathbb{E}[\operatorname{Var}(D_{i,\ell} \mid S_\ell) \operatorname{Var}(W_\ell \mid S_\ell)]} \\
&\lesssim \frac{\mathbb{E}[\operatorname{Var}(D_{i,\ell} \mid S_\ell) \sqrt{\operatorname{Var}(W_\ell \mid S_\ell)}]}{\mathbb{E}[\operatorname{Var}(D_{i,\ell} \mid S_\ell) \operatorname{Var}(W_\ell \mid S_\ell)]} = O(1)
\end{aligned} \tag{C.1.31}$$

uniformly over w. Thus, Assumption C.1.1, and the fact that the treatments and distances are defined on sets whose supports are bounded by a constant, imply that

$$\begin{aligned}
\frac{1}{n} \int_{\overline{\mathcal{W}}} \mathbb{E}\Bigg[\lambda'(w \mid S_q)\Bigg(\left(\frac{n+1}{n}\right) \partial_w \mathbb{E}[Y_i \mid W_q = w, S_q] \\
- \partial_w \mathbb{E}[Y_q \mid W_q = w, S_q]\Bigg)\Bigg]\, \mathrm{d}w = O(n^{-1})\,.
\end{aligned} \tag{C.1.32}$$

Hence, we obtain the representation

$$\overline{\theta}_n = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid S_q) \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, S_q]]\, \mathrm{d}w\, \mathrm{d}\delta + o(n^{-1/2})\,, \tag{C.1.33}$$

by plugging the bounds (C.1.30) and (C.1.32) into (C.1.27). The convexity of the weights is verified by the statement (C.1.29), completing the proof. ■

C.1.2.2 Proof of Theorem 3.4.1

By the same steps used to derive the representation (C.1.22), the parameter $\overline{\theta}_n^\star$ admits the representation

$$\overline{\theta}_n^\star = \frac{\sum_{i=1}^n \mathbb{E}[Y_i \tilde{\Delta}_i^\star]}{\sum_{i=1}^n \mathbb{E}[(\tilde{\Delta}_i^\star)^2]} \tag{C.1.34}$$

where

$$\tilde{\Delta}_i^\star = \sum_{j \neq i} \left(D_{i,j}^\star - \frac{1}{n} \sum_{k \neq j} D_{k,j}^\star \right) W_j^\star - \frac{1}{n} \sum_{k \neq i} D_{k,i}^\star W_i^\star \,. \tag{C.1.35}$$

Observe that Assumptions 3.2.1 and 3.4.1 imply that $(D_{i,j}^\star)_{i,j \in [n]}$ satisfies a representation of the form (C.1.10), with latent coordinates $\overline{S}_i$, and that the random variables $(\overline{S}_i, W_i^\star)_{i=1}^n$ are i.i.d., and satisfy the condition

$$\begin{aligned}
\mathbb{E}[W_i^\star \mid \overline{S}_i] &= \mathbb{E}[\mathbb{E}[W_i^\star \mid X_i, \overline{S}_i] \mid \overline{S}_i] \\
&= \mathbb{E}[\mathbb{E}[W_i \mid X_i, \overline{S}_i] \mid S_i] - \mathbb{E}[\mathbb{E}[W_i \mid X_i] \mid \overline{S}_i] \\
&= \mathbb{E}[\mathbb{E}[W_i \mid X_i] \mid \overline{S}_i] - \mathbb{E}[\mathbb{E}[W_i \mid X_i] \mid \overline{S}_i] = 0 \,,
\end{aligned} \tag{C.1.36}$$

almost surely. Now, define

$$\tilde{D}_{i,q}^\star = D_{i,q}^\star - \mathbb{E}[D_{i,q}^\star \mid S_q] \tag{C.1.37}$$

and observe that Assumption 3.4.3 implies that

$$\begin{aligned}
\mathbb{E}[D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}] \mid H_{i,q}, S_q] &= \mathbb{E}[D_{i,q} \mid H_{i,q}, S_q] - \mathbb{E}[D_{i,q} \mid H_{i,q}] \\
&= \psi_n(S_q) \\
&= \mathbb{E}[\psi_n(S_q) \mid S_q] \\
&= \mathbb{E}[\mathbb{E}[D_{i,q} \mid H_{i,j}, S_q] - \mathbb{E}[D_{i,q} \mid H_{i,q}] \mid S_q] \\
&= \mathbb{E}[D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}] \mid S_q] \,.
\end{aligned} \tag{C.1.38}$$

Consequently, we have that

$$\begin{aligned}
\tilde{D}_{i,q}^\star &= D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}] - \mathbb{E}[D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}] \mid S_q] \\
&= D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}] - \mathbb{E}[D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}] \mid H_{i,q}, S_q] \\
&= D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}, S_q] = D_{i,q} - \mathbb{E}[D_{i,q} \mid H_{i,q}, \overline{S}_q] \,,
\end{aligned} \tag{C.1.39}$$

almost surely. Thus, the conditions of Lemma C.1.1 are satisfied with $D^\star_{i,j}$ taking the role of $M_{i,j}$ and $W^\star_i$ taking the role of W_i. Hence, we obtain the expansion

$$\begin{aligned}\overline{\theta}^\star_n &= \frac{\mathbb{E}[Y_i\tilde{D}^\star_{i,q}W^\star_q] - \mathbb{E}[Y_i\tilde{D}^\star_{\ell,q}W^\star_q]}{\mathbb{E}[\mathrm{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\,\mathrm{Var}(W^\star_q \mid \overline{S}_q)]} \\ &\quad + \frac{1}{n}\frac{\mathbb{E}[Y_i\tilde{D}^\star_{\ell,q}W^\star_q] - \mathbb{E}[Y_i\tilde{D}^\star_{\ell,i}W^\star_i]}{\mathbb{E}[\mathrm{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\,\mathrm{Var}(W^\star_q \mid \overline{S}_q)]} + \frac{1}{n^2}\frac{\mathbb{E}[Y_i\tilde{D}^\star_{i,q}W^\star_q]}{\mathbb{E}[\mathrm{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\,\mathrm{Var}(W^\star_q \mid \overline{S}_q)]} \\ &\quad + \frac{n+1}{n^2}\frac{\mathbb{E}[\mathbb{E}[Y_iW^\star_q \mid G_{\ell,q},\overline{S}_q]\mathbb{E}[D^\star_{\ell,q} \mid H_{\ell,q},\overline{S}_q]]}{\mathbb{E}[\mathrm{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\,\mathrm{Var}(W^\star_q \mid \overline{S}_q)]} \\ &\quad - \frac{1}{n}\frac{\mathbb{E}[\mathbb{E}[Y_qW^\star_q \mid H_{\ell,q},\overline{S}_q]\mathbb{E}[D^\star_{\ell,q} \mid H_{\ell,q},\overline{S}_q]]}{\mathbb{E}[\mathrm{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\,\mathrm{Var}(W^\star_q \mid \overline{S}_q)]} + O(n^{-2})\ .\end{aligned} \tag{C.1.40}$$

Now, observe that Assumption 3.4.2 implies that

$$W_q \perp\!\!\!\perp D_{\ell,q} \mid S_q, H_{\ell,q}, X_q \quad \text{and} \quad \mathbb{E}[W_q \mid X_q] = \mathbb{E}[W_q \mid S_q, H_{\ell,q}] \tag{C.1.41}$$

respectively. Thus, Lemma C.1.2 implies that

$$\begin{aligned}\mathbb{E}[Y_i\tilde{D}^\star_{i,q}W^\star_q] &= \int_{\mathcal{D}}\int_{\mathcal{W}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W_q \geq w\}W^\star_q \mid \overline{S}_q]\mathbb{E}[\mathbb{I}\{D_{i,q} \geq \delta\}\tilde{D}^\star_{i,q} \mid H_{i,q},\overline{S}_q]\} \\ &\qquad \partial^2_{\delta,w}\mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, H_{i,q},\overline{S}_q]]\,\mathrm{d}w\,\mathrm{d}\delta\ , \\ \mathbb{E}[Y_i\tilde{D}_{\ell,q}W^\star_q] &= \int_{\mathcal{D}}\int_{\mathcal{W}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W_q \geq w\}W^\star_q \mid \overline{S}_q]\mathbb{E}[\mathbb{I}\{D_{\ell,q} \geq \delta\}\tilde{D}^\star_{\ell,q} \mid H_{i,q},\overline{S}_q] \\ &\qquad \partial^2_{\delta,w}\mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, H_{i,q},\overline{S}_q]]\,\mathrm{d}w\,\mathrm{d}\delta\ , \\ \mathbb{E}[Y_iW^\star_q \mid H_{i,q}, X_q, \overline{S}_q] &= \int_{\mathcal{W}} \mathbb{E}[\mathbb{I}\{W_q \geq w\}W^\star q \mid \overline{S}_q]\partial_w\mathbb{E}[Y_i \mid W^\star_q = w, H_{i,q},\overline{S}_q]\,\mathrm{d}w, \quad \text{and} \\ \mathbb{E}[Y_qW^\star_q \mid \overline{S}_q] &= \int_{\mathcal{W}} \mathbb{E}[\mathbb{I}\{W^\star_q \geq w\}W^\star_q \mid \overline{S}_q]\partial_w\mathbb{E}[Y_q \mid W^\star_q = w, H_{i,q},\overline{S}_q]\,\mathrm{d}w\ ,\end{aligned} \tag{C.1.42}$$

respectively. Consequently, by plugging each of the representations (C.1.42) into the expression (C.1.40), we find that

$$\begin{aligned}\overline{\theta}^\star_n &= \int_{\overline{\mathcal{D}}}\int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid H_{\ell,q},\overline{S}_q)(\partial^2_{\delta,w}\mathbb{E}[Y_\ell \mid D_{i,q} = \delta, W_q = w, H_{\ell,q},\overline{S}_q] \\ &\qquad - \partial^2_{\delta,w}\mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, H_{\ell,q},\overline{S}_q])]\,\mathrm{d}w\,\mathrm{d}\delta \\ &\quad + \frac{1}{n}\int_{\overline{\mathcal{D}}}\int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid H_{\ell,q},\overline{S}_q)(\partial^2_{\delta,w}\mathbb{E}[Y_i \mid D_{\ell,q} = \delta, H_{\ell,q},\overline{S}_q]\end{aligned} \tag{C.1.43}$$

$$
\begin{aligned}
&\qquad\qquad\qquad - \partial^2_{\delta,w}\mathbb{E}[Y_q \mid D_{\ell,q} = \delta, W_q = w, H_{\ell,q}, \overline{S}_q])]]\,\mathrm{d}w\,\mathrm{d}\delta \\
&+ \frac{1}{n^2}\int_{\overline{\mathcal{D}}}\int_{\overline{\mathcal{W}}}\mathbb{E}[\lambda(\delta, w \mid H_{\ell,q}, \overline{S}_q)\partial^2_{\delta,w}\mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, H_{\ell,q}, \overline{S}_q]]\,\mathrm{d}w\,\mathrm{d}\delta \\
&+ \frac{1}{n}\int_{\overline{\mathcal{W}}}\mathbb{E}\Big[\lambda'(w \mid H_{\ell,q}, \overline{S}_q)\Big(\Big(\frac{n+1}{n}\Big)\partial_w\mathbb{E}[Y_i \mid W_q = w, H_{\ell,q}, \overline{S}_q] \\
&\qquad\qquad\qquad - \partial_w\mathbb{E}[Y_q \mid W_q = w, H_{\ell,q}, \overline{S}_q]\Big)\Big]\,\mathrm{d}w + O(n^{-2})\ ,
\end{aligned}
$$

where

$$
\lambda(\delta, w \mid H_{\ell,q}, \overline{S}_q) = \frac{\operatorname{Cov}(\mathbb{I}\{D_{\ell,q} \geq \delta\}, D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)\operatorname{Cov}(\mathbb{I}\{W_\ell \geq w\}, W_\ell \mid \overline{S}_q)}{\mathbb{E}[\operatorname{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\operatorname{Var}(W^\star_q \mid \overline{S}_q)]} \quad \text{and} \tag{C.1.44}
$$

$$
\lambda'(w \mid X_q, H_{\ell,q}, \overline{S}_q) = \frac{\mathbb{E}[\mathbb{E}[D^\star_{\ell,q} \mid H_{\ell,q}, \overline{S}_q]\operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)}{\mathbb{E}[\operatorname{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\operatorname{Var}(W^\star_q \mid \overline{S}_q)]}\ ,
$$

respectively. Thus, it suffices to show that all but the first quantity in (C.1.43) are sufficiently small.

To this end, observe that Lemma C.1.3 implies that

$$
\begin{aligned}
\int_{\overline{\mathcal{W}}}\operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)\,\mathrm{d}w &= \operatorname{Var}(W_q \mid X_q, \overline{S}_q) \quad \text{and} \qquad (\text{C.1.45}) \\
\int_{\overline{\mathcal{D}}}\operatorname{Cov}(\mathbb{I}\{D_{\ell,q} \geq \delta\}, D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)\,\mathrm{d}\delta &= \operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)\ .
\end{aligned}
$$

Thus, it holds that

$$
\begin{aligned}
&\mathbb{E}\left[\int_{\overline{\mathcal{D}}}\int_{\overline{\mathcal{W}}}\lambda(\delta, w \mid H_{\ell,q}, \overline{S}_q)\,\mathrm{d}w\,\mathrm{d}\delta\right] \qquad (\text{C.1.46}) \\
&= \frac{\mathbb{E}\left[\int_{\overline{\mathcal{D}}}\int_{\overline{\mathcal{W}}}\mathbb{E}[\operatorname{Cov}(\mathbb{I}\{D_{\ell,q} \geq \delta\}, D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)\operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)]\,\mathrm{d}w\,\mathrm{d}\delta\right]}{\mathbb{E}[\operatorname{Var}(D^\star_{\ell,q} \mid \overline{S}_q)\operatorname{Var}(W^\star_q \mid \overline{S}_q)]} \\
&= \frac{\mathbb{E}\left[\mathbb{E}[\mathbb{E}[\int_{\overline{\mathcal{D}}}\operatorname{Cov}(\mathbb{I}\{D_{\ell,q} \geq \delta\}, D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)\,\mathrm{d}\delta \mid \overline{S}_q]\mathbb{E}[\int_{\overline{\mathcal{W}}}\operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)\,\mathrm{d}w \mid \overline{S}_q]]\right]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q]\operatorname{Var}(W_q \mid \overline{S}_q)]} \\
&= \frac{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q]\operatorname{Var}(W_q \mid \overline{S}_q)]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q]\operatorname{Var}(W_q \mid \overline{S}_q)]} = 1\ ,
\end{aligned}
$$

where the second equality follows from several applications of Fubini. Likewise, Lemma C.1.3 implies that

$$
\lambda(\delta, w \mid H_{\ell,q}, \overline{S}_q) \geq 0\ , \tag{C.1.47}
$$

almost surely. Hence, Assumption C.1.3 implies that

$$\int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{\ell,q}, \overline{S}_q) \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, H_{\ell,q}, \overline{S}_q]] \, \mathrm{d}w \, \mathrm{d}\delta = o(n^{-1/2}) \, , \tag{C.1.48}$$

$$\frac{1}{n} \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid X_q, H_{\ell,q}, \overline{S}_q)(\partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{\ell,q} = \delta, W_q = w, H_{\ell,q}, \overline{S}_q] - \partial^2_{\delta,w} \mathbb{E}[Y_q \mid D_{\ell,q} = \delta, W_q = w, H_{\ell,q}, \overline{S}_q])]] \, \mathrm{d}w \, \mathrm{d}\delta = O(n^{-1}) \, ,$$

and

$$\frac{1}{n^2} \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{\ell,q}, \overline{S}_q) \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,q} = \delta, W_q = w, H_{\ell,q}, \overline{S}_q]] \, \mathrm{d}w \, \mathrm{d}\delta = O(n^{-2}) \, , \tag{C.1.49}$$

respectively, and it so remains to bound the fourth term in (C.1.43).

Observe that we can decompose

$$\lambda'(w \mid H_{\ell,q}, \overline{S}_q) = \frac{\mathbb{E}[D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q] \operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} - \frac{\mathbb{E}[D_{\ell,q} \mid H_{\ell,q}] \operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} \, . \tag{C.1.50}$$

Lemma C.1.3 and the fact that the distances $D_{\ell,q}$ are positive imply that both terms in (C.1.50) are positive almost surely. Moreover, it holds that

$$\begin{aligned} &\frac{\mathbb{E}[\mathbb{E}[D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q] \operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} \\ &\lesssim \frac{\mathbb{E}[\mathbb{E}[\mathbb{E}[D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q] \mid \overline{S}_q] \sqrt{\operatorname{Var}(W_q \mid \overline{S}_q)}]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} \\ &\lesssim \frac{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \sqrt{\operatorname{Var}(W_q \mid \overline{S}_q)}]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} = O(1) \end{aligned} \tag{C.1.51}$$

uniformly in w, where the first inequality follows from Lemma C.1.3 and Cauchy-Schwarz and the second inequality follows from Assumption C.1.4. Analogously, it holds that

$$\frac{\mathbb{E}[\mathbb{E}[D_{\ell,q} \mid H_{\ell,q}] \operatorname{Cov}(\mathbb{I}\{W_q \geq w\}, W_q \mid \overline{S}_q)]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]}$$

$$
\begin{aligned}
&\leq \frac{\mathbb{E}[\mathbb{E}[\mathbb{E}[\mathbb{E}[D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q] \mid H_{\ell,q}] \mid \overline{S}_q] \sqrt{\operatorname{Var}(W_q \mid \overline{S}_q)}]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} \\
&\lesssim \frac{\mathbb{E}[\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid H_{\ell,q}] \mid \overline{S}_q] \sqrt{\operatorname{Var}(W_q \mid \overline{S}_q)}]}{\mathbb{E}[\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q) \mid \overline{S}_q] \operatorname{Var}(W_q \mid \overline{S}_q)]} \\
&\lesssim \frac{\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)]}{\mathbb{E}[\operatorname{Var}(D_{\ell,q} \mid H_{\ell,q}, \overline{S}_q)]} = O(1) && \text{(C.1.52)}
\end{aligned}
$$

uniformly in w, where the first inequality follows from Lemma C.1.3 and Cauchy-Schwarz and the second and third inequalities follow from Assumption C.1.4. Thus, by Assumption C.1.3 and the fact that the treatments and distances are defined on sets whose supports are bounded by a constant, we find that

$$
\begin{aligned}
\frac{1}{n} \int_{\overline{\mathcal{W}}} \mathbb{E}\bigg[\lambda'(w \mid H_{\ell,q}, \overline{S}_q)\bigg(\left(\frac{n+1}{n}\right) \partial_w \mathbb{E}[Y_i \mid W_q = w, H_{\ell,q}, \overline{S}_q] \\
- \partial_w \mathbb{E}[Y_q \mid W_q = w, H_{\ell,q}, \overline{S}_q]\bigg)\bigg] \, \mathrm{d}w = O(n^{-1}) \, .
\end{aligned} \tag{C.1.53}
$$

Hence, we obtain the representation

$$
\overline{\theta}_n^\star = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid H_j, S_j) \partial^2_{\delta,w} \, \mu(\delta, w \mid H_j, S_j)] \, \mathrm{d}w \, \mathrm{d}\delta + o(n^{-1/2}) \, , \tag{C.1.54}
$$

by plugging the bounds (C.1.48) and (C.1.53) into (C.1.43). The convexity of the weights is verified by the statements (C.1.46) and (C.1.47), completing the proof. ■

C.1.3 Proofs for Supporting Lemmas

C.1.3.1 Proof of Lemma C.1.1

We begin with the equality (C.1.13). Consider the decomposition

$$
\hat{\Xi}_i = \sum_{j \neq i} \left(M_{i,j} - \frac{1}{n} \sum_{k \neq j} M_{k,j} \right) W_j - \frac{1}{n} \sum_{k \neq i} M_{k,i} W_i = Q_i^{(0)} + Q^{(1)} - Q_i^{(2)} + Q^{(3)} \, , \tag{C.1.55}
$$

where

$$Q_i^{(0)} = \sum_{j\neq i}\left(M_{i,j} - \frac{1}{n-1}\sum_{k\neq j} M_{k,j}\right) W_j\ , \quad Q^{(1)} = \frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k\notin\{i,j\}} M_{k,j}W_j \tag{C.1.56}$$

$$Q_i^{(2)} = \frac{1}{n}\sum_{j\neq i} M_{j,i}W_i\ , \quad \text{and} \qquad Q^{(3)} = \frac{1}{n(n-1)}\sum_{j\neq i} M_{i,j}W_j\ . \tag{C.1.57}$$

We can evaluate

$$\begin{aligned}
\mathbb{E}[Y_iQ_i^{(0)}] &= \sum_{j\neq i}\mathbb{E}[Y_i\tilde{M}_{i,j}W_j] - \frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\mathbb{E}[Y_i\tilde{M}_{k,j}W_j] \\
&= \frac{(n-2)}{n-1}\sum_{j\neq i}\mathbb{E}[Y_i\tilde{M}_{i,j}W_j] - \frac{1}{n-1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\mathbb{E}[Y_i\tilde{M}_{k,j}W_j] \\
&= (n-2)\left(\mathbb{E}[Y_i\tilde{M}_{i,q}W_q] - \mathbb{E}[Y_i\tilde{M}_{\ell,q}W_q]\right)
\end{aligned} \tag{C.1.58}$$

and

$$\mathbb{E}[Y_i(Q^{(1)} - Q_i^{(2)} + Q^{(3)})] = \frac{n-2}{n}\mathbb{E}[Y_iM_{q,\ell}W_\ell] - \frac{n-1}{n}\mathbb{E}[Y_iM_{\ell,i}W_i] + \frac{1}{n}\mathbb{E}[Y_iM_{i,\ell}W_\ell]\ , \tag{C.1.59}$$

respectively. Hence, we have that

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[Y_i\hat{\Xi}_i] &= \frac{n-2}{n}\left(\mathbb{E}\left[Y_i\tilde{M}_{i,\ell}W_\ell\right] - \mathbb{E}\left[Y_i\tilde{M}_{q,\ell}W_\ell\right]\right) \\
&\quad + \frac{n-2}{n^2}\mathbb{E}[Y_iM_{q,\ell}W_\ell] - \frac{n-1}{n^2}\mathbb{E}[Y_iM_{\ell,i}W_i] + \frac{1}{n^2}\mathbb{E}[Y_iM_{i,\ell}W_\ell]
\end{aligned} \tag{C.1.60}$$

as required.

Next, we verify the inequality (C.1.14). Observe that

$$\begin{aligned}
\mathbb{E}[Q_i^{(0)}Q_i^{(1)}] &= \frac{1}{n(n-1)}\sum_{j\neq i}\sum_{j'\neq i}\sum_{k'\notin\{i,j'\}}\mathbb{E}\left[\left(M_{i,j} - \frac{1}{n-1}\sum_{k\neq j}M_{k,j}\right)M_{k',j'}^\top W_jW_j'\right] \\
&= \frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k'\notin\{i,j\}}\mathbb{E}\left[\left(\tilde{M}_{i,j} - \frac{1}{n-1}\sum_{k\neq j}\tilde{M}_{k,j}\right)M_{k',j}\operatorname{Var}(W_j \mid S_j)\right] \\
&= -\frac{1}{n(n-1)^2}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\mathbb{E}[\tilde{M}_{k,j}\tilde{M}_{k,j}\operatorname{Var}(W_j \mid S_j)]
\end{aligned}$$

$$= \frac{n-2}{n(n-1)}\mathbb{E}[\tilde{M}_{i,\ell}^2 \operatorname{Var}(W_\ell \mid S_\ell)] \tag{C.1.61}$$

as the variables W_j are mutually independent and satisfy the restrictions (C.1.11) and the variables $M_{i,j}$ and $M_{\ell,j}$ are independent conditional on S_j. We can evaluate

$$\mathbb{E}[Q_i^{(0)}Q_i^{(2)}] = 0 \quad \text{and} \quad \mathbb{E}[Q_i^{(0)}Q^{(3)}] = \frac{n-2}{n(n-1)}\mathbb{E}[\tilde{M}_{i,\ell}^2 \operatorname{Var}(W_\ell \mid S_\ell)] \tag{C.1.62}$$

analogously. As a consequence, it holds that

$$\begin{aligned}
&\mathbb{E}[\hat{\Xi}_i^2] - \operatorname{Var}(Q_i^{(0)}) \\
&\lesssim \frac{1}{n^2(n-1)^2}\sum_{j\neq i}^{n}\mathbb{E}\left[\sum_{k\notin\{i,j\}} M_{k,j}^2 W_j^2\right] + \frac{1}{n^2}\mathbb{E}\left[\sum_{j\neq i} M_{j,i}^2 W_i^2\right] \\
&+ \frac{1}{n^2(n-1)^2}\sum_{j\neq i}\mathbb{E}\left[M_{i,j}^2 W_j^2\right] + \frac{n-2}{n(n-1)}\mathbb{E}[\tilde{M}_{i,\ell}^2 \operatorname{Var}(W_\ell \mid S_\ell)] = O(n^{-1})\,, \quad \text{(C.1.63)}
\end{aligned}$$

by the Cauchy-Schwarz inequality and the fact that the supports of the variables W_j and $M_{i,j}$ are bounded by a constant. To conclude the proof, observe that we can evaluate

$$\operatorname{Var}(Q_i^{(0)}) = \mathbb{E}\left[\left(\sum_{j\neq i}\left(M_{i,j} - \frac{1}{n-1}\sum_{k\neq j} M_{k,j}\right)W_j\right)^2\right] = (n-2)\mathbb{E}\left[\tilde{M}_{i,\ell}^2 \operatorname{Var}(W_\ell \mid S_\ell)\right]\,, \tag{C.1.64}$$

by the same steps as before. Hence, the inequality (C.1.14) follows from (C.1.63) and (C.1.64). ■

C.1.3.2 Proof of Lemma C.1.2

The result is based on two applications of the following Lemma, obtained from a method of argument developed in Kolesár and Plagborg-Møller (2024).

Lemma C.1.4. *Fix the random variables X, Y, and U. If the distribution of X, conditioned on U, has support contained in the bounded set $\mathcal{X}$ in $\mathbb{R}$ almost surely, Y is real-valued, and the conditional expectation $\mathbb{E}[Y \mid X = x, Z]$ is locally absolutely continuous and uniformly bounded on $\mathcal{X}$, then the equality*

$$\operatorname{Cov}(Y, X \mid U) = \int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X \geq x\}(X - \mathbb{E}[X \mid U]) \mid U]\, \partial_x \mathbb{E}[Y \mid X = x, U]\, \mathrm{d}x \tag{C.1.65}$$

holds almost surely.

Observe that the equality (C.1.16) follows immediately from Lemma C.1.4. To verify the second equality, observe that

$$\mathbb{E}[Y(Z-\mathbb{E}[Z\mid U])(W-\mathbb{E}[W\mid U])] = \mathbb{E}[Y(Z-\mathbb{E}[Z\mid U,W])(W-\mathbb{E}[W\mid U])] \tag{C.1.66}$$

$$= \mathbb{E}[\mathrm{Cov}(Y,Z\mid U,W)(W-\mathbb{E}[W\mid U])], \tag{C.1.67}$$

as W is independent of Z conditional on U. Lemma C.1.4 implies that

$$\mathrm{Cov}(Y,Z\mid U,W) = \int_{\mathcal{Z}} \mathbb{E}[\mathbb{I}\{Z\geq z\}(Z-\mathbb{E}[Z\mid U,W])\mid U]\,\partial_z\mathbb{E}[Y\mid Z=z,U,W]\,\mathrm{d}z\,. \tag{C.1.68}$$

Thus, we can evaluate

$$\begin{aligned}
&\mathbb{E}[Y(Z-\mathbb{E}[Z\mid U])(W-\mathbb{E}[W\mid U])]\\
&=\mathbb{E}\left[\left(\int_{\mathcal{Z}}\mathbb{E}[\mathbb{I}\{Z\geq z\}(Z-\mathbb{E}[Z\mid U,W])\mid U,W]\,\partial_z\mathbb{E}[Y\mid Z=z,U,W]\,\mathrm{d}z\right)(W-\mathbb{E}[W\mid U])\right]\\
&=\int_{\mathcal{Z}}\mathbb{E}\left[\mathbb{E}[\mathbb{I}\{Z\geq z\}(Z-\mathbb{E}[Z\mid U,W])\mid U,W](W-\mathbb{E}[W\mid U])\,\partial_z\mathbb{E}[Y\mid Z=z,U,W]\right]\,\mathrm{d}z\\
&=\int_{\mathcal{Z}}\mathbb{E}\left[\mathbb{E}[\mathbb{I}\{Z\geq z\}(Z-\mathbb{E}[Z\mid U])\mid U]\mathbb{E}\left[(W-\mathbb{E}[W\mid U])\,\partial_z\mathbb{E}[Y\mid Z=z,U,W]\mid U\right]\right]\,\mathrm{d}z\,,
\end{aligned} \tag{C.1.69}$$

where the second equality follows from Fubini's Theorem and the third equality follows from the fact that W is independent of Z conditional on U. Lemma C.1.4 again implies that

$$\begin{aligned}
&\mathbb{E}\left[\partial_z\mathbb{E}[Y\mid Z=z,U,W](W-\mathbb{E}[W\mid U])\mid U\right]\\
&=\mathrm{Cov}(\partial_z\mathbb{E}[Y\mid Z=z,U,W],W\mid U)\\
&=\int_{\mathcal{W}}\mathbb{E}[\mathbb{I}\{W\geq w\}(W-\mathbb{E}[W\mid U])\mid U]\partial^2_{z,w}\mathbb{E}[Y\mid Z=z,W=w,U]\mid U]\,\mathrm{d}w\,.
\end{aligned} \tag{C.1.70}$$

By plugging (C.1.70) into (C.1.69), and applying Fubini's Theorem, we obtain

$$\mathbb{E}[Y(Z-\mathbb{E}[Z\mid U])(W-\mathbb{E}[W\mid U])]$$

$$= \int_{\mathcal{Z}} \int_{\mathcal{W}} \mathbb{E}[\mathbb{E}[\mathbb{I}\{W \geq w\}(W - \mathbb{E}[W \mid U]) \mid U]\mathbb{E}[\mathbb{I}\{Z \geq z\}(Z - \mathbb{E}[Z \mid U]) \mid U] \\ \partial^2_{z,w}\mathbb{E}[Y \mid Z = z, W = w, U]] \,\mathrm{d}w \,\mathrm{d}z \;, \tag{C.1.71}$$

as required. ∎

C.1.3.3 Proof of Lemma C.1.3

First, observe that

$$\begin{aligned}
&\mathrm{Cov}(\mathbb{I}\{Z \geq z\}, Z \mid U) \\
&= \mathbb{E}[\mathbb{I}\{Z \geq z\}(Z - \mathbb{E}[Z \mid U]) \mid U] \\
&= \mathrm{Var}(\mathbb{I}\{Z \geq z\} \mid U)(\mathbb{E}[(Z - \mathbb{E}[Z \mid U]) \mid Z \geq z, U] - \mathbb{E}[(Z - \mathbb{E}[Z \mid U]) \mid Z < z, U]) \geq 0 \;,
\end{aligned} \tag{C.1.72}$$

as required. In turn, by Fubini's theorem, it holds that

$$\begin{aligned}
&\int_{\mathcal{Z}} \mathrm{Cov}(\mathbb{I}\{Z \geq z\}, Z \mid U) \,\mathrm{d}z \\
&= \mathbb{E}\left[\int_{\mathcal{Z}} \mathbb{I}\{Z \geq z\}(Z - \mathbb{E}[Z \mid U]) \,\mathrm{d}z \mid U\right] = \mathrm{Var}(Z \mid U) \;,
\end{aligned} \tag{C.1.73}$$

as required. ∎

C.1.3.4 Proof of Lemma C.1.4

Define the conditionally centered random variable

$$\tilde{X} = X - \mathbb{E}[X \mid U] \;. \tag{C.1.74}$$

The fact that $\mathbb{E}[\tilde{X} \mid U] = 0$ implies that

$$\mathbb{E}[\mathbb{I}\{X \geq x\}\tilde{X} \mid U] = -\mathbb{E}[\mathbb{I}\{X < x\}\tilde{X} \mid U] \tag{C.1.75}$$

almost surely for each x in $\mathcal{X}$. Thus, choosing an arbitrary point x_0 in $\mathcal{X}$, we obtain

$$\begin{aligned}
&\int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X \geq x\}\tilde{X} \mid U]\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x \\
&\qquad = \int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X \geq x \geq x_0\}\tilde{X} \mid U]\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x
\end{aligned}$$

$$- \int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X < x < x_0\}\tilde{X} \mid U]\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x \;. \quad \text{(C.1.76)}$$

Now, we can evaluate

$$\begin{aligned}
&\int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X \geq x \geq x_0\}\tilde{X} \mid U]\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x \\
&= \mathbb{E}\left[\int_{\mathcal{X}} \mathbb{I}\{X \geq x \geq x_0\}\tilde{X}\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x \mid Z\right] \\
&= \mathbb{E}\left[\mathbb{I}\{X \geq x_0\}\tilde{X}(\mathbb{E}[Y \mid X, U] - \mathbb{E}[Y \mid X = x_0, U]) \mid Z\right] \qquad \text{(C.1.77)}
\end{aligned}$$

almost surely, where the first equality follows from Fubini's Theorem and the second equality follows from the Fundamental Theorem of Calculus, both of which are justified by the fact that the domain of X and the derivative $\partial_x \mathbb{E}[Y \mid X = x, U]$ are uniformly bounded, almost surely. By an identical argument we have that

$$\begin{aligned}
&\int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X < x < x_0\}\tilde{X} \mid U]\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x \\
&= \mathbb{E}\left[\mathbb{I}\{X < x_0\}\tilde{X}(\mathbb{E}[Y \mid X = x_0, U] - \mathbb{E}[Y \mid X, U]) \mid U\right] \qquad \text{(C.1.78)}
\end{aligned}$$

almost surely. Plugging (C.1.77) and (C.1.78) into (C.1.76) , we find that

$$\begin{aligned}
&\int_{\mathcal{X}} \mathbb{E}[\mathbb{I}\{X \geq x\}\tilde{X} \mid U]\partial_x \mathbb{E}[Y \mid X = x, U] \,\mathrm{d}x \\
&= \mathbb{E}\left[\tilde{X}\mathbb{E}[Y \mid X, U] \mid U\right] = \mathrm{Cov}(Y, X \mid U) \;, \qquad \text{(C.1.79)}
\end{aligned}$$

as required. ■

C.2. PROOFS FOR IDENTIFICATION RESULTS

In this Appendix, we give proofs for the corollaries stated in Section 3.4. Throughout, let the sets

$$\mathcal{D}(H_{i,j}, S_j) \quad \text{and} \quad \mathcal{W}(\overline{S}_j) \qquad \text{(C.2.1)}$$

denote the support sets of the proximity measure $D_{i,j}$, conditional on $H_{i,j}$ and S_j, and the treatment W_j, conditional on $\overline{S}_j$, respectively. Define the set

$$\mathcal{F} = \mathcal{F}(H_{i,j}, \overline{S}_j) = \mathcal{D}(H_{i,j}, S_j) \times \mathcal{W}(\overline{S}_j) \;. \qquad \text{(C.2.2)}$$

with an abuse of notation.

C.2.1 Proof of Corollary 3.4.1

Observe that the choice (3.3.4) is uniquely-defined on the event that $D_{i,j} = \delta$. Thus, Assumption 3.3.1 implies that, in this case, the potential outcome

$$Y_{i,j}(\delta, w) = F(Z_i, Z_j(S_j^{(i)}(\delta), w), Z_{-i,j}) \,, \quad \text{where} \quad Z_j(w, s) = (w, s, U_j) \,, \tag{C.2.3}$$

is uniquely-defined. Observe that Assumption 3.4.2 implies that the collections Z_i and $Z_{-i,j}$ are independent of treatment W_j. Thus, for each (δ, w) in $\mathcal{F}$, it holds that

$$\begin{aligned}
&\mathbb{E}[Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j] \\
&= \mathbb{E}[F(Z_i, Z_j(\xi_j^{(i)}(\delta), w), Z_{-i,j}) \mid D_{i,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j] \\
&= \mathbb{E}[F(Z_i, Z_j(\xi_j^{(i)}(\delta), w), Z_{-i,j}) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j] \\
&= \mathbb{E}[Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j] \,.
\end{aligned} \tag{C.2.4}$$

Hence, we can evaluate

$$\begin{aligned}
\theta_n^\star &= \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \partial_{\delta,w}^2 \, \mathbb{E}[Y_{i,j} \mid D_{i,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j]] \, \mathrm{d}w \, \mathrm{d}\delta \\
&= \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \partial_{\delta,w}^2 \, \mathbb{E}[Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, W_j = w, H_{i,j}, \overline{S}_j]] \, \mathrm{d}w \, \mathrm{d}\delta \\
&= \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \partial_{\delta} \, \mathbb{E}[\partial_w \, Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j]] \, \mathrm{d}w \, \mathrm{d}\delta \,,
\end{aligned} \tag{C.2.5}$$

where the first two equalities follows by definition and the third equality follows by (C.2.4) and the Leibniz integral rule. Observe that, by the definition of the weight function $\lambda(\delta, w \mid H_{i,j}, S_j)$ and the convention used to define the operators ∂_δ, ∂_w, and $\partial_{\delta,w}^2$, the final equality in (C.2.5) only requires the application of the equality (C.2.4) at points (δ, w) in $\mathcal{F}$. ■

C.2.2 Proof of Corollary 3.4.2

Fix a scalar δ in $\mathcal{D}(H_{i,j}, S_j)$. Observe that the level set

$$\mathcal{S}^{(i)}(\delta) = \{s : D_n(S_i, s) = \delta\} \,. \tag{C.2.6}$$

is non-empty by definition. As the function $\overline{D}(\cdot)$ is continuous and monotone, by Assumption 3.3.2, the set

$$\{r \geq 0 : \overline{D}(\cdot) = \delta\} \tag{C.2.7}$$

is a closed interval $[r_-, r_+]$. Now, observe that

$$S_j + \alpha(S_i - S_j) \in \mathcal{S}^{(i)}(\delta) \quad \text{if and only if} \quad |1-\alpha|\|S_i - S_j\| \in [r_-, r_+]\;. \tag{C.2.8}$$

If the distance $\|S_i - S_j\| = 0$, then every α is feasible, and $\alpha_j^{(i)}(\delta)$ is uniquely defined as 0. On the event that $\|S_i - S_j\| > 0$, we have that

$$\arg\min_{\alpha} \{|\alpha| : S_j + \alpha(S_i - S_j) \in \mathcal{S}^{(i)}(\delta)\} = 1 - \frac{r^\star}{\|S_i - S_j\|}\;, \tag{C.2.9}$$

where

$$r^\star \in \arg\min_{r\in[r_-,r_+]} \left|1 - \frac{r}{\|S_i - S_j\|}\right| = r^\star \in \arg\min_{r\in[r_-,r_+]} |r - \|S_i - S_j\||\;. \tag{C.2.10}$$

Observe that $r^\star$ is defined as the projection of the point $\|S_i - S_j\|$ onto the closed interval $[r_-, r_+]$, which is unique. Thus, the minimizer of the problem (C.2.9) is unique, and so the counterfactual coordinate

$$S_j^{(i)}(\delta) = S_j + \alpha_j^{(i)}(\delta)(S_i - S_j) \tag{C.2.11}$$

is uniquely defined. Hence, the potential outcome

$$Y_{i,j}(\delta, w) = F(Z_i, Z_j(S_j^{(i)}(\delta), w), Z_{-i,j})\;, \tag{C.2.12}$$

is uniquely defined.

As a consequence, for each (δ, w) in $\mathcal{F}$, Assumption 3.4.4 implies that

$$\mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j] = \mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid H_{i,j}, \overline{S}_j]\;. \tag{C.2.13}$$

Hence, we can evaluate

$$\theta_n^\star = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j)\partial_\delta\, \mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_j]]\, \mathrm{d}w\, \mathrm{d}\delta$$

$$
\begin{aligned}
&= \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \partial_\delta \, \mathbb{E}[\partial_w \, Y_{i,j}(\delta, w) \mid H_{i,j}, \overline{S}_j]] \, \mathrm{d}w \, \mathrm{d}\delta \\
&= \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \mathbb{E}[\partial^2_{\delta,w} \, Y_{i,j}(\delta, w) \mid H_{i,j}, \overline{S}_j]] \, \mathrm{d}w \, \mathrm{d}\delta \\
&= \int_{\mathcal{D}} \int_{\mathcal{W}} \mathbb{E}[\lambda(\delta, w \mid H_{i,j}, \overline{S}_j) \partial^2_{\delta,w} \, Y_{i,j}(\delta, w)] \, \mathrm{d}w \, \mathrm{d}\delta \;, \qquad \text{(C.2.14)}
\end{aligned}
$$

where the first equality follows by Corollary 3.4.1, the second equality follows by (C.2.13), the third equality follows by the Leibniz integral rule, and the final equality follows by the law of iterated expectations. Again, by the definition of the weight function $\lambda(\delta, w \mid H_{i,j}, \overline{S}_j)$ and the operators ∂_δ and ∂_w, the third equality in (C.2.14) only requires the application of the equality (C.2.13) at points δ in $\mathcal{D}(H_{i,j}, S_j)$. ■

C.3. Proofs for Asymptotic Results

In this Appendix, we give proofs for the theorems stated in Section 3.5 and Section 3.6, respectively. These results hold under additional regularity conditions, stated in Appendix C.3.1. Proofs are given in the subsequent subsections. Proofs for all supporting Lemmas are given in Appendix C.4. We use the notation

$$
\tilde{D}^\star_{i,\ell} = \tilde{D}^\star(S_i, S_\ell) = D_{i,\ell} - \mathbb{E}[D_{i,\ell} \mid H_{i,\ell}, S_\ell] \qquad \text{(C.3.1)}
$$

throughout.

C.3.1 Additional Regularity Conditions

C.3.1.1 Theorems 3.5.1 and 3.5.2

We begin by giving the additional moment bounds needed for Theorems 3.5.1 and 3.5.2. Throughout, we let $\mathcal{A}_{i,j}$ denote the event that $A_{i,j} = 1$ and let $\tilde{\mathcal{A}}_{i,j}$ denote its complement. Likewise, define the indicator

$$
B_{i,j} = \mathbb{I}\{P\{\overline{S}_k \in L^{(1)}_n(\overline{S}_i, \overline{S}_j) \mid \overline{S}_i, \overline{S}_j\} > 0\} \qquad \text{(C.3.2)}
$$

where the set $L^{(1)}_n(\overline{S}_i, \overline{S}_j)$ is defined in Assumption 3.5.2. Define the events $\mathcal{B}_{i,j}$ and $\tilde{\mathcal{B}}_{i,j}$, accordingly.

Assumption C.3.1. *Define the conditional expectations*

$$\overline{F}(Z_i, Z_j) = \mathbb{E}[Y_i \mid Z_i, Z_j] \quad \text{and} \quad \tilde{F}(Z_i, Z_j) = \mathbb{E}[Y_i \mid Z_i, Z_j] - \mathbb{E}[Y_i \mid Z_i] \, . \tag{C.3.3}$$

The means and variances of each of the conditional moments (C.3.3) are bounded by a constant almost surely. Moreover, for each $a, b \in \{0, 1\}$, the moments

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[\overline{F}(Z_i, Z_j)^2 \mid A_{i,j} = a, Z_i]] \, , && \mathbb{E}[\mathbb{E}[\overline{F}(Z_i, Z_j)^2 \mid A_{i,j} = a, Z_j]] \, , \\ &\mathbb{E}[\mathbb{E}[\overline{F}(Z_i, Z_j)^2 \mid A_{k,j} = a, Z_k]] \, , && \mathbb{E}[\mathbb{E}[\tilde{F}(Z_i, Z_j)^2 \mid A_{i,j} = a, Z_j]] \, , \\ &\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i, Z_j)^2 \mid A_{k,j} = a, Z_k]] \, , && \mathbb{E}[\tilde{F}(Z_i, Z_k)^2 \mid A_{i,j} = a, A_{i,k} = 1] \, , \\ \text{and} \quad &\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i, Z_k)^2 \mid A_{i,j} = a, Z_j, Z_k] \mid B_{j,k} = b] \end{aligned} \tag{C.3.4}$$

are each bounded by a constant almost surely.

This condition is relatively innocuous, and states that, in the asymptotic sequences under consideration, the moments (C.3.4) do not grow without bound. We note that this condition does not necessarily imply that the variance of the outcomes is bounded by a constant. This is important in settings in which, for instance, each unit has an increasing number of neighbors whose influence on a given unit's outcome is on the order of a constant. We go through the trouble of enumerating the moment bounds enforced by Assumption C.3.1 because the proof of Theorem 3.5.2 is based on constructing distributions that satisfy the conditions of Theorem 3.5.1, but where the distributions of the quantities

$$\mathbb{E}[Y_i \mid Z_i, Z_j] \quad \text{and} \quad \mathbb{E}[Y_i \mid Z_i, Z_j] - \mathbb{E}[Y_i \mid Z_i] \tag{C.3.5}$$

are unbounded.

C.3.1.2 Theorems 3.6.1 and 3.6.2

Next, we state the three additional conditions needed for Theorems 3.6.1 and 3.6.2. The first condition is a mild strengthening of Assumption 3.5.2.

Assumption C.3.2 (Extended Latent Spatial Structure)**.** *Define the set*

$$L_n^{(2)}(\overline{S}_i, \overline{S}_j) = L_n^{(1)}(\overline{S}_i) \cap L_n^{(1)}(\overline{S}_j) \, . \tag{C.3.6}$$

It holds that

$$P\{P\{\overline{S}_k \in L_n^{(2)}(\overline{S}_i, \overline{S}_j) \mid \overline{S}_i, \overline{S}_j\} > 0\} = O(\rho_n)\ . \tag{C.3.7}$$

Recall that the second part of the relation (3.5.6), stated in Assumption 3.5.2, ensures that the fraction of pairs of units i, j that could feasibly be proximate to a third unit k is of order $O(\rho_n)$. Assumption C.3.2 is a further iteration of this idea. That is, the fraction of pairs of units i, j that *could* each be proximate to units that are proximate to k is of order $O(\rho_n)$. As before, this condition is satisfied in settings where the latent proximity indicators are generated by the specification (3.5.10), and the measure of the latent factors is doubling.

The second condition is necessary for ensuring that the leading term in our asymptotic approximation to the estimator $\hat{\theta}_n^\star$ is non-degenerate.

Assumption C.3.3 (Non-Degeneracy)**.** *Define the conditional expectation*

$$\Psi(Z_i, Z_j) = \mathbb{E}[\tilde{F}(Z_k, Z_i)\tilde{D}_{k,j}^\star \mid Z_i, Z_j]\ . \tag{C.3.8}$$

The lower bound

$$\mathrm{Var}\left(\begin{pmatrix} \Psi(Z_i, Z_j)W_j^\star + \Psi(Z_j, Z_i)W_i^\star \\ 2\mathbb{E}[\tilde{D}_{k,i}\tilde{D}_{k,j}^\star \mid Z_i, Z_j]W_i^\star W_j^\star \end{pmatrix} \mid A_{i,j} = 1\right) \succeq c\rho_n^2 \mathbf{I}_2 \tag{C.3.9}$$

holds, where $\preceq$ denotes the Loewner order on matrices.

To unpack Assumption C.3.3, consider the case that the outcomes are mean independent of the states of non-proximate units, in the sense that

$$\text{if} \quad A_{i,j} = 0 \quad \text{then} \quad \mathbb{E}[Y_i \mid Z_i, Z_j] - \mathbb{E}[Y_i \mid Z_i] = 0 \tag{C.3.10}$$

almost surely. In this case, Assumption C.3.3 reduces to the lower bound

$$\mathrm{Var}\left(\tilde{\Psi}(Z_j, Z_i)W_j^\star + \tilde{\Psi}(Z_i, Z_j)W_i^\star \mid A_{i,j}\right) \succeq c\mathbf{I}_2\ , \tag{C.3.11}$$

where

$$\tilde{\Psi}(Z_i, Z_j) = \mathbb{E}\left[\begin{pmatrix} \tilde{F}_{k,i} \\ \tilde{D}_{k,i}^\star W_i^\star \end{pmatrix} \tilde{D}_{k,j}^\star \mid A_{k,j}, Z_i, Z_j\right]\ . \tag{C.3.12}$$

Consider, in turn, the closely related, but even simpler, lower bound

$$\mathrm{Var}(\mathbb{E}[\tilde{F}_{k,i} D_{k,j} \mid A_{k,j}, Z_i, Z_j] \mid A_{i,j} = 1) \geq c \,. \tag{C.3.13}$$

In effect, the lower bound (C.3.13) asks that, when i is proximate to j *and* k, the probability that k is also proximate to j is on the order of a constant. To see this, observe that in the latter event, i.e., when k is proximate to both i and j, the term $\tilde{F}(Z_k, Z_i) D_{k,j}$ can be on the order of a constant.

In other words, Assumption C.3.3 can be interpreted as stipulating that when i and j are proximate, the proportion of j's neighbors that are also proximate to i is on the order of a constant. Clustering, in this sense, is widely observed in social network data and arises in models where connections are generated by the specifications analogous to (3.5.10), as well as many other standard models of network formation (see e.g., Jackson and Rogers 2007 and Graham 2016 for further discussion). However, Assumption C.3.3 also asks that there no cancellations that cause the eigenvalues of the variance covariance matrix in (C.3.11) to be much smaller than the variance in (C.3.13).

The final condition enforces a slightly stronger set of moment bounds. Define the indicator

$$E_{i,j} = \mathbb{I}\{P\{\overline{S}_k \in L_n^{(2)}(\overline{S}_i, \overline{S}_j) \mid \overline{S}_i, \overline{S}_j\} > 0\} \,,$$

where the sets $L_n^{(2)}(S_i, S_j)$ is defined in Assumption C.3.2. Define the events $\mathcal{E}_{i,j}$, and $\tilde{\mathcal{E}}_{i,j}$, accordingly.

Assumption C.3.4. *For each $b, b', e \in \{0, 1\}$, the moment*

$$\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^4 \mid \mathcal{A}_{4,3}, Z_1, Z_2, Z_3] \mid B_{3,1} = b, B_{3,2} = b', Z_1, Z_2] \mid E_{1,2} = e] \tag{C.3.14}$$

is bounded by a constant.

Assumption C.3.4 again holds in the setting where the quantities (C.3.5) are bounded by constants almost surely.

C.3.2 Proof of Theorem 3.5.1

We continue to use the notation introduced in the proof of Theorem 3.4.1. For the sake of simplicity, we restrict attention to the case that data W_i, $D_{i,j}$, X_i, and $H_{i,j}$ have been unconditionally centered. The general case, that accounts for the inclusion of intercepts in the estimators $\hat{\pi}_n$ and $\hat{\gamma}_n$, follows from the same argument, but the expansions applied below are more tedious to verify. Throughout, we let $\overline{M} = M - \mathbb{E}[M]$ denote the unconditionally centered version of the random variable M.

Our objective is to give an asymptotic approximation to the estimator

$$\hat{\theta}_n^\star = \arg\min_\theta \min_\alpha \left\{ \sum_{i=1}^n (Y_i - \alpha - \theta \cdot \hat{\Delta}_i^\star)^2 \right\} . \tag{C.3.15}$$

By the Frisch-Waugh-Lovell and Projection Theorems, we can write

$$\hat{\theta}_n^\star = \frac{\sum_{i=1}^n Y_i \tilde{\Delta}_i^\star}{\sum_{i=1}^n (\tilde{\Delta}_i^\star)^2} , \tag{C.3.16}$$

where

$$\tilde{\Delta}_i^\star = \sum_{j\neq i} (\hat{D}_{i,j}^\star - \frac{1}{n}\sum_{k\neq j} \hat{D}_{k,j}^\star)\hat{W}_j^\star - \frac{1}{n}\sum_{j\neq i} \hat{D}_{j,i}^\star \hat{W}_i^\star . \tag{C.3.17}$$

We successively evaluate the denominator and numerator of the representation (C.3.16).

First, we evaluate the denominator. With a mild abuse of notation, we let

$$\pi_n = \arg\min_\pi \mathbb{E}\left[(\overline{W}_i - \pi^\top \overline{X}_i)^2\right] \quad \text{and} \quad \gamma_n = \arg\min_\gamma \mathbb{E}\left[(\overline{D}_{i,\ell} - \gamma^\top \overline{H}_{i,\ell})^2\right] \tag{C.3.18}$$

denote the population best linear predictors associated with the two nuisance parameters and

$$\hat{\pi}_{0,n} = \pi_n - \hat{\pi}_n \quad \text{and} \quad \hat{\gamma}_{0,n} = \gamma_n - \hat{\gamma}_n \tag{C.3.19}$$

denote their respective estimation errors. Consider the decomposition

$$\begin{aligned}\tilde{\Delta}_i^\star &= \sum_{j\neq i}(D_{i,j}^\star - \frac{1}{n}\sum_{k\neq j} D_{k,j}^\star)W_j^\star - \frac{1}{n}\sum_{j\neq i} D_{j,i}^\star W_i^\star \\ &+ \hat{\gamma}_{0,n}^\top \left(\sum_{j\neq i}(\overline{H}_{i,j} - \frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})W_j^\star - \frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}W_i^\star \right) \\ &+ \hat{\gamma}_{0,n}^\top \left(\sum_{j\neq i}(\overline{H}_{i,j} - \frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})\overline{X}_j^\top - \frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}\tilde{X}_i^\top \right) \hat{\pi}_{0,n} .\end{aligned} \tag{C.3.20}$$

The leading order behavior of the denominator of (C.3.16) is determined by the first term in

(C.3.20). To evaluate this term, consider the further decomposition

$$\begin{aligned}
&\sum_{j\neq i}(D^\star_{i,j} - \frac{1}{n}\sum_{k\neq j} D^\star_{k,j})W^\star_j - \frac{1}{n}\sum_{j\neq i} D^\star_{j,i}W^\star_i \\
&= \sum_{j\neq i} \tilde{D}^\star_{i,j}W^\star_j - \frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\tilde{D}^\star_{k,j}W^\star_j + \frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k\neq j} D^\star_{k,j}W^\star_j - \frac{1}{n}\sum_{j\neq i} D^\star_{j,i}W^\star_i \ ,
\end{aligned} \tag{C.3.21}$$

where we have used the fact that Assumption 3.4.2 implies that

$$\tilde{D}^\star_{i,j} = D_{i,j} - \mathbb{E}[D_{i,j} \mid H_{i,j}, S_j] = D^\star_{i,j} - \mathbb{E}[D^\star_{i,j} \mid S_j] \ . \tag{C.3.22}$$

The following Lemma quantifies the contributions of the various terms in (C.3.21).

Lemma C.3.1. *If the conditions of Theorem 3.5.1 hold, then:*

(i) *It holds that*

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n\left(\sum_{j\neq i}\tilde{D}^\star_{i,j}W^\star_j\right)^2 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j}\tilde{D}^\star_{i,j}\tilde{D}^\star_{i,k}W^\star_jW^\star_k \\
&\quad + \mathbb{E}[\mathrm{Var}(D_{i,\ell} \mid H_{i,j}, S_\ell)\,\mathrm{Var}(W_\ell \mid \overline{S}_\ell)] + O_p(n^{-1/2}\rho_n) \\
&= \mathbb{E}[\mathrm{Var}(D_{i,\ell} \mid H_{i,j}, S_\ell)\,\mathrm{Var}(W_\ell \mid \overline{S}_\ell)] + O_p(\rho_n^{3/2})
\end{aligned} \tag{C.3.23}$$

and

$$\frac{1}{n^2}\sum_{i=1}^n\left(\frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\tilde{D}^\star_{k,j}W^\star_j\right)^2 = O_p(n^{-1}\rho_n) \ ,$$

respectively.

(ii) *It holds that*

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n\left(\frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k\neq j}D^\star_{k,j}W^\star_j\right)^2 &= O(n^{-2}) \ , \quad \textit{and} \\
\frac{1}{n^2}\sum_{i=1}^n\left(\frac{1}{n}\sum_{j\neq i}D^\star_{j,i}W^\star_i\right)^2 &= O(n^{-1}\rho_n) \ ,
\end{aligned} \tag{C.3.24}$$

respectively.

As a consequence, we obtain the representation

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}(D^{\star}_{i,j}-\frac{1}{n}\sum_{k\neq j}D^{\star}_{k,j})W^{\star}_j-\frac{1}{n}\sum_{j\neq i}D^{\star}_{j,i}W^{\star}_i\right)^2 \qquad \text{(C.3.25)}\\
&=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j}\tilde{D}^{\star}_{i,j}\tilde{D}^{\star}_{i,k}W^{\star}_jW^{\star}_k+\mathbb{E}[\operatorname{Var}(D_{i,\ell}\mid H_{i,j},S_\ell)\operatorname{Var}(W_\ell\mid\overline{S}_\ell)]+O_p(n^{-1/2}\rho_n)\\
&=\mathbb{E}[\operatorname{Var}(D_{i,\ell}\mid H_{i,\ell},S_\ell)\operatorname{Var}(W_\ell\mid\overline{S}_\ell)]+O(\rho_n^{3/2})\ ,
\end{aligned}$$

by plugging the bounds stated in Lemma C.3.1 into the decomposition (C.3.21) and applying Cauchy-Schwarz.

In turn, we handle the second two terms in (C.3.20) through the application of the following Lemma.

Lemma C.3.2. *Assume that the covariates X_i and $H_{i,j}$ have lengths q and p, respectively. Let $\hat{\pi}_{n,k}$ and $\hat{\gamma}_{n,k}$ denote the kth components of the nuisance parameter estimators $\hat{\pi}_n$ and $\hat{\gamma}_n$. Define $\pi_{n,k}$ and $\gamma_{n,k}$ accordingly. If the conditions of Theorem 3.5.1 hold, then:*

(i) *It holds that*

$$\hat{\pi}_{n,k}-\pi_{n,k}=O_p(n^{-1/2}) \quad \textit{and} \quad \hat{\gamma}_{n,k}-\gamma_{n,k}=O_p(n^{-1/2}\kappa_{n,k}^{-1}\rho_n)\ , \qquad \text{(C.3.26)}$$

respectively.

(ii) *Define the quantities*

$$\begin{aligned}
J_i&=\sum_{j\neq i}(\overline{H}_{i,j}-\frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})W^{\star}_j-\frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}W^{\star}_i \quad \textit{and} \qquad \text{(C.3.27)}\\
M_i&=\sum_{j\neq i}(\overline{H}_{i,j}-\frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})\overline{X}_j^{\top}-\frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}\overline{X}_i^{\top}\ .
\end{aligned}$$

Let $J_{i,r}$ denote the rth component of J_i and $M_{i,r}=(M_{i,r,s})_{s=1}^{q}$ denote the rth column of M_i. It holds that

$$\frac{1}{n^2}\sum_{i=1}^{n}J_{i,s}J_{i,r}=O_p(\kappa_{n,s}^{1/2}\kappa_{n,r}^{1/2}) \quad \textit{and} \quad \frac{1}{n^2}\sum_{i=1}^{n}M_{i,s,r}M_{i,s',r'}=O_p(n\kappa_{n,s}^{1/2}\kappa_{n,s'}^{1/2}) \qquad \text{(C.3.28)}$$

respectively.

In particular, observe that Lemma C.3.2, a union bound, and Assumption 3.5.3 imply that

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{\gamma}_{0,n}^{\top}\left(\sum_{j\neq i}(\overline{H}_{i,j}-\frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})W_j^{\star}-\frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}W_i^{\star}\right)\right)^2\\
&=\sum_{s=1}^{p}\sum_{r=1}^{p}(\hat{\gamma}_{n,s}-\gamma_{n,s})\left(\frac{1}{n^2}\sum_{i=1}^{n}J_{i,s}J_{i,r}\right)(\hat{\gamma}_{n,r}-\gamma_{n,r})\\
&=\sum_{s=1}^{p}\sum_{r=1}^{p}O_p(n^{-1}\kappa_{n,s}^{-1/2}\kappa_{n,r}^{-1/2}\rho_n^2)=O_p(n^{-1}\rho_n)\;. \qquad \text{(C.3.29)}
\end{aligned}$$

Likewise, by an analogous argument, we can evaluate

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{\gamma}_{0,n}^{\top}\left(\sum_{j\neq i}(\overline{H}_{i,j}-\frac{1}{n}\sum_{k\neq j}H_{k,j})\overline{X}_j^{\top}-\frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}\overline{X}_i^{\top}\right)\hat{\pi}_{0,n}\right)^2\\
&=\sum_{s=1}^{p}\sum_{s'=1}^{p}\sum_{r=1}^{q}\sum_{r'=1}^{q}(\hat{\gamma}_{n,s}-\gamma_{n,s})(\hat{\gamma}_{n,s'}-\gamma_{n,s'})\\
&\qquad\qquad(\hat{\pi}_{n,r}-\pi_{n,r})(\hat{\pi}_{n,r'}-\pi_{n,r'})\left(\frac{1}{n^2}\sum_{i=1}^{n}M_{i,s,r}M_{i,s',r'}\right)\\
&=\sum_{s=1}^{p}\sum_{s'=1}^{p}\sum_{r=1}^{q}\sum_{r'=1}^{q}O_p(n^{-1}(\kappa_{n,s}\kappa_{n,s'})^{-1/2}\rho_n^2)=O_p(n^{-1}\rho_n)\;. \qquad \text{(C.3.30)}
\end{aligned}$$

Thus, we can conclude that

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}(\tilde{\Delta}_i^{\star})^2\\
&=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j}\tilde{D}_{i,j}^{\star}\tilde{D}_{i,k}^{\star}W_jW_k+\mathbb{E}[\mathrm{Var}(D_{i,\ell}\mid H_{i,j},S_\ell)\,\mathrm{Var}(W_\ell\mid\overline{S}_\ell)]+O_p(n^{-1/2}\rho_n)\\
&=\mathbb{E}[\mathrm{Var}(D_{i,\ell}\mid H_{i,\ell},S_\ell)\,\mathrm{Var}(W_\ell\mid\overline{S}_\ell)]+O_p(\rho_n^{3/2})\;, \qquad \text{(C.3.31)}
\end{aligned}$$

by squaring the decomposition (C.3.20) within an expectation, applying Cauchy-Schwarz, and plugging in the bounds (C.3.25), (C.3.29), and (C.3.30).

Next, we evaluate the numerator of (C.3.16). Here, we consider the decomposition

$$\frac{1}{n^2}\sum_{i=1}^{n}Y_i\tilde{\Delta}_i^{\star}=\frac{1}{n^2}\sum_{i=1}^{n}Y_i\left(\sum_{j\neq i}(D_{i,j}^{\star}-\frac{1}{n}\sum_{k\neq j}D_{k,j}^{\star})W_j^{\star}-\frac{1}{n}\sum_{j\neq i}D_{j,i}^{\star}W_i^{\star}\right) \qquad \text{(C.3.32)}$$

$$+ \hat{\gamma}_{0,n}^\top \left(\frac{1}{n^2} \sum_{i=1}^n Y_i \left(\sum_{j \neq i} (\overline{H}_{i,j} - \frac{1}{n} \sum_{k \neq j} \overline{H}_{k,j}) W_j^\star - \frac{1}{n} \sum_{j \neq i} \overline{H}_{j,i} W_i^\star \right) \right)$$
$$+ \hat{\gamma}_{0,n}^\top \left(\frac{1}{n^2} \sum_{i=1}^n Y_i \left(\sum_{j \neq i} (\overline{H}_{i,j} - \frac{1}{n} \sum_{k \neq j} \overline{H}_{k,j}) \overline{X}_j^\top - \frac{1}{n} \sum_{j \neq i} \overline{H}_{j,i} \overline{X}_i^\top \right) \right) \hat{\pi}_{0,n} .$$

The leading order behavior of this quantity is determined by the first term in (C.3.32). To evaluate this term, we again consider the further decomposition

$$\begin{aligned} &\frac{1}{n^2} \sum_{i=1}^n Y_i \left(\sum_{j \neq i} (D_{i,j}^\star - \frac{1}{n} \sum_{k \neq j} D_{k,j}^\star) W_j^\star - \frac{1}{n} \sum_{j \neq i} D_{j,i}^\star W_i^\star \right) \\ &= \frac{n-2}{n-1} \frac{1}{n^2} \sum_{i=1}^n Y_i \check{\Delta}_i^\star - \frac{1}{n-1} \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} Y_i \check{\Delta}_{i,j}^\star \\ &+ \frac{1}{n^2} \sum_{i=1}^n Y_i (Q_i^{(1)} - Q_i^{(2)} + Q_i^{(3)}) , \end{aligned} \tag{C.3.33}$$

where

$$\check{\Delta}_i^\star = \sum_{j \neq i} \tilde{D}_{i,j}^\star W_j^\star , \qquad \check{\Delta}_{i,j}^\star = \sum_{k \notin \{i,j\}} \tilde{D}_{k,j}^\star W_j^\star , \tag{C.3.34}$$
$$Q_i^{(1)} = \frac{1}{n(n-1)} \sum_{j \neq i} \sum_{k \notin \{i,j\}} D_{k,j}^\star W_j^\star , \qquad Q_i^{(2)} = \frac{1}{n-1} \sum_{j \neq i} D_{j,i}^\star W_i^\star , \quad \text{and}$$
$$Q_i^{(3)} = \frac{1}{n(n-1)} \sum_{j \neq i} D_{i,j}^\star W_j^\star ,$$

respectively. The following Lemma quantifies the contributions of the various terms in (C.3.33).

Lemma C.3.3. *Define the conditional expectation*

$$\tilde{F}(Z_i, Z_k) = \mathbb{E}[Y_i \mid Z_k, Z_i] - \mathbb{E}[Y_i \mid Z_i] . \tag{C.3.35}$$

If the conditions of Theorem 3.4.1 hold, then:

***(i)** It holds that*

$$\frac{1}{n^2} \sum_{i=1}^n Y_i \check{\Delta}_i^\star = \mathbb{E}[Y_i \tilde{D}_{i,j}^\star W_j] + \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{F}(Z_i, Z_k) \tilde{D}_{i,j}^\star W_j^\star + O_p(n^{-1/2} \rho_n) \quad \textit{and} \tag{C.3.36}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{F}(Z_i,Z_k)\tilde{D}^{\star}_{i,j}W^{\star}_j = O_p(\rho_n^{3/2})\ , \tag{C.3.37}$$

respectively.

(ii) *It holds that*

$$\frac{1}{n-1}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}Y_i\check{\Delta}^{\star}_{i,j} = \mathbb{E}[Y_i\tilde{D}^{\star}_{\ell,j}W^{\star}_j] + O_p(n^{-1/2}\rho_n) \quad \textit{and} \tag{C.3.38}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}Y_i(Q_i^{(1)} - Q_i^{(2)} + Q_i^{(3)}) = \frac{1}{n}\mathbb{E}[Y_iD^{\star}_{q,\ell}W_\ell] - \frac{1}{n}\mathbb{E}[Y_iD^{\star}_{\ell,i}W_i] \tag{C.3.39}$$
$$+\frac{1}{n^2}\mathbb{E}[Y_iD^{\star}_{i,\ell}W_\ell]) + O_p(n^{-1}\rho_n^{1/2})\ ,$$

respectively.

Consequently, Lemma C.3.3, as well as the bounds (C.1.48), (C.1.49), and (C.1.53), imply that

$$\frac{1}{n^2}\sum_{i=1}^{n}Y_i\left(\sum_{j\neq i}(D^{\star}_{i,j} - \frac{1}{n}\sum_{k\neq j}D^{\star}_{k,j})W^{\star}_j - \frac{1}{n}\sum_{j\neq i}D^{\star}_{j,i}W^{\star}_i\right) = \mathbb{E}[Y_i\tilde{D}^{\star}_{i,\ell}W^{\star}_\ell] + O_p(\rho_n^{3/2})\ . \tag{C.3.40}$$

In turn, we handle the second two terms in (C.3.32) through the application of the following Lemma.

Lemma C.3.4. *If the conditions of Theorem 3.4.1 hold, then*

$$\frac{1}{n^2}\sum_{i=1}^{n}Y_iJ_{i,s} = O_p(\kappa_{n,s}) \quad \textit{and} \quad \frac{1}{n^2}\sum_{i=1}^{n}Y_iM_{i,s,r} = O_p(n^{1/2}\kappa_{n,s})\ , \tag{C.3.41}$$

respectively.

In particular, Lemma C.3.2 and Lemma C.3.4 imply that

$$\hat{\gamma}_{0,n}^{\top}\left(\frac{1}{n^2}\sum_{i=1}^{n}Y_i\left(\sum_{j\neq i}(\overline{H}_{i,j} - \frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})W^{\star}_j - \frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}W^{\star}_i\right)\right) = O_p(n^{-1/2}\rho_n) \quad \text{and} \tag{C.3.42}$$

$$\hat{\gamma}_{0,n}^{\top}\left(\frac{1}{n^2}\sum_{i=1}^{n}Y_i\left(\sum_{j\neq i}(\overline{H}_{i,j}-\frac{1}{n}\sum_{k\neq j}\overline{H}_{k,j})\overline{X}_j^{\top}-\frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}\overline{X}_i^{\top}\right)\right)\hat{\pi}_{0,n}=O_p(n^{-1/2}\rho_n)\ ,$$

respectively. Thus, we obtain

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}Y_i\tilde{\Delta}_i^{\star}&=\mathbb{E}[Y_i\tilde{D}_{i,\ell}^{\star}W_\ell^{\star}]+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{F}(Z_i,Z_k)\tilde{D}_{i,j}^{\star}W_j^{\star}+O_p(n^{-1/2}\rho_n)\\&=\mathbb{E}[Y_i\tilde{D}_{i,\ell}^{\star}W_\ell^{\star}]+O_p(\rho_n^{3/2})\ .\end{aligned}\tag{C.3.43}$$

by plugging the bounds (C.3.40) and (C.3.42) into the decomposition (C.3.32).

To conclude the proof, observe that Assumption C.1.4 and Assumption 3.5.2 imply that

$$\begin{aligned}&\mathbb{E}[\operatorname{Var}(D_{i,\ell}\mid H_{i,\ell},S_\ell)\operatorname{Var}(W_\ell\mid\overline{S}_\ell)]\\&\gtrsim\mathbb{E}[\mathbb{E}[D_{i,\ell}\mid H_{i,\ell},S_\ell]\operatorname{Var}(W_\ell\mid\overline{S}_\ell)]\\&=\mathbb{E}[\mathbb{E}[D_{i,\ell}\mid Z_\ell]\operatorname{Var}(W_\ell\mid\overline{S}_\ell)]\gtrsim\rho_n\ .\end{aligned}\tag{C.3.44}$$

Thus, the representations (C.1.42), (C.3.31), (C.3.43), and a Taylor expansion, imply that

$$\begin{aligned}\hat{\theta}_n^{\star}&=\frac{\sum_{i=1}^{n}Y_i\tilde{\Delta}_i^{\star}}{\sum_{i=1}^{n}(\tilde{\Delta}_i^{\star})^2}\\&=\overline{\theta}_n^{\star}+O_p(\rho_n^{3/2}\mathbb{E}[\operatorname{Var}(D_{i,\ell}\mid G_{i,\ell},S_\ell)\operatorname{Var}(W_i\mid X_\ell,S_\ell)]^{-1})\\&=\int_{\mathcal{D}}\int_{\mathcal{W}}\mathbb{E}[\lambda(\delta,w\mid G_{i,j},X_j,\overline{S}_j)\partial_{\delta,w}^2\,\mu(\delta,w\mid G_{i,j},X_j,\overline{S}_j)]\,\mathrm{d}w\,\mathrm{d}\delta+O_p(\rho_n^{1/2})\ ,\end{aligned}\tag{C.3.45}$$

where the final equality follows by the bound (C.3.44) and Theorem 3.4.1, completing the proof. ■

C.3.3 Proof of Theorem 3.5.2

To ease exposition, we write $\mathcal{P}_n=\mathcal{P}_n(\rho_n)$. Moreover, for the sake of simplicity, we restrict attention to the case that ρ_n^{-1} is an integer that divides n. The more general result can be obtained by the same argument, at the cost of more involved notation. The desired bound follows from an application of Le Cam's two-point method, stated as follows.

Lemma C.3.5 (Equation 15.4, Wainwright, 2019)**.** *Let X be a random variable, valued on* *the domain $\mathcal{X}$ and drawn according to a distribution P in the class $\mathcal{P}$. Let $\theta(P)$ denote a real-valued parameter on $\mathcal{P}$. Let the distributions P_0 and P_1 in $\mathcal{P}$ satisfy $|\theta(P_0) - \theta(P_1)| \geq 2\delta$ for some positive constant δ. For any measurable function $\hat{\theta} : \mathcal{X} \mapsto \mathbb{R}$, it holds that*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[(\hat{\theta}(X) - \theta(P))^2 \right] \geq \frac{\delta^2}{2} (1 - \|P_0 - P_1\|_{\mathsf{TV}}) ,$$

where $\|\cdot\|_{\mathsf{TV}}$ denotes the total variation norm on probability measures.

To apply Lemma C.3.5, we require some further notation. Observe that each distribution P^n in the class $\mathcal{P}_n$ induces a distribution $\hat{P}^n$ for the observable data $(Y_i, W_i, D_i)_{i=1}^n$. In turn, define the collection $J_n = (W_i, D_i)_{i=1}^n$ and let $\hat{P}^n_{|J_n}$ denote the distribution of the outcomes $(Y_i)_{i=1}^n$, conditioned on J_n, under the base distribution $\hat{P}^n$.

With this in place, we begin by specifying two distributions, P_0^n and P_1^n, in $\mathcal{P}_n$. In both cases, the treatments W_i are i.i.d. Bernoulli with success rate $1/2$. Likewise, in both cases, the coordinates S_i are mutually independent, independent of the treatments, and uniformly distributed on the set $\{1, \dots, \rho_n^{-1}\}$. Again, in both cases, the distances $D_{i,j}$ and latent proximity indicators are given by

$$A_{i,j} = D_{i,j} = \mathbb{I}\{S_i = S_j\} . \tag{C.3.46}$$

Now, again in both cases, the collection $(U_i)_{i=1}^n$ are real-valued, i.i.d., independent of the treatments W_i and coordinates ξ_i, and have a standard Gaussian distribution. Consequently, it holds that

$$P\{A_{i,j} = 1 \mid S_i\} = \rho_n . \tag{C.3.47}$$

It remains to specify the distribution of the outcomes. Under P_0^n, the outcomes Y_i are generated by

$$Y_i = U_i + \sum_{j \neq i} D_{i,j} U_j . \tag{C.3.48}$$

In turn, under P_1^n, the outcomes are generated by

$$Y_i = \psi_n \left(\left(W_i - \frac{1}{2} \right) + \sum_{j \neq i} D_{i,j} \left(W_j - \frac{1}{2} \right) \right) + \left(U_i + \sum_{j \neq i} D_{i,j} U_j \right) , \tag{C.3.49}$$

for a sequence ψ_n that will be specified at a later point in the proof. The specification of P_0^n and P_1^n is now complete. It can be verified that both distributions are elements of $\mathcal{P}_n$.

Observe that, in both cases, as W_j and $D_{i,j}$ are binary, and their variances, conditioned on S_j, are constant, the weight function $\lambda(\delta, w \mid \overline{S}_j)$ is equal to zero almost surely if δ or w is equal to zero, and is equal to one otherwise. Thus, the parameter $\hat{\theta}^\star(\hat{P}^n)$ can be expressed as the double difference

$$\hat{\theta}^\star(\hat{P}^n) = \mathbb{E}_{\hat{P}^n}[(\mu(1,1 \mid \overline{S}_j) - \mu(1,0 \mid \overline{S}_j)) - (\mu(0,1 \mid \overline{S}_j) - \mu(0,0 \mid \overline{S}_j))] \,, \quad \text{(C.3.50)}$$

where we recall that $\mu(\delta, w \mid \overline{S}_j) = \mathbb{E}[Y_i \mid D_{i,j} = \delta, W_j = w, \overline{S}_j]$ for each $\delta, w \in \{0,1\}$. Hence, we can evaluate

$$\hat{\theta}^\star(\hat{P}_0^n) = 0 \quad \text{and} \quad \hat{\theta}^\star(\hat{P}_1^n) = \psi_n \,, \quad \text{(C.3.51)}$$

respectively

Thus, it remains to bound the total variation distance between P_0^n and P_1^n. We require some further notation. The random variables

$$B_k = \sum_{i=1}^n \mathbb{I}\{S_i = k\} \quad \text{and} \quad V_k = \sum_{i=1}^n \mathbb{I}\{S_i = k\} \left(W_j - \frac{1}{2}\right) \quad \text{(C.3.52)}$$

measure the number of units whose coordinate is equal to k and the aggregate, centered treatments associated with those units, respectively. In turn, we let $\mathsf{N}^b(m, v)$ denote the distribution of the b-vector $(X, \ldots, X)$, where X has the distribution $\mathsf{N}(m, v)$. With this in place, observe that, conditional on J_n, both P_0^n and P_1^n can be expressed as product distributions though

$$\hat{P}_{0|J_n}^n = \prod_{k=1}^{\rho_n^{-1}} \mathsf{N}^{B_k}(0, B_k) \quad \text{and} \quad \hat{P}_{1|J_n}^n = \prod_{k=1}^{\rho_n^{-1}} \mathsf{N}^{B_k}(\psi_n V_k, B_k) \,, \quad \text{(C.3.53)}$$

respectively. Consequently, we obtain the bound

$$\begin{aligned}
\|\hat{P}_0^n - \hat{P}_1^n\|_{\mathsf{TV}}^2 &\leq \frac{1}{2} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_0^n, \hat{P}_1^n\right) \\
&\leq \frac{1}{2} \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{0|J_n}^n, \hat{P}_{1|J_n}^n\right)\right] \\
&\leq \frac{1}{2} \sum_{k=1}^{\rho_n^{-1}} \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(\mathsf{N}^{B_k}(0, B_k), \mathsf{N}^{B_k}(\psi_n V_k, B_k)\right) \mid B_k \neq 0\right] P\{B_k \neq 0\}
\end{aligned}$$

$$\leq \frac{1}{2}\sum_{k=1}^{\rho_n^{-1}} \mathbb{E}\left[\mathsf{D}_{\mathrm{KL}}\left(\mathsf{N}(0, B_k), \mathsf{N}(\psi_n V_k, B_k)\right) \mid B_k \neq 0\right] \ , \tag{C.3.54}$$

where $\mathsf{D}_{\mathrm{KL}}(\cdot, \cdot)$ denotes the KL-divergence between probability measures, the first inequality follows from Pinsker's inequality (e.g., Lemma 15.2 of Wainwright, 2019), the second inequality follows from the joint-convexity of the KL-divergence (e.g., Proposition 2.2.11 of Duchi, 2024), the third inequality follows from the tensorization of the KL-divergence over product distributions (e.g., Equation 15.11 of Wainwright, 2019), and the final inequality follows from the fact that the KL divergence is invariant under invertible transformations. Now, observe that, we can evaluate

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{D}_{\mathrm{KL}}\left(\mathsf{N}(0, B_k), \mathsf{N}(\psi_n V_k, B_k)\right) \mid B_k \neq 0\right] \\
&= \psi_n^2 \mathbb{E}[V_k^2 B_k^{-1} \mid B_k \neq 0] \\
&= \psi_n^2 \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{I}\{S_i = k\}\left(W_j - \frac{1}{2}\right)\right)^2 \mid S_1, \ldots, S_n\right] B_k^{-1} \mid B_k \neq 0\right] \\
&= \frac{\psi_n^2}{4}\mathbb{E}\left[B_k B_k^{-1} \mid B_k \neq 0\right] = \frac{\psi_n^2}{4}
\end{aligned} \tag{C.3.55}$$

where the first equality follows from the standard characterization of the KL-divergence between two Gaussian distributions with the same variance, the second equality follows by the assumed independence between the treatments and the coordinates. Thus, we find that

$$\|\hat{P}_0^n - \hat{P}_1^n\|_{\mathsf{TV}}^2 \leq \frac{1}{8}\rho_n^{-1}\psi_n^2 \tag{C.3.56}$$

by plugging the bound (C.3.55) into (C.3.54). Hence, by making the choice $\psi_n = \rho_n^{1/2}$, we can conclude that

$$\mathfrak{M}_n(\rho_n) \gtrsim \rho_n^{1/2} \tag{C.3.57}$$

by applying Lemma C.3.5 with the distributions P_0^n and P_1^n, as required. ∎

C.3.4 Proof of Theorem 3.6.1

We apply the following Lemma at several points in the argument.

Lemma C.3.6. *Define the sequences*

$$\sigma_{n,1}^2 = \mathbb{E}\left[\left(\mathbb{E}[\tilde{F}(Z_1, Z_3)\tilde{D}^\star_{1,2} \mid Z_2, Z_3]W_2^\star + \mathbb{E}[\tilde{F}(Z_1, Z_2)\tilde{D}^\star_{1,3} \mid Z_2, Z_3]W_3^\star\right)^2\right] , \tag{C.3.58}$$

$$\sigma_{n,2}^2 = 4\mathbb{E}\left[\left(\mathbb{E}[\tilde{D}_{1,3}\tilde{D}^\star_{1,2} \mid Z_2, Z_3]W_3^\star W_2^\star\right)^2\right] , \quad \textit{and}$$
$$\varsigma_n = 4\mathbb{E}\left[\mathbb{E}[\tilde{F}(Z_1, Z_3)\tilde{D}^\star_{1,2} \mid Z_2, Z_3]\mathbb{E}[\tilde{D}^\star_{1,2}W_2^\star\tilde{D}^\star_{1,3} \mid Z_2, Z_3]W_2^\star W_3^\star\right] , \tag{C.3.59}$$

and the matrix

$$\Sigma_n = \begin{pmatrix} \sigma_{n,1}^2 & \varsigma_n \\ \varsigma_n & \sigma_{n,2}^2 \end{pmatrix} . \tag{C.3.60}$$

Under the conditions of Theorem 3.6.1, it holds that

$$c\rho_n^3\mathbf{I}_2 \preceq \Sigma_n \preceq C\rho_n^3\mathbf{I}_2 , \tag{C.3.61}$$

where $\preceq$ denotes the Loewner order on matrices.

Proof. The sequence of inequalities (C.3.61) is equivalent to the assertion that

$$\xi^\top\Sigma_n\xi \asymp \rho_n^3 \tag{C.3.62}$$

uniformly over the collection of real-valued vectors $\xi = (\xi_1, \xi_2)^\top$ that satisfy $\xi_1^2 + \xi_2^2 = 1$. The upper bound $\xi^\top\hat{\Sigma}_n\xi \lesssim \rho_n^3$ follows immediately from the the bounds (C.4.134) and (C.4.185) stated in the Proof of Lemma C.4.5 and the Proof of Part (iii) of Lemma C.4.7, respectively. The lower bound $\xi^\top\hat{\Sigma}_n\xi \gtrsim \rho_n^3$ follows immediately from Assumptions 3.5.2 and C.3.3, respectively. ■

The result is an application of the following central limit theorem, obtained by applying a general result concerning the Gaussian approximation of degenerate U-statistics, due to Liu et al. (2025).

Lemma C.3.7. *Under the conditions of Theorem 3.6.2, it holds that*

$$\frac{\sqrt{2}}{n^2}\sum_{i=1}^n\sum_{j\neq i}\sum_{k\notin\{i,j\}}\Sigma_n^{-1/2}\begin{pmatrix}\tilde{F}(Z_i, Z_k)\tilde{D}^\star_{i,j}W_j^\star \\ \tilde{D}^\star_{i,k}\tilde{D}^\star_{i,j}W_j^\star W_k^\star\end{pmatrix} \xrightarrow{d} \mathsf{N}(\mathbf{0}_2, \mathbf{I}_2) , \tag{C.3.63}$$

where Σ_n is defined in the statement of Lemma C.3.6.

Recall the expansions

$$\frac{1}{n^2}\sum_{i=1}^{n} Y_i\tilde{\Delta}_i^\star = \mathbb{E}[Y_i\tilde{D}_{i,\ell}W_\ell^\star] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i,Z_k)\tilde{D}_{i,j}^\star W_j^\star + O_p(n^{-1/2}\rho_n)\ , \tag{C.3.64}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}(\tilde{\Delta}_i^\star)^2 = \mathbb{E}[(\tilde{D}_{i,\ell}^\star W_\ell^\star)^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{D}_{i,k}^\star W_k^\star\tilde{D}_{i,j}^\star W_j^\star + O_p(n^{-1/2}\rho_n)\ , \tag{C.3.65}$$

stated as (C.3.43) and (C.3.31) in the Proof of Theorem 3.5.1. Thus, we obtain

$$\frac{\sqrt{2}}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\Sigma_n^{-1/2}\begin{pmatrix} Y_i\tilde{\Delta}_i^\star - \mathbb{E}[\tilde{D}_{i,\ell}W_\ell^\star] \\ (\tilde{\Delta}_i^\star)^2 - \mathbb{E}[(\tilde{D}_{i,\ell}^\star W_\ell^\star)^2]\end{pmatrix} \xrightarrow{d} \mathsf{N}(\mathbf{0}_2,\mathbf{I}_2)\ , \tag{C.3.66}$$

by Lemmas C.3.6 and C.3.7 and the continuous mapping theorem.

Now, recall the representation

$$\hat{\theta}_n^\star = \frac{\frac{1}{n^2}\sum_{i=1}^{n} Y_i\tilde{\Delta}_i^\star}{\frac{1}{n^2}\sum_{i=1}^{n}(\tilde{\Delta}_i^\star)^2}\ , \tag{C.3.67}$$

obtained in the proof of Theorem 3.5.1. By an application of the Delta method, the approximation (C.3.66) implies that

$$\sigma_n^{-1}(\hat{\theta}_n^\star - \theta_n^\star) \xrightarrow{d} \mathsf{N}(0,1) \tag{C.3.68}$$

where

$$\begin{aligned}\sigma_n^2 &= \frac{1}{2\mathbb{E}[(\tilde{D}_{i,\ell}^\star W_\ell^\star)^2]^2}\left(\sigma_{n,1}^2 - 2\varsigma_n\theta_n^\star + \sigma_{n,2}^2(\theta_n^\star)^2\right) \\ &= \frac{1}{2\mathbb{E}[(\tilde{D}_{i,j}^\star W_j^\star)^2]^2}\operatorname{Var}(\mathbb{E}[\tilde{\varepsilon}_{k,i}\tilde{D}_{k,j}^\star \mid Z_i,Z_j]W_j^\star + \mathbb{E}[\tilde{\varepsilon}_{k,j}\tilde{D}_{k,i}^\star \mid Z_j,Z_i]W_i^\star)\ .\end{aligned} \tag{C.3.69}$$

We conclude the proof by noting that the relation $\sigma_n^{-1} \asymp \rho_n^{1/2}$ follows immediately from Lemma C.3.6 and the bound

$$\mathbb{E}[(\tilde{D}_{i,j}^\star W_j^\star)^2] \gtrsim \rho_n\ , \tag{C.3.70}$$

implied by Assumptions 3.5.2 and C.1.4, as required. ■

C.3.5 Proof of Theorem 3.6.2

The result follows from an application of the following lemma, due to Hoeffding (1952).

Lemma C.3.8 (Theorem 17.2.3, Lehmann and Romano, 2022)**.** *Suppose that $\mathcal{X}^n$ has distribution P_n in $\mathcal{X}_n$ and $\mathbf{G}_n$ is a finite group of transformations from $\mathcal{X}_n$ to $\mathcal{X}_n$. Let G_n be a random variable that is uniform on $\mathbf{G}_n$. Also, let G_n' have the same distribution as G_n, with X^n, G_n, and G_n' mutually independent. Define the randomization distribution*

$$\hat{R}_n(t) = M_n^{-1} \sum_{g \in \mathbf{G}_n} \mathbb{I}\{T_n(gX^n) \leq t\} , \tag{C.3.71}$$

where M_n is the cardinality of $\mathbf{G}_n$. Suppose, under P_n, it holds that

$$(T_n(G_n X^n), T_n(G_n' X_n)) \xrightarrow{d} (T, T') , \tag{C.3.72}$$

where T and T' are independent, each with common c.d.f. $R(\cdot)$. Then, under P_n, it holds that

$$\hat{R}_n(t) \xrightarrow{p} R(t) , \tag{C.3.73}$$

for every t that is a continuity point of $R(\cdot)$.

To apply this strategy to our setting, we require some intermediate results. First, we use the following central limit theorem. We omit the proof of this result, as it follows from an argument that is identical to the proof of Lemma C.3.7.

Lemma C.3.9. *Fix any scalar θ. Define the random variable*

$$\tilde{\varepsilon}_\theta(Z_i, Z_j) = \tilde{F}(Z_i, Z_j) - \theta \tilde{D}_{i,j} W_j^\star \tag{C.3.74}$$

and the sequences

$$\varphi_{n,1}^2(\theta) = 2\,\mathrm{Var}\left(\mathbb{E}[\tilde{\varepsilon}_\theta(Z_i, Z_j)\tilde{D}_{1,2}^\star \mid Z_2, Z_3] W_2^\star\right) \quad \text{and} \quad \tilde{\Sigma}_n(\theta) = \begin{pmatrix} \varphi_{n,1}^2(\theta) & 0 \\ 0 & \sigma_{n,2}^2 \end{pmatrix} \tag{C.3.75}$$

respectively. Let $V = (V_i)_{i=1}^n$ and $V' = (V_i')_{i=1}^n$ denote two independent collections of i.i.d.

Rademacher random variables. Under the conditions of Theorem 3.6.2, it holds that

$$\frac{\sqrt{2}}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\begin{pmatrix}\tilde{\Sigma}_n^{-1/2}(\theta) & \mathbf{0}_{2\times 2}\\ \mathbf{0}_{2\times 2} & \tilde{\Sigma}_n^{-1/2}(\theta)\end{pmatrix}\begin{pmatrix}\tilde{\varepsilon}_\theta(Z_i,Z_k)\tilde{D}^\star_{i,j}W^\star_jV_j\\ \tilde{D}^\star_{i,k}\tilde{D}^\star_{i,j}W^\star_jW^\star_kV_jV_k\\ \tilde{\varepsilon}_\theta(Z_i,Z_k)\tilde{D}^\star_{i,j}W^\star_jV'_j\\ \tilde{D}^\star_{i,k}\tilde{D}^\star_{i,j}W^\star_jW^\star_kV'_jV'_k\end{pmatrix}\xrightarrow{d}\mathsf{N}(\mathbf{0}_4,\mathbf{I}_4)\ , \tag{C.3.76}$$

respectively.

Moreover, we make use of the following expansions, which follow from arguments identical to the proofs of Lemmas C.3.1 to C.3.4, respectively. Again, we have omitted the details for reasons of space.

Lemma C.3.10. *Let $V=(V_i)_{i=1}^n$ denote an independent collection of i.i.d. Rademacher random variables. Define the variable*

$$\tilde{\Delta}^\star_i(V)=\sum_{j\neq i}(\hat{D}^\star_{i,j}-\frac{1}{n}\hat{D}^\star_{k,j})\hat{W}^\star_jV_j-\frac{1}{n}\sum_{j\neq i}\hat{D}^\star_{j,i}\hat{W}^\star_i\ . \tag{C.3.77}$$

Fix any scalar θ. Under the conditions of Theorem 3.5.1, it holds that

$$\frac{1}{n^2}\sum_{i=1}^{n}(Y_i-\theta\tilde{\Delta}^\star_i)\tilde{\Delta}^\star_i(V)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\varepsilon_\theta(Z_i,Z_k)\tilde{D}^\star_{i,j}W^\star_jV_j+O_p(n^{-1/2}\rho_n)\quad\textit{and} \tag{C.3.78}$$

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}(\tilde{\Delta}^\star_i(V))^2&=\mathbb{E}[(\tilde{D}^\star_{i,\ell}W^\star_\ell)^2]\\&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,k}W^\star_kV_k\tilde{D}^\star_{i,j}W^\star_jV_j+O_p(n^{-1/2}\rho_n)\ ,\end{aligned}$$

respectively.

With these results in place, consider the statistic

$$\hat{\phi}^\star_n(V,\theta)=\frac{\frac{1}{n^2}\sum_{i=1}^{n}(Y_i-\theta\tilde{\Delta}^\star_i)\tilde{\Delta}^\star_i(V)}{\frac{1}{n^2}\sum_{i=1}^{n}(\tilde{\Delta}^\star_i(V))^2} \tag{C.3.79}$$

and observe that Lemmas C.3.9 and C.3.10, the continuous mapping theorem, and the Delta method imply that

$$\begin{pmatrix} \varphi_n^{-1/2}(\theta)\hat{\phi}_n^\star(V,\theta) \\ \varphi_n^{-1/2}(\theta)\hat{\phi}_n^\star(V',\theta) \end{pmatrix} \xrightarrow{d} \mathsf{N}(\mathbf{0}_2, \mathbf{I}_2) , \tag{C.3.80}$$

where

$$\varphi_n^{-1/2}(\theta) = \frac{1}{\mathbb{E}[(\tilde{D}_{i,j}^\star W_j^\star)^2]^2} \operatorname{Var}(\mathbb{E}[\tilde{\varepsilon}_\theta(Z_k, Z_i)\tilde{D}_{k,j}^\star \mid Z_i, Z_j]W_j^\star) . \tag{C.3.81}$$

Thus, Lemma C.3.8 implies that

$$\frac{1}{2^n} \sum_{v\in\{-1,1\}^n} \mathbb{I}\{\varphi_n^{-1/2}(\theta)\hat{\phi}_n^\star(v,\theta) \leq x\} \xrightarrow{p} \Phi(x) , \tag{C.3.82}$$

where $\Phi(x)$ denotes the standard normal cumulative distribution function. Now, observe that

$$\hat{\phi}_n^\star(V) = \hat{\phi}_n^\star(V, \hat{\theta}_n^\star) \tag{C.3.83}$$

by the Frisch-Waugh-Lovell Theorem, and so

$$\hat{R}_n(x) = \frac{1}{2^n} \sum_{v\in\{-1,1\}^n} \mathbb{I}\{\varphi_n^{-1/2}\hat{\phi}_n^\star(v) \leq x\} = \frac{1}{2^n} \sum_{v\in\{-1,1\}^n} \mathbb{I}\{\varphi_n^{-1/2}(\theta_n^\star)\hat{\phi}_n^\star(v, \hat{\theta}_n^\star) \leq x\} \xrightarrow{p} \Phi(x) \tag{C.3.84}$$

by Theorem 3.5.1 and Slutsky's Lemma. Observe that the relation

$$\rho_n \leq \sigma_n^2 \leq 2\varphi_n^2 \lesssim \rho_n^2 \tag{C.3.85}$$

follows immediately from Lemma C.3.6, the bound (C.3.70), and Cauchy-Schwarz, as required. ∎

C.4. Proofs for Lemmas Supporting Appendix C.3

Throughout the proofs stated in this section, we apply results concerning the decomposition and asymptotic approximation of symmetric statistics. In particular, at several points, our characterization of the limiting behavior of various statistics is facilitated by Hoeffding's decomposition (Hoeffding, 1948; Efron and Stein, 1981). We adopt the following somewhat nonstandard formulation, considered in Lachièze-Rey and Peccati (2017), based on the use of stochastic differences.

Lemma C.4.1 (Theorem 2.2, Lachièze-Rey and Peccati, 2017)**.** *Let $X = (X_i)_{i=1}^n$ denote a sequence of independent random variables valued in $\mathcal{X}$. Let X' denote an independent copy of X. Construct $X^{(i)}$ by replacing X_i with X_i' in X, leaving all other entries unchanged. For each i, define the difference operator*

$$\nabla_i f(X) = f(X) - f(X^{(i)}) \tag{C.4.1}$$

on the set of all measurable functions $f : \mathcal{X}^n \to \mathbb{R}$. If the moment $\mathbb{E}[f(X)^2]$ is finite, then the representation

$$f(X) = \mathbb{E}[f(X)] + \sum_{m=1}^{n} \sum_{1 \le i_1 < \cdots < i_m < n} f^{(m)}(X_{i_1}, ..., X_{i_m}) \ , \quad \text{where} \tag{C.4.2}$$

$$f^{(m)}(X_{i_1}, ..., X_{i_m}) = (-1)^m \mathbb{E}[\nabla_{i_1} \cdots \nabla_{i_m} f(X', X) \mid X_{i_1}, ..., X_{i_m}] \ ,$$

holds. Moreover, all 2^n terms appearing on the right-hand side of (C.4.2) are mean-zero and uncorrelated.

When applied to U-statistics, Lemma C.4.1 has the following, widely-known effect.

Lemma C.4.2 (Lemmas 5.1.5 A and 5.2.1 A, Serfling, 1980)**.** *Let $X = (X_i)_{i=1}^n$ denote an i.i.d. sequence of random variables valued in $\mathcal{X}$. Suppose that the function $f : \mathcal{X}^b \to \mathbb{R}$ is symmetric in its arguments and that the statistic $f(X_1, \ldots, X_b)$ is mean-zero.*
(i) *It holds that*

$$F_n = \frac{1}{\binom{n}{b}} \sum_{1 \le i_1 \le \cdots \le i_b \le n} f(X_{i_1}, \ldots, X_{i_b}) = \sum_{m=1}^{b} \binom{b}{m} F_{n,m} \ , \quad \text{where} \tag{C.4.3}$$

$$F_{n,m} = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 \le \cdots \le i_m \le n} f^{(m)}(X_{i_1}, \ldots, X_{i_m}) \tag{C.4.4}$$

and the quantities $f^{(m)}(\cdot)$ are defined in Lemma C.4.1.
(ii) *The decomposition*

$$\operatorname{Var}(F_n) = \binom{n}{b}^{-1} \sum_{m=1}^{b} \binom{b}{m} \binom{n-b}{b-m} \operatorname{Var}(E[f(X_1, \ldots, X_n \mid X_1, \ldots, X_m]) \tag{C.4.5}$$

holds.

At some points, we will make use of the following two, more exotic results. The first is a version of a second order Efron-Stein inequality, due to Bobkov et al. (2019).

Lemma C.4.3 (Theorem 1.8, Bobkov et al., 2019). *Let $X = (X_i)_{i=1}^n$ denote a collection of independent random variances. Let X' denote an independent copy of X. Construct $X^{(i)}$ by replacing X_i with X'_i in X, leaving all other entries unchanged. Construct $X^{(i,j)}$ analogously. If $f(X)$ has a finite variance and satisfies the equality*

$$\mathbb{E}[f(X) \mid X_i] = \mathbb{E}[f(X)] \tag{C.4.6}$$

almost surely for each i in $[n]$, then the inequality

$$\operatorname{Var}(f(X)) \leq \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}[\mathbb{E}[(f(X) - f(X^{(i)})) - (f(X^{(j)}) - f(X^{(i,j)})) \mid X]^2] \tag{C.4.7}$$

holds

The second is a Berry-Esseen type central limit theorem for degenerate U-statistics, due to Liu et al. (2025). The result stated in Liu et al. (2025) is given for U-statistics with kernels and degeneracies of arbitrary orders. The version that follows has been adapted to suit our purposes.

Lemma C.4.4 (Theorem 2.2, Liu et al., 2025). *Let $X = (X_i)_{i=1}^n$ denote an i.i.d. sequence of random variables valued in $\mathcal{X}$. Let X' denote an independent copy of X. Suppose that the symmetric function $f : \mathcal{X}^3 \to \mathbb{R}$ satisfies the equality*

$$\mathbb{E}[f(X_1, X_2, X_3) \mid X_1] = 0 \tag{C.4.8}$$

almost surely. Then, it holds that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|P\left\{\frac{1}{\sigma_n}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}(f(X_i, X_j, X_k) - \mathbb{E}[f(X_1, X_2, X_2)]) \leq x\right\} - \Phi(x)\right| \\
&\lesssim \sqrt{\frac{1}{n} + \frac{\mathbb{E}[(f^{(3)}(X_1, X_2, X_3))^2]}{n\mathbb{E}[(f^{(2)}(X_1, X_2))^2]}} + \frac{\sqrt{\frac{1}{n}\mathbb{E}[(f^{(2)}(X_1, X_2))^4] + \mathbb{E}[(g^{(2)}(X_1, X_2))^2]}}{\mathbb{E}[(f^{(2)}(X_1, X_2))^2]},
\end{aligned} \tag{C.4.9}$$

where

$$\sigma_n^2 = \frac{\binom{3}{2}^2}{\binom{n}{2}}\mathbb{E}[(f^{(2)}(X_1,X_2))^2]\,, \quad g^{(2)}(x_1,x_2) = \mathbb{E}[f^{(2)}(X_2,x_1)f^{(2)}(X_1,x_2)]\,, \tag{C.4.10}$$

the quantities $f^{(m)}(\cdot)$ *are defined in Lemma C.4.1, and* $\Phi(\cdot)$ *denotes the standard normal c.d.f.*

C.4.1 Proof of Lemma C.3.1

C.4.1.1 Part (i)

We begin by verifying the first bound stated in (C.3.23). Consider the decomposition

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}\tilde{D}^\star_{i,j}W^\star_j\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}(\tilde{D}^\star_{i,j})^2(W^\star_j)^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,j}\tilde{D}^\star_{i,k}W^\star_jW^\star_k\,. \tag{C.4.11}$$

We bound the two terms in (C.4.11) through the application of the following Lemma.

Lemma C.4.5. *If the conditions of Theorem 3.5.1 hold, then*

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}(\tilde{D}^\star_{i,j})^2(W^\star_j) - \mathbb{E}[\mathrm{Var}(D_{i,\ell}\mid H_{i,\ell},S_\ell)\,\mathrm{Var}(W_\ell\mid\overline{S}_\ell)] = O_p(n^{-1/2}\rho_n) \quad \textit{and} \tag{C.4.12}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,j}\tilde{D}^\star_{i,k}W^\star_jW^\star_k = O_p(\rho_n^{3/2}) \tag{C.4.13}$$

hold, respectively.

Consequently, applying Lemma C.4.5 to the representation (C.4.11), we find that

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}\tilde{D}^\star_{i,j}W^\star_j\right)^2 - \mathbb{E}[\mathrm{Var}(D_{i,\ell}\mid H_{i,\ell},S_\ell)\,\mathrm{Var}(W_\ell\mid\overline{S}_\ell)] = O_p(\rho_n^{3/2})\,. \tag{C.4.14}$$

as required. Next, we consider second bound stated in (C.3.23). Here, we consider the decomposition

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\tilde{D}^\star_{k,j}W^\star_j\right)^2 \tag{C.4.15}$$
$$= \frac{1}{n^2(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}\sum_{k'\neq j}\tilde{D}^\star_{k,j}\tilde{D}^\star_{k',j}(W^\star_j)^2 + \frac{n-2}{n^2(n-1)^2}\sum_{j=1}^{n}\sum_{j'\neq j}\sum_{k\neq j}\sum_{k'\neq j'}\tilde{D}^\star_{k,j}\tilde{D}^\star_{k',j'}W^\star_jW^\star_{j'}$$
$$= \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}(\tilde{D}^\star_{i,j})^2(W^\star_j)^2 + \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{j,k}(W^\star_k)^2$$
$$+ \frac{n-2}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{D}^\star_{j,i}\tilde{D}^\star_{i,j}W^\star_jW^\star_i + \frac{2(n-2)}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{j,i}W^\star_kW^\star_i$$
$$+ \frac{n-2}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{k\neq i}\sum_{\ell\notin\{i,k\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{i,\ell}W^\star_kW^\star_\ell$$
$$+ \frac{n-2}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,\ell\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{j,\ell}W^\star_kW^\star_\ell\ .$$

Observe that Lemma C.4.5 implies that

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}(\tilde{D}^\star_{i,j})^2(W^\star_j)^2 = n^{-1}\mathbb{E}[\mathrm{Var}(D_{i,\ell}\mid H_{i,\ell},S_\ell)\,\mathrm{Var}(W_\ell\mid \overline{S}_\ell)] + O_p(n^{-3/2}\rho_n) \tag{C.4.16}$$
$$= O_p(n^{-1}\rho_n^{1/2})$$

and

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{k\neq i}\sum_{\ell\notin\{i,k\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{i,\ell}W^\star_kW^\star_\ell = O_p(n^{-1}\rho_n^{3/2})\ , \tag{C.4.17}$$

respectively. The remaining four terms are then handled through the following Lemma.

Lemma C.4.6. *If the conditions of Theorem 3.5.1 hold, then*

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{D}^\star_{j,i}\tilde{D}^\star_{i,j}W^\star_jW^\star_i = O_p(n^{-3/2}\rho_n^{1/2})\ , \tag{C.4.18}$$
$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{j,k}(W^\star_k)^2 = O_p(n^{-1}\rho_n^{1/2})\ , \tag{C.4.19}$$
$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{D}^\star_{i,k}\tilde{D}^\star_{j,i}W^\star_kW^\star_i = O_p(n^{-3/2}\rho_n^{1/2})\ , \quad \textit{and} \tag{C.4.20}$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,\ell\}}\tilde{D}^{\star}_{i,k}\tilde{D}^{\star}_{j,\ell}W^{\star}_{k}W^{\star}_{\ell} = O(n^{-1}\rho_n^{1/2})\ , \tag{C.4.21}$$

hold, respectively.

Thus, combining (C.4.16), (C.4.17), and Lemma C.4.6, we find that

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\tilde{D}^{\star}_{k,j}W^{\star}_{j}\right)^2 = O_p(n^{-1}\rho_n^{1/2})\ , \tag{C.4.22}$$

as required. ∎

C.4.1.2 Part (ii)

We begin by verifying the first bound in (C.3.24). By a decomposition analogous to (C.4.15), we have that

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}&\left(\frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k\neq j}D^{\star}_{k,j}W^{\star}_{j}\right)^2 \\ &= \frac{1}{n^4(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}\sum_{k'\neq j}D^{\star}_{k,j}D^{\star}_{k',j}(W^{\star}_{j})^2 \\ &+ \frac{n-2}{n^4(n-1)^2}\sum_{j=1}^{n}\sum_{j'\neq j}\sum_{k\neq j}\sum_{k'\neq j'}D^{\star}_{k,j}D^{\star}_{k',j'}W^{\star}_{j}W^{\star}_{j'} = O(n^{-2})\ ,\end{aligned} \tag{C.4.23}$$

where the bound follows from the fact that the random variables $D_{i,j}$ and W_j are bounded by constants almost surely. Next, we verify the second bound in (C.3.24). Observe that, by writing

$$g(Z_1,Z_2,Z_3) = \sum_{\pi\in\Pi_3}D^{\star}_{\pi(2),\pi(1)}D^{\star}_{\pi(3),\pi(1)}(W^{\star}_{\pi(1)})^2\ , \tag{C.4.24}$$

the term

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j\neq i}D^{\star}_{j,i}W^{\star}_{i}\right)^2 &= \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i}D^{\star}_{j,i}D^{\star}_{k,i}(W^{\star}_{i})^2 \\ &= \frac{\binom{n}{3}}{n^4}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}g(Z_i,Z_j,Z_k)\end{aligned} \tag{C.4.25}$$

can be recognized as a scaled U-statistic of order 3. Thus, Lemma C.4.1 implies that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{3}} \sum_{1\le i<j<k\le n} g(Z_i, Z_j, Z_k)\right) \lesssim n^{-1} \operatorname{Var}(g(Z_i, Z_j, Z_k)) \,. \tag{C.4.26}$$

We can evaluate

$$\begin{aligned} \operatorname{Var}(g(Z_i, Z_j, Z_k)) &\lesssim \mathbb{E}[(D^\star_{2,1} D^\star_{3,1})^2 (W^\star_1)^4] & \text{(C.4.27)} \\ &\lesssim \mathbb{E}[(D^\star_{2,1})^4] \lesssim \mathbb{E}[D^4_{2,1}] = O(\rho_n) & \text{(C.4.28)} \end{aligned}$$

by Cauchy-Schwarz and Assumption 3.5.2. Hence, we obtain

$$\begin{aligned} \left| \frac{1}{n^2} \sum_{i=1}^n \left(\frac{1}{n} \sum_{j\neq i} D^\star_{j,i} W^\star_i \right)^2 \right| &\le \left| \frac{1}{n} \mathbb{E}[D^\star_{j,i} D^\star_{k,i} (W^\star_i)^2] \right| + O_p(n^{-3/2} \rho_n^{1/2}) & \text{(C.4.29)} \\ &\lesssim n^{-1} \mathbb{E}[(D^\star_{j,i})^2] + O_p(n^{-3/2} \rho_n^{1/2}) = O(n^{-1} \rho_n) \,, & \text{(C.4.30)} \end{aligned}$$

by Chebyshev's inequality, Cauchy-Schwarz, and Assumption 3.5.2, as required. ■

C.4.2 Proof of Lemma C.3.2

C.4.2.1 Part (i)

We being by evaluating the estimator $\hat{\pi}_n$. Observe that we can express

$$\hat{\pi}_n = \left(\frac{1}{n} \sum_{i=1}^n \overline{X}_i \overline{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \overline{X}_i \overline{W}_i \right) \tag{C.4.31}$$

by the Projection Theorem. As the dimension of covariate vector X_j is fixed, and each component is bounded by a constant, Chebyshev's inequality and a union bound imply that

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \overline{X}_i \overline{X}_i^\top - \operatorname{Var}(X_i) \right\|_\infty &= O_p(n^{-1/2}) \quad \text{and} & \text{(C.4.32)} \\ \left\| \frac{1}{n} \sum_{i=1}^n \overline{X}_i \overline{W}_i - \operatorname{Cov}(X_i, W_i) \right\|_\infty &= O_p(n^{-1/2}) \end{aligned}$$

respectively. Consequently, as the Projection Theorem implies that $\pi_n = \operatorname{Var}(X_i)^{-1} \operatorname{Cov}(X_i, W_i)$, a Taylor expansion, the fact that variance of each component of X_i is bounded below by a constant, and Assumption 3.5.1 give the bound

$$\|\hat{\pi}_n - \pi_n\|_\infty = O_p(n^{-1/2}\|\operatorname{Var}(X_i)^{-1}\|_{\text{op}}) \\ + O_p(n^{-1/2}\|\operatorname{Var}(X_i)^{-1}\|_{\text{op}}^2\|\operatorname{Cov}(X_iW_i)\|_2) = O_p(n^{-1/2}) ,$$

as required.

Next, we evaluate the estimator $\hat{\gamma}_n$. Again, by Assumption 3.5.1 and the Projection Theorem, we can write

$$\hat{\gamma}_n = \hat{\Sigma}_{H,H}^{-1}\hat{\Sigma}_{H,D} \quad \text{and} \quad \gamma_n = \Sigma_{H,H}^{-1}\Sigma_{H,D} \tag{C.4.33}$$

where

$$\hat{\Sigma}_{H,H} = \frac{1}{n-1}\sum_{i \neq j} \overline{H}_{i,j}\overline{H}_{i,j}^\top , \quad \Sigma_{H,H} = \mathbb{E}[\overline{H}_{i,j}\overline{H}_{i,j}^\top] , \tag{C.4.34}$$

and

$$\hat{\Sigma}_{H,D} = \frac{1}{n-1}\sum_{i \neq j} \overline{H}_{i,j}\overline{D}_{i,j} , \quad \Sigma_{H,D} = \mathbb{E}[\overline{H}_{i,j}\overline{D}_{i,j}] , \tag{C.4.35}$$

respectively. Define the perturbations

$$\Gamma_{H,H} = \hat{\Sigma}_{H,H} - \Sigma_{H,H} \quad \text{and} \quad \Gamma_{H,D} = \hat{\Sigma}_{H,D} - \Sigma_{H,D} . \tag{C.4.36}$$

Now, observe that, for any invertible matrices B and C, it holds that

$$(B + C)^{-1} = B^{-1} - B^{-1}C(B + C)^{-1} . \tag{C.4.37}$$

Consequently, we have that

$$\hat{\Sigma}_{H,H}^{-1} = \Sigma_{H,H}^{-1} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\hat{\Sigma}_{H,H}^{-1} \tag{C.4.38}$$

and thereby

$$\begin{aligned} \hat{\gamma}_n - \gamma_n &= \Sigma_{H,H}^{-1}\Gamma_{H,D} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\hat{\Sigma}_{H,H}^{-1}\Sigma_{H,D} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\hat{\Sigma}_{H,H}^{-1}\Gamma_{H,D} \\ &= \Sigma_{H,H}^{-1}\Gamma_{H,D} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\gamma_n \\ &+ \Sigma_{H,H}^{-1}\Gamma_{H,H}(\hat{\Sigma}_{H,H}^{-1} - \Sigma_{H,H}^{-1})\Sigma_{H,D} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\hat{\Sigma}_{H,H}^{-1}\Gamma_{H,D} \\ &= \Sigma_{H,H}^{-1}\Gamma_{H,D} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\gamma_n \\ &+ \Sigma_{H,H}^{-1}\Gamma_{H,H}\Sigma_{H,H}^{-1}\Gamma_{H,H}\hat{\Sigma}_{H,H}^{-1}\Sigma_{H,D} - \Sigma_{H,H}^{-1}\Gamma_{H,H}\hat{\Sigma}_{H,H}^{-1}\Gamma_{H,D} , \end{aligned} \tag{C.4.39}$$

where the final equality follows from the identity

$$\hat{\Sigma}_{H,H}^{-1} - \Sigma_{H,H}^{-1} = -\Sigma_{H,H}^{-1}\Gamma_{H,H}\Sigma_{H,H}^{-1} \,. \tag{C.4.40}$$

The result follows by giving suitable bounds for the kth component of each of the four terms in (C.4.39).

In order to do this, we make repeated use of the inequalities

$$\Gamma_{H,H}^{k,\ell} = O_p(n^{-1/2}\min\{\kappa_{n,k},\kappa_{n,\ell}\}) \quad \text{and} \quad \Gamma_{H,D}^{k} = O_p(n^{-1/2}\rho_n) \tag{C.4.41}$$

where $\Gamma_{H,H}^{k,\ell}$ and $\Gamma_{H,D}^{k}$ are the k,ℓth and kth components of $\Gamma_{H,H}$ and $\Gamma_{H,D}$, respectively. To verify the first inequality in (C.4.41), observe that Lemma C.4.1 implies that

$$\operatorname{Var}\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\overline{H}_{i,j}^{k}\overline{H}_{i,j}^{\ell}\right) \tag{C.4.42}$$

$$\begin{aligned}
&\lesssim n^{-1}\operatorname{Var}(\mathbb{E}[\overline{H}_{i,j}^{(k)}\overline{H}_{i,j}^{(\ell)} \mid \overline{S}_j] + \mathbb{E}[\overline{H}_{i,j}^{(k)}\overline{H}_{i,j}^{(\ell)} \mid \overline{S}_i]) + n^{-2}\operatorname{Var}(\overline{H}_{i,j}^{(k)}\overline{H}_{i,j}^{(\ell)}) \\
&\lesssim n^{-1}\min\left\{P\{H_{i,j}^{(k)} \neq 0 \mid \overline{S}_j\}, P\{H_{i,j}^{(\ell)} \neq 0 \mid \overline{S}_j\}\right\}^2 \\
&+ n^{-1}\min\left\{P\{H_{i,j}^{(k)} \neq 0 \mid \overline{S}_i\}, P\{H_{i,j}^{(\ell)} \neq 0 \mid \overline{S}_i\}\right\}^2 \lesssim n^{-1}\min\{\kappa_{n,k},\kappa_{n,\ell}\}^2 \,,
\end{aligned} \tag{C.4.43}$$

where the last inequality follows by Assumption 3.5.3. Thus, the first inequality in (C.4.41) follows from Chebyshev's inequality. By an analogous argument, Lemma C.4.1 implies that

$$\begin{aligned}
&\operatorname{Var}\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\overline{H}_{i,j}^{k}\overline{D}_{i,j}\right) \\
&\lesssim n^{-1}\min\left\{P\{H_{i,j}^{(k)} \neq 0 \mid \overline{S}_i\}, P\{D_{i,j} \neq 0 \mid \overline{S}_i\}\right\}^2 \\
&+ n^{-1}\min\left\{P\{H_{i,j}^{(k)} \neq 0 \mid \overline{S}_j\}, P\{D_{i,j} \neq 0 \mid \overline{S}_j\}\right\}^2 = O(n^{-1}\rho_n^2) \,,
\end{aligned} \tag{C.4.44}$$

and so the second inequality in (C.4.41) also follows from Chebyshev's inequality.

With this in place, we now turn to bounding the kth component of the first term in (C.4.39). Let e_k denote the kth basis vector in $\mathbb{R}^p$. Observe that

$$\|e_k^\top\Sigma_{H,H}^{-1}\|_1 = \sum_{\ell=1}^{p}|(\Sigma_{H,H}^{-1})_{k,\ell}| = \sum_{\ell=1}^{p}\frac{|K_{k,\ell}^{-1}|}{\sqrt{\operatorname{Var}(H_{i,j}^{k})\operatorname{Var}(H_{i,j}^{\ell})}}$$

$$\lesssim \frac{1}{\kappa_{n,k}} \sum_{\ell=1}^{p} \sqrt{\frac{\kappa_{n,\ell}}{\kappa_{n,\ell}}} |K_{k,\ell}^{-1}| = O(\kappa_{n,k}^{-1}) \tag{C.4.45}$$

by Assumption 3.5.3. Thus, we have that

$$|e_k^\top \Sigma_{H,H}^{-1} \Gamma_{H,D}| \leq \|e_k^\top \Sigma_{H,H}^{-1}\|_1 \|\Gamma_{H,D}\|_\infty = O_p(n^{-1/2} \kappa_{n,k}^{-1} \rho_n) \ . \tag{C.4.46}$$

by the second inequality in (C.4.41) and (C.4.45).

Next, we consider the second term in (C.4.39). To this end, observe that

$$|\gamma_{n,k}| = |e_k^\top \Sigma_{H,H}^{-1} \Sigma_{H,D}| \leq \|e_k^\top \Sigma_{H,H}^{-1}\| \|\Sigma_{H,D}\|_\infty = O(\kappa_{n,k}^{-1} \rho_n) \tag{C.4.47}$$

for each k. Likewise, we can evaluate

$$\begin{aligned} |(\Gamma_{H,H}\gamma_n)_\ell| &= \sum_{k=1}^{p} |\Gamma_{H,H}^{\ell,k} \gamma_{n,\ell}| \leq \sum_{k=1}^{p} |\Gamma_{H,H}^{\ell,k}| |\gamma_{n,\ell}| \\ &= \sum_{k=1}^{p} O_p(n^{-1/2} \min\{\kappa_{n,k}, \kappa_{n,\ell}\} \kappa_{n,k}^{-1} \rho_n) = O_p(n^{-1/2} \rho_n) \ , \end{aligned} \tag{C.4.48}$$

by the first inequality in (C.4.41) and the inequality (C.4.47). Consequently, we have that

$$|e_k^\top \Sigma_{H,H}^{-1} \Gamma_{H,H} \gamma_n| \leq \|e_k^\top \Sigma_{H,H}^{-1}\|_1 \|\Gamma_{H,H}\gamma_n\|_\infty = O_p(n^{-1/2} \kappa_{n,k}^{-1} \rho_n) \ , \tag{C.4.49}$$

by the bound (C.4.45).

In turn, we consider the third term in (C.4.39), given by

$$R_n^{(1)} = \Sigma_{H,H}^{-1} \Gamma_{H,H} \Sigma_{H,H}^{-1} \Gamma_{H,H} \hat{\Sigma}_{H,H}^{-1} \Sigma_{H,D} \ . \tag{C.4.50}$$

Define the matrix

$$\sigma_n = \mathsf{diag}(\sqrt{\mathrm{Var}(H_{i,j}^1)}, \ldots, \sqrt{\mathrm{Var}(H_{i,j}^p)}) \ . \tag{C.4.51}$$

and observe that we can write

$$R_n^{(1)} = \sigma_n^{-1} K^{-1} \tilde{\Gamma}_{H,H} K^{-1} \tilde{\Gamma}_{H,H} (K + \tilde{\Gamma}_{H,H})^{-1} \tilde{\Gamma}_{H,D} \ , \tag{C.4.52}$$

where

$$\tilde{\Gamma}_{H,H} = \sigma_n^{-1} \Gamma_{H,H} \quad \text{and} \quad \tilde{\Sigma}_{H,G} = \sigma_n^{-1} \Sigma_{H,D} \ . \tag{C.4.53}$$

Observe that Assumption 3.5.3 and the first inequality in (C.4.41) imply that

$$\|\tilde{\Gamma}_{H,H}\|_\infty = O_p(n^{-1/2}) \ , \quad \|K^{-1}\|_\infty = O_p(1) \ , \quad \text{and} \quad \|(K^{-1} + \tilde{\Gamma}_{H,H})^{-1}\|_\infty = O_p(1) \ , \tag{C.4.54}$$

respectively. In turn, we have that

$$\|e_k^\top \sigma_n^{-1} K^{-1}\|_1 = O(\kappa_{n,k}^{-1}) \quad \text{and} \quad \|\tilde{\Sigma}_{H,D}\|_\infty = O(\rho_n^{1/2}) \tag{C.4.55}$$

by the same steps used to establish (C.4.48) and the fact that $\rho_n \lesssim \kappa_{n,k}$, respectively. Hence, we obtain

$$\begin{aligned} \|e_k^\top R_n^{(1)}\|_1 &= \|e_k^\top \sigma_n^{-1} K^{-1}\|_1 \|\tilde{\Gamma}_{H,H}\|_\infty \|K^{-1}\|_\infty \|\tilde{\Gamma}_{H,H}\|_\infty \|(K^{-1} + \tilde{\Gamma}_{H,H})^{-1}\|_\infty \|\tilde{\Sigma}_{H,D}\|_\infty \\ &= O_p(n^{-1}\kappa_{n,k}^{-1}\rho_n^{1/2}) = o_p(n^{-1/2}\kappa_{n,k}^{-1}\rho_n) \end{aligned} \tag{C.4.56}$$

by plugging the bounds (C.4.54) and (C.4.55) into (C.4.52).

Finally, we consider the fourth term in (C.4.39), given by

$$R_n^{(2)} = \Sigma_{H,H}^{-1} \Gamma_{H,H} \hat{\Sigma}_{H,H}^{-1} \Gamma_{H,D} \ . \tag{C.4.57}$$

Observe that we can write

$$R_n^{(2)} = \sigma_n^{-1} K^{-1} \tilde{\Gamma}_{H,H} (K + \tilde{\Gamma}_{H,H})^{-1} \tilde{\Gamma}_{H,D} \tag{C.4.58}$$

where

$$\tilde{\Gamma}_{H,D} = \sigma_n^{-1} \Gamma_{H,D} \tag{C.4.59}$$

Observe that the second inequality in (C.4.41) and the fact that $\rho_n \lesssim \kappa_{n,k}$ imply that

$$\|\tilde{\Gamma}_{H,D}\|_\infty = O_p(n^{-1/2}\rho_n^{1/2}) \ . \tag{C.4.60}$$

Consequently, we obtain the bound

$$\begin{aligned} \|e_k^\top R_n^{(1)}\|_1 &= \|e_k^\top \sigma_n^{-1} K^{-1}\|_1 \|\tilde{\Gamma}_{H,H}\|_\infty \|(K + \tilde{\Gamma}_{H,H})^{-1}\|_\infty \|\tilde{\Gamma}_{H,D}\|_\infty \\ &= O_p(n^{-1}\kappa_{n,k}^{-1}\rho_n^{1/2}) = o_p(n^{-1/2}\kappa_{n,k}^{-1}\rho_n) \end{aligned} \tag{C.4.61}$$

by plugging the bounds (C.4.54), (C.4.55) and (C.4.55) into (C.4.58). Hence, we obtain the bound

$$\hat{\gamma}_{n,k} - \gamma_{n,k} = O_p(n^{-1}\kappa_{n,k}^{-1}\rho_n^{1/2}) \tag{C.4.62}$$

by plugging the bounds (C.4.46), (C.4.49), (C.4.56), and (C.4.61) into the decomposition (C.4.39), as required. ■

C.4.2.2 Part (ii)

Observe that Cauchy-Schwarz implies that

$$\left|\frac{1}{n^2}\sum_{i=1}^{n} J_{i,s}J_{i,r}\right| \leq \sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n} J_{i,s}^2\right)\left(\frac{1}{n^2}\sum_{i=1}^{n} J_{i,r}^2\right)} \ . \tag{C.4.63}$$

Thus, to verify the first bound, it suffices to show that

$$\mathbb{E}[J_{i,s}^2] = O(n\kappa_{n,s}) \ , \tag{C.4.64}$$

for each s, as this implies that

$$\frac{1}{n^2}\sum_{i=1}^{n} J_{i,s}^2 = O_p(\kappa_{n,s}) \tag{C.4.65}$$

by Markov's inequality. Thus, to ease notation, and without loss, we can drop the subscript s from $A_{i,s}$ and $\kappa_{n,s}$ and treat the quantity $\overline{H}_{i,j}$ as a scalar-valued random variable. Consider the decomposition

$$J_i = \sum_{j\neq i} \tilde{H}_{i,j}W_j^\star - \frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\tilde{H}_{k,j}W_j^\star + \frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k\neq j}\overline{H}_{k,j}W_j^\star - \frac{1}{n}\sum_{j\neq i}\overline{H}_{k,i}W_i^\star \tag{C.4.66}$$

where

$$\tilde{G}_{i,j} = \tilde{H}_{i,j} - \mathbb{E}[\tilde{H}_{i,j} \mid \overline{S}_j] \ . \tag{C.4.67}$$

Observe that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{j\neq i}\tilde{H}_{i,j}W_j^\star\right)^2\right] &= \sum_{j\neq i}\sum_{j'\neq i}\mathbb{E}[\tilde{H}_{i,j}\tilde{H}_{i,j'}W_j^\star W_{j'}^\star] \\ &= \sum_{j\neq i}\mathbb{E}[\tilde{H}_{i,j}^2(W_j^\star)^2] \lesssim n\mathbb{E}[H_{i,j}^2] = O(n\kappa_n) \ ,\end{aligned} \tag{C.4.68}$$

where the second equality follows from Assumption 3.4.2 and the final bound follows from Assumption 3.5.1. Analogously, we can evaluate

$$\mathbb{E}\left[\left(\frac{1}{n-1}\sum_{j\neq i}\sum_{k\neq j}\tilde{H}_{k,j}W_j^\star\right)^2\right] = \frac{1}{(n-1)^2}\sum_{j\neq i}\sum_{k\neq j}\sum_{j'\neq i}\sum_{k'\neq j'}\mathbb{E}[\tilde{H}_{k,j}\tilde{H}_{k',j'}W_j^\star W_{j'}^\star]$$
$$= \frac{1}{(n-1)^2}\sum_{j\neq i}\sum_{k\neq j}\mathbb{E}[\tilde{H}_{k,j}^2(W_j^\star)^2] = O(\kappa_n)\ , \quad \text{(C.4.69)}$$

as well as

$$\mathbb{E}\left[\left(\frac{1}{n(n-1)}\sum_{j\neq i}\sum_{k\neq j}\overline{H}_{k,j}W_j^\star\right)^2\right]$$
$$= \frac{1}{n^2(n-1)^2}\sum_{j\neq i}\sum_{k\neq j}\sum_{k'\neq j}\mathbb{E}[\overline{H}_{k,j}\overline{H}_{k',j}(W_j^\star)^2] = O(n^{-1}\kappa_n) \quad \text{(C.4.70)}$$

and

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{j\neq i}\overline{H}_{j,i}W_i^\star\right)^2\right] = \frac{1}{n^2}\sum_{j\neq i}\sum_{j'\neq i}\mathbb{E}[\overline{H}_{j,i}\overline{H}_{j',i}(W_i^\star)^2] = O(\kappa_n)\ , \quad \text{(C.4.71)}$$

respectively. Hence, the moment bound (C.4.64) follows by squaring the decomposition (C.4.66) within an expectation, applying Cauchy-Schwarz, and plugging in the bounds (C.4.68) through (C.4.71).

The proof of the second bound has the same structure. In particular, Cauchy-Schwarz implies that

$$\left|\frac{1}{n^2}\sum_{i=1}^n M_{i,s,r}M_{i,s',r'}\right| \lesssim \sqrt{\left(\frac{1}{n^2}\sum_{i=1}^n M_{i,s,r}^2\right)\left(\frac{1}{n^2}\sum_{i=1}^n M_{i,s',r'}^2\right)}\ . \quad \text{(C.4.72)}$$

Thus, it suffices to show that

$$\mathbb{E}\left[M_{i,s,r}^2\right] = O(n^2\kappa_{n,s}) \quad \text{(C.4.73)}$$

for each s, r, as this implies that

$$\frac{1}{n^2}\sum_{i=1}^n M_{i,s,r}^2 = O_p(n\kappa_{n,s}) \quad \text{(C.4.74)}$$

by Markov's inequality. Thus, again, to ease notation, we can drop the subscripts s and r and treat the covariates $\overline{H}_{i,j}$ and $\overline{X}_j$ as scalar-valued random variables, without loss. Here, it suffices to consider the decomposition

$$M_i = \sum_{j \neq i} \overline{H}_{i,j} \overline{X}_j - \frac{1}{n} \sum_{j=1}^{n} \sum_{k \neq j} \overline{H}_{k,j} \overline{X}_j \ . \tag{C.4.75}$$

Observe that we can evaluate

$$\mathbb{E}\left[\left(\sum_{j \neq i} \overline{H}_{i,j} \overline{X}_j\right)^2\right] = \sum_{j \neq i} \sum_{j' \neq i} \mathbb{E}[\overline{H}_{i,j} \overline{H}_{i,j'} \overline{X}_j \overline{X}_{j'}] \lesssim n^2 \mathbb{E}[\overline{H}_{i,j}^2 \overline{X}_j^2] = O(n^2 \kappa_n) \ , \tag{C.4.76}$$

and

$$\begin{aligned} &\mathbb{E}\left[\left(\frac{1}{n} \sum_{j=1}^{n} \sum_{k \neq j} \overline{H}_{k,j} \overline{X}_j\right)^2\right] \\ &= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \sum_{k \neq j} \sum_{k' \neq j'} \mathbb{E}[\overline{H}_{k,j} \overline{X}_j \overline{H}_{k',j'} \overline{X}_{j'}] \lesssim n^2 \mathbb{E}[\overline{H}_{i,j}^2 \overline{X}_j^2] = O(n^2 \kappa_n) \ , \end{aligned} \tag{C.4.77}$$

by Cauchy-Schwarz and Assumption 3.5.1. Hence, the moment bound (C.4.73) follows by squaring the decomposition (C.4.75) within an expectation, applying Cauchy-Schwarz to remove the cross-terms, and plugging in the bounds (C.4.76) and (C.4.77). ∎

C.4.3 Proof of Lemma C.3.3

C.4.3.1 Part (i)

Define the conditional expectations

$$\tilde{F}(Z_i) = \mathbb{E}[Y_i \mid Z_i] \quad \text{and} \quad \tilde{F}(Z_i, Z_j) = \mathbb{E}[Y_i \mid Z_i, Z_j] - \tilde{F}(Z_i) \tag{C.4.78}$$

and the associated residual

$$R_i = Y_i - \tilde{F}(Z_i) - \sum_{j \neq i} \tilde{F}(Z_i, Z_j) \ , \tag{C.4.79}$$

for each unit i in $[n]$. Consider the decomposition

$$\frac{1}{n^2}\sum_{i=1}^{n} Y_i \check{\Delta}_i = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} Y_i \tilde{D}^{\star}_{i,j} W^{\star}_j = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} R_i \tilde{D}^{\star}_{i,j} W^{\star}_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[Y_i \mid Z_i, Z_j]\tilde{D}^{\star}_{i,j} W^{\star}_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i, Z_k)\tilde{D}^{\star}_{i,j} W^{\star}_j \,. \tag{C.4.80}$$

We evaluate each term, in succession, through the application of the following Lemma.

Lemma C.4.7. *If the conditions of Theorem 3.5.1 hold, then:*

(i) *It holds that*

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} R_i \tilde{D}^{\star}_{i,j} W^{\star}_j = O_p(n^{-1/2}\rho_n) \,. \tag{C.4.81}$$

(ii) *It holds that*

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[Y_i \mid Z_i, Z_j]\tilde{D}^{\star}_{i,j} W^{\star}_j = \mathbb{E}[Y_i \tilde{D}^{\star}_{i,j} W^{\star}_j] + O_p(\rho_n n^{-1/2}) \,. \tag{C.4.82}$$

(iii) *It holds that*

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i, Z_k)\tilde{D}^{\star}_{i,j} W^{\star}_j = O_p(\rho_n^{3/2}) \,. \tag{C.4.83}$$

Consequently, by applying each part of Lemma C.4.7 to the representation (C.4.80), we find that

$$\frac{1}{n^2}\sum_{i=1}^{n} Y_i \check{\Delta}_i - \mathbb{E}[Y_i \tilde{D}^{\star}_{i,j} W^{\star}_j] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i, Z_k)\tilde{D}^{\star}_{i,j} W^{\star}_j + O_p(n^{-1/2}\rho_n) \tag{C.4.84}$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i, Z_k)\tilde{D}^{\star}_{i,j} W^{\star}_j = O_p(\rho_n^{3/2}) \,. \tag{C.4.85}$$

as required. ■

C.4.3.2 Part (ii)

We now turn to the terms

$$\frac{1}{n-1}\frac{1}{n^2}\sum_{i=1}\sum_{j\neq i} Y_i\check{\Delta}^\star_{i,j} \quad \text{and} \quad \frac{1}{n^2}\sum_{i=1}\sum_{j\neq i} Y_i(Q_i^{(1)} - Q_i^{(2)} + Q_i^{(3)}) \,. \tag{C.4.86}$$

Observe that each term can be decomposed analogously to (C.4.80). That is, we have that

$$\begin{aligned}\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i} Y_i\check{\Delta}^\star_{i,j} &= \frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}} R_i\tilde{D}^\star_{k,j}W^\star_j + \frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \mathbb{E}[Y_i \mid Z_i, Z_j]\tilde{D}^\star_{k,j}W^\star_j \\ &+ \frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i, Z_\ell)\tilde{D}^\star_{k,j}W^\star_j \,.\end{aligned} \tag{C.4.87}$$

Likewise, we can write

$$\begin{aligned}\frac{n-1}{n^3}\sum_{i=1}\sum_{j\neq i} Y_iQ_i^{(1)} &= \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} R_iD^\star_{k,j}W^\star_j + \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \mathbb{E}[Y_i \mid Z_i, Z_j]D^\star_{k,j}W^\star_j \\ &+ \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i, Z_\ell)D^\star_{k,j}W^\star_j\end{aligned} \tag{C.4.88}$$

as well as

$$\begin{aligned}\frac{n-1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i} Y_iQ_i^{(2)} &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i} R_iD^\star_{j,i}W^\star_i + \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[Y_i \mid Z_i, Z_j]D^\star_{j,i}W^\star_i \\ &+ \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i, Z_\ell)D^\star_{j,i}W^\star_i\end{aligned} \tag{C.4.89}$$

and

$$\begin{aligned}\frac{n-1}{n^3}\sum_{i=1}\sum_{j\neq i} Y_iQ_i^{(3)} &= \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i} R_iD^\star_{i,j}W^\star_j + \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[Y_i \mid Z_i, Z_j]D^\star_{i,j}W^\star_j \\ &+ \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i, Z_\ell)D^\star_{i,j}W^\star_j \,,\end{aligned} \tag{C.4.90}$$

respectively. We evaluate each term through the application of the following lemma.

Lemma C.4.8. *If the conditions of Theorem 3.5.1 hold, then:*

(i) *It holds that*

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} R_i\tilde{D}^{\star}_{k,j}W^{\star}_j = O_p(n^{-1}\rho_n)\ , \quad \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} R_i D^{\star}_{k,j}W^{\star}_j = O_p(n^{-3/2}\rho_n) \tag{C.4.91}$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i} R_i D^{\star}_{j,i}W^{\star}_i = O_p(n^{-1}\rho_n)\ , \quad \text{and} \quad \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i} R_i D^{\star}_{i,j}W^{\star}_j = O_p(n^{-5/2}\rho_n)\ ,$$

respectively.

(ii) *It holds that*

$$\frac{1}{n^2(n-1)}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \mathbb{E}[Y_i \mid Z_i, Z_j]\tilde{D}^{\star}_{k,j}W^{\star}_j = \mathbb{E}[Y_i\tilde{D}_{\ell,q}W^{\star}_q] + O_p(n^{-1/2}\rho_n)\ , \tag{C.4.92}$$

$$\frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \mathbb{E}[Y_i \mid Z_i, Z_j]D^{\star}_{k,j}W^{\star}_j = \frac{1}{n}\mathbb{E}[Y_i D^{\star}_{\ell,q}W^{\star}_q] + O_p(n^{-3/2}\rho_n^{1/2})\ ,$$

$$\frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[Y_i \mid Z_i, Z_j]D^{\star}_{j,i}W^{\star}_i = \frac{1}{n}\mathbb{E}[Y_i D^{\star}_{\ell,i}W^{\star}_i] + O_p(n^{-3/2}\rho_n^{1/2})\ ,$$

$$\frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[Y_i \mid Z_i, Z_j]D^{\star}_{i,j}W^{\star}_j = \frac{1}{n^2}\mathbb{E}[Y_i D^{\star}_{i,\ell}W^{\star}_\ell] + O_p(n^{-5/2}\rho_n^{1/2})$$

respectively.

(iii) *It holds that*

$$\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i,Z_\ell)\tilde{D}^{\star}_{k,j}W^{\star}_j = O_p(n^{-1/2}\rho_n)\ , \tag{C.4.93}$$

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i,Z_\ell)D^{\star}_{k,j}W^{\star}_j = O_p(n^{-1}\rho_n^{1/2})\ ,$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i,Z_\ell)D^{\star}_{j,i}W^{\star}_i = O_p(n^{-1}\rho_n^{1/2})\ , \quad \text{and}$$

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i,Z_\ell)D^{\star}_{i,j}W^{\star}_j = O_p(n^{-3/2}\rho_n^{1/2})$$

respectively.

By plugging each of the bounds from Lemma C.4.8 into the decompositions (C.4.87) though (C.4.90), we obtain

$$\frac{1}{n-1}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} Y_i\check{\Delta}_{i,j} = \mathbb{E}[Y_i\tilde{D}_{\ell,j}W_j] + O_p(n^{-1/2}\rho_n)\ , \quad \text{and} \tag{C.4.94}$$

$$\frac{1}{n^2}\sum_{i=1}^{n} Y_i(Q_i^{(1)} - Q_i^{(2)} + Q_i^{(3)}) = \frac{1}{n}\mathbb{E}[Y_i\overline{D}_{q,\ell}W_\ell] - \frac{1}{n}\mathbb{E}[Y_i\overline{D}_{\ell,i}W_i] + \frac{1}{n^2}\mathbb{E}[Y_i\overline{D}_{i,\ell}W_\ell] + O_p(n^{-1}\rho_n^{1/2})\ , \tag{C.4.95}$$

verifying the representations (C.3.38) and (C.3.39), respectively. ■

C.4.4 Proof of Lemma C.3.4

We continue to use the notation introduced in the proof of Lemma C.3.3. To verify the bounds, we consider the decompositions

$$\frac{1}{n^2}\sum_{i=1}^{n} Y_iJ_{i,s} = \frac{1}{n^2}\sum_{i=1}^{n} R_iJ_{i,s} + \frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)J_{i,s} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)J_{i,s} \quad \text{and}$$

$$\frac{1}{n^2}\sum_{i=1}^{n} Y_iM_{i,s,r} = \frac{1}{n^2}\sum_{i=1}^{n} R_iM_{i,s,r} + \frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)M_{i,s,r} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)M_{i,s,r}\ . \tag{C.4.96}$$

Bounds for each term are given in the following Lemma.

Lemma C.4.9. *If the conditions of Theorem 3.5.1 hold, then:*

(i) *It holds that*

$$\frac{1}{n^2}\sum_{i=1}^{n} R_iJ_{i,s} = O_p((\rho_n\kappa_{n,s})^{1/2}n^{-1/2}) \quad \textit{and} \quad \frac{1}{n^2}\sum_{i=1}^{n} R_iM_{i,s,r} = O_p((\rho_n\kappa_{n,s})^{1/2})\ . \tag{C.4.97}$$

(ii) *It holds that*

$$\frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)J_{i,s} = O_p(\kappa_{n,s}) \quad \textit{and} \quad \frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)M_{i,s,r} = O_p(n^{1/2}\kappa_{n,s})\ . \tag{C.4.98}$$

***(iii)** It holds that*

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)J_{i,s}=O_p(\kappa_n) \quad \textit{and} \quad \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)M_{i,s,r}=O_p(n^{1/2}\kappa_n)\ . \tag{C.4.99}$$

Thus, we can conclude that

$$\frac{1}{n^2}\sum_{i=1}^{n}Y_iJ_{i,s}=O_p(\kappa_{n,s}) \quad \text{and} \quad \frac{1}{n^2}\sum_{i=1}^{n}Y_iM_{i,s,r}=O_p(n^{1/2}\kappa_{n,s})\ , \tag{C.4.100}$$

by plugging each of the bounds from Lemma C.4.9 into (C.4.96). ∎

C.4.5 Proof of Lemma C.3.7

Let $\xi=(\xi_1,\xi_2)$ be any real-valued vector that satisfies $\xi_1^2+\xi_2^2=1$. By the Cramér-Wold device, it suffices to show that

$$\frac{\sqrt{2}}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\overline{\sigma}_n^{-1}(\xi)(\xi_1\tilde{F}(Z_i,Z_k)+\xi_2\tilde{D}^\star_{i,k}W^\star_k)\tilde{D}^\star_{i,j}W^\star_j\xrightarrow{d}\mathsf{N}(0,1)\ , \tag{C.4.101}$$

where

$$\overline{\sigma}_n^2(\xi)=\xi_1^2\sigma_{n,1}^2+2\xi_1\xi_2\varsigma_n+\xi_2^2\sigma_{n,2}^2\ . \tag{C.4.102}$$

To this end, observe that, by writing

$$f_\xi(Z_1,Z_2,Z_3)=\sum_{\pi\in\Pi_3}(\xi_1\tilde{F}(Z_{\pi(1)},Z_{\pi(3)})+\xi_2\tilde{D}^\star(Z_{\pi(1)},Z_{\pi(3)})W^\star_{\pi(3)})\tilde{D}^\star(S_{\pi(1)},S_{\pi(2)})W^\star_{\pi(2)}\ , \tag{C.4.103}$$

the quantity

$$\begin{aligned}&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}(\xi_1\tilde{F}(Z_i,Z_k)+\xi_2\tilde{D}^\star_{i,k}W^\star_k)\tilde{D}^\star_{i,j}W^\star_j\\&=\frac{(n-1)(n-2)}{6n}\left(\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}f_\xi(Z_i,Z_j,Z_k)\right)\end{aligned} \tag{C.4.104}$$

can be recognized as scaled U-statistic of order 3. Moreover, by the equalities (C.4.121), (C.4.177), and (C.4.178), the U-statistic (C.4.104) is degenerate of order one. In turn, by the equalities (C.4.129) and (C.4.180), the second term in Hoeffding decomposition of

(C.4.104) is given by

$$f_\xi^{(2)}(Z_2, Z_3) = \mathbb{E}[(\xi_1 \tilde{F}(Z_1, Z_3) + \xi_2 \tilde{D}^\star_{1,3} W^\star_3)\tilde{D}^\star_{1,2} \mid Z_2, Z_3] W^\star_2 \\ + \mathbb{E}[(\xi_1 \tilde{F}(Z_1, Z_2) + \xi_2 \tilde{D}^\star_{1,2} W^\star_2)\tilde{D}^\star_{1,3} \mid Z_2, Z_3] W^\star_3 \ . \tag{C.4.105}$$

Thus, Lemma C.4.4 implies that

$$\sup_{x\in\mathbb{R}} \left| P\left\{ \frac{1}{\overline{\sigma}_n(\xi)} \frac{\sqrt{\binom{n}{2}}}{\binom{3}{2}} \frac{1}{\binom{n}{3}} \sum_{1\le i<j<k\le n} f_\xi(Z_i, Z_j, Z_k) \le x \right\} - \Phi(x) \right| \\ \lesssim \sqrt{\frac{1}{n} + \frac{\mathbb{E}[(f_\xi^{(3)}(Z_1, Z_2, Z_3))^2]}{n\mathbb{E}[(f_\xi^{(2)}(Z_1, Z_2))^2]}} + \frac{\sqrt{\frac{1}{n}\mathbb{E}[(f_\xi^{(2)}(Z_1, Z_2))^4] + \mathbb{E}[(g_\xi^{(2)}(Z_1, Z_2))^2]}}{\mathbb{E}[(f_\xi^{(2)}(Z_1, Z_2))^2]} \ , \tag{C.4.106}$$

where $f_\xi^{(3)}(Z_1, Z_2, Z_3)$ is the third term in Hoeffding decomposition of (C.4.104), and

$$g_\xi^{(2)}(z_1, z_2) = \mathbb{E}[f_\xi^{(2)}(Z_3, z_1) f_\xi^{(2)}(Z_3, z_2)] \ , \tag{C.4.107}$$

respectively. Observe that Lemma C.3.6 implies that

$$\overline{\sigma}_n^2(\xi) = \mathbb{E}[(f_\xi^{(2)}(X_1, X_2))^2] \gtrsim \rho_n^3 \tag{C.4.108}$$

Upper bounds for each of the other three quantities appearing in (C.4.106) are given by the following Lemma.

Lemma C.4.10. *Under the conditions of Theorem 3.6.2, the bounds*

$$\mathbb{E}[(f_\xi^{(3)}(Z_1, Z_2, Z_3))^2] = O(\rho_n^2) \ , \\ \mathbb{E}[(f_\xi^{(2)}(Z_1, Z_2))^4] = O(\rho_n^5) \ , \quad \textit{and} \\ \mathbb{E}[(g_\xi^{(2)}(Z_1, Z_2))^2] = o(\rho_n^6) \tag{C.4.109}$$

hold.

Plugging (C.4.108) and each of the bounds from Lemma C.4.10 into (C.4.106), we obtain

$$\frac{1}{\overline{\varphi}_n(\xi)} \frac{\sqrt{\binom{n}{2}}}{\binom{3}{2}} \frac{1}{\binom{n}{3}} \sum_{1\le i<j<k\le n} f_\xi(Z_i, Z_j, Z_k) \xrightarrow{d} \mathsf{N}(0,1) \ . \tag{C.4.110}$$

Consequently, as

$$\frac{\sqrt{\binom{n}{2}}}{\binom{3}{2}}\frac{1}{\binom{n}{3}} = \frac{\sqrt{2}}{(n-2)\sqrt{n(n-1)}} = \frac{\sqrt{2}}{n^2} + O(n^{-3}) \tag{C.4.111}$$

we can conclude that

$$\frac{\sqrt{2}}{\overline{\sigma}_n(\xi)}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}(\xi_1\tilde{F}(Z_i,Z_k)+\xi_2\tilde{D}^\star_{i,k}W^\star_k)\tilde{D}^\star_{i,j}W^\star_j \xrightarrow{d} \mathsf{N}(0,1)\ , \tag{C.4.112}$$

by the continuous mapping theorem, as required. ■

C.4.6 Proof of Lemma C.4.5

First, we verify the bound (C.4.12). By writing

$$g(Z_1,Z_2) = \sum_{\pi\in\Pi_2}\tilde{D}^\star(S_{\pi(1)},S_{\pi(2)})^2(W^\star_{\pi(2)})^2\ , \tag{C.4.113}$$

where we recall that Π_m denotes the set of permutations on $[m]$, the term

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}(\tilde{D}^\star_{i,j})^2(W^\star_j) = \frac{n-1}{2n}\left(\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j>i}g(Z_i,Z_j)\right) \tag{C.4.114}$$

can be recognized as a scaled U-statistic of order 2. Thus, Part (ii) of Lemma C.4.2 implies that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j>i}g(Z_i,Z_j)\right) \lesssim n^{-1}\operatorname{Var}(\mathbb{E}[g(Z_1,Z_2)\mid Z_1]) + n^{-2}\operatorname{Var}(g(Z_1,Z_2))\ . \tag{C.4.115}$$

Observe that

$$\begin{aligned}\operatorname{Var}(g(Z_1,Z_2)) &= \operatorname{Var}((\tilde{D}^\star_{2,1})^2(W^\star_1)^2+(\tilde{D}^\star_{1,2})^2(W^\star_2)^2)\\ &\lesssim \mathbb{E}[(\tilde{D}^\star_{2,1})^4(W^\star_1)^4]\\ &\lesssim \mathbb{E}[D^4_{2,1}+\mathbb{E}[D(S_2,S_1)\mid H_{2,1},S_1]^4] \lesssim \rho_n\mathbb{E}[D^4_{2,1}\mid\mathcal{A}_{2,1}] = O(\rho_n)\ ,\end{aligned} \tag{C.4.116}$$

where the first two inequalities follow from Cauchy-Schwarz, the third inequality follows from Assumption 3.5.2, and the final bound follows from the fact that the treatments and distances are bounded buy a constant. In turn, we can evaluate

$$\begin{aligned}
&\operatorname{Var}(\mathbb{E}[g(Z_1, Z_2) \mid Z_1]) \\
&\lesssim \mathbb{E}[\mathbb{E}[\tilde{D}^\star(S_2, S_1)^2 \mid Z_1]^2] + \mathbb{E}[\mathbb{E}[\tilde{D}^\star(S_1, S_2)^2] \mid Z_1]^2] \\
&\lesssim \mathbb{E}[\mathbb{E}[D(S_2, S_1) \mid Z_1]^2 + \mathbb{E}[\mathbb{E}[D(S_2, S_1) \mid H_{2,1}, S_1] \mid Z_1]^2] \\
&+ \mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid Z_1]^2 + \mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2] \mid Z_1]^2] \\
&\lesssim \rho_n^2(\mathbb{E}[\mathbb{E}[D(S_2, S_1) \mid \mathcal{A}_{2,1}, Z_1]^2 + \mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid \mathcal{A}_{2,1}, Z_1]^2]) \\
&+ \rho_n^2 \mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2] \mid \mathcal{A}_{2,1}, Z_1]^2] \\
&+ \mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2] \mid \tilde{\mathcal{A}}_{2,1}, Z_1]^2] \lesssim O(\rho_n^2) \,, \qquad \text{(C.4.117)}
\end{aligned}$$

by Cauchy-Schwarz and Assumption 3.5.2. Thus, the bound (C.4.115) and Chebyshev's inequality imply that

$$\begin{aligned}
&\frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} (\tilde{D}^\star_{i,j})^2 (W^\star_j)^2 - \frac{n-1}{n} \mathbb{E}[\operatorname{Var}(D_{i,j} \mid H_{i,j}, S_j) \operatorname{Var}(W_i \mid S_j)] \\
&\qquad = O_p(\rho_n n^{-1/2} + \rho_n^{1/2} n^{-1}) = O_p(\rho_n n^{-1/2}) \,, \qquad \text{(C.4.118)}
\end{aligned}$$

as required.

Next, we verify the bound (C.4.13). Proceeding in the same way, by writing

$$h(Z_1, Z_2, Z_3) = \sum_{\pi \in \Pi_3} \tilde{D}^\star(S_{\pi(1)}, S_{\pi(2)}) \tilde{D}^\star(S_{\pi(1)}, S_{\pi(3)}) W^\star_{\pi(2)} W^\star_{\pi(3)} \,, \qquad \text{(C.4.119)}$$

the term

$$\frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{D}^\star_{i,j} \tilde{D}^\star_{i,k} W^\star_j W^\star_k = \frac{(n-1)(n-2)}{6n} \left(\frac{1}{\binom{n}{3}} \sum_{1 \leq i < j < k \leq n} h(Z_i, Z_j, Z_k) \right) \qquad \text{(C.4.120)}$$

can be recognized as a scaled U-statistic order 3. Now, by Assumption 3.4.2, it holds that

$$\mathbb{E}[\tilde{D}^\star_{1,2} \tilde{D}^\star_{1,3} W^\star_2 W^\star_3 \mid Z_i] = 0 \qquad \text{(C.4.121)}$$

for each i in $\{1,2,3\}$, and so

$$\operatorname{Var}(\mathbb{E}[h(Z_1,Z_2,Z_3) \mid Z_1]) = 0\ . \tag{C.4.122}$$

Consequently, Part (ii) of Lemma C.4.2 implies that

$$\begin{aligned}&\operatorname{Var}\left(\frac{(n-1)(n-2)}{6n}\left(\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h(X_i,X_j,X_k)\right)\right)\\&\lesssim \operatorname{Var}(\mathbb{E}[h(Z_1,Z_2,Z_3) \mid Z_2,Z_3]) + n^{-1}\operatorname{Var}(h(Z_1,Z_2,Z_3))\ .\end{aligned} \tag{C.4.123}$$

Observe that

$$\begin{aligned}&\operatorname{Var}(h(Z_1,Z_2,Z_3))\\&\lesssim \mathbb{E}[(\tilde{D}^\star(S_1,S_2)\tilde{D}^\star(S_1,S_3))^2]\\&\lesssim \mathbb{E}[(D(S_1,S_2)D(S_1,S_3))^2] + \mathbb{E}[D(S_1,S_2)^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2]\\&+ \mathbb{E}[\mathbb{E}[D(S_1,S_2) \mid H_{1,2},S_2]^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2]\ ,\end{aligned} \tag{C.4.124}$$

by Cauchy-Schwarz. We can evaluate

$$\mathbb{E}[(D(S_1,S_2)D(S_1,S_3))^2] \lesssim \rho_n^2\mathbb{E}[(D(S_1,S_2)D(S_1,S_3))^2 \mid \mathcal{A}_{1,2},\mathcal{A}_{1,3}] = O(\rho_n^2) \tag{C.4.125}$$

by Assumption 3.5.2. Likewise, Assumption 3.5.2 implies that

$$\begin{aligned}&\mathbb{E}[D(S_1,S_2)^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2]\\&\lesssim \rho_n^2\mathbb{E}[D(S_1,S_2)^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2 \mid \mathcal{A}_{1,2},\mathcal{A}_{1,3}]\\&+ \rho_n\mathbb{E}[D(S_1,S_2)^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2 \mid \mathcal{A}_{1,2},\tilde{\mathcal{A}}_{1,3}] = O(\rho_n^2)\ ,\end{aligned} \tag{C.4.126}$$

and analogously

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[D(S_1,S_2) \mid H_{1,2},S_2]^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2]\\&\lesssim \rho_n^2\mathbb{E}[\mathbb{E}[D(S_1,S_2) \mid H_{1,2},S_2]^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2 \mid \mathcal{A}_{1,2},\mathcal{A}_{1,3}]\\&+ \rho_n\mathbb{E}[\mathbb{E}[D(S_1,S_2) \mid H_{1,2},S_2]^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2 \mid \mathcal{A}_{1,2},\tilde{\mathcal{A}}_{1,3}]\\&+ \mathbb{E}[\mathbb{E}[D(S_1,S_2) \mid H_{1,2},S_2]^2\mathbb{E}[D(S_1,S_3) \mid H_{1,3},S_3]^2 \mid \tilde{\mathcal{A}}_{1,2},\tilde{\mathcal{A}}_{1,3}] = O(\rho_n^2)\ .\end{aligned} \tag{C.4.127}$$

Thus, we have that

$$\mathrm{Var}(h(Z_1, Z_2, Z_3)) = O(\rho_n^2) \; . \tag{C.4.128}$$

In turn, observe that

$$\mathbb{E}[\tilde{D}^\star_{1,2}\tilde{D}^\star_{1,3}W^\star_2 W^\star_3 \mid Z_1, Z_2] = 0 \quad \text{and} \quad \mathbb{E}[\tilde{D}^\star_{1,2}\tilde{D}^\star_{1,3}W^\star_2 W^\star_3 \mid Z_1, Z_3] = 0 \; , \tag{C.4.129}$$

respectively. As a consequence, we can evaluate

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[h(Z_1, Z_2, Z_3) \mid Z_2, Z_3]) \\
&\lesssim \mathrm{Var}(\mathbb{E}[\tilde{D}^\star(S_1, S_2)\tilde{D}^\star(S_1, S_3) \mid Z_2, Z_3]W_2 W_3) \\
&\leq \mathbb{E}[\mathbb{E}[\tilde{D}^\star(S_1, S_2)\tilde{D}^\star(S_1, S_3) \mid Z_2, Z_3]^2] \\
&\lesssim \mathbb{E}[\mathbb{E}[D(S_1, S_2)D(S_1, S_3) \mid S_2, S_3]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]D(S_1, S_3) \mid S_2, S_3]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]\mathbb{E}[D(S_1, S_3) \mid H_{1,3}, S_3] \mid S_2, S_3]^2] \; .
\end{aligned} \tag{C.4.130}$$

Observe that

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[D(S_1, S_2)D(S_1, S_3) \mid S_2, S_3]^2] \\
&\leq P\{\mathcal{B}_{2,3}\}\mathbb{E}[(P\{\mathcal{A}_{1,2}\}\mathbb{E}[D(S_1, S_2)D(S_1, S_3) \mid S_2, S_3, \mathcal{A}_{1,2}])^2 \mid \mathcal{B}_{2,3}] = O(\rho_n^3) \; ,
\end{aligned} \tag{C.4.131}$$

by Assumption 3.5.2. In turn, as $\tilde{\mathcal{A}}_{1,2} \subseteq \tilde{\mathcal{B}}_{2,3} \cap \mathcal{A}_{1,3}$ we have that

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid G_{1,2}, S_2]D(S_1, S_3) \mid S_2, S_3]^2] \\
&\lesssim \rho_n^3\mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]D(S_1, S_3) \mid \mathcal{A}_{1,3}, S_2, S_3]^2 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^2\mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]D(S_1, S_3) \mid \mathcal{A}_{1,3}, S_2, S_3]^2 \mid \tilde{\mathcal{B}}_{2,3}] = O(\rho_n^3)
\end{aligned} \tag{C.4.132}$$

by Assumption 3.5.2. Likewise, we have that

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]\mathbb{E}[D(S_1, S_3) \mid H_{1,3}, S_3] \mid S_2, S_3]^2] \\
&\lesssim \rho_n^4\mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]\mathbb{E}[D(S_1, S_3) \mid H_{1,3}, S_3] \mid \mathcal{A}_{1,3}, \mathcal{A}_{2,3}, S_2, S_3]^2] \\
&+ \rho_n^2\mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]\mathbb{E}[D(S_1, S_3) \mid H_{1,3}, S_3] \mid \tilde{\mathcal{A}}_{1,3}, \mathcal{A}_{2,3}, S_2, S_3]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]\mathbb{E}[D(S_1, S_3) \mid H_{1,3}, S_3] \mid \tilde{\mathcal{A}}_{1,3}, \tilde{\mathcal{A}}_{2,3}, S_2, S_3]^2] = O(\rho_n^4) \; .
\end{aligned} \tag{C.4.133}$$

Thus, we obtain the bound

$$\mathrm{Var}(\mathbb{E}[h(Z_1, Z_2, Z_3) \mid Z_2, Z_3]) = O(\rho_n^3) \, . \tag{C.4.134}$$

Consequently, plugging (C.4.128) and (C.4.134) into (C.4.123), we find that

$$\mathrm{Var}\left(\frac{(n-1)(n-2)}{6n}\left(\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h(Z_i, Z_j, Z_k)\right)\right) = O(\rho_n^3) \, . \tag{C.4.135}$$

The representation (C.4.120) and Chebyshev's inequality, in turn, imply that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{D}^\star_{i,j}\tilde{D}^\star_{i,k}W^\star_j W^\star_k = O(\rho_n^{3/2}) \, , \tag{C.4.136}$$

as required. ∎

C.4.7 Proof of Lemma C.4.6

We continue to use the notation introduced in the proofs of Lemma C.3.3 and Lemma C.4.5. Each bound follows by demonstrating that each term can be expressed as a U-statistic, and applying Lemma C.4.2. In particular, observe that we can write

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i} \tilde{D}^\star_{j,i}\tilde{D}^\star_{i,j}W^\star_j W^\star_i = \frac{\binom{n}{2}}{n^3}\frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} h_0(Z_i, Z_j) \tag{C.4.137}$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{D}^\star_{i,k}\tilde{D}^\star_{j,k}(W^\star_k)^2 = \frac{\binom{n}{3}}{n^3}\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h_1(Z_i, Z_j, Z_k) \, , \tag{C.4.138}$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{j,k\}} \tilde{D}^\star_{i,j}\tilde{D}^\star_{j,\ell}W^\star_j W^\star_\ell = \frac{\binom{n}{3}}{n^3}\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h_2(Z_i, Z_j, Z_k) \, , \quad \text{and} \tag{C.4.139}$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,\ell\}} \tilde{D}^\star_{i,k}\tilde{D}^\star_{j,\ell}W^\star_k W^\star_\ell = \frac{\binom{n}{4}}{n^3}\frac{1}{\binom{n}{4}}\sum_{1\le i<j<k<\ell\le n} h_3(Z_i, Z_j, Z_k, Z_\ell) \, , \tag{C.4.140}$$

where

$$h_0(Z_1, Z_2) = \sum_{\pi\in\Pi_2} \tilde{D}^\star(S_{\pi(2)}, S_{\pi(1)})\tilde{D}^\star(S_{\pi(1)}, S_{\pi(2)})W^\star_{\pi(2)}W^\star_{\pi(1)} \tag{C.4.141}$$

$$h_1(Z_1, Z_2, Z_3) = \sum_{\pi \in \Pi_3} \tilde{D}^\star(S_{\pi(1)}, S_{\pi(3)}) \tilde{D}^\star(S_{\pi(2)}, S_{\pi(3)}) (W^\star_{\pi(3)})^2 ,$$
$$h_2(Z_1, Z_2, Z_3) = \sum_{\pi \in \Pi_3} \tilde{D}^\star(S_{\pi(1)}, S_{\pi(2)}) \tilde{D}^\star(S_{\pi(2)}, S_{\pi(3)}) W^\star_{\pi(2)} W^\star_{\pi(3)} , \quad \text{and}$$
$$h_3(Z_1, Z_2, Z_3, Z_4) = \sum_{\pi \in \Pi_4} \tilde{D}^\star(S_{\pi(1)}, S_{\pi(3)}) \tilde{D}^\star(S_{\pi(2)}, S_{\pi(4)}) W^\star_{\pi(3)} W^\star_{\pi(4)} ,$$

respectively. We begin by considering the term (C.4.137). Part (ii) of Lemma C.4.1 and the law of total variance imply that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} h_0(Z_i, Z_j) \right) \lesssim n^{-1} \operatorname{Var}(h_0(Z_i, Z_j)) . \tag{C.4.142}$$

We can evaluate

$$\begin{aligned} \operatorname{Var}(h_0(Z_i, Z_j)) &\lesssim \mathbb{E}[\tilde{D}^\star(S_2, S_1)^2 \tilde{D}^\star(S_1, S_2)^2 (W_2^\star)^2 (W_1^\star)^2] \\ &\lesssim \mathbb{E}[\tilde{D}^\star(S_2, S_1)^4] \lesssim \mathbb{E}[D(S_2, S_1)^4] \lesssim \rho_n \end{aligned} \tag{C.4.143}$$

by Cauchy-Schwarz and Assumption 3.5.2. Consequently, we find that

$$\operatorname{Var}\left(\frac{1}{n^3} \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}^\star_{j,i} \tilde{D}^\star_{i,j} W_j^\star W_i^\star \right) = O(n^{-3} \rho_n) \tag{C.4.144}$$

and so Chebyshev's inequality gives

$$\frac{1}{n^3} \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}^\star_{j,i} \tilde{D}^\star_{i,j} W_j^\star W_i^\star = O_p(n^{-3/2} \rho_n^{1/2}) , \tag{C.4.145}$$

which verifies the bound (C.4.19).

Next, we consider the term (C.4.138). Observe that

$$\mathbb{E}[\tilde{D}^\star_{1,3} \tilde{D}^\star_{2,3} (W_3^\star)^2 \mid Z_i] = 0 , \tag{C.4.146}$$

holds for each i in $\{1, 2, 3\}$, as the treatments are independent of the latent coordinates and the latent coordinates are i.i.d. Thus, Part (ii) of Lemma C.4.2 and the law of total variance imply that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h_1(Z_1,Z_2,Z_3)\right) \lesssim n^{-2}\operatorname{Var}(h_1(Z_1,Z_2,Z_3))\,. \tag{C.4.147}$$

Observe that the bound

$$\operatorname{Var}(h_1(X_1,X_2,X_3)) \lesssim \mathbb{E}[(\tilde{D}^\star_{1,3}\tilde{D}^\star_{2,3})^2 W_3^4] \lesssim \sqrt{\mathbb{E}[(\tilde{D}^\star_{1,3})^4]\mathbb{E}[(\tilde{D}^\star_{2,3})^4]} = O(\rho_n) \tag{C.4.148}$$

follows from Cauchy-Schwarz and Assumption 3.5.2. Consequently, we find that

$$\operatorname{Var}\left(\frac{1}{n^3}\sum_{i=1}^n\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{D}^\star_{i,k}\tilde{D}^\star_{j,k}(W^\star_k)^2\right) = O(n^{-2}\rho_n) \tag{C.4.149}$$

and so Chebyshev's inequality gives

$$\frac{1}{n^3}\sum_{i=1}^n\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{D}^\star_{i,k}\tilde{D}^\star_{j,k}(W^\star_k)^2 = O_p(n^{-1}\rho_n^{1/2})\,, \tag{C.4.150}$$

which verifies the bound (C.4.19).

Next, we consider the term (C.4.139). In this case, observe that both

$$\mathbb{E}[\tilde{D}^\star_{1,2}\tilde{D}^\star_{2,3}W^\star_2W^\star_3 \mid Z_i] = 0 \quad \text{and} \quad \mathbb{E}[\tilde{D}^\star_{1,2}\tilde{D}^\star_{2,3}W^\star_2W^\star_3 \mid Z_i, Z_j] = 0 \tag{C.4.151}$$

hold for each i and j in $\{1,2,3\}$, by the same reasoning as before. Thus, Part (ii) of Lemma C.4.2 again implies

$$\operatorname{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h_2(Z_1,Z_2,Z_3)\right) \lesssim n^{-3}\operatorname{Var}(h_2(Z_1,Z_2,Z_3))\,. \tag{C.4.152}$$

We can evaluate

$$\begin{aligned}\operatorname{Var}(h_2(X_1,X_2,X_3)) &\lesssim \mathbb{E}[(\tilde{D}^\star_{1,2}\tilde{D}^{\star}{}_{2,3})^2(W^\star_2)^2(W^\star_3)^2] \\ &\lesssim \sqrt{\mathbb{E}[(\tilde{D}^\star_{1,2})^4]\mathbb{E}[(\tilde{D}^\star_{2,3})^4]} = O(\rho_n)\end{aligned} \tag{C.4.153}$$

by Cauchy=Schwarz and Assumption 3.5.2. Thus, by plugging (C.4.153) into (C.4.152) and applying Chebyshev's inequality, we find that

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{j,k\}} \tilde{D}^\star_{i,j}\tilde{D}^\star_{j,\ell}W^\star_j W^\star_\ell = O_p(n^{3/2}\rho_n^{1/2}) \;, \tag{C.4.154}$$

which verifies the bound (C.4.20).

Finally, we consider the term (C.4.140). Here, we have that

$$\mathbb{E}[\tilde{D}^\star_{1,3}\tilde{D}^\star_{2,4}W^\star_3 W^\star_4 \mid Z_i, Z_j, Z_k] = 0 \tag{C.4.155}$$

holds for any i, j, and k in $\{1,2,3,4\}$. Thus, Part (ii) of Lemma C.4.2 implies that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{4}}\sum_{1\leq i<j<k<\ell\leq n} h_2(Z_1,Z_2,Z_3,Z_4)\right) \lesssim n^{-4}\operatorname{Var}(h_3(Z_1,Z_2,Z_3,Z_4)) \;. \tag{C.4.156}$$

We can again evaluate

$$\begin{aligned}\operatorname{Var}(h_3(Z_1,Z_2,Z_3,XZ4)) &\lesssim \mathbb{E}[(\tilde{D}^\star_{1,3}\tilde{D}^\star_{2,4})^2(W^\star_3)^2(W^\star_4)^2] \\ &\lesssim \sqrt{\mathbb{E}[(\tilde{D}^\star_{1,3})^4]\mathbb{E}[(\tilde{D}^\star_{2,4})^4]} = O(\rho_n) \;. \end{aligned}\tag{C.4.157}$$

Hence, by plugging (C.4.157) into (C.4.156), and applying the representation (C.4.140) and Chebyshev's inequality, we find that

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,\ell\}} \tilde{D}^\star_{i,k}\tilde{D}^\star_{j,\ell}W^\star_k W^\star_\ell = O(n^{-1}\rho_n^{1/2}) \;, \tag{C.4.158}$$

which verifies the bound (C.4.21) and completes the proof. ■

C.4.8 Proof of Lemma C.4.7

We continue to use the notation introduced in the proofs of Lemma C.3.3 and Lemma C.4.5. Moreover, we adopt the short-hand

$$F_i(Z) = F_i(Z_j, Z_{-j}) = F(Z_i, Z_{-i}) \quad \text{and} \quad \overline{F}(Z_i,Z_j) = \mathbb{E}[Y_i \mid Z_i, Z_j] \tag{C.4.159}$$

throughout.

C.4.8.1 Part (i)

The result is obtained by applying the following bound

Lemma C.4.11. *If the conditions of Theorem 3.5.1 hold, then* $\mathbb{E}[R_i^2] = O_p(\rho_n)$.

In particular, Cauchy-Schwarz implies that

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j \neq i} \tilde{D}^{\star}_{i,j} W^{\star}_j \right) R_i \right| \leq \frac{1}{\sqrt{n}} \sqrt{ \left(\frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j \neq i} \tilde{D}^{\star}_{i,j} W^{\star}_j \right)^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} R_i^2 \right) } . \tag{C.4.160}$$

In turn, Lemma C.4.11 and Markov's inequality imply that

$$\frac{1}{n} \sum_{i=1}^{n} R_i^2 = O_p(\rho_n) \tag{C.4.161}$$

and Lemma C.4.5 and Assumption 3.5.2 imply that

$$\frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j \neq i} \tilde{D}^{\star}_{i,j} W^{\star}_j \right)^2 = \mathbb{E}[\operatorname{Var}(D_{i,j} \mid H_{i,j}, S_j) \operatorname{Var}(W_j \mid \overline{S}_j)] + O_p(\rho_n^{3/2}) = O_p(\rho_n) . \tag{C.4.162}$$

Consequently, the bound (C.4.160) implies that

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j \neq i} \tilde{D}^{\star}_{i,j} W^{\star}_j \right) R_i = O_p(n^{-1/2} \rho_n) \tag{C.4.163}$$

as required. ■

C.4.8.2 Part (ii)

Observe that, by writing

$$f_1(Z_1, Z_2) = \sum_{\pi \in \Pi_2} \overline{F}(Z_{\pi(1)}, Z_{\pi(2)}) \tilde{D}^{\star}(S_{\pi(1)}, S_{\pi(2)}) W^{\star}_{\pi(2)} , \tag{C.4.164}$$

the quantity

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \overline{F}(Z_i, Z_j) \tilde{D}^{\star}_{i,j} W^{\star}_j = \frac{n-1}{2n} \left(\frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j > i} f_1(Z_i, Z_j) \right) \tag{C.4.165}$$

can be recognized as scaled U-statistic of order 2. Thus, Part (ii) of Lemma C.4.2 implies that

$$\begin{aligned}&\operatorname{Var}\left(\frac{n-1}{2n}\left(\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j>i}f_1(Z_i,Z_j)\right)\right)\\&\lesssim n^{-1}\operatorname{Var}(\mathbb{E}[f_1(Z_1,Z_2)\mid Z_1])+n^{-2}\operatorname{Var}(f_1(Z_1,Z_2))\ .\end{aligned}\tag{C.4.166}$$

We successively bound the two variances appearing in (C.4.166).

First, observe that

$$\begin{aligned}\operatorname{Var}(\mathbb{E}[f_1(Z_1,Z_2)\mid Z_1])&=\operatorname{Var}(\mathbb{E}[\overline{F}(Z_1,Z_2)\tilde{D}^\star_{1,2}W^\star_2+\overline{F}(Z_2,Z_1)\tilde{D}^\star_{2,1}W^\star_1\mid Z_1])\\&\lesssim\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\tilde{D}^\star_{1,2}W^\star_2\mid Z_1]^2]+\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)\tilde{D}^\star_{2,1}W^\star_1\mid Z_1]^2]\ ,\end{aligned}\tag{C.4.167}$$

by Cauchy-Schwarz. We can evaluate

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\tilde{D}^\star_{1,2}W^\star_2\mid Z_1]^2]\\&\lesssim\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)D_{1,2}W^\star_2\mid Z_1]^2]+\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\mathbb{E}[D_{1,2}\mid H_{1,2},S_2]W^\star_2\mid Z_1]^2]\\&\lesssim\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)D_{1,2}W^\star_2\mid\mathcal{A}_{2,1},Z_1]^2]+\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\mathbb{E}[D_{1,2}\mid H_{1,2},S_2]W^\star_2\mid\mathcal{A}_{2,1},Z_1]^2]\\&+\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\mathbb{E}[D_{1,2}\mid H_{1,2},S_2]W^\star_2\mid\tilde{\mathcal{A}}_{2,1},Z_1]^2]\\&\lesssim\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)^2\mid\mathcal{A}_{2,1},Z_1]]+\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)^2\mid\tilde{\mathcal{A}}_{2,1},Z_1]]=O(\rho_n^2)\ ,\end{aligned}\tag{C.4.168}$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. We can evaluate

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)\tilde{D}^\star_{2,1}W^\star_1\mid Z_1]^2]\\&\lesssim\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)D_{2,1}W^\star_1\mid Z_1]^2]+\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)\mathbb{E}[D_{2,1}\mid H_{2,1},S_1]W^\star_1\mid Z_1]^2]\\&\lesssim\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)D_{2,1}W^\star_1\mid\mathcal{A}_{2,1},Z_1]^2]+\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)\mathbb{E}[D_{2,1}\mid H_{2,1},S_1]W^\star_1\mid\mathcal{A}_{2,1},Z_1]^2]\\&+\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)\mathbb{E}[D_{2,1}\mid H_{2,1},S_1]W^\star_1\mid\tilde{\mathcal{A}}_{2,1},Z_1]^2]\\&\lesssim\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)^2\mid\mathcal{A}_{2,1},Z_1]]+\rho_n^2\mathbb{E}[\mathbb{E}[\overline{F}(Z_2,Z_1)^2\mid\tilde{\mathcal{A}}_{2,1},Z_1]]=O(\rho_n^2)\ ,\end{aligned}\tag{C.4.169}$$

again by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Consequently, we find that

$$\operatorname{Var}(\mathbb{E}[f_1(Z_1,Z_2)\mid Z_1])=O(\rho_n^2)\tag{C.4.170}$$

by plugging (C.4.168) and (C.4.169) into (C.4.167). In turn, observe that

$$\begin{aligned}
&\mathrm{Var}(f_1(Z_1, Z_2)) \\
&\lesssim \mathbb{E}[\overline{F}(Z_1, Z_2)^2 (\tilde{D}^\star_{1,2})^2 (W^\star_2)^2] \\
&\lesssim \mathbb{E}[\overline{F}(Z_1, Z_2)^2 D^2_{1,2} (W^\star_2)^2] \\
&+ \mathbb{E}[\overline{F}(Z_1, Z_2)^2 D_{1,2} \mathbb{E}[D_{1,2} \mid H_{1,2}, S_2] (W^\star_2)^2] \\
&+ \mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mathbb{E}[D_{1,2} \mid G_{1,2}, S_2]^2 (W^\star_2)^2] \\
&\lesssim \rho_n \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mid \mathcal{A}_{1,2}, Z_2]] + \rho_n \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mid \tilde{\mathcal{A}}_{1,2}, Z_2]] = O(\rho_n) \ , \qquad \text{(C.4.171)}
\end{aligned}$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Thus, by plugging (C.4.170) and (C.4.171) into (C.4.166), we obtain

$$\mathrm{Var}\left(\frac{n-1}{2n} \left(\frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j>i} f_1(Z_i, Z_j) \right) \right) = O(\rho_n^2 n^{-1} + \rho_n n^{-2}) \qquad \text{(C.4.172)}$$

Hence, as

$$\mathbb{E}[\overline{F}(Z_i, Z_j) \tilde{D}^\star_{i,j} W^\star_j] = \mathbb{E}[Y_i \tilde{D}^\star_{i,j} W^\star_j] \qquad \text{(C.4.173)}$$

we can conclude that

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \overline{F}(Z_i, Z_j) \tilde{D}^\star_{i,j} W^\star_j - \mathbb{E}[Y_i \tilde{D}_{i,j} W_j] = O_p(\rho_n n^{-1/2}) \ , \qquad \text{(C.4.174)}$$

by applying Chebyshev's inequality and the representation (C.4.165). ■

C.4.8.3 Part (iii)

Observe that, by writing

$$f_2(Z_1, Z_2, Z_3) = \sum_{\pi \in \Pi_3} \tilde{F}(Z_{\pi(1)}, Z_{\pi(3)}) \tilde{D}^\star(S_{\pi(1)}, S_{\pi(2)}) W^\star_{\pi(2)} \ , \qquad \text{(C.4.175)}$$

the quantity

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{F}(Z_i, Z_k) \tilde{D}^\star_{i,j} W^\star_j = \frac{(n-1)(n-2)}{6n} \left(\frac{1}{\binom{n}{3}} \sum_{1 \leq i < j < k \leq n} f_2(Z_i, Z_j, Z_k) \right) \qquad \text{(C.4.176)}$$

can be recognized as scaled U-statistic of order 3. We can evaluate

$$\mathbb{E}[\tilde{F}(Z_i, X_k)\tilde{D}^\star_{i,j}W^\star_j \mid Z_i] = 0 \quad \text{and} \quad \mathbb{E}[\tilde{F}(Z_i, X_k)\tilde{D}^\star_{i,j}W^\star_j \mid Z_k] = 0 \,, \tag{C.4.177}$$

by Assumption 3.4.2. Moreover, it holds that

$$\begin{aligned}\mathbb{E}[\tilde{F}(Z_i, Z_k)\tilde{D}^\star_{i,j}W^\star_j \mid Z_j] &= \mathbb{E}[\mathbb{E}[\tilde{F}(Z_i, Z_k) \mid Z_i, Z_j]\tilde{D}^\star_{i,j} \mid Z_j]W^\star_j \\ &= \mathbb{E}[\mathbb{E}[\tilde{F}(Z_i, Z_k) \mid Z_i]\tilde{D}^\star_{i,j} \mid Z_j]W^\star_j = 0 \,,\end{aligned} \tag{C.4.178}$$

by the definition (C.4.78). Consequently, Part (ii) of Lemma C.4.2 implies that

$$\begin{aligned}&\operatorname{Var}\left(\frac{(n-1)(n-2)}{6n}\left(\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} f_2(Z_i, Z_j, Z_k)\right)\right) \\ &\lesssim \operatorname{Var}(\mathbb{E}[f_2(Z_1, Z_2, Z_3) \mid Z_2, Z_3]) + n^{-1}\operatorname{Var}(f_2(Z_1, Z_2, Z_3)) \,.\end{aligned} \tag{C.4.179}$$

We successively bound the two variances appearing in (C.4.179).

First, observe that

$$\begin{aligned}&\mathbb{E}[\tilde{F}(Z_1, Z_3)\tilde{D}^\star(S_1, S_2)W^\star_2 \mid Z_1, Z_3] = \mathbb{E}[\tilde{D}^\star(S_1, S_2)W^\star_2 \mid Z_1]\tilde{F}(Z_1, Z_3) = 0 \quad \text{and} \\ &\mathbb{E}[\tilde{F}(Z_1, Z_3)\tilde{D}^\star(S_1, S_2)W^\star_2 \mid Z_1, Z_2] = \mathbb{E}[\tilde{F}(Z_1, Z_3) \mid Z_1]\tilde{D}^\star(S_1, S_2)W^\star_2 = 0 \,.\end{aligned} \tag{C.4.180}$$

Thus, we obtain the bound

$$\begin{aligned}\operatorname{Var}(\mathbb{E}[f_2(Z_1, Z_2, Z_3) \mid Z_2, Z_3]) &\lesssim \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W^\star_2 \mid Z_2, Z_3]^2] \\ &\quad + \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W^\star_2 \mid Z_2, Z_3]^2] \,.\end{aligned} \tag{C.4.181}$$

Observe that $\tilde{\mathcal{A}}_{1,3} \subseteq \tilde{B}_{2,3} \cap \mathcal{A}_{1,2}$ and that on the event $\tilde{\mathcal{A}}_{1,3}$, we have that

$$\tilde{F}(Z_1, Z_3) = O(\rho_n) \tag{C.4.182}$$

uniformly almost surely, by Assumption 3.5.2. Thus, we can evaluate

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W^\star_2 \mid Z_2, Z_3]^2] \\ &\lesssim \rho_n\mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W^\star_2 \mid Z_2, Z_3]^2 \mid \mathcal{B}_{2,3}]\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star \mid Z_2, Z_3]^2 \mid \tilde{\mathcal{B}}_{2,3}] \\
&\lesssim \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^2 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^2 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^2 \mid \tilde{\mathcal{B}}_{2,3}] \\
&\lesssim \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^2 \mid \mathcal{A}_{1,2}, Z_2, Z_3] \mid \mathcal{B}_{2,3}] + O(\rho_n^3) = O(\rho_n^3) \, , \qquad \text{(C.4.183)}
\end{aligned}
$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Similarly, we can evaluate

$$
\begin{aligned}
&\mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid Z_2, Z_3]^2] \\
&\lesssim \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^2 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \tilde{\mathcal{A}}_{1,2}, Z_2, Z_3]^2 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^2 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^2 \mid \tilde{\mathcal{B}}_{2,3}] \\
&+ \rho_n^2 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \mathcal{A}_{1,3}, Z_2, Z_3]^2 \mid \tilde{\mathcal{B}}_{2,3}] \\
&+ \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \tilde{\mathcal{A}}_{1,3}, \tilde{\mathcal{A}}_{1,2}, Z_2, Z_3]^2 \mid \tilde{\mathcal{B}}_{2,3}] \\
&\lesssim \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^2 \mid \mathcal{A}_{1,2}, Z_2, Z_3] \mid \mathcal{B}_{2,3}] + \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^2 \mid \tilde{\mathcal{A}}_{1,2}, Z_2, Z_3] \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^4 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^2 \mid \mathcal{A}_{1,3}, Z_2, Z_3] \mid \tilde{\mathcal{B}}_{2,3}] + O(\rho_n^3) = O(\rho_n^3) \, , \qquad \text{(C.4.184)}
\end{aligned}
$$

by Assumption 3.5.2 and Assumption C.3.1. Consequently, we find that

$$
\mathrm{Var}(\mathbb{E}[f_2(Z_1, Z_2, Z_3) \mid Z_2, Z_3]) = O(\rho_n^3) \, , \qquad \text{(C.4.185)}
$$

by plugging (C.4.183) and (C.4.184) into (C.4.181). In turn, we can evaluate

$$
\begin{aligned}
&\mathrm{Var}(f(Z_1, Z_2, Z_3)) \\
&\lesssim \mathbb{E}[(\tilde{F}(Z_1, Z_3)\tilde{D}^\star(S_1, S_2)W_2^\star)^2] \\
&\lesssim \mathbb{E}[(\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star)^2] \\
&+ \mathbb{E}[\tilde{F}(Z_1, Z_3)^2 D(S_1, S_2)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2](W_2^\star)^2] \\
&+ \mathbb{E}[(\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star)^2] \\
&\lesssim \rho_n^2 \mathbb{E}[(\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star)^2 \mid \mathcal{A}_{1,3}, \mathcal{A}_{1,2}] \\
&+ \rho_n \mathbb{E}[(\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star)^2 \mid \tilde{\mathcal{A}}_{1,3}, \mathcal{A}_{1,2}] \\
&+ \rho_n^2 \mathbb{E}[\overline{F}^{(1)}(Z_1, Z_3)^2 D(S_1, S_2)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2](W_2^\star)^2 \mid \mathcal{A}_{1,3}, \mathcal{A}_{1,2}] \\
&+ \rho_n \mathbb{E}[\overline{F}^{(1)}(Z_1, Z_3)^2 D(S_1, S_2)\mathbb{E}[D(S_1, S_2| \mid H_{1,2}, S_2](W_2^\star)^2 \mid \tilde{\mathcal{A}}_{1,3}, \mathcal{A}_{1,2}] \\
&+ \rho_n^2 \mathbb{E}[(\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star)^2 \mid \mathcal{A}_{1,3}, \mathcal{A}_{1,2}]
\end{aligned}
$$

$$
\begin{aligned}
&+ \rho_n \mathbb{E}[(\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star)^2 \mid \tilde{\mathcal{A}}_{1,3}, \mathcal{A}_{1,2}] \\
&+ \mathbb{E}[(\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star)^2 \mid \tilde{\mathcal{A}}_{1,2}] \\
&\lesssim \rho_n^2 \mathbb{E}[\tilde{F}(Z_1, Z_3)^2 \mid \mathcal{A}_{1,3}, \mathcal{A}_{1,2}] + \rho_n^3 \mathbb{E}[\tilde{F}(Z_1, Z_3)^2 \mid \mathcal{A}_{1,3}, \tilde{\mathcal{A}}_{1,2}] + O(\rho_n^2) = O(\rho_n^2) \,,
\end{aligned} \tag{C.4.186}
$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. By plugging (C.4.184) and (C.4.186) into (C.4.179), we find that

$$
\operatorname{Var}\left(\frac{(n-1)(n-2)}{6n}\left(\frac{1}{\binom{n}{3}} \sum_{1 \leq i<j<k \leq n} f_2(Z_i, Z_j, Z_k)\right)\right) = O(\rho_n^3) \,. \tag{C.4.187}
$$

Hence, Chebyshev's inequality and the representation (C.4.176) imply that

$$
\frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{F}(Z_i, Z_k)\tilde{D}_{i,j}^\star W_j^\star = O_p(\rho_n^{3/2}) \,, \tag{C.4.188}
$$

as required. ■

C.4.9 Proof of Lemma C.4.8

We continue to use the notation introduced in the proofs of Lemmas C.3.3, C.4.5, and C.4.7.

C.4.9.1 Part (i)

Each bound follows from an application of Lemma C.4.11. In particular, we can evaluate

$$
\left|\frac{1}{n^3} \sum_{i=1}^n \sum_{j \neq i} \sum_{k \notin \{i,j\}} R_i \tilde{D}_{k,j}^\star W_j^\star\right| \leq \frac{1}{n} \sqrt{\left(\frac{1}{n^3} \sum_{i=1}^n \left(\sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{D}_{k,j}^\star W_j^\star\right)^2\right)\left(\frac{1}{n} \sum_{i=1}^n R_i^2\right)} \,, \tag{C.4.189}
$$

$$
\left|\frac{1}{n^4} \sum_{i=1}^n \sum_{j \neq i} \sum_{k \notin \{i,j\}} R_i D_{k,j}^\star W_j^\star\right| \leq \frac{1}{n^{3/2}} \sqrt{\left(\frac{1}{n^4} \sum_{i=1}^n \left(\sum_{j \neq i} \sum_{k \notin \{i,j\}} D_{k,j}^\star W_j^\star\right)^2\right)\left(\frac{1}{n} \sum_{i=1}^n R_i^2\right)} \,,
$$

$$
\left|\frac{1}{n^3} \sum_{i=1}^n \sum_{j \neq i} R_i D_{j,i}^\star W_i^\star\right| \leq \frac{1}{n} \sqrt{\left(\frac{1}{n^3} \sum_{i=1}^n \left(\sum_{j \neq i} D_{j,i}^\star W_i^\star\right)^2\right)\left(\frac{1}{n} \sum_{i=1}^n R_i^2\right)} \,, \quad \text{and}
$$

$$\left|\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i} R_i D^\star_{i,j} W^\star_j\right| \leq \frac{1}{n^{5/2}}\sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i} D^\star_{i,j} W^\star_j\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} R_i^2\right)}$$

by Cauchy-Schwarz. Recall that

$$\frac{1}{n}\sum_{i=1}^{n} R_i^2 = O_p(\rho_n) \tag{C.4.190}$$

by Markov's inequality and Lemma C.4.11. In turn, we can evaluate

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n^3}\sum_{i=1}^{n}\left(\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{D}^\star_{k,j} W^\star_j\right)^2\right] &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{j'\neq i}\sum_{k\notin\{i,j\}}\sum_{k'\notin\{i,j\}} \mathbb{E}[\tilde{D}^\star_{k,j}\tilde{D}^\star_{k',j'} W^\star_j W^\star_{j'}] \\ &= \frac{1}{n^2}\sum_{k\notin\{i,j\}}\sum_{k'\notin\{i,j\}} \mathbb{E}[\tilde{D}^\star_{k,j}\tilde{D}^\star_{k',j}(W^\star_j)^2] = O(\rho_n)\ ,\end{aligned} \tag{C.4.191}$$

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n^4}\sum_{i=1}^{n}\left(\sum_{j\neq i}\sum_{k\notin\{i,j\}} D^\star_{k,j} W^\star_j\right)^2\right] &= \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{j'\neq i}\sum_{k\notin\{i,j\}}\sum_{k'\notin\{i,j\}} \mathbb{E}[D^\star_{k,j} D^\star_{k',j'} W^\star_j W^\star_{j'}] \\ &= \frac{1}{n^2}\sum_{k\notin\{i,j\}}\sum_{k'\notin\{i,j\}} \mathbb{E}[D^\star_{k,j} D^\star_{k',j}(W^\star_j)^2] = O(\rho_n)\ ,\end{aligned}$$

$$\mathbb{E}\left[\frac{1}{n^3}\sum_{i=1}^{n}\left(\sum_{j\neq i} D^\star_{j,i} W^\star_i\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{j'\neq i} \mathbb{E}[D^\star_{j,i} D^\star_{j',i}(W^\star_i)^2] = O(\rho_n)\ , \quad \text{and}$$

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i} D^\star_{i,j} W^\star_j\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[(D^\star_{i,j})^2 (W^\star_j)^2] = O(\rho_n)\ .$$

Thus, the desired bounds follow by applying Markov's inequality, using the bounds (C.4.191), and by plugging the resultant bounds and (C.4.190) into the inequalities (C.4.189). ■

C.4.9.2 Part (ii)

We continue to use the notation adopted in the proof of Lemma C.4.7. Observe that we can write

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \overline{F}(Z_i,Z_j)\tilde{D}^\star_{k,j} W^\star_j = \frac{\binom{n}{3}}{n^3}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n} f_1(Z_i,Z_j,Z_k)\ , \tag{C.4.192}$$

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\overline{F}(Z_i,Z_j)D^\star_{k,j}W^\star_j = \frac{\binom{n}{3}}{n^4}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n} f_3(Z_i,Z_j,Z_k)\ , \tag{C.4.193}$$

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\overline{F}(Z_i,Z_j)D^\star_{j,i}W^\star_i = \frac{\binom{n}{2}}{n^3}\frac{1}{\binom{n}{2}}\sum_{1\leq i<j\leq n} f_3(Z_i,Z_j)\ , \quad \text{and} \tag{C.4.194}$$

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\overline{F}(Z_i,Z_j)D^\star_{i,j}W^\star_j = \frac{\binom{n}{2}}{n^4}\frac{1}{\binom{n}{2}}\sum_{1\leq i<j\leq n} f_4(Z_i,Z_j) \tag{C.4.195}$$

where

$$\begin{aligned}
f_1(Z_i,Z_j,Z_k) &= \sum_{\pi\in\Pi_3}\overline{F}(Z_{\pi(1)},Z_{\pi(2)})\tilde{D}^\star(S_{\pi(3)},S_{\pi(2)})W^\star_{\pi(2)}\ ,\\
f_2(Z_i,Z_j,Z_k) &= \sum_{\pi\in\Pi_3}\overline{F}(Z_{\pi(1)},Z_{\pi(2)})D^\star(S_{\pi(3)},S_{\pi(2)})W^\star_{\pi(2)}\ ,\\
f_3(Z_i,Z_j) &= \sum_{\pi\in\Pi_2}\overline{F}(Z_{\pi(1)},Z_{\pi(2)})D^\star(S_{\pi(2)},S_{\pi(1)})W^\star_{\pi(1)}\ , \quad \text{and}\\
f_4(Z_i,Z_j) &= \sum_{\pi\in\Pi_2}\overline{F}(Z_{\pi(1)},Z_{\pi(2)})D^\star(S_{\pi(1)},S_{\pi(2)})W^\star_{\pi(2)}\ .
\end{aligned} \tag{C.4.196}$$

We successively apply Lemma C.4.2 to bound the variance of each term.

We begin by considering the term (C.4.192). Lemma C.4.2 and the law of total variance imply that

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n} f_1(Z_i,Z_j,Z_k)\right)\\
&\lesssim n^{-1}\,\mathrm{Var}(\mathbb{E}[f_1(Z_1,Z_2,Z_3)\mid Z_1]) + n^{-2}\,\mathrm{Var}(f_1(Z_1,Z_2,Z_3))\ .
\end{aligned} \tag{C.4.197}$$

Observe that

$$\mathbb{E}[\overline{F}(Z_i,Z_j)\tilde{D}^\star_{k,j}W^\star_j\mid Z_i] = 0 \quad \text{and} \quad \mathbb{E}[\overline{F}(Z_i,Z_j)\tilde{D}^\star_{k,j}W^\star_j\mid Z_j] = 0 \tag{C.4.198}$$

respectively. We can evaluate

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[f_1(Z_1,Z_2,Z_3)\mid Z_3])\\
&\lesssim \mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\tilde{D}^\star(S_3,S_2)W^\star_2\mid Z_3]^2]\\
&\lesssim \mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)D(S_3,S_2)W^\star_2\mid Z_3]^2]\\
&+\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)\mathbb{E}[D(S_3,S_2)\mid H_{3,2},S_2]W^\star_2\mid Z_3]^2]
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \rho_n^2 \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2) D(S_3, S_2) W_2^\star \mid \mathcal{A}_{3,2}, Z_3]^2] \\
&+ \rho_n^2 \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2) \mathbb{E}[D(S_3, S_2) \mid H_{3,2}, S_2] W_2^\star \mid \mathcal{A}_{3,2}, Z_3]^2] \\
&+ \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2) \mathbb{E}[D(S_3, S_2) \mid H_{3,2}, S_2] W_2^\star \mid \tilde{\mathcal{A}}_{3,2}, Z_3]^2] \\
&\lesssim \rho_n^2 \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mid \mathcal{A}_{3,2}, Z_3]] + \rho_n^2 \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mid \tilde{\mathcal{A}}_{3,2}, Z_3]] = O(\rho_n^2)
\end{aligned}
\tag{C.4.199}
$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Likewise, we can evaluate

$$
\begin{aligned}
&\mathrm{Var}(f_1(Z_1, Z_2, Z_3)) \\
&\lesssim \mathbb{E}[\overline{F}(Z_1, Z_2)^2 \tilde{D}^\star(S_3, S_2)^2 (W_2^\star)^2] \\
&\lesssim \mathbb{E}[\overline{F}(Z_1, Z_2)^2 D(S_3, S_2)^2 (W_2^\star)^2] \\
&+ \mathbb{E}[\overline{F}(Z_1, Z_2)^2 D(S_3, S_2) \mathbb{E}[D(S_3, S_2) \mid H_{3,2}, S_2] (W_2^\star)^2] \\
&+ \mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mathbb{E}[D(S_3, S_2) \mid H_{3,2}, S_2]^2 (W_2^\star)^2] \\
&\lesssim \rho_n \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mid \mathcal{A}_{2,3}, Z_3]] + \rho_n^2 \mathbb{E}[\mathbb{E}[\overline{F}(Z_1, Z_2)^2 \mid \tilde{\mathcal{A}}_{2,3}, Z_3]] = O(\rho_n) \ ,
\end{aligned}
\tag{C.4.200}
$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Thus, by plugging (C.4.199) and (C.4.200) into the bound (C.4.197), and applying Chebyshev's inequality, we obtain

$$
\frac{1}{n^3} \sum_{i=1} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \overline{F}(Z_i, Z_j) \tilde{D}^\star_{k,j} W_j^\star = O_p(n^{-1/2} \rho_n) \ , \tag{C.4.201}
$$

verifying the first representation in (C.4.92).

Next, we consider the terms (C.4.193) through (C.4.195). Lemma C.4.2 and the law of total variance imply that

$$
\mathrm{Var}\left(\frac{1}{\binom{n}{3}} \sum_{1 \le i < j < k \le n} f_2(Z_i, Z_j, Z_k) \right) \lesssim n^{-1} \, \mathrm{Var}(f_1(Z_1, Z_2, Z_3)) \ , \tag{C.4.202}
$$

$$
\mathrm{Var}\left(\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} f_3(Z_i, Z_j) \right) \lesssim n^{-1} \, \mathrm{Var}(f_3(Z_1, Z_2)) \ , \quad \text{and} \tag{C.4.203}
$$

$$
\mathrm{Var}\left(\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} f_4(Z_i, Z_j) \right) \lesssim n^{-1} \, \mathrm{Var}(f_4(Z_1, Z_2)) \ , \tag{C.4.204}
$$

respectively. We can evaluate

$$
\begin{aligned}
&\mathrm{Var}(f_1(Z_1, Z_2, Z_3) \\
&\lesssim \mathbb{E}[\overline{F}(Z_1, Z_2)^2 D^\star(S_3, S_2)^2 (W_2^\star)^2]
\end{aligned}
$$

$$
\begin{aligned}
&\lesssim \mathbb{E}[\overline{F}(Z_1,Z_2)^2 D(S_3,S_2)^2 (W_2^\star)^2] \\
&+ \mathbb{E}[\overline{F}(Z_1,Z_2)^2 D(S_3,S_2)\mathbb{E}[D(S_3,S_2) \mid G_{3,2}](W_2^\star)^2] \\
&+ \mathbb{E}[\overline{F}(Z_1,Z_2)^2 \mathbb{E}[D(S_3,S_2) \mid G_{3,2}]^2 (W_2^\star)^2] \\
&\lesssim \rho_n \mathbb{E}[\mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)^2 \mid \mathcal{A}_{2,3}, Z_3]] \\
&+ \rho_n \mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)^2 \mid \mathcal{A}_{1,2}, Z_2]] + \mathbb{E}[\mathbb{E}[\overline{F}(Z_1,Z_2)^2 \mid \tilde{\mathcal{A}}_{1,2}, Z_2]] = O(\rho_n)\ , \qquad \text{(C.4.205)}
\end{aligned}
$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. The bounds

$$
\mathrm{Var}(f_3(Z_1,Z_2)) = O(\rho_n) \quad \text{and} \quad \mathrm{Var}(f_4(Z_1,Z_2)) = O(\rho_n)\ , \qquad \text{(C.4.206)}
$$

follow from identical arguments. Thus, Chebyshev's inequality implies that

$$
\frac{1}{n^4}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \overline{F}(Z_i,Z_j) D^\star_{k,j} W^\star_j = \frac{1}{n}\mathbb{E}[Y_i D^\star_{\ell,q} W^\star_q] + O_p(n^{-3/2}\rho_n^{1/2})\ , \qquad \text{(C.4.207)}
$$

$$
\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i} \overline{F}(Z_i,Z_j) D^\star_{j,i} W^\star_i = \frac{1}{n}\mathbb{E}[Y_i D^\star_{\ell,i} W^\star_i] + O_p(n^{-3/2}\rho_n^{1/2})\ , \quad \text{and} \qquad \text{(C.4.208)}
$$

$$
\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i} \overline{F}(Z_i,Z_j) D^\star_{j,i} W^\star_i = \frac{1}{n^2}\mathbb{E}[Y_i D^\star_{i,\ell} W^\star_\ell] + O_p(n^{-5/2}\rho_n^{1/2})\ , \qquad \text{(C.4.209)}
$$

verifying the second though fourth representations in (C.4.92) and completing the proof. ■

C.4.9.3 Part (iii)

Observe that each of the terms on the left-hand side of (C.4.93) has expectation equal to zero. Thus, it will suffice to bound their variances and apply Chebyshev's inequality. We begin by considering the term

$$
\begin{aligned}
&\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j\}} \tilde{F}(Z_i,Z_\ell)\tilde{D}^\star_{k,j} W^\star_j \\
&= \frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i,Z_k)\tilde{D}^\star_{k,j} W^\star_j + \frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,k\}} \tilde{F}(Z_i,Z_\ell)\tilde{D}^\star_{k,j} W^\star_j\ . \qquad \text{(C.4.210)}
\end{aligned}
$$

Observe that we can write

$$\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{F}(Z_i,Z_k)\tilde{D}^{\star}_{k,j}W^{\star}_j=\frac{\binom{n}{3}}{n^3}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}g_1(Z_i,Z_j,Z_k)\quad\text{and}\tag{C.4.211}$$

$$\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,k\}}\tilde{F}(Z_i,Z_\ell)\tilde{D}^{\star}_{k,j}W^{\star}_j=\frac{\binom{n}{4}}{n^3}\frac{1}{\binom{n}{4}}\sum_{1\leq i<j<k<\ell\leq n}g_2(Z_i,Z_j,Z_k,Z_\ell)\;,\tag{C.4.212}$$

where

$$g_1(Z_i,Z_j,Z_k)=\sum_{\pi\in\Pi_3}\tilde{F}(Z_{\pi(1)},Z_{\pi(3)})\tilde{D}^{\star}(S_{\pi(3)},S_{\pi(2)})W^{\star}_{\pi(2)}\quad\text{and}\tag{C.4.213}$$
$$g_2(Z_i,Z_j,Z_k,Z_\ell)=\sum_{\pi\in\Pi_4}\tilde{F}(Z_{\pi(1)},Z_{\pi(4)})\tilde{D}^{\star}(S_{\pi(3)},S_{\pi(2)})W^{\star}_{\pi(2)}\;,$$

respectively.

First, we evaluate the term (C.4.211). Observe that Lemma C.4.2 and the law of total variance imply

$$\begin{aligned}&\operatorname{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}g_1(Z_i,Z_j,Z_k)\right)\\&\lesssim n^{-1}\operatorname{Var}(\mathbb{E}[g_1(Z_i,Z_j,Z_k)\mid Z_i])+n^{-2}\operatorname{Var}(g_1(Z_i,Z_j,Z_k))\;.\end{aligned}\tag{C.4.214}$$

Observe that

$$\mathbb{E}[\tilde{F}(Z_i,Z_k)\tilde{D}^{\star}_{k,j}W^{\star}_j\mid Z_i]=0\quad\text{and}\quad\mathbb{E}[\tilde{F}(Z_i,Z_k)\tilde{D}^{\star}_{k,j}W^{\star}_j\mid Z_k]=0\;,\tag{C.4.215}$$

respectively. Thus, we can evaluate

$$\begin{aligned}&\operatorname{Var}(\mathbb{E}[g_1(Z_i,Z_j,Z_k)\mid Z_i])\\&\lesssim\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i,Z_k)\tilde{D}^{\star}_{k,j}W^{\star}_j\mid Z_j]^2]\\&\lesssim\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i,Z_k)D_{k,j}W^{\star}_j\mid Z_j]^2]+\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i,Z_k)\mathbb{E}[D_{k,j}\mid G_{k,j},S_j]W^{\star}_j\mid Z_j]^2]\\&\lesssim\rho_n^2\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i,Z_k)^2\mid\mathcal{A}_{k,j},Z_j]]+\rho_n^2\mathbb{E}[\mathbb{E}[\tilde{F}(Z_i,Z_k)^2\mid\tilde{\mathcal{A}}_{k,j},Z_j]]=O(\rho_n^2)\end{aligned}\tag{C.4.216}$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Similarly, we can evaluate

$$\begin{aligned}\operatorname{Var}(g_1(Z_i, Z_j, Z_k)) &\lesssim \rho_n \mathbb{E}[(\tilde{F}(Z_i, Z_k)\tilde{D}^\star_{k,j} W^\star_j)^2 \mid \mathcal{A}_{i,k}] \\ &\quad + \mathbb{E}[(\tilde{F}(Z_i, Z_k)\tilde{D}^\star_{k,j} W^\star_j)^2 \mid \tilde{\mathcal{A}}_{i,k}] \\ &\lesssim \rho_n \mathbb{E}[\tilde{F}(Z_i, Z_k)^2 \mid \mathcal{A}_{i,k}] + O(\rho_n) = O(\rho_n) \end{aligned} \tag{C.4.217}$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Consequently, Chebyshev's inequality implies

$$\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}} \tilde{F}(Z_i, Z_k)\tilde{D}^\star_{k,j} W^\star_j = O_p(n^{-1/2}\rho_n)\ . \tag{C.4.218}$$

Second, we evaluate the term (C.4.212). Observe that

$$\mathbb{E}[\tilde{F}(Z_1, Z_4)\tilde{D}^\star(S_3, S_2)W^\star_2 \mid Z_i, Z_j] = 0 \tag{C.4.219}$$

for any pair i, j in $\{1, 2, 3, 4\}$. Thus, Lemma C.4.2 and the law of total variance imply that

$$\begin{aligned}&\operatorname{Var}\left(\frac{1}{\binom{n}{4}}\sum_{1\le i<j<k<\ell\le n} g_2(Z_i, Z_j, Z_k, Z_\ell)\right) \\ &\lesssim n^{-3}\operatorname{Var}(\mathbb{E}[g_2(Z_i, Z_j, Z_k, Z_\ell) \mid Z_i, Z_j, Z_k]) + n^{-4}\operatorname{Var}(g_2(Z_i, Z_j, Z_k, Z_\ell))\ . \end{aligned} \tag{C.4.220}$$

Observe that

$$\mathbb{E}[\tilde{F}(Z_1, Z_4)\tilde{D}^\star(S_3, S_2)W^\star_2 \mid Z_i, Z_j, Z_k] = 0 \tag{C.4.221}$$

for each collection i, j, and k in $\{1, 2, 3, 4\}$ other than $\{2, 3, 4\}$. Hence, we can evaluate

$$\begin{aligned}&\operatorname{Var}(\mathbb{E}[g_2(Z_i, Z_j, Z_k, Z_\ell) \mid Z_i, Z_j, Z_k]) \\ &\lesssim \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_4) \mid Z_4]^2 \tilde{D}^\star_{3,2})^2 (W^\star_2)^2] \\ &\lesssim \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_4) \mid \mathcal{A}_{1,4}, Z_4]^2 D^2_{3,2}(W^\star_2)^2 \mid \mathcal{A}_{2,3}] \\ &+ \rho_n^2 \mathbb{E}[D^2_{3,2}(W^\star_2)^2 \mid \mathcal{A}_{2,3}] \\ &+ \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_4) \mid \mathcal{A}_{1,4}, Z_4]^2 D_{3,2}\mathbb{E}[D_{3,2} \mid H_{3,2}, S_2](W^\star_2)^2 \mid \mathcal{A}_{2,3}] \\ &+ \rho_n^2 \mathbb{E}[\mathbb{E}[D_{3,2}\mathbb{E}[D_{3,2} \mid H_{3,2}, S_2](W^\star_2)^2 \mid \mathcal{A}_{2,3}] \\ &+ \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_4) \mid \mathcal{A}_{1,4}, Z_4]^2 \mathbb{E}[D_{3,2} \mid H_{3,2}, S_2]^2 (W^\star_2)^2 \mid \mathcal{A}_{2,3}] \\ &+ \rho_n^2 \mathbb{E}[\mathbb{E}[\mathbb{E}[D_{3,2} \mid H_{3,2}, S_2]^2 (W^\star_2)^2 \mid \mathcal{A}_{2,3}] \end{aligned}$$

$$+ \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_4) \mid Z_4]^2 \mathbb{E}[D_{3,2} \mid H_{3,2}, S_2]^2 (W_2^\star)^2 \mid \tilde{\mathcal{A}}_{2,3}]$$
$$\lesssim \rho_n^3 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_4)^2 \mid \mathcal{A}_{1,4}, Z_4]] + O(\rho_n^2) = O(\rho_n^2) \tag{C.4.222}$$

by Cauchy-Schwarz, Assumption 3.5.2, Assumption C.3.1. Likewise, we can evaluate

$$\mathrm{Var}(g_2(Z_i, Z_j, Z_k, Z_\ell))$$
$$\lesssim \rho_n \mathbb{E}[\tilde{F}(Z_1, Z_4)^2 \tilde{D}^\star(S_3, S_2)^2 (W_2^\star)^2 \mid \mathcal{A}_{1,4}] + O(\rho_n) = O(\rho_n) \tag{C.4.223}$$

by Assumption 3.5.2 and Assumption C.3.1. Consequently, Chebyshev's inequality implies that

$$\frac{1}{n^3} \sum_{i=1} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \sum_{\ell \notin \{i,j,k\}} \tilde{F}(Z_i, Z_\ell) \tilde{D}_{k,j}^\star W_j^\star = O_p(n^{-1/2} \rho_n) \, . \tag{C.4.224}$$

Hence, the first representation in (C.4.93) follows by plugging (C.4.218) and (C.4.224) into (C.4.210).

Next, we consider the term

$$\frac{1}{n^4} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \sum_{\ell \notin \{i,j\}} \tilde{F}(Z_i, Z_\ell) D_{k,j}^\star W_j^\star$$
$$= \frac{1}{n^4} \sum_{i=1} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{F}(Z_i, Z_k) D_{k,j}^\star W_j^\star + \frac{1}{n^4} \sum_{i=1} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \sum_{\ell \notin \{i,j,k\}} \tilde{F}(Z_i, Z_\ell) D_{k,j}^\star W_j^\star \, . \tag{C.4.225}$$

As before, we can write

$$\frac{1}{n^4} \sum_{i=1} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \tilde{F}(Z_i, Z_k) D_{k,j}^\star W_j^\star = \frac{\binom{n}{3}}{n^4} \frac{1}{\binom{n}{3}} \sum_{1 \le i < j < k \le n} \overline{g}_1(Z_i, Z_j, Z_k) \quad \text{and} \tag{C.4.226}$$

$$\frac{1}{n^4} \sum_{i=1} \sum_{j \neq i} \sum_{k \notin \{i,j\}} \sum_{\ell \notin \{i,j,k\}} \tilde{F}(Z_i, Z_\ell) D_{k,j}^\star W_j^\star = \frac{\binom{n}{4}}{n^4} \frac{1}{\binom{n}{4}} \sum_{1 \le i < j < k < \ell \le n} \overline{g}_2(Z_i, Z_j, Z_k, Z_\ell) \, , \tag{C.4.227}$$

where

$$\overline{g}_1(Z_i, Z_j, Z_k) = \sum_{\pi \in \Pi_3} \tilde{F}(Z_{\pi(1)}, Z_{\pi(3)}) D^\star(S_{\pi(3)}, S_{\pi(2)}) W_{\pi(2)}^\star \quad \text{and} \tag{C.4.228}$$
$$\overline{g}_2(Z_i, Z_j, Z_k, Z_\ell) = \sum_{\pi \in \Pi_4} \tilde{F}(Z_{\pi(1)}, Z_{\pi(4)}) D^\star(S_{\pi(3)}, S_{\pi(2)}) W_{\pi(2)}^\star \, ,$$

respectively. First, we consider the term (C.4.226). Lemma C.4.2 and the law of total variance imply that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n}\overline{g}_1(Z_i,Z_j,Z_k)\right)\lesssim n^{-1}\operatorname{Var}(\overline{g}_1(Z_i,Z_j,Z_k))\,. \tag{C.4.229}$$

We can evaluate

$$\operatorname{Var}(\overline{g}_1(Z_i,Z_j,Z_k))\lesssim \rho_n\mathbb{E}[\tilde{F}(Z_1,Z_3)^2D^\star(S_3,S_2)^2(W_2^\star)^2\mid \mathcal{A}_{1,3}]+O(\rho_n)=O(\rho_n) \tag{C.4.230}$$

by Assumption 3.5.2 and Assumption C.3.1. Thus, Chebyshev's inequality and (C.4.226) imply that

$$\frac{1}{n^4}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\tilde{F}(Z_i,Z_k)D^\star_{k,j}W^\star_j=O_p(n^{-1}\rho_n^{-1/2})\,. \tag{C.4.231}$$

Second, we consider the term (C.4.227). Observe that

$$\mathbb{E}[\tilde{F}(Z_1,Z_4)D^\star(S_3,S_2)W_2\mid X_i]=0 \tag{C.4.232}$$

for each i in $\{1,2,3,4\}$. Thus, Lemma C.4.2 and the law of total variance imply that

$$\operatorname{Var}\left(\frac{1}{\binom{n}{4}}\sum_{1\le i<j<k<\ell\le n}\overline{g}_2(Z_i,Z_j,Z_k,Z_\ell)\right)\lesssim n^{-2}\operatorname{Var}(\overline{g}_2(Z_i,Z_j,Z_k,Z_\ell))\,. \tag{C.4.233}$$

We can evaluate

$$\operatorname{Var}(\overline{g}_2(Z_i,Z_j,Z_k,Z_\ell))\lesssim \rho_n\mathbb{E}[\tilde{F}(Z_1,Z_4)^2D^\star(S_3,S_2)^2(W_2^\star)^2\mid \mathcal{A}_{1,4}]+O(\rho_n)=O(\rho_n) \tag{C.4.234}$$

by Assumption 3.5.2 and Assumption C.3.1. Thus, Chebyshev's inequality and (C.4.226) imply that

$$\frac{1}{n^4}\sum_{i=1}\sum_{j\neq i}\sum_{k\notin\{i,j\}}\sum_{\ell\notin\{i,j,k\}}\tilde{F}(Z_i,Z_\ell)D^\star_{k,j}W^\star_j=O_p(n^{-1}\rho_n^{-1/2})\,. \tag{C.4.235}$$

Hence, the second representation in (C.4.93) follows by plugging (C.4.231) and (C.4.235) into (C.4.225).

Next, we consider the term

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}}\tilde{F}(Z_i,Z_\ell)D^\star_{j,i}W^\star_i\ . \tag{C.4.236}$$

Observe that we can write

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}}\tilde{F}(Z_i,Z_\ell)D^\star_{j,i}W^\star_i = \frac{\binom{n}{3}}{n^3}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n} g_3(Z_1,Z_j,Z_k) \tag{C.4.237}$$

where

$$g_3(Z_1,Z_j,Z_k) = \sum_{\pi\in\Pi_3}\tilde{F}(Z_{\pi(1)},Z_{\pi(3)})D^\star(S_{\pi(2)},S_{\pi(1)})W^\star_{\pi(1)}\ . \tag{C.4.238}$$

It holds that that

$$\mathbb{E}[\tilde{F}(Z_1,Z_3)D^\star(S_2,S_1)W^\star_1 \mid X_i] = 0 \tag{C.4.239}$$

for each i in $\{1,2,3\}$. Thus, Lemma C.4.2 and the law of total variance imply that

$$\mathrm{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n} g_3(Z_1,Z_j,Z_k)\right) \lesssim n^{-2}\,\mathrm{Var}(g_3(Z_1,Z_j,Z_k))\ . \tag{C.4.240}$$

We can evaluate

$$\mathrm{Var}(g_3(Z_1,Z_j,Z_k)) \lesssim \rho_n\mathbb{E}[\tilde{F}(Z_1,Z_3)^2D^\star(S_2,S_1)^2(W^\star_1)^2 \mid \mathcal{A}_{1,3}] + O(\rho_n) = O(\rho_n) \tag{C.4.241}$$

by Assumption 3.5.2 and Assumption C.3.1. Hence, Chebyshev's inequality implies that

$$\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}}\tilde{F}(Z_i,Z_\ell)D^\star_{j,i}W^\star_i = O_p(n^{-1}\rho_n^{1/2})\ , \tag{C.4.242}$$

verifying the third representation in (C.4.93).

Finally, we consider the term

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}}\tilde{F}(Z_i,Z_\ell)D^\star_{i,j}W^\star_j\ . \tag{C.4.243}$$

Observe that we can write

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}}\tilde{F}(Z_i,Z_\ell)D^{\star}_{i,j}W^{\star}_j=\frac{\binom{n}{3}}{n^4}\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}g_4(Z_1,Z_j,Z_k) \tag{C.4.244}$$

where

$$g_4(Z_1,Z_j,Z_k)=\sum_{\pi\in\Pi_3}\tilde{F}(Z_{\pi(1)},Z_{\pi(3)})\overline{D}(S_{\pi(1)},S_{\pi(2)})W^{\star}_{\pi(2)}\ . \tag{C.4.245}$$

Lemma C.4.2 and the law of total variance imply that

$$\mathrm{Var}\left(\frac{1}{\binom{n}{3}}\sum_{1\leq i<j<k\leq n}g_4(Z_1,Z_j,Z_k)\right)\lesssim n^{-1}\,\mathrm{Var}(g_4(Z_1,Z_j,Z_k))\ . \tag{C.4.246}$$

We can evaluate

$$\mathrm{Var}(g_4(Z_1,Z_j,Z_k))\lesssim\rho_n\mathbb{E}[\tilde{F}(Z_1,Z_3)^2D^{\star}(S_1,S_2)^2(W^{\star}_2)^2\mid\mathcal{A}_{1,3}]+O(\rho_n)=O(\rho_n) \tag{C.4.247}$$

by Assumption 3.5.2 and Assumption C.3.1. Hence, Chebyshev's inequality implies that

$$\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\notin\{i,j\}}\tilde{F}(Z_i,Z_\ell)D^{\star}_{i,j}W^{\star}_j=O_p(n^{-3/2}\rho_n^{1/2})\ , \tag{C.4.248}$$

verifying the fourth representation in (C.4.93) and completing the proof. ■

C.4.10 Proof of Lemma C.4.9

We continue to use the notation introduced in the proofs of Lemmas C.3.3, C.4.5, and C.4.7.

C.4.10.1 Part (i)

Observe that

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}R_iJ_{i,s}\right|\leq\frac{1}{\sqrt{n}}\sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}J_{i,s}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}R_i^2\right)}\quad\text{and} \tag{C.4.249}$$

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}R_iM_{i,s,r}\right|\leq\frac{1}{\sqrt{n}}\sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}M_{i,s,r}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}R_i^2\right)}$$

by Cauchy-Schwarz. Recall that

$$\frac{1}{n}\sum_{i=1}^{n} R_i^2 = O_p(\rho_n) \tag{C.4.250}$$

by Markov's inequality and Lemma C.4.11. Moreover, we have that

$$\frac{1}{n^2}\sum_{i=1}^{n} J_{i,s}^2 = O_p(\kappa_{n,s}) \quad \text{and} \quad \frac{1}{n^2}\sum_{i=1}^{n} M_{i,r,s}^2 = O_p(n\kappa_{n,s}) \tag{C.4.251}$$

by Lemma C.3.2. Hence, we can conclude that

$$\frac{1}{n^2}\sum_{i=1}^{n} R_i J_{i,s} = O_p((\rho_n\kappa_{n,s})^{1/2}n^{-1/2}) \quad \text{and} \quad \frac{1}{n^2}\sum_{i=1}^{n} R_i M_{i,r,s} = O_p((\rho_n\kappa_{n,s})^{1/2}) \ , \tag{C.4.252}$$

as required. ■

C.4.10.2 Part (ii)

Observe that

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)J_{i,s}\right| \leq \frac{1}{\sqrt{n}}\sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}J_{i,s}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{F}(Z_i)^2\right)} \quad \text{and} \tag{C.4.253}$$

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)M_{i,s,r}\right| \leq \frac{1}{\sqrt{n}}\sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}M_{i,s,r}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{F}(Z_i)^2\right)}$$

by Cauchy-Schwarz. Moreover, as Assumption C.3.1 implies that

$$\mathbb{E}[\tilde{F}(Z_i)^2] = O(1) \ , \tag{C.4.254}$$

Markov's inequality implies that

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{F}(Z_i)^2 = O(1) \ . \tag{C.4.255}$$

Thus, by applying the bounds (C.4.251), we can conclude

$$\frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)J_{i,s} = O_p(n^{-1/2}\kappa_{n,s}^{1/2}) = O_p(\kappa_{n,s})\ . \tag{C.4.256}$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)M_{i,s,r} = O_p(\kappa_{n,s}^{1/2}) = O_p(n^{1/2}\kappa_{n,s})\ , \tag{C.4.257}$$

as required. ■

C.4.10.3 Part (iii)

Observe that

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)J_{i,s}\right| \leq \sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}J_{i,s}^2\right)\left(\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}\tilde{F}(Z_i,Z_j)\right)^2\right)} \tag{C.4.258}$$

and

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)M_{i,s,r}\right| \leq \sqrt{\left(\frac{1}{n^2}\sum_{i=1}^{n}M_{i,s,r}^2\right)\left(\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}\tilde{F}(Z_i,Z_j)\right)^2\right)} \tag{C.4.259}$$

by Cauchy-Schwarz. We can evaluate

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}\tilde{F}(Z_i,Z_j)\right)^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{j'\neq i}\mathbb{E}[\tilde{F}(Z_i,Z_j)\tilde{F}(Z_i,Z_{j'})] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}[\tilde{F}(Z_i,Z_j)^2] \\ &= \rho_n\mathbb{E}[\tilde{F}(Z_i,Z_j)^2 \mid \mathcal{A}_{i,j}] + O(\rho_n) = O(\rho_n)\ ,\end{aligned} \tag{C.4.260}$$

where the second equality follows from the definition (C.4.78) and the bound follows from Assumptions 3.5.2 and C.3.1. Consequently, Markov's inequality implies that

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j\neq i}\tilde{F}(Z_i,Z_j)\right)^2 = O_p(\rho_n)\ . \tag{C.4.261}$$

Thus, by applying the bounds (C.4.251), we can conclude

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\tilde{F}(Z_i,Z_j)J_{i,s}=O_p(\rho_n^{1/2}\kappa_n^{1/2})=O_p(\kappa_{n,s})\;. \tag{C.4.262}$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n}\tilde{F}(Z_i)M_{i,s,r}=O_p(\rho_n^{1/2}n^{1/2}\kappa_{n,s}^{1/2})=O_p(n^{1/2}\kappa_{n,s})\;, \tag{C.4.263}$$

as required. ■

C.4.11 Proof of Lemma C.4.10

To verify the first bound, observe that

$$\begin{aligned}\mathrm{Var}(f_\xi^{(3)}(Z_1,Z_2,Z_3)) &\leq \mathrm{Var}(f_\xi(Z_1,Z_2,Z_3))\\ &\leq \mathrm{Var}(\tilde{F}(Z_1,Z_3)\tilde{D}^\star_{1,2}W^\star_2)+\mathrm{Var}(\tilde{D}^\star\tilde{D}^\star_{1,2}W^\star_2)=O(\rho_n^2)\;,\end{aligned} \tag{C.4.264}$$

where the first inequality follows from Lemma C.4.1, the second inequality follows from the definition of $f_\xi(Z_1,Z_2,Z_3)$, and the third inequality follows from the bounds (C.4.186) and (C.4.128) stated in Part (iii) of the Proof of Lemma C.4.7 and the Proof of Lemma C.4.5, respectively.

The second inequality follows from an argument similar to the bounds (C.4.134) and (C.4.185), respectively. In particular, observe that

$$\mathbb{E}[(f^{(2)}(Z_2,Z_2))^4]\lesssim\mathbb{E}[(\mathbb{E}[\tilde{F}(Z_1,Z_3)\tilde{D}^\star_{1,2}\mid Z_2,Z_3]W^\star_2)^4] \tag{C.4.265}$$

$$+\,\mathbb{E}[(\mathbb{E}[\tilde{D}^\star_{1,3}W^\star_3\mid Z_2,Z_3]W^\star_2)^4]\;. \tag{C.4.266}$$

We provide the details for the verification of the bound

$$\mathbb{E}[(\mathbb{E}[\tilde{F}(Z_1,Z_3)\tilde{D}^\star_{1,2}\mid Z_2,Z_3]W^\star_2)^4]=O(\rho_n^5) \tag{C.4.267}$$

as the obtaining the same bound for the second term in (C.4.266) follows from ana analogous argument, in the same sense that the bound (C.4.185) follows from an argument analogous to the bound (C.4.134). To this end, observe that

$$\mathbb{E}[(\mathbb{E}[\tilde{F}(Z_1,Z_3)\tilde{D}^\star_{1,2}\mid Z_2,Z_3]W^\star_2)^4]\lesssim\mathbb{E}[\mathbb{E}[\tilde{F}(Z_1,Z_3)D(S_1,S_2)W^\star_2\mid Z_2,Z_3]^4]$$

$$+ \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid Z_2, Z_3]^4] \,. \tag{C.4.268}$$

We can evaluate

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star \mid Z_2, Z_3]^4] \\
&\lesssim \rho_n^5 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^4 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^4 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)D(S_1, S_2)W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^4 \mid \tilde{\mathcal{B}}_{2,3}] \\
&\lesssim \rho_n^5 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^4 \mid \mathcal{A}_{1,2}, Z_2, Z_3] \mid \mathcal{B}_{2,3}] + O(\rho_n^6) = O(\rho_n^5) \,,
\end{aligned} \tag{C.4.269}$$

by Cauchy-Schwarz, Assumption 3.5.2, and Assumption C.3.1. Similarly, we can evaluate

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid G_{1,2}, S_2]W_2^\star \mid Z_2, Z_3]^4] \\
&\lesssim \rho_n^5 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^4 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \tilde{\mathcal{A}}_{1,2}, Z_2, Z_3]^4 \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^4 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \mathcal{A}_{1,2}, Z_2, Z_3]^4 \mid \tilde{\mathcal{B}}_{2,3}] \\
&+ \rho_n^4 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \mathcal{A}_{1,3}, Z_2, Z_3]^4 \mid \tilde{\mathcal{B}}_{2,3}] \\
&+ \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)\mathbb{E}[D(S_1, S_2) \mid H_{1,2}, S_2]W_2^\star \mid \tilde{\mathcal{A}}_{1,3}, \tilde{\mathcal{A}}_{1,2}, Z_2, Z_3]^4 \mid \tilde{\mathcal{B}}_{2,3}] \\
&\lesssim \rho_n^5 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^4 \mid \mathcal{A}_{1,2}, Z_2, Z_3] \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^5 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^4 \mid \tilde{\mathcal{A}}_{1,2}, Z_2, Z_3] \mid \mathcal{B}_{2,3}] \\
&+ \rho_n^8 \mathbb{E}[\mathbb{E}[\tilde{F}(Z_1, Z_3)^4 \mid \mathcal{A}_{1,3}, Z_2, Z_3] \mid \tilde{\mathcal{B}}_{2,3}] + O(\rho_n^5) = O(\rho_n^5) \,,
\end{aligned} \tag{C.4.270}$$

by Assumption 3.5.2 and Assumption C.3.4. Consequently, we find that

$$\mathbb{E}[(\mathbb{E}[\tilde{F}(Z_1, Z_3)\tilde{D}_{1,2}^\star \mid Z_2, Z_3]W_2^\star)^4] = O(\rho_n^5) \,, \tag{C.4.271}$$

by plugging (C.4.269) and (C.4.270) into (C.4.268).

Finally, we verify the third bound in (C.4.109). Observe that

$$\begin{aligned}
&\mathbb{E}[g_\xi^{(2)}(Z_1, Z_2)^2] \\
&= \mathbb{E}[\mathbb{E}[f_\xi^{(2)}(Z_3, Z_1)f_\xi^{(2)}(Z_3, Z_2) \mid Z_1, Z_2]^2] \\
&\lesssim \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)\tilde{D}_{4,3}^\star \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)\tilde{D}_{4,3}^\star \mid S_3, Z_2](W_3^\star)^2 \mid Z_1, Z_2]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{D}_{4,1}^\star W_1^\star \tilde{D}_{4,3}^\star \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)\tilde{D}_{4,3}^\star \mid S_3, Z_2](W_3^\star)^2 \mid Z_1, Z_2]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{D}_{4,1}^\star W_1^\star \tilde{D}_{4,3}^\star \mid S_3, Z_1]\mathbb{E}[\tilde{D}_{4,2}^\star W_2^\star \tilde{D}_{4,3}^\star \mid S_3, Z_2](W_3^\star)^2 \mid Z_1, Z_2]^2] \,,
\end{aligned} \tag{C.4.272}$$

by Cauchy-Schwarz and Assumption 3.4.2. We give the details for verifying the bound

$$\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)\tilde{D}^{\star}_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)\tilde{D}^{\star}_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] = o(\rho_n^6) \tag{C.4.273}$$

as the argument supporting the same bound for the other two terms is identical. To this end, we have that

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)\tilde{D}^{\star}_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)\tilde{D}^{\star}_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] \\
&\lesssim \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)D_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)D_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)D_{4,3} \mid S_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)\mathbb{E}[D_{4,3} \mid H_{4,3}, S_3] \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)\mathbb{E}[D_{4,3} \mid H_{4,3}, S_3] \mid S_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)\mathbb{E}[D_{4,3} \mid H_{4,3}, S_3] \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] ,
\end{aligned} \tag{C.4.274}$$

by Cauchy-Schwarz. In turn, we give the details for the verification that

$$\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)D_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)D_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] = o(\rho_n^6) . \tag{C.4.275}$$

An argument with an identical structure will verify the same bound for the second two terms in (C.4.274), in the same manner that the bound (C.4.270) is obtained from an argument with the same structure as the bound (C.4.269). To do this, recall the definition of the event $\mathcal{E}_{i,j}$, given at the beginning of Appendix C.3, and observe that $P\{\mathcal{E}_{i,j}\} = O(\rho_n)$ by Assumption C.3.2. We can evaluate

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)D_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)D_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2] \\
&\lesssim \rho_n\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)D_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)D_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2 \mid \mathcal{E}_{1,2}] \\
&+ \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)D_{4,3} \mid S_3, Z_1]\mathbb{E}[\tilde{F}(Z_4, Z_2)D_{4,3} \mid S_3, Z_2](W^{\star}_3)^2 \mid Z_1, Z_2]^2 \mid \tilde{\mathcal{E}}_{1,2}] \\
&\lesssim \rho_n^7\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \mathcal{B}_{3,1}, \mathcal{B}_{3,2}, Z_1, Z_2] \mid \mathcal{E}_{1,2}] \\
&+ \rho_n^7\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \mathcal{B}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2]^2 \mid \mathcal{E}_{1,2}] \\
&+ \rho_n^5\mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1]
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2] \mid \mathcal{E}_{1,2}] \\
&+ \rho_n^6 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \mathcal{B}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] \\
&+ \rho_n^4 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] \\
&\lesssim \rho_n^7 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^4 \mid \mathcal{A}_{4,3}, Z_3, Z_1, Z_2] \mid \mathcal{B}_{3,1}, \mathcal{B}_{3,2}, Z_1, Z_2] \mid \mathcal{E}_{1,2}] \\
&+ \rho_n^8 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \mid \mathcal{B}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2]^2 \mid \mathcal{E}_{1,2}] \\
&+ \rho_n^7 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \mathcal{B}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] \\
&+ \rho_n^4 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] + O(\rho_n^7) \\
&= \rho_n^4 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] + O(\rho_n^7) \ ,
\end{aligned}
\tag{C.4.276}
$$

through several applications of Assumption 3.5.2 and Assumption C.3.4. In turn, we can evaluate

$$
\begin{aligned}
&\rho_n^4 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \\
&\qquad\qquad \mathbb{E}[\tilde{F}(Z_4, Z_2)^2 \mid \mathcal{A}_{4,3}, Z_3, Z_2] \mid \tilde{\mathcal{B}}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] \\
&\lesssim \rho_n^4 \mathbb{E}[\mathbb{E}[\mathbb{E}[\tilde{F}(Z_4, Z_1)^4 \mid \mathcal{A}_{4,3}, Z_3, Z_1] \mid \tilde{\mathcal{B}}_{3,1}, \tilde{\mathcal{B}}_{3,2}, Z_1, Z_2] \mid \tilde{\mathcal{E}}_{1,2}] = o(\rho_n^6)
\end{aligned}
\tag{C.4.277}
$$

by Assumption 3.5.2 and Assumption C.3.4. The bound (C.4.275) is thus verified by plugging (C.4.275) into (C.4.276), completing the proof. ■

C.4.12 Proof of Lemma C.4.11

The result follows from an application of Lemma C.4.3. Throughout, we adopt the shorthand

$$
F_i(Z) = F_i(Z_j, Z_{-j}) = F(Z_i, Z_{-i}) \ . \tag{C.4.278}
$$

Now, observe that as

$$
\mathbb{E}[R_i \mid Z_i] = 0 \tag{C.4.279}
$$

we have that

$$\mathbb{E}[R_i^2] = \mathbb{E}[\mathbb{E}[R_i^2 \mid Z_i]] = \mathbb{E}[\mathrm{Var}(R_i \mid Z_i)] \ . \tag{C.4.280}$$

Thus, as the equality

$$\mathbb{E}[R_i \mid Z_i, Z_j] = 0 \tag{C.4.281}$$

holds almost surely, by definition, Lemma C.4.3 implies that the inequality

$$\begin{aligned}
&\mathbb{E}[\mathrm{Var}(R_i \mid Z_i)] \\
&\leq \frac{1}{2}\sum_{j\neq i}^{n}\sum_{k\notin\{i,j\}} \mathbb{E}[\mathbb{E}[(F(Z_i, Z_{-i}) - F(Z_i, Z_{-i}^{(j)})) - (F(Z_i, Z_{-i}^{(k)}) - F(Z_i, Z_{-i}^{(j,k)})) \mid Z]^2] \\
&\lesssim n^2\mathbb{E}[\mathbb{E}[\nabla_{j,k}^2 F(Z_i, Z_{-i}) \mid Z^{(j,k)}]^2]
\end{aligned} \tag{C.4.282}$$

holds, where we recall the notation introduced in Assumption C.3.1. Now, observe that we can evaluate

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\nabla_{j,k}^2 F(Z_i, Z_{-i}) \mid Z^{(j,k)}]^2] \\
&\lesssim \rho_n^4\mathbb{E}[\mathbb{E}[\nabla_{j,k}F(Z_i, Z_{-i}) \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}]^2] \\
&+ \rho_n^2\mathbb{E}[\mathbb{E}[\nabla_{j,k}F(Z_i, Z_{-i}) \mid A_{i,j} = 1, A_{i,k} = 0, Z^{(j,k)}]^2] \\
&+ \mathbb{E}[\mathbb{E}[\nabla_{j,k}F(Z_i, Z_{-i}) \mid A_{i,j} = 1, A_{i,k} = 0, Z^{(j,k)}]^2] = O(n^{-2}\rho_n)
\end{aligned} \tag{C.4.283}$$

by Assumption C.3.1. By plugging (C.4.283) into (C.4.282), we find that

$$\mathbb{E}[R_i^2] = O(\rho_n) \ , \tag{C.4.284}$$

completing the proof. ■

C.5. Additional Results and Discussion

C.5.1 Extensions to Theorem 3.2.1

In this appendix, we outline several extensions to the baseline decomposition given in Section 3.2.

C.5.1.1 Direct Effects

In many applications, the regression specification (3.2.1) is augmented to additionally include the covariate W_i; that is, through the specification

$$Y_i = \alpha + \tau \cdot W_i + \theta \cdot \Delta_i + \varepsilon_i \; . \tag{C.5.1}$$

The coefficient on the covariate W_i is often interpreted as an estimate of the "direct effect" of the treatment W_i on Y_i.[1] The decomposition stated in Theorem 3.2.1 holds for this augmented specification.

In particular, consider the augmented loss function

$$L_n(\tau, \theta) = \min_{\alpha} \; \sum_{i=1}^{n} (Y_i - \alpha - \tau \cdot W_i - \theta \cdot \Delta_i)^2 \tag{C.5.2}$$

associated with the specification (C.5.1). By the Frisch-Waugh-Lovell Theorem, we can write

$$\begin{aligned} \mathbb{E}[L_n(\tau, \theta)] &= \mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \tau\tilde{W}_i - \theta \cdot \tilde{\Delta}_i)^2\right] \; , \\ \text{where} \quad \tilde{W}_i &= W_i - \frac{1}{n}\sum_{j=1}^{n} W_j \quad \text{and} \\ \tilde{\Delta}_i &= \sum_{j \neq i}\left(D_{i,j} - \frac{1}{n}\sum_{k \neq j} D_{k,j}\right) W_j - \frac{1}{n}\sum_{k \neq i} D_{k,i} W_i \; . \end{aligned} \tag{C.5.3}$$

Observe that

$$\mathbb{E}[\tilde{\Delta}_i \tilde{W}_i] = -\frac{n-1}{n}\mathbb{E}[D_{i,j} W_j^2] \tag{C.5.4}$$

and that

$$\mathrm{Var}(\tilde{W}_i) = \frac{n-1}{n}\,\mathrm{Var}(W_i) \; . \tag{C.5.5}$$

[1] See Li and Wager (2022) for discussion of estimators of direct effects in models similar to the settings considered in this paper.

Thus, by a second application of the Frisch-Waugh-Lovell Theorem, it holds that

$$\min_{\tau} \mathbb{E}[L_n(\tau,\theta)] = \mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \theta\cdot\tilde{\Delta}'_i)^2\right] ,$$
$$\text{where} \quad \tilde{\Delta}'_i = \sum_{j\neq i}\left(D_{i,j} - \frac{1}{n}\sum_{k\neq j} D_{k,j}\right) W_j - \left(\frac{1}{n}\sum_{k\neq i} D_{k,i} - \frac{\mathbb{E}[D_{i,j}W_j^2]}{\mathrm{Var}(W_i)}\right) W_i \,. \tag{C.5.6}$$

Hence, by Projection Theorem, it holds that

$$\overline{\theta}_n = \frac{\sum_{i=1}^{n}\mathbb{E}[Y_i\tilde{\Delta}'_i]}{\sum_{i=1}^{n}\mathbb{E}[(\tilde{\Delta}'_i)^2]} \,. \tag{C.5.7}$$

The representation (C.5.7) is very similar to the representation (C.1.22) considered in the proof of Theorem 3.2.1. The only difference is through the second term in (C.5.6), which will contribute slightly different lower order terms when expressed as a representation analogous to (C.1.23). It can be shown that these lower order terms are smaller than $o(n^{-1/2})$ under the same regularity conditions used for Theorem 3.2.1.

C.5.1.2 Monotone Transformations

In some settings, the proximity measure of interest enters through a monotonic function $Q(D_{i,j})$; that is, thorough the specification

$$Y_i = \alpha + \theta\cdot Q_i + \varepsilon_i \quad \text{where} \quad Q_i = \sum_{j\neq i} Q(D_{i,j})W_j \,. \tag{C.5.8}$$

For instance, when $D_{i,j}$ is a function of geographic distance, researchers often set $Q(\cdot)$ as a binary indicator that $D_{i,j}$ is an element of some pre-defined interval; see, e.g, Miguel and Kremer (2004), Feyrer et al. (2017), Egger et al. (2022b), Myers and Lanahan (2022), and Muralidharan et al. (2023). In other cases, the function $Q(\cdot)$ often depends on various elasticities fixed at values chosen by referencing pre-existing literature; see e.g., Kovak (2013), Autor et al. (2013), Donaldson and Hornbeck (2016), and Adao et al. (2025).

It follows immediately from Theorem 3.2.1 that, in this case, we have that

$$\overline{\theta}_n = \int_{\overline{\mathcal{Q}}}\int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(q,w \mid S_j)\partial^2_{q,w}\mathbb{E}[Y_i \mid Q_i = q, W_j = w, S_j]\Big]\,\mathrm{d}w\,\mathrm{d}q + o(n^{-1/2}) \,, \tag{C.5.9}$$

where the weights are defined analogously to (3.2.8) and $\overline{\mathcal{Q}}$ is some interval containing the support of $Q(D_{i,j})$. In cases where the function $Q(\cdot)$ is invertible (e.g., $Q(D_{i,j})$ is a smooth decreasing function of a distance $D_{i,j}$), it follows immediately that

$$\overline{\theta}_n = \int_{\overline{\mathcal{Q}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(q, w \mid S_j) \partial^2_{q,w} \mu(Q^{-1}(q), w \mid S_j)\Big] \, \mathrm{d}w \, \mathrm{d}q + o(n^{-1/2}) \ . \tag{C.5.10}$$

That is, in this case, the parameter $\overline{\theta}_n$ is still a convex average of the parameter $\partial^2_{q,w}\mu(\delta, w \mid S_j)$, but now the weights in this average depend on the function $Q(\cdot)$.

C.5.1.3 Dyadic Information

Network linkages are often modeled through specifications of the form

$$D_{i,j} = D_n(S_i, S_j, \varepsilon_{i,j}) \ , \tag{C.5.11}$$

where the variables $\varepsilon_{i,j}$ are i.i.d. and distributed independently of the coordinates S_i. See, for instance, the latent position or graphon network models considered by, e.g., Hoff et al. (2002) and Li and Wager (2022), respectively. Indeed, by the Aldous-Hoover representation theorem, any exchangeable array admits such a representation (Aldous, 1981; Hoover, 1979). See Menzel (2021) for discussion.

Under the assumption that the treatments W_i are independent of the error terms $\varepsilon_{i,j}$, the decomposition Theorem 3.2.1 holds for this setting, unchanged. In particular, Theorem 3.2.1 only requires that the treatment W_j is independent of the proximity measure $D_{i,j}$, conditional on the coordinate S_j. So long as this continues to hold, the theorem will be unchanged.

We impose Assumption 3.2.1 because the variables $\varepsilon_{i,j}$ would complicate the construction of potential outcomes. In particular, when considering interventions on the proximity measure $D_{i,j}$, while holding S_i fixed, we would need to specify whether we are changing the value of the coordinate S_j or the error term $\varepsilon_{i,j}$. Likewise, incorporating the variables $\varepsilon_{i,j}$ would require augmenting the Assumption 3.3.1 to accommodate dyadically indexed variables, which would substantially complicate our theoretical analysis.

C.5.1.4 Bipartite Data

In many applications, treatments and outcomes are measured for different types of units. In particular, suppose that there are n units and m treatment nodes. Each unit i is associated with an outcome Y_i. Each node k is associated with a treatment W_k. Each unit i and node k

is associated with the measure $D_{i,k}$. A prominent instance of this setting are cases where treatments are assigned to industries and outcomes are measured at geographic units of aggregation. See, for instance, Kovak (2013), Autor et al. (2013), Adao et al. (2019), and Borusyak et al. (2022b). Here, the treatments W_k might be changes to the productivity of industry k and $D_{i,k}$ might be the share of employment of city i in industry k.

Researchers often report coefficients from regression specifications of the form

$$Y_i = \alpha + \theta \cdot \Delta_i + \varepsilon_i \quad \text{and} \quad \Delta_i = \sum_{j=1}^{m} D_{j,k} W_k \tag{C.5.12}$$

A result very similar to Theorem 3.2.1 holds for this setting. That is, such regressions identify convex averages of the parameters

$$\partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,k} = \delta, W_k = w, R_k] \tag{C.5.13}$$

Thus, "shift-share" specifications of the form (C.5.12) can be interpreted as measuring the effect of changing the measure $D_{i,k}$ on the effect of the treatment W_k on the outcome Y_i. We conclude this subsection by giving a formal statement and proof for this result.

Causal interpretations of the parameter (C.5.13) are susceptible to the same sort of confounding considered in the main text. For instance, suppose that the variable $D_{i,k}$ denotes the share of employment in location i that belongs to industry k and that the variable $G_{i,k}$ denotes the share of imports to location i from industry k. If the productivity of industry k increases, wages in location i may increase through $D_{i,k}$, due to increased local labor demand, or through $G_{i,k}$, due to cheaper intermediates.

Approaches analogous to those considered in the main text can be used to disentangle these channels. In particular, the variable of interest $D_{i,k}$ can be residualized on the auxiliary channel $G_{i,k}$, and the residuals can be used in the place of $D_{i,k}$ in the specification (C.5.12). Estimators with this structure can be shown to be consistent through arguments analogous to those used to establish Theorem 3.5.1. We note, however, that obtaining a result analogous to the Gaussian approximation result given in Theorem 3.6.1 would be a more significant undertaking, as the central limit theorem from Liu et al. (2025) is not immediately applicable.

We now state and prove a version of Theorem 3.2.1 for settings with bipartite data. We consider the following assumptions, which mirror Assumptions 3.2.1 to 3.2.3.

Assumption C.5.1. *Each unit i and node k are associated with coordinates S_i and R_k, respectively. There exists a function $D_n(\cdot,\cdot)$, valued on the non-negative, real-valued set $\mathcal{D}$,* *such that the representation*

$$D_{i,k} = D_n(S_i, R_k) \ , \tag{C.5.14}$$

holds almost surely for each unit i and node k.

Assumption C.5.2. *The replicates $(R_k, W_k)_{k=1}^m$ and $(S_i)_{i=1}^n$ are i.i.d., and are independent of each other. Moreover, the orthogonality condition*

$$\mathbb{E}[W_i \mid R_i] = 0 \tag{C.5.15}$$

holds almost surely.

Assumption C.5.3. *For any units i and j, and node k, the map*

$$(\delta, w) \mapsto \mathbb{E}[Y_i \mid D_{j,k} = \delta, W_j = w, R_j] \tag{C.5.16}$$

is twice continuously differentiable, almost surely. Moreover, it holds that

$$\partial^2_{\delta,w} \, \mathbb{E}[Y_i \mid D_{j,k} = \delta, W_k = w, R_k = s] = o(n^{-1/2}) \tag{C.5.17}$$

uniformly over each δ, w, and s in their respective domains.

Theorem C.5.1. *Assume that the treatments and proximity measures have support contained in intervals $\overline{\mathcal{W}}$ and $\overline{\mathcal{D}}$ of constant length and that the outcomes are identically distributed. Consider the loss function*

$$L_n(\theta) = \min_{\alpha} \sum_{i=1}^{n} (Y_i - \alpha - \theta \cdot \Delta_i)^2, \quad \text{where} \quad \Delta_i = \sum_{j=1}^{m} D_{i,k} W_k \ . \tag{C.5.18}$$

Under Assumptions C.5.1 to C.5.3, the parameter $\overline{\theta}_n$ that minimizes $\mathbb{E}[L_n(\theta)]$ admits the representation

$$\overline{\theta}_n = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(\delta, w \mid R_j) \, \partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,k} = \delta, W_k = w, R_k]\Big] \, \mathrm{d}w \, \mathrm{d}\delta + o(n^{-1/2}) \ , \tag{C.5.19}$$

where the weights

$$\lambda(\delta, w \mid R_j) = \frac{\mathrm{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j} \mid R_j) \, \mathrm{Cov}\left(\mathbb{I}\{W_j \geq w\}, W_j \mid R_j\right)}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid R_j) \, \mathrm{Var}(W_j \mid R_j)]} \tag{C.5.20}$$

are convex.

Proof. By the Frisch-Waugh-Lovell and Projection Theorems, we can write

$$\overline{\theta}_n = \frac{\sum_{i=1}^n \mathbb{E}[Y_i \tilde{\Delta}_i]}{\sum_{i=1}^n \mathbb{E}[(\tilde{\Delta}_i)^2]} \,, \tag{C.5.21}$$

where

$$\tilde{\Delta}_i = \sum_{k=1}^m D_{i,k} W_k - \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^m D_{j,k} W_k = \sum_{k=1}^m (D_{i,k} - \frac{1}{n} \sum_{j=1}^n D_{j,k}) W_k \,. \tag{C.5.22}$$

Observe that

$$\mathbb{E}[Y_i \tilde{\Delta}_i] = m \left(1 - \frac{1}{n}\right) \left(\mathbb{E}[Y_i \tilde{D}_{i,k} W_k] - \mathbb{E}[Y_i \tilde{D}_{j,k} W_k]\right) \tag{C.5.23}$$

and

$$\mathbb{E}[(\tilde{\Delta}_i)^2] = m \left(1 - \frac{1}{n}\right) \mathbb{E}[\tilde{D}_{i,k}^2 W_k^2] \,. \tag{C.5.24}$$

Thus, we obtain the representation

$$\overline{\theta}_n = \frac{\mathbb{E}[Y_i \tilde{D}_{i,k} W_k] - \mathbb{E}[Y_i \tilde{D}_{j,k} W_k]}{\mathbb{E}[\mathrm{Var}(D_{i,k} \mid R_k) \, \mathrm{Var}(R_k \mid W_k)]} \,. \tag{C.5.25}$$

Consequently, by Lemma C.1.2, we obtain the representation

$$\begin{aligned} \overline{\theta}_n = \int_{\overline{\mathcal{D}}} \int_{\overline{\mathcal{W}}} \mathbb{E}\Big[\lambda(\delta, w \mid R_j) \Big(\partial^2_{\delta,w} \mathbb{E}[Y_i \mid D_{i,k} = \delta, W_k = w, R_k] \\ - \mathbb{E}[Y_i \mid D_{j,k} = \delta, W_k = w, R_k]\Big)\Big] \, \mathrm{d}w \, \mathrm{d}\delta \end{aligned} \tag{C.5.26}$$

where the weights $\lambda(\delta, w \mid R_j)$ are defined in (C.5.20). The weights are convex by Lemma C.1.3. Thus, the representation (C.5.19) follows from Assumption C.5.3, as required. ■

C.5.2 Extensions to Assumption 3.3.2

In this Appendix, we detail two extensions to Assumption 3.3.2.

C.5.2.1 Discontinuities

Assumption 3.3.2 stipulates that

$$D_{i,j} = D(S_i, S_j) = \overline{D}(\|S_i - S_j\|) \tag{C.5.27}$$

for some function $\overline{D}(\cdot)$ that is continuous and monotone. Here, we outline how to extend this assumption to treat cases where the function $\overline{D}(\cdot)$ has discontinuities. Observe that, in our formal arguments, Assumption 3.3.2 is only applied once, in Corollary 3.4.2 to ensure that the counterfactual coordinate $S_j^{(i)}(\delta)$ is uniquely defined. Here, the restriction that $\overline{D}(\cdot)$ is continuous is only used to ensure that the level set $\mathcal{S}^{(i)}(\delta)$ is closed.

Thus, to handle the discontinuous case, let

$$\mathcal{S}^{(i)}(\delta, \varepsilon) = \{s : \|s - s'\| \leq \varepsilon, s' \in \mathsf{cl}(\mathcal{S}^{(i)}(\delta))\}. \tag{C.5.28}$$

denote the ε enlargement of the closure of the set

$$\mathcal{S}^{(i)}(\delta) = \{s : D_n(S_i, s) = \delta\} \ . \tag{C.5.29}$$

Define the counterfactual coordinate

$$\begin{aligned} &S_j^{(i)}(\delta, \varepsilon) = S_j + \alpha_j^{(i)}(\delta, \varepsilon)(S_i - S_j) \ , \quad \text{where} \\ &\alpha_j^{(i)}(\delta, \varepsilon) = \arg\min_{\alpha} \{|\alpha| : S_j + \alpha(S_i - S_j) \in \mathcal{S}^{(i)}(\delta, \varepsilon)\} \ . \end{aligned} \tag{C.5.30}$$

With this construction, we can define the potential outcome through the limit

$$Y_{i,j}(\delta, w) = \lim_{\varepsilon \downarrow 0} F(Z_i, Z_j(w, S_j^{(i)}(\delta, \varepsilon)), Z_{-i,j}) \ . \tag{C.5.31}$$

Observe that the construction agrees with the construction stated in the main text for the continuous case, so long as the function F is continuous in S_j.

For the sake of concreteness, consider the case that

$$D_{i,j} = \mathbb{I}\{\|S_i - S_j\| \leq \phi\} \tag{C.5.32}$$

for some constant radius ϕ. The level set $\mathcal{S}^{(i)}(1)$ is closed, and so it suffices to consider the construction of the counterfactual coordinate $S_j^{(i)}(0)$. In this case, the coordinate $S_j^{(i)}(\delta, \varepsilon)$ moves toward S_i on the line $S_j + \alpha(S_i - S_j)$ as α increases toward one, achieving

$\|S_i - S_j\| = \phi$ as $\varepsilon \to 0$.

C.5.2.2 Several Distances

In some applications, proximity measures of interest are functions of many features of the units under consideration. Commuting and migration probabilities, for instance, are often modeled as being functions of geographic distance, as well as wages and amenity values (Monte et al., 2018). We sketch three distinct approaches for handling settings with this structure.

Monotone Index: The first approach is based on the assumption that we can decompose the coordinate S_i into the components $S_i = (S_{i,0}, S_{i,1})$. For instance, suppose that $D_{i,j}$ denotes the probability of commuting from location i to location j. In this case, the component $S_{i,0}$ might denote the geographic coordinate of location i and $S_{i,1}$ might denote the average utility associated with working in location i.

We consider the following alternative to Assumption 3.3.2.

Assumption C.5.4 (Extended Index Structure)**.** *The coordinate $S_{i,1}$ is scalar-valued and the proximity measure $D_{i,j}$ admits the representation*

$$D_{i,j} = D(S_i, S_j) = \overline{D}(S_i, S_{j,0}, S_{j,1}) \tag{C.5.33}$$

where the function $\overline{D}(S_i, S_{j,0}, \cdot)$ is monotone.

That is, the proximity measure of interest $D_{i,j}$ is monotone, almost surely, in some component $S_{j,1}$ of the coordinate S_j. For instance, the probability of commuting from S_i to S_j might be decreasing in the utility associated with working in location j, conditional on location j's geographic location and the features of the location i. In this case, we can define

$$\begin{aligned} &\mathcal{S}_j^{(i)}(\delta) = \{s : D_n(S_i, (S_{j,1}, s)) = \delta\} \,, \qquad\qquad\qquad\qquad (\text{C.5.34}) \\ &S_{j,1}^{(i)}(\delta) = S_{j,1} + \alpha_j^{(i)}(\delta) \,, \quad \text{and} \quad \alpha_j^{(i)}(\delta) = \arg\min_{\alpha} \{|\alpha| : S_{j,0} + \alpha \in \mathcal{S}_j^{(i)}(\delta)\} \,, \end{aligned}$$

as before. Potential outcomes can then be defined by

$$Y_{i,j}(\delta, w) = F(Z_i, Z_j(w, S_{j,1}^{(i)}(\delta)), Z_{-i,j}) \,, \quad \text{where} \quad Z_j(w, s) = (w, S_{j,0}, s, U_j) \,. \tag{C.5.35}$$

it is clear that, so long as (C.5.34) is non-empty, the potential outcome (C.5.35) is uniquely defined.

In this setting, identification of averages of spillover proximity gradients is most credible in the case that the coordinates $S_{i,0}$ are observed. In this case, Assumption 3.4.4 should be replaced by the condition that

$$\mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, S_{i,0}, \overline{S}_j] = \mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid H_{i,j}, S_{i,0}, \overline{S}_j] \,. \tag{C.5.36}$$

That is, we can more transparently compare changes in spillovers induced by changes to the index $S_{1,i}$ by holding constant the features $S_{i,0}$ that might otherwise change the value of the proximity measure $D_{i,j}$.

Monotone Norm: The second approach is closely related and again based on the decomposition $S_i = (S_{i,0}, S_{i,1})$. Here, we consider the following extension to Assumption 3.3.2.

Assumption C.5.5 (Conditional Monotone Factor Structure)**.** *There exists a norm $\|\cdot\|$ such that*

$$D_{i,j} = D(S_i, S_j) = \overline{D}(\|S_{0,i} - S_{0,j}\|, S_{i,1}, S_{j,1}) \tag{C.5.37}$$

for some function $\overline{D}(\cdot)$ that is continuous and non-increasing in its first entry.

In other words, we assume that the proximity measures of interest still depend monotonically on a distance $\|S_{0,i} - S_{0,j}\|$. For instance, the probability of commuting from S_i to S_j might be decreasing in the commuting time between location i and location j, conditional on all other features of the two locations. In this case, we can define

$$\begin{aligned} \mathcal{S}^{(i)}(\delta) &= \{s : D_n(S_i, (s, S_{j,1})) = \delta\} \,, \\ S_{j,0}^{(i)}(\delta) &= S_{j,0} + \alpha_j^{(i)}(\delta)(S_{i,0} - S_{j,0}) \,, \quad \text{and} \\ \alpha_j^{(i)}(\delta) &= \arg\min_{\alpha} \{|\alpha| : S_{j,0} + \alpha(S_{i,0} - S_{j,0}) \in \mathcal{S}^{(i)}\} \,. \end{aligned} \tag{C.5.38}$$

and, thus, potential outcomes can be defined by

$$Y_{i,j}(\delta, w) = F(Z_i, Z_j(w, S_{j,0}^{(i)}(\delta)), Z_{-i,j}) \,, \quad \text{where} \quad Z_j(w, s) = (w, s, S_{j,1}U_j) \,. \tag{C.5.39}$$

By arguments analogous to those used in the proof of Corollary 3.4.2, the potential outcome (C.5.39) is uniquely defined whenever $\mathcal{S}^{(i)}(\delta)$ is non-empty.

As before, identification is most credible when the features $S_{i,1}$ are observed. In this case, Assumption 3.4.4 should be replaced by the condition that

$$\mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, H_{i,j}, \overline{S}_{i,1}, \overline{S}_j] = \mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid H_{i,j}, \overline{S}_{i,1}, \overline{S}_j] \,. \tag{C.5.40}$$

Here, we compare changes in spillovers induced by changes to the distance $\|S_{0,i} - S_{0,j}\|$ by holding constant the features $\overline{S}_{i,1}$ of the unit i that might otherwise change the value of the proximity measure $D_{i,j}$.

Invariance: The final approach is best suited for settings where the researcher is not willing to pose a concrete interpretation for the coordinate S_i. For instance, in Examples 5 and 6, although the coordinate S_i is probably best understood as collecting the geographic coordinates, wages, and amenity values of the location i, at this level of abstraction, this interpretation is contestable. This ambiguity can be resolved by imposing the following restriction.

Assumption C.5.6 (Stable Pairwise Proximity Spillovers)**.** *It holds that*

$$Y_{i,j}(\delta, w) = F(Z_i, Z_j(w, s), Z_{-i,j}), \quad \textit{for all} \quad s \in \mathcal{S}_j^{(i)}(\delta) \,. \tag{C.5.41}$$

This assumption is the analogue of the classical "Stable Unit Treatment Value Assumption" (SUTVA), due to Cox (1958) and Rubin (1980), adapted to our setting. In words, Assumption C.5.6 imposes the restriction that unit j's location S_j only modifies the influence of the treatment W_j on the outcome Y_i though $D_{i,j}$. That is, the "direction" of the spillover from j into i makes no difference, all that matters is the "distance" $D_{i,j}$.

Like SUTVA, Assumption C.5.6 places strong restrictions on the structure of interference. In particular, Assumption C.5.6 effectively rules out endogenous spillovers, in the sense of Manski (1993). That is, as the coordinate S_j varies over the level set $\mathcal{S}^{(i)}(\delta)$, unit j's proximity to units other than i changes. Assumption C.5.6 stipulates that these changes have no influence on the spillover effect of j on i. Of course, standard models of economic geography and trade feature endogenous spillovers (Allen and Arkolakis, 2014; Monte et al., 2018). On the other hand, approaches to estimating spillover effects based on well-posed exposure mappings rule out endogenous spillovers as well (Hudgens and Halloran, 2008; Manski, 2013; Li and Wager, 2022).

In sum, Assumption C.5.6 provides a clear, albeit strong, tool for reasoning about the potential outcomes $Y_{i,j}(\delta, w)$ in settings where the the coordinates S_i lack a concrete interpretation or the proximity measure of interest $D_{i,j}$ cannot be represented in a way that is amenable to the application of Assumption 3.3.2, Assumption C.5.4, or Assumption C.5.5. When applying Assumption C.5.6 it is important make it plain that endogenous spillover effects are assumed to be negligible.

C.5.3 Differences Between Residualized and Long Spillover Regressions

In this appendix, we compare regressions

$$Y_i = \alpha + \theta \cdot \Delta_i^\star + \varepsilon_i \quad \text{and} \tag{C.5.42}$$
$$Y_i = \alpha + \theta \cdot \Delta_i + \xi \cdot \Gamma_i + \varepsilon_i \ , \tag{C.5.43}$$

where

$$\Delta_i = \sum_{j \neq i} D_{i,j} W_j \ , \quad \Delta_i^\star = \sum_{j \neq i} (D_{i,j} - \mathbb{E}[D_{i,j} \mid G_{i,j}]) W_j \ , \quad \text{and} \quad \Gamma_i = \sum_{j \neq i} G_{i,j} W_j \ . \tag{C.5.44}$$

In particular, we outline the conditions under which the population coefficient on the covariate Δ_i in the "long" regression (C.5.43) is equal to the population coefficient on $\Delta_i^\star$ in the "short" regression (C.5.42).

To this end, consider the loss function

$$L_n^{\text{long}}(\theta, \xi) = \min_\alpha \sum_{i=1}^n (Y_i - \alpha - \theta \cdot \Delta_i - \xi \cdot \Gamma_i) \tag{C.5.45}$$

associated with the specification (C.5.42). By the Frisch-Waugh-Lovell Theorem, we can write

$$\mathbb{E}[L_n^{\text{long}}(\theta, \xi)] = \mathbb{E}\left[\sum_{i=1}^n \left((Y_i - \tilde{Y}_n) - \theta \cdot \tilde{\Delta}_i - \xi \cdot \tilde{\Delta}_i \right)^2 \right] \ , \tag{C.5.46}$$

where $\tilde{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ denotes the centered outcome,

$$\tilde{\Delta}_i = \sum_{j \neq i} \left(D_{i,j} - \frac{1}{n} \sum_{k \neq j} D_{k,j} \right) W_j - \frac{1}{n} \sum_{k \neq i} D_{k,i} W_i \ , \quad \text{and} \tag{C.5.47}$$
$$\tilde{\Gamma}_i = \sum_{j \neq i} \left(G_{i,j} - \frac{1}{n} \sum_{k \neq j} G_{k,j} \right) W_j - \frac{1}{n} \sum_{k \neq i} G_{k,i} W_i \ .$$

Thus, by the Projection Theorem and a second application of the Frisch-Waugh-Lovell Theorem, the parameter $\overline{\theta}_n^{\text{long}}$ that minimizes the expected loss (C.5.46) admits the representation

$$\overline{\theta}_n^{\text{long}} = \frac{\mathbb{E}[Y_i \tilde{\Delta}'_i]}{\mathbb{E}[(\tilde{\Delta}'_i)^2]} \quad \text{where} \quad \tilde{\Delta}'_i = \sum_{j \neq i} \left(D'_{i,j} - \frac{1}{n} \sum_{k \neq j} D'_{k,j} \right) W_j - \frac{1}{n} \sum_{k \neq i} D'_{k,i} W_i \ , \tag{C.5.48}$$

$$D'_{i,j} = D_{i,j} - \xi' \cdot G_{i,j} \ , \quad \text{and} \quad \xi' = \frac{\operatorname{Cov}(\tilde{\Delta}_i, \tilde{\Gamma}_i)}{\operatorname{Var}(\tilde{\Gamma}_i)} \ .$$

By comparing the representation (C.5.48) to the representation (C.1.34) stated in the Proof of Theorem 3.4.1, we find that the condition

$$\xi' \cdot G_{i,j} = \mathbb{E}[D_{i,j} \mid G_{i,j}] \tag{C.5.49}$$

is necessary and sufficient for the population coefficient on Δ_i in the regression (C.5.43) to coincide with the population coefficient on $\Delta_i^\star$ in the regression (C.5.42).

We conclude this appendix by showing that the condition (C.5.49) holds, up to a small error term, if and only if the conditional expectation $\mathbb{E}[D_{i,j} \mid G_{i,j}]$ is linear and the treatments W_j are homoskedastic, in the sense that their variances conditional on the variables $\overline{S}_j$ are constant. To see this, observe that

$$\begin{aligned} \operatorname{Cov}(\tilde{\Delta}_i, \tilde{\Gamma}_i) &= \sum_{j \neq i} \operatorname{Cov}\left(\left(D_{i,j} - \frac{1}{n} \sum_{k \neq j} D_{k,j} \right) W_j, \left(G_{i,j} - \frac{1}{n} \sum_{k \neq j} G_{k,j} \right) W_j \right) \\ &\quad + \frac{1}{n} \sum_{k \neq i} \sum_{q \neq i} \operatorname{Cov}(D_{k,i} W_i, G_{k,i} W_i) \\ &= (n-1) \mathbb{E}[\operatorname{Cov}(D_{i,j}, G_{i,j} \mid S_j) \operatorname{Var}(W_j \mid \overline{S}_j)] \ , \end{aligned} \tag{C.5.50}$$

where we have applied Assumptions 3.2.1, 3.4.1, and 3.4.2. Likewise, we can evaluate

$$\operatorname{Var}(\tilde{\Gamma}_i) = (n-2) \mathbb{E}[\operatorname{Var}(G_{i,j} \mid \overline{S}_j) \operatorname{Var}(W_j \mid \overline{S}_j)] + O(n^{-1}) \tag{C.5.51}$$

by Lemma C.1.1. Thus, we have that

$$\xi' \cdot G_{i,j} = \frac{\mathbb{E}[\operatorname{Cov}(D_{i,j}, G_{i,j} \mid S_j) \operatorname{Var}(W_j \mid \overline{S}_j)]}{\mathbb{E}[\operatorname{Var}(G_{i,j} \mid \overline{S}_j) \operatorname{Var}(W_j \mid \overline{S}_j)]} + O(n^{-2}) \tag{C.5.52}$$

by (C.5.50) and (C.5.51). Now, observe that Assumption 3.4.3 and the assumption that

$\mathbb{E}[D_{i,j} \mid G_{i,j}]$ is linear imply that we can write

$$\mathbb{E}[D_{i,j} \mid G_{i,j}] = \frac{\mathbb{E}[\mathrm{Cov}(D_{i,j}, G_{i,j} \mid S_j))]}{\mathbb{E}[\mathrm{Var}(G_{i,j} \mid \overline{S}_j))]} G_{i,j} \ . \tag{C.5.53}$$

Hence, so long as the quantities $\mathrm{Cov}(D_{i,j}, G_{i,j} \mid S_j))]$ and $\mathrm{Var}(G_{i,j} \mid \overline{S}_j))]$ are non-constant, the condition (C.5.49) holds if any only if $\mathrm{Var}(W_j \mid \overline{S}_j)$ is constant almost surely.

C.5.4 Alternative Approaches for Addressing Treatment Endogeneity

In this Appendix, we illustrate how to augment the framework proposed in the main text to accommodate settings where endogeneity in treatment assignment is addressed with instrumental variable or difference-in-differences research designs.

C.5.4.1 Instrumental Variables

First, we demonstrate how to adapt our estimation framework to treat settings based on instrumental variable research designs. For the sake of simplicity, we assume that we observe an instrumental variable M_i, that is also valued on $\{0, 1\}$ and is some deterministic function of the latent variable $\overline{S}_i$. For the sake of simplicity, we focus attention on the case without covariates X_i or auxiliary measures of proximity $G_{i,j}$, as the results stated here will generalize to that setting in the same way that Theorem 3.2.1 generalizes to Theorem 3.4.1.

We consider a distinct two-step procedure, analogous to the standard cross-sectional two-stage least-squares estimator. In the first stage, we compute an estimate $\hat{\pi}_n(m)$ of the conditional expectation

$$\pi_n(m) = \mathbb{E}[W_i \mid M_i] \ . \tag{C.5.54}$$

In the second stage, we consider the regression specification

$$Y_i = \alpha + \theta \cdot \hat{\Delta}_i^{\mathrm{IV}} + \varepsilon_i \ , \quad \text{where} \quad \hat{\Delta}_i^{\mathrm{IV}} = \sum_{j \neq i} D_{i,j} \hat{W}_j^{\mathrm{IV}} \quad \text{and} \quad \hat{W}_j^{\mathrm{IV}} = \hat{\pi}_n(M_j) - \frac{1}{n} \sum_{i=1}^{n} W_j \tag{C.5.55}$$

denotes the centered, predicted value of the treatment W_j conditional on the instrument M_j. We impose the condition

$$M_j \perp\!\!\!\perp S_j \ . \tag{C.5.56}$$

Observe that this condition can be interpreted as (or implied by) an instrumental variable exogeneity condition.

In this section, we sketch the derivation of a result analogous to Corollary 3.4.1 for this two-stage estimator. For the sake of space, we omit many of the formal details of this argument. Moreover, we leave considerations concerning consistency and inference to further work. We consider the infeasible loss function

$$L_n^{\mathrm{IV}}(\theta) = \min_{\alpha} \sum_{i=1}^{n} (Y_i - \alpha - \theta \cdot \Delta_i^{\mathrm{IV}})^2 \ , \tag{C.5.57}$$

where

$$\Delta_i^{\mathrm{IV}} = \sum_{j \neq i} D_{i,j} W_j^{\mathrm{IV}} \quad \text{and} \quad W_j^{\mathrm{IV}} = \pi_n(M_j) - \mathbb{E}[W_j] \ . \tag{C.5.58}$$

Under regularity conditions analogous to those imposed in Theorem 3.2.1, and by the same steps used in the proof of that result, the exogeneity condition (C.5.56) implies that the population minimizer $\overline{\theta}_n^{\mathrm{IV}}$ of the risk $\mathbb{E}[L_n^{\mathrm{IV}}(\theta)]$ can be expressed as

$$\begin{aligned}
\overline{\theta}_n^{\mathrm{IV}} &= \frac{\mathbb{E}[Y_i(D_{i,j} - \mathbb{E}[D_{i,j} \mid S_j])(\mathbb{E}[W_j \mid M_j] - \mathbb{E}[W_j])]}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid S_j)(\mathbb{E}[W_j \mid M_j] - \mathbb{E}[W_j])^2]} + o(n^{-1/2}) \\
&= \frac{\mathrm{Var}(M_j)}{\mathrm{Cov}(W_j, M_j)} \frac{\mathbb{E}[Y_i(D_{i,j} - \mathbb{E}[D_{i,j} \mid S_j])(M_j - \mathbb{E}[M_j])]}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid S_j)(M_j - \mathbb{E}[M_j])^2]} + o(n^{-1/2}) \\
&= \frac{\mathbb{E}[Y_i(D_{i,j} - \mathbb{E}[D_{i,j} \mid S_j])(M_j - \mathbb{E}[M_j])]}{\mathbb{E}[\mathrm{Var}(D_{i,j} \mid S_j) \, \mathrm{Cov}(W_j, M_j)]} + o(n^{-1/2}) \ ,
\end{aligned} \tag{C.5.59}$$

where we have used the fact that the instrument M_j is binary in the second equality and the exogeneity condition (C.5.56) again in the third equality. By the argument applied in the proof of Lemma C.1.2, we can express the parameter (C.5.59) as

$$\overline{\theta}_n^{\mathrm{IV}} = \int_{\overline{\mathcal{D}}} \mathbb{E}\left[\lambda(\delta \mid S_j) \left(\frac{(M_j - \mathbb{E}[M_j]) \partial_\delta \mathbb{E}[Y_i \mid D_{i,j} = \delta, S_j]}{\mathrm{Cov}(M_j, W_j)} \right)\right] \mathrm{d}\delta \tag{C.5.60}$$

$$\begin{aligned}
&= \int_{\overline{\mathcal{D}}} \mathbb{E}\Bigg[\lambda(\delta \mid S_j) \\
&\qquad \partial_\delta \left(\frac{\mathbb{E}[Y_i \mid D_{i,j} = \delta, M_j = 1, S_j] - \mathbb{E}[Y_i \mid D_{i,j} = \delta, M_j = 0, S_j]}{\mathbb{E}[W_j \mid M_j = 1] - \mathbb{E}[W_j \mid M_j = 0]} \right)\Bigg] \mathrm{d}\delta \ ,
\end{aligned} \tag{C.5.61}$$

where

$$\lambda(\delta \mid S_j) = \frac{\mathrm{Cov}(\mathbb{I}\{D_{i,j} \geq \delta\}, D_{i,j})}{\mathrm{Var}(D_{i,j} \mid S_j)} \tag{C.5.62}$$

and we have again used the fact that the instruments are binary to obtain the second equality.

Now, observe that the parameter

$$\frac{\mathbb{E}[Y_i \mid D_{i,j} = \delta, M_j = 1, S_j] - \mathbb{E}[Y_i \mid D_{i,j} = \delta, M_j = 0, S_j]}{\mathbb{E}[W_j \mid M_j = 1] - \mathbb{E}[W_j \mid M_j = 0]} \tag{C.5.63}$$

can be interpreted as the usual Wald instrumental variables estimand, associated with the treatment W_j, the instrument M_j, and the outcome

$$\mathbb{E}[Y_i \mid D_{i,j} = \delta, S_j] \ . \tag{C.5.64}$$

Thus, under standard exogeneity, exclusion, relevance, and monotonicity conditions, the parameter (C.5.63) can be expressed as a convex average of the spillover effect

$$\mathbb{E}[\partial_w Y_{i,j}(\delta, w) \mid D_{i,j} = \delta, S_j, W_j(1) > W_j(0)] \tag{C.5.65}$$

on the population of compilers (see, for instance, Imbens and Angrist (1994) and Angrist et al. (2000)).

C.5.4.2 Parallel Trends

Next, we sketch how adapt our framework to settings based on difference-in-differences designs. For the sake of simplicity, we assume that there are two time periods and that the treatment is binary. We assume that all units are untreated at time zero. Let $Y_{i,t}$ denote the outcome observed at period t and

$$Y_i^{(1)} = Y_{i,1} - Y_{i,0} \tag{C.5.66}$$

denote the first difference in the outcome for unit i.

Suppose that each unit i is associated with an additional error term $U_i^{(1)}$ and that the outcomes and first differences are generated by the structural equations

$$\begin{aligned} Y_{i,1} &= F_1(Z_i, Z_{-i}) \ , \quad \text{where} \quad Z_i = (W_i, S_i, U_i) \ , \quad \text{and} \\ Y_i^{(1)} &= F^{(1)}(Z_i^{(1)}, Z_{-i}^{(1)}) \ , \quad \text{where} \quad Z_i^{(1)} = (W_i, S_i, U_i^{(1)}) \ . \end{aligned} \tag{C.5.67}$$

This is a natural extension of Assumption 3.3.1. Here, the unobservable terms U_i and $U_i^{(1)}$ determine differences in period one values $Y_{i,1}$ and trends $Y_i^{(1)}$, respectively. In general, these terms can be distinct, but correlated. We define the potential trend and outcome

$$\begin{aligned} Y_{i,j}^{(1)}(\delta, w) &= F^{(1)}(Z_i, Z_j(w, S_j^{(i)}(\delta)), Z_{-i,j}) \quad \text{and} \\ Y_{i,j,1}(\delta, w) &= F_1(Z_i^{(1)}, Z_j^{(1)}(w, S_j^{(i)}(\delta)), Z_{-i,j}^{(1)}) \ , \end{aligned} \tag{C.5.68}$$

as before.

In this setting, a suitable parallel trends condition is given by

$$\mathbb{E}[W_j \mid \overline{S}_j^{(1)}] = \mathbb{E}[W_j], \quad \text{where} \quad \overline{S}_j^{(1)} = (S_j, U_j^{(1)}) \ . \tag{C.5.69}$$

In particular, the restriction (C.5.69) says that any of the factors S_j and $U_j^{(1)}$ that might contribute to larger or smaller values of the potential trend $Y_i^{(1)}(\delta, 0)$ are mean-independent of the treatments W_j. Observe that this condition is directly analogous to the condition (3.4.2) considered in the main text, where the outcomes are replaced by the first differences.

Indeed, under this assumption, by Theorem 3.4.1, the regression

$$Y_i^{(1)} = \alpha + \theta \cdot \Delta_i^\star + \varepsilon_i \quad \text{where} \quad \Delta^\star = \sum_{j \neq i} D_{i,j}(W_j - \mathbb{E}[W_j]) \tag{C.5.70}$$

identifies a convex average of the parameters

$$\begin{aligned} &\partial_\delta(\mathbb{E}[Y_{i,j}^{(1)}(\delta, 1) \mid D_{i,j} = \delta, W_j = 1, \overline{S}_j] - \mathbb{E}[Y_{i,j}^{(1)}(\delta, 0) \mid D_{i,j} = \delta, W_j = 0, \overline{S}_j]) \\ &= \partial_\delta(\mathbb{E}[Y_{i,j}^{(1)}(\delta, 1) \mid D_{i,j} = \delta, \overline{S}_j] - \mathbb{E}[Y_{i,j}^{(1)}(\delta, 0) \mid D_{i,j} = \delta, \overline{S}_j]) \\ &= \partial_\delta(\mathbb{E}[Y_{i,j,1}(\delta, 1) \mid D_{i,j} = \delta, \overline{S}_j] - \mathbb{E}[Y_{i,j,1}(\delta, 0) \mid D_{i,j} = \delta, \overline{S}_j]) \ , \end{aligned} \tag{C.5.71}$$

reproducing a version of Corollary 3.4.1. In sum, if the outcome in the class of regressions under consideration takes the form of a first difference, then the mean-independence condition (C.5.69) can be interpreted as a parallel trends assumption and the class of estimators proposed in the main text can be applied, unchanged.

C.5.5 Addressing Endogenous Measures of Proximity with Instrumental Variables

In this appendix, we illustrate how to adapt the framework proposed in the main text to accommodate settings where endogeneity in a measure of proximity is addressed with an instrumental variable. We assume that, for each pair of units i, j, we observe a proximity measure

$$M_{i,j} = M(\overline{S}_i, \overline{S}_j) \tag{C.5.72}$$

valued on $\{0, 1\}$. For the sake of simplicity, we focus attention on the case without covariates X_i or auxiliary measures of proximity $G_{i,j}$.

We consider a two-step procedure, analogous to the standard cross-sectional two-stage least-squares estimator. In the first stage, we compute an estimate $\hat{\gamma}_n^{\mathrm{IV}}(m)$ of the conditional expectation

$$\gamma_n^{\mathrm{IV}}(m) = \mathbb{E}[D_{i,j} \mid M_{i,j}] \,. \tag{C.5.73}$$

In the second stage, we consider the regression specification

$$Y_i = \alpha + \theta \cdot \hat{\Delta}_i^{\mathrm{IV}} + \varepsilon_i \,, \quad \text{where} \quad \hat{\Delta}_i^{\mathrm{IV}} = \sum_{j \neq i} \hat{D}_{i,j}^{\mathrm{IV}} W_j \quad \text{and} \quad \hat{D}_{i,j}^{\mathrm{IV}} = \hat{\gamma}_n^{\mathrm{IV}}(M_{i,j}) \tag{C.5.74}$$

denotes the predicted value of the proximity measure treatment $D_{i,j}$ conditional on the instrument $M_{i,j}$. For the sake of simplicity, we assume that the treatments satisfy the condition

$$\mathbb{E}[W_j \mid \overline{S}_j] = 0 \tag{C.5.75}$$

almost surely. That is, in effect, they arise from a randomized experiment. Further extensions to settings where the treatments satisfy Assumption 3.4.2, or the conditions considered in Appendix C.5.4, follow from the same arguments.

To justify the specification (C.5.74), we require the following analogue to Assumption 3.4.3.

Assumption C.5.7 (Additively Separable Heterogeneity)**.** *There exists a function $\psi_n^{IV}(\cdot)$ such that*

$$\mathbb{E}[D_{i,j} \mid M_{i,j}, S_j] = \gamma_n^{IV}(M_{i,j}) + \psi_n^{IV}(S_j) \tag{C.5.76}$$

almost surely.

Like Assumption 3.4.3, Assumption C.5.7 stipulates that any unobserved heterogeneity in the relationship between the proximity measure and the instrument is additively separable.

In this section, we sketch the derivation of a result analogous to Corollary 3.4.2 for the two-stage estimator (C.5.74). Like in Appendix C.5.4, we omit many of the formal details of this argument and leave considerations concerning consistency and inference to further work. We consider the infeasible loss function

$$L_n^{\mathrm{IV}}(\theta) = \min_{\alpha} \sum_{i=1}^{n} (Y_i - \alpha - \theta \cdot \Delta_i^{\mathrm{IV}})^2 \ , \tag{C.5.77}$$

where

$$\Delta_i^{\mathrm{IV}} = \sum_{j \neq i} D_{i,j}^{\mathrm{IV}} W_j \quad \text{and} \quad D_{i,j}^{\mathrm{IV}} = \pi_n(M_j) - \mathbb{E}[W_j] \ . \tag{C.5.78}$$

Observe that Assumption C.5.7 implies that

$$\mathbb{E}[D_{i,j} \mid M_{i,j}] - \mathbb{E}[\mathbb{E}[D_{i,j} \mid M_{i,j}] \mid S_j] = \mathbb{E}[D_{i,j} \mid M_{i,j}, S_j] - \mathbb{E}[D_{i,j} \mid S_j] \ . \tag{C.5.79}$$

Thus, under regularity conditions analogous to those imposed in Theorem 3.2.1, and by the same steps used in the proof of that result, the condition (C.5.75) implies that the population minimizer $\overline{\theta}_n^{\mathrm{IV}}$ of the risk $\mathbb{E}[L_n^{\mathrm{IV}}(\theta)]$ can be expressed as

$$\overline{\theta}_n^{\mathrm{IV}} = \frac{\mathbb{E}[Y_i(\mathbb{E}[D_{i,j} \mid M_{i,j}, \overline{S}_j] - \mathbb{E}[D_{i,j} \mid \overline{S}_j])W_j]}{\mathbb{E}[(\mathbb{E}[D_{i,j} \mid M_{i,j}, \overline{S}_j] - \mathbb{E}[D_{i,j} \mid \overline{S}_j])^2 \operatorname{Var}(W_j \mid \overline{S}_j)]} + o(n^{-1/2}) \ . \tag{C.5.80}$$

Observe that, because the instrument $M_{i,j}$ is binary, we can write

$$\begin{aligned} \mathbb{E}[D_{i,j} \mid M_{i,j}, \overline{S}_j] - \mathbb{E}[D_{i,j} \mid \overline{S}_j] &= \eta(\overline{S}_j)(M_{i,j} - \mathbb{E}[M_{i,j} \mid \overline{S}_j]) \ , \quad \text{where} \\ \eta(\overline{S}_j) &= \mathbb{E}[D_{i,j} \mid M_{i,j} = 1, \overline{S}_j] - \mathbb{E}[D_{i,j} \mid M_{i,j} = 1, \overline{S}_j] \ . \end{aligned} \tag{C.5.81}$$

Thus, we have that

$$\overline{\theta}_n^{\mathrm{IV}} = \frac{\mathbb{E}[\eta(\overline{S}_j) Y_i W_j (M_{i,j} - \mathbb{E}[M_{i,j} \mid \overline{S}_j])]}{\mathbb{E}[\eta(\overline{S}_j)^2 (M_{i,j} - \mathbb{E}[M_{i,j} \mid \overline{S}_j])^2 \operatorname{Var}(W_j \mid \overline{S}_j)]} + o(n^{-1/2}) \ . \tag{C.5.82}$$

Moreover, again by the fact that the instrument $M_{i,j}$ is binary, we can evaluate

$$\begin{aligned} \operatorname{Cov}(Y_i W_j, M_{i,j} \mid \overline{S}_j) &= \operatorname{Var}(M_{i,j} \mid \overline{S}_j)(\mathbb{E}[Y_i W_j \mid M_{i,j} = 1, \overline{S}_j] \\ &\qquad - \mathbb{E}[Y_i W_j \mid M_{i,j} = 0, \overline{S}_j]) \\ &= \operatorname{Var}(M_{i,j} \mid \overline{S}_j) \eta(\overline{S}_j) \beta(\overline{S}_j) \ , \end{aligned} \tag{C.5.83}$$

where

$$\beta(\overline{S}_j) = \frac{\mathbb{E}[Y_i W_j \mid M_{i,j} = 1, \overline{S}_j] - \mathbb{E}[Y_i W_j \mid M_{i,j} = 0, \overline{S}_j]}{\mathbb{E}[D_{i,j} \mid M_{i,j} = 1, \overline{S}_j] - \mathbb{E}[D_{i,j} \mid M_{i,j} = 0, \overline{S}_j]} \ . \tag{C.5.84}$$

Consequently, we obtain the representation

$$\overline{\theta}_n^{\mathrm{IV}} = \frac{\mathbb{E}[\eta^2(\overline{S}_j)\operatorname{Var}(M_{i,j} \mid \overline{S}_j)\beta(\overline{S}_j)]}{\mathbb{E}[\eta(\overline{S}_j)^2\operatorname{Var}(M_{i,j} \mid \overline{S}_j)\operatorname{Var}(W_j \mid \overline{S}_j)]} + o(n^{-1/2}) . \tag{C.5.85}$$

Observe that the parameter (C.5.84) can be interpreted as the usual Wald instrumental variables estimand, associated with the treatment $D_{i,j}$, the instrument $M_{i,j}$, and the outcome $Y_i W_j$. Thus, under standard exogeneity, exclusion, relevance, and monotonicity conditions (Imbens and Angrist, 1994), we can write

$$\beta(\overline{S}_j) = \int_{\overline{D}} \lambda^{\mathrm{IV}}(\delta \mid \overline{S}_j)\mathbb{E}[\partial_\delta Y_{i,j}(\delta, W_j)W_j \mid \overline{S}_j, D_{i,j}(1) > D_{i,j}(0)]\,\mathrm{d}\delta , \tag{C.5.86}$$

where the weights $\lambda^{\mathrm{IV}}(\delta \mid \overline{S}_j)$ are convex (see Angrist et al. (2000) for details). Hence, by plugging (C.5.86) into (C.5.85) and applying an argument very similar to the proof of Lemma C.1.2, we can conclude that

$$\overline{\theta}_n^{\mathrm{IV}} = \int_{\overline{D}}\int_{\overline{W}} \lambda^{\mathrm{IV}}(\delta, w \mid \overline{S}_j)\mathbb{E}[\partial^2_{\delta,w} Y_{i,j}(\delta, W_j) \mid \overline{S}_j, D_{i,j}(1) > D_{i,j}(0)]\,\mathrm{d}\delta\,\mathrm{d}w + o(n^{-1/2}) . \tag{C.5.87}$$

where the weights $\lambda^{\mathrm{IV}}(\delta, w \mid \overline{S}_j)$ are convex.

C.5.6 Additional Discussion Concerning Assumption 3.5.4

In this Appendix, we give addition details concerning Assumption 3.5.4. Recall that Assumption 3.5.4 stipulates that the bounds

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k}F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}]] = O(n^{-2}\rho_n^{-3}) , && \text{(C.5.88)}\\
&\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k}F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 1, A_{i,k} = 0, Z^{(j,k)}]] = O(n^{-2}\rho_n^{-1}) , \quad \text{and} &&\\
&\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k}F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 0, A_{i,k} = 0, Z^{(j,k)}]] = O(n^{-2}\rho_n) &&
\end{aligned}$$

hold, where

$$\nabla_j f(Z) = f(Z) - f(Z^{(j)}) \quad \text{and} \quad \nabla^2_{j,k} f(Z) = \nabla_j \nabla_k f(Z) , \tag{C.5.89}$$

denote difference operators, Z' denotes an independent copy of Z, $Z^{(j)}$ is constructed by replacing Z_j with Z'_j in Z, and $Z^{(j,k)}$ is constructed analogously, by replacing Z_j and Z_k

with Z'_j and Z'_k in Z.

To unpack the restrictions implied by these conditions, observe that the quantity

$$P\{|\nabla^2_{j,k}F(Z_i, Z_{-i})| \geq C \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}\} \tag{C.5.90}$$

can be interpreted as the measuring the proportion of pairs of units j, k, that are proximate to a unit i, for which the curvature

$$\nabla^2_{j,k}F(Z_i, Z_{-i}) \ , \tag{C.5.91}$$

i.e., the change in dependence of i on j, induced by changes in k, is larger than a constant C. The bound (C.5.88) and Chebyshev's inequality imply that

$$\begin{aligned} &\mathbb{E}\left[P\{|\nabla^2_{j,k}F(Z_i, Z_{-i})| \geq C \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}\}\right] \\ &\lesssim \mathbb{E}[\mathbb{E}[(\nabla^2_{j,k}F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}]] = O(n^{-2}\rho_n^{-3}) \end{aligned} \tag{C.5.92}$$

In other words, if $\rho_n \lesssim n^{-2/3}$, then this proportion can be on the order of a constant for all pairs of units proximate to each unit i. For larger values of ρ_n, the proportion decreases. In particular, in general, each unit i will have $n^2\rho_n^2$ pairs of neighbors. The bound (C.5.92) says that ρ_n^{-1} of these pairs can be associated with curvature on the order of a constant.

We conclude showing that Assumption 3.5.4 holds in settings where

$$Y_i = F_i\left(W_i, \frac{1}{n\rho_n}\sum_{j\neq i} D_{i,j}W_j\right) \ , \tag{C.5.93}$$

where F_i is some unit-specific function. That is, each unit's outcome depends on its own treatment as well as the proportion of its "neighbors" that are treated. We have used the denominator $n\rho_n$, rather than $\sum_{j=1}^n D_{i,j}$, for the sake of simplicity. We assume that $F_i(\cdot)$ is twice continuously differentiable, with second derivative uniformly bounded by the sequence B_n. A closely related model is considered in Li and Wager (2022).

In particular, define the differences $R_j = D_{i,j}W_j - D'_{i,j}W'_j$, where $D'_{i,j}$ and W'_j are constructed with Z'_j. We have that

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[(\nabla^2_{j,k}F(Z_i, Z_{-i}))^2 \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}]] \\ &\leq n^{-4}\rho_n^{-4}B_n^2\mathbb{E}[\mathbb{E}[(R_jR_k)^2 \mid A_{i,j} = 1, A_{i,k} = 1, Z^{(j,k)}]] \end{aligned} \tag{C.5.94}$$

by two applications of the mean-value theorem. Thus, the bound (C.5.88) holds so long as the curvature bound

$$B_n \lesssim n\rho_n^{1/2} \tag{C.5.95}$$

holds. The second two bounds in (C.5.88) hold by construction.

C.6. Data and Additional Applications

In this Appendix, we detail our treatment of the data associated with Example 4 and give additional applications to data from Examples 5 and 6.

C.6.1 Bloom et al. (2013)

First, we give details of our treatment of the data associated with Example 4. Results are given in Section 3.7. We use data obtained from the replication package associated with Lucking et al. (2019), who increase the coverage of the data considered in Bloom et al. (2013).[2]

We follow the steps used in Bloom et al. (2013) and Lucking et al. (2019) to construct the inputs used in our analysis. There are two, main raw data components. The first components are the Compustat Fundamentals Annual and Compustat Segments datasets. From these sources, we assemble a cleaned panel of measurements of R&D expenditure, sales, market value, and product market composition for a sample of U.S. public companies. All measures of sales, value, and R&D expenditure are deflated to 1996 U.S. dollars. All sample restrictions and data cleaning steps follow Lucking et al. (2019). For each firm, we aggregate these measures over five year periods, spanning 1986 to 2005. We use the observations of the shares of sales across four digit manufacturing industries to construct our measure of product market proximity for each five year time period.

The second component is data from the NBER patent data project (Hall et al., 2001). With these data, we measure each firm's patenting activity, across USPTO technology classes, and total patent citations. Again, all sample restrictions and data cleaning steps follow Lucking et al. (2019). We aggregate these measures to the same five-year periods. We use the observations of the shares of patents across USPTO technology classes to construct our measure of technological proximity for each five year time period. Additionally, following

[2]These data are available on Nicholas Bloom's website, at the link https://www.dropbox.com/scl/fi/fmtqtpuv41l6c0p30ybop/spillovers_rep.zip?rlkey=pu0bsc3scesewb3yt9y7juh9m&dl=0.

Bloom et al. (2013), use these data to construct a measure of geographic proximity, in an analogous way. In particular, we compute the share of each firm's patents filed by inventors across a set of locations. The geographic proximity measure is then given by the uncentered correlation between these shares.

For the estimates reported in Table 3.1, we restrict attention to the same subset of firms considered in Lucking et al. (2019). Moreover, for each outcome, we omit firms associated with observations of the treatment variable and outcome variable are equal to zero, as both measures enter in logs. For the resultant samples, all missing proximity measures are imputed to zero.

C.6.2 Hornbeck and Moretti (2024)

Next, we detail an additional application, associated with Example 5. Recall that Hornbeck and Moretti (2024) are interested in characterizing how the spillover effects of shocks to one cities' total factor productivity (TFP) on another cities' labor market outcomes depends on the propensity to migrate between the two cities. Hornbeck and Moretti (2024) use a simple structural model to address this objective. Here, we outline how to apply the methodology developed in this paper to this problem.

C.6.2.1 Data

We use two sources of data. First, we use data obtained from the replication package associated with Hornbeck and Moretti (2024).[3] We use four types of data from this source as inputs into our analysis. First, we obtain measurements of the number of workers in each United States Metropolitan Statistical Area (MSA) that are employed in each two digit industry in 1980. These data were originally obtained by Hornbeck and Moretti (2024) though an IPUMS census extract. Second, we use measurements of the pairwise distances between the centroids of each MSA. Third, we obtain measurements of the bilateral migration shares between each pair of MSAs for 1975 to 1980. That is, this is this the proportion, of individuals who lived in MSA i in 1975 and migrated to another MSA, that live in MSA j in 1980. Fourth, we use a set of MSA level variables, associated with the years 1980 and 1990. These variables are the TFP, total employment, average earnings, and average home value associated with each MSA.

[3]These data were obtained from Richard Hornbeck's website, at the link https://www.dropbox.com/scl/fi/66qqjxtlzg3xmgwxjx5mi/TFP_HM_May2022_ReplicationFiles.rar?rlkey=yf4dje4zhpw3erqb3es8pp99z&dl=0.

Second, we use data the 1997 Bureau of Economic Analysis Historical Benchmark Input-Output Table.[4] These data report input output tables between two-digit industries in the United States.

C.6.2.2 Input-Output Flows

To the best of our knowledge, no input-output tables across MSAs are publicly available for the relevant time period. Thus, we impute shares with a simple model. Our approach is based on first constructing industry-specific gravity matrices, whose bilateral terms combine exponential distance frictions with the destination-specific intermediate-input demand implied by an input-output table. For each origin–industry pair, we row-normalize the bilateral terms to obtain destination-choice shares (the fraction of shipments from origin i in industry j going to destination i'). We then average across industries using the origin's employment mix to form the origin–destination share matrix. This, highly stylized, approach is closely related to standard models of structural gravity and market access (Redding and Venables, 2004; Head and Mayer, 2011).

Formally, let $Q_{i,j}$ denote the number of workers in MSA i that are employed in industry j. Collect these data into the matrix Q. Let $I_{j,j'}$ denote the share of industry j' inputs that are produced by industry j. Collect these data into the matrix I. Let $\mathsf{Dist}_{i,i'}$ denote the distance between MSAs i and i' in miles. We use the Distance kernel

$$\tau_{i,i'} = \exp(-\lambda \mathsf{Dist}_{i,i'}) \;, \tag{C.6.1}$$

where we chose the parameter λ so that the 90th percentile of $\tau_{i,i'}$ over pairs of MSAs $i \neq i'$ is equal to 0.01. We estimate intermediate demand by

$$\mathsf{ID} = QI^\top \;. \tag{C.6.2}$$

That is, the i, j component of ID denotes destination j's demand for inputs from industry g. Industry-specific destination choice shares are then constructed by

$$\pi^{(j)}_{i,i} = \frac{\tau_{i,i'} \mathsf{ID}_{i',j}}{\sum_k \tau_{i,k} \mathsf{ID}_{k,j}} \;. \tag{C.6.3}$$

[4] These data were obtained from the BEA's webpage, at the link `https://www.bea.gov/industry/historical-benchmark-input-output-tables`, on May 12th, 2025.

We then aggregate across industries by employment mix through

$$\tilde{\Pi}_{i,i'} = \sum_j S_{i,j} \pi_{i,i}^{(j)} \quad \text{where} \quad S_{i,j} = Q_{i,j} / \sum_k Q_{i,k} \,. \tag{C.6.4}$$

Our final estimate $\Pi_{i,i'}$ of the proportion of intermediates purchased by MSA i that are produced in MSA i' is obtained by row-normalizing the shares (C.6.4).

Figure C.1 displays the joint distributions of the trade flow, bilateral distance, and migration probability between each pair of distinct MSAs. By construction, pairs of MSAs that are geographically close are closely connected by trade. However, for pairs of MSAs that are relatively close, there is a considerable amount of variation in trade flows, driven by heterogeneity in intermediate demand and employment. Likewise, pairs of MSAs that are closely connected by trade have large migration flows and the migration flows for pairs of MSAs that have negligible trade are also negligible.

C.6.2.3 Analysis and Results

In this application, the treatment W_i denotes the TFP growth in MSA i between 1980 and 1990. The proximity measure of interest $D_{i,j}$ is the share of people who live in MSA j in 1975 and live in a different MSA in 1980 that live in MSA j in 1980. That is, $D_{i,j}$ measures the probability that a migrant from MSA i relocates to MSA j. We consider three choices of outcomes Y_i: the differences between the total employment, averages wages, and average home values between 1980 and 1990. We use the logarithm of each outcome variable. Results are displayed in Table C.1.

The first row displays the realized value of the estimator $\hat{\theta}_n$ associated with each outcome, where neither the treatment nor proximity measures have been residualized. To address endogeneity in the treatments, in the second row, we residualize the treatments with the values of each MSA's total factor productivity, as well as each of the outcome variables, in 1980.[5] The sign of the coefficient changes for each outcome.

Now, our interest is in characterizing how changes to the propensity to migrate between two MSAs change the spillover effects of TFP growth. The hypothesis is that, when productivity growth increases labor demand in MSA j, migration from MSA i is encouraged, reducing labor supply and increasing wages. Migration probabilities $D_{i,j}$ depend on many features of both the MSAs i and j. As a consequence, Assumption 3.3.2 is not immediately

[5] Hornbeck and Moretti (2024) consider "shift-share" instruments, based on exposures to industry level changes to productivity, stock market returns, export markets, and patenting activity. We do not consider these instruments because they are spatially correlated. That is an analogue of Assumption 3.4.2 would be violated.

Table C.1: Application to Productivity Spillover Estimation

		Employment	Wages	Home Value
Unadjusted		0.618	0.392	0.229
		(1.556)	(0.880)	(3.219)
Resid. Treatment		-0.619	-0.244	-0.336
		(1.186)	(0.878)	(3.087)
Resid. Proximity	Distance	-1.456	-0.325	-0.151
		(1.205)	(0.787)	(3.214)
	Distance + IO	-1.507	-0.311	-0.078
		(1.204)	(0.807)	(3.292)
Number of MSAs		193	193	193

Notes: Table C.1 displays values of the estimator $\hat{\theta}_n^\star$ for the application to data from Hornbeck and Moretti (2024), using the choices for the outcome variable. The standard error $\hat{\sigma}_n$, defined in Algorithm 2, is displayed in parentheses below each estimate. Further details are given in Appendix C.6.2. In each case, the treatment variable W_i denotes the total factor productivity growth for MSA i between 1980 and 1990 and the proximity measures $D_{i,j}$ are probability that a migrant from MSA i relocates to MSA j between 1975 and 1980. The outcomes are differences between 1980 and 1990, and enter in logs. In the first row, neither the treatments nor proximity measures have been residualized. In the second row, the treatments are residualized using the values of total factor productivity and the three outcomes in 1980. In the remaining rows, the proximity measure has been residualized using various measures of proximity, in addition to the covariates used to residualize the treatments. The estimators $\hat{\pi}_n(\cdot)$ and $\hat{\gamma}_n(\cdot)$ are both obtained from linear regressions.

applicable. Instead, we apply Assumption C.5.4, stated in Appendix C.5.2. In particular, we assume that probability of migrating from i to j is a function of each MSA's geographic coordinates $S_{i,0}$ and $S_{j,0}$, as well as the utilities $S_{i,1}$ and $S_{j,1}$ associated with living and working in the respective MSAs. To apply Assumption C.5.4, we assume that the probability of migrating from i to j is monotonically decreasing in the utility $S_{j,1}$ associated with living in MSA j. Thus, the potential outcome $Y_{i,j}(\delta, w)$ denotes the value of MSA i's outcome associated with the intervention that sets TFP growth in MSA j to w and changes the probability of migrating from i to j to δ by changing the utility $S_{j,1}$.

To operationalize this assumption, it is essential to control for differences in other features of the unit i that might otherwise affect the value of the migration probability $D_{i,j}$. See Appendix C.5.2 for further discussion. Accordingly, the third row of Table C.1 displays revised estimates, where the migration probabilities have been residualized with flexible functions of the distance between MSAs.[6] This change decreases the coefficients associated with employment and wages and increases the coefficient associated with home values.

Moreover, migration probabilities are likely to be associated with other channels that

[6]In particular, we residualize the migration probabilities with indicators that the distance between the two MSAs is less than the 0.5th, 1st, 2.5th, and 5th percentiles of the distribution of the distance between MSAs.

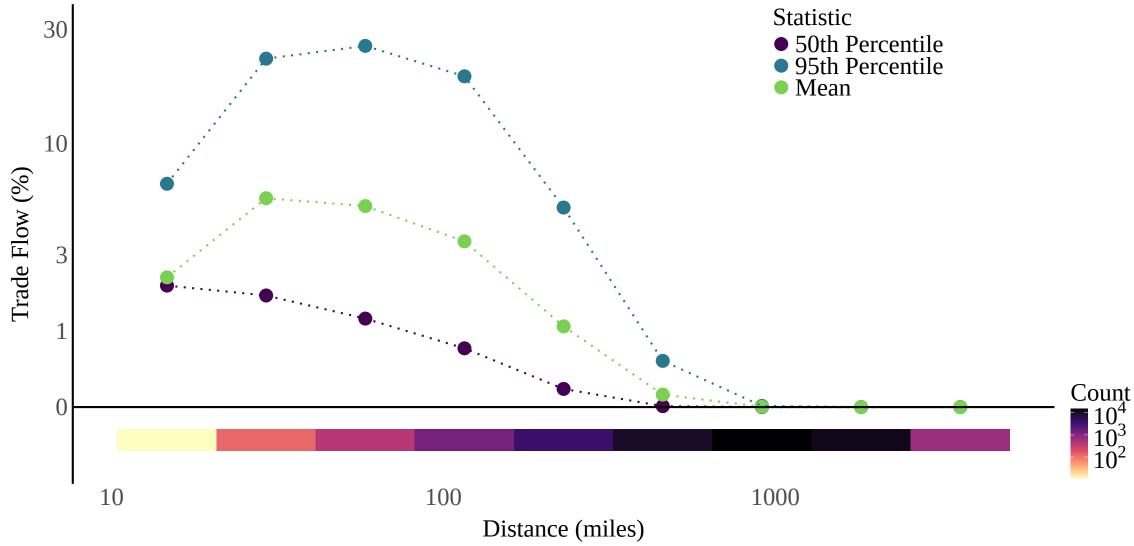
otherwise mediate spillover effects. For instance, MSAs that are closely connected by migration are likely to be closely connected by trade, and TFP growth in MSA i might increase demand for intermediate inputs produced by city i, increasing labor demand and wages. To mitigate this source of confounding, in the fourth row of Table C.1 we additionally control for flexible functions of the origin-destination trade flows constructed in the previous section.[7] The changes in the estimates are more modest.

As would be expected, we find a negative association between migration probabilities and the spillover effect of TFP changes on employment. However, in contrast to estimates obtained in Hornbeck and Moretti (2024), we find negative associations with wages and home values as well. We caution that all estimates are statistically insignificant at conventional levels. Greater power might be attained if analogous exercises could be implemented at lower levels of geographic aggregation, such as counties or commuting zones.

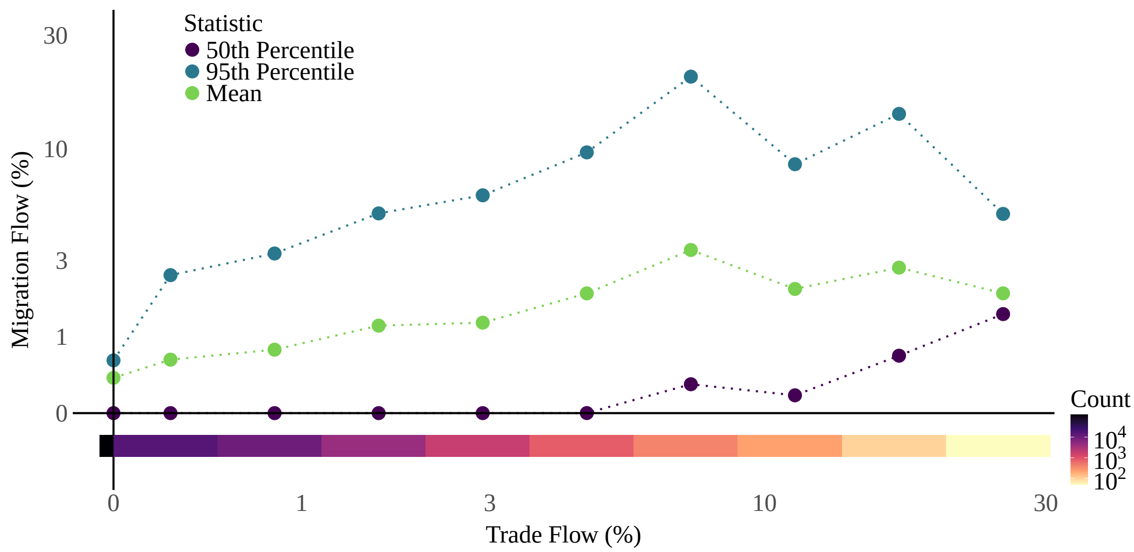
[7]We additionally residualize the migration probabilities with indicators that share of goods that used in MSA i are produced in MSA j greater than the 99.5th, 99th, 97.5th, and 95th percentiles of the distribution of this quantity.

Figure C.1: Trade, Migration, and Distance

Panel A: Trade and Distance



Panel B: Trade and Migration



Notes: Figure C.1 displays features of the joint distributions of the trade flow, bilateral distance, and migration probability between each pair of distinct MSAs. The measures are computed used data obtained from the replication package associated with Hornbeck and Moretti (2024). The x-axes measure distance between MSAs in miles and the proportion of intermediates purchased by MSA i that are produced in MSA i', respectively, and have been partitioned into bins. The y-axes display the proportion of intermediates purchased by MSA i that are produced in MSA i' and the proportion of workers from MSA i who migrate to MSA i', respectively. Details on the construction of these measures are given in Appendix C.6.2. The means, 50th percentiles, and 90th percentiles of each statistic within each bin are displayed using green, purple, and blue dots and are measured relative to the y-axis. A heatmap measuring the number of pairs of units in each bin is displayed below both the x-axis.

C.6.3 Franklin et al. (2024)

Finally, we detail our application associated with Example 6, using data from Franklin et al. (2024). Franklin et al. (2024) develop a model of urban spatial equilibrium in which the labor market response to a public works program depends on commuting flows. They use this model to argue that the the indirect effects of the program are larger for pairs of neighborhoods that are closely connected by commuting. In this section, we detail how to apply the methodology developed in this paper to this problem.

C.6.3.1 Data

We use data obtained from the replication package associated with Franklin et al. (2024).[8] Franklin et al. (2024) consider a public works program implemented in Addis Ababa, Ethiopia. The program provided guaranteed public work to targeted households and was randomized across neighborhoods (referred to as "Woredas"). The treatment W_j indicates that neighborhood j has been assigned to the program. The outcome Y_i denotes the log average earnings in neighborhood i. The proximity measure of interest $D_{i,j}$ is the pretreatment probability that a worker who works in neighborhood i lives in neighborhood j. Figure C.2 displays the joint distribution between these commuting probabilities and the walking distance between each pair of neighborhoods.

C.6.3.2 Results

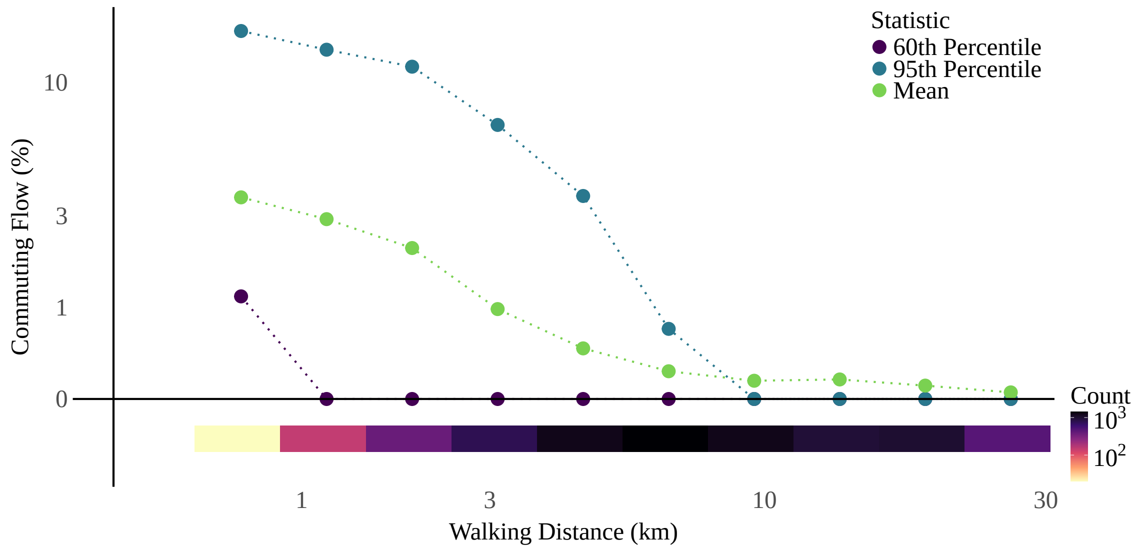
The first row of the first column of Table C.2 displays the coefficient $\hat{\theta}_n$ obtained from the regression specification

$$Y_i = \alpha + \theta \cdot \Delta_i + \varepsilon_i \ , \quad \text{where} \quad \Delta_i = \sum_{j \neq i} D_{i,j}(W_j - \mathbb{E}[W_j]) \ . \tag{C.6.5}$$

The second row of this column gives an analogous estimate, where the treatments have been residualized on indicators for a higher level of geographic aggregation (referred to as a "sub-city"). A very similar specification is reported in Franklin et al. (2024). The main difference is that they include each neighborhood's own treatment W_i in the covariate Δ_i with the weight $D_{i,i} = 1$. In Franklin et al. (2024), this specification approximates the reduced form of a model of spatial equilibrium.

[8]This replication package is available from ICPSR at the link `https://www.openicpsr.org/openicpsr/project/195483/version/V1/view`.

Figure C.2: Commuting and Distance



Notes: Figure C.2 displays features of the joint distributions of the commuting flows and bilateral distance for the sample of woredas considered in Franklin et al. (2024). The x-axis measures the walking distance between each pair of woredas, and has been partitioned into bins. The y-axis displays the probability that a worker who lives in woreda i commutes to woreda j. The means, 60th percentiles, and 90th percentiles of each statistic within each bin are displayed using green, purple, and blue dots and are measured relative to the y-axis. A heatmap measuring the number of pairs of units in each bin is displayed below both the x-axis.

Table C.2: Application to Public Works Spillover Estimation

		Commuting	Dist. $<$ 2 km
Unadjusted		0.396	0.011
		(0.439)	(0.062)
Resid. Treatment		0.542	0.018
		(0.539)	(0.057)
Resid. Proximity	Distance	0.510	
		(0.481)	
	Commuting		0.002
			(0.056)
Number of Woredas		90	90

Notes: Table C.2 displays values of the estimator $\hat{\theta}_n^\star$ for the application to data from Franklin et al. (2024). The standard error $\hat{\sigma}_n$, defined in Algorithm 2, is displayed in parentheses below each estimate. Further details are given in Appendix C.6.3. In each case, the treatment variable W_i indicates assignment to treatment and the outcome Y_i measures the log average wage. We consider two proximity measures of interest: the probability of commuting from Woreda j to Woreda i and an indicator that the walking distance between the two Woredas is less than two kilometers. In the first row, neither the treatments nor proximity measures have been residualized. In the second row, the treatments are residualized using sub-city indicators. In the remaining rows, the proximity measure of interest has been residualized using various additional measures of proximity, in addition to the covariates used to residualize the treatments. The estimators $\hat{\pi}_n(\cdot)$ and $\hat{\gamma}_n(\cdot)$ are both obtained from linear regressions.

We have shown that such specifications admit a nonparametric interpretation, as targeting the effect of the proximity measure $D_{i,j}$ on the intensity of the spillover effect of W_j on Y_i. The hypothesis is that individuals who participate in the program no longer commute to other neighborhoods, reducing their private sector labor supply and increasing wages.

Observe that, like the application considered in Appendix C.6.2, commuting probabilities will not satisfy Assumption 3.3.2. Instead, we again apply Assumption C.5.4. That is, we assume that commuting probabilities are functions of neighborhoods' locations and expected wages, and that the probability of commuting from j to i is decreasing in the expected wage in neighborhood j. Thus, the potential outcome $Y_{i,j}(\delta, w)$ denotes the value of neighborhood i's post-treatment wages associated with the intervention that sets j's treatment status to w and changes the probability of commuting from j to i by changing the pre-treatment wage in neighborhood j. To operationalize this assumption, we must control for other features of the unit i that might otherwise affect the commuting probability $D_{i,j}$. Thus, the third row of the first column displays estimates where the commuting probabilities have been residualized using flexible functions of the walking distance between neighborhoods. All three estimates are insignificant at conventional levels.

The second column of Table C.2 displays analogous results, where the proximity measure of interest is an indicator that the walking distance is smaller than two kilometers. Here, as well, all results are insignificant at conventional levels.

Bibliography

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., and Venkatesh, A. (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, 106(12):3730–3764.

Abou-Moustafa, K. and Szepesvári, C. (2019). An exponential efron-stein inequality for ℓ_q stable learning rules. In *Algorithmic Learning Theory*, pages 31–63. PMLR.

Acemoglu, D., Akcigit, U., and Kerr, W. (2016a). Networks and the macroeconomy: An empirical exploration. *Nber macroeconomics annual*, 30(1):273–335.

Acemoglu, D., Autor, D., Dorn, D., Hanson, G. H., and Price, B. (2016b). Import competition and the great us employment sag of the 2000s. *Journal of Labor Economics*, 34(S1):S141–S198.

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13(none):1000–1034.

Adao, R., Arkolakis, C., and Esposito, F. (2025). General equilibrium effects in space: Theory and measurement. Technical report, National Bureau of Economic Research.

Adao, R., Kolesár, M., and Morales, E. (2019). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics*, 134(4):1949–2010.

Adermon, A., Lindahl, M., and Palme, M. (2021). Dynastic human capital, inequality, and intergenerational mobility. *American Economic Review*, 111(5):1523–1548.

Adler, R. J. and Taylor, J. E. (2009). *Random fields and geometry*. Springer.

Agarwal, N., Ashlagi, I., Azevedo, E., Featherstone, C. R., and Karaduman, Ö. (2019). Market failure in kidney exchange. *American Economic Review*, 109(11):4026–4070.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Ahrens, A., Hansen, C. B., Schaffer, M. E., and Wiemann, T. (2024). ddml: Double/debiased machine learning in stata. *The Stata Journal*, 24(1):3–45.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Allcott, H. and Keniston, D. (2018). Dutch disease or agglomeration? the local economic effects of natural resource booms in modern america. *The Review of Economic Studies*, 85(2):695–731.

Allen, T. and Arkolakis, C. (2014). Trade and the topography of the spatial economy. *The Quarterly Journal of Economics*, 129(3):1085–1140.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249.

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527.

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, pages 1277–1366. Elsevier.

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 600–607. Cambridge University Press.

Arcones, M. A. (1995). A Bernstein-type inequality for U-statistics and U-processes. *Statistics & Probability Letters*, 22(3):239–247.

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *The Annals of Probability*, pages 1494–1542.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40 – 79.

Arnold, D., Dobbie, W., and Hull, P. (2020). Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research.

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*, pages 1912–1947.

Arque-Castells, P. and Spulber, D. F. (2022). Measuring the private and social returns to r&d: Unintended spillovers versus technology markets. *Journal of Political Economy*, 130(7):1860–1918.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In *Readings in industrial economics: Volume two: Private enterprise and state intervention*, pages 219–236. Springer.

Athey, S., Chetty, R., Imbens, G., and Kang, H. (2020). Estimating treatment effects using multiple surrogates: the role of the surrogate score and the surrogate index. *arXiv:1603.09326*.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *Journal of Econometrics*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica*, 90(1):347–365.

Auerbach, E. and Tabord-Meehan, M. (2025). The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*.

Austern, M. and Zhou, W. (2020). Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*.

Autor, D. H., Dorn, D., and Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American economic review*, 103(6):2121–2168.

Bach, P., Kurz, M. S., Chernozhukov, V., Spindler, M., and Klaassen, S. (2024). DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56.

Balboni, C., Bandiera, O., Burgess, R., Ghatak, M., and Heil, A. (2022). Why do people stay poor? *The Quarterly Journal of Economics*, 137(2):785–844.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236):1260799.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of statistics*, pages 2055–2085.

Bates, S., Hastie, T., and Tibshirani, R. (2023). Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, pages 1–12.

Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350.

Beaman, L., Karlan, D., Thuysbaert, B., and Udry, C. (2023a). Data and code for: Selection into credit markets: Evidence from agriculture in mali. Econometric Society replication files for Econometrica.

Beaman, L., Karlan, D., Thuysbaert, B., and Udry, C. (2023b). Selection into credit markets: Evidence from agriculture in mali. *Econometrica*, 91(5):1595–1627.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. *The Annals of Statistics*, 46(6B):3643.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Beraja, M., Kao, A., Yang, D. Y., and Yuchtman, N. (2023). Ai-tocracy. *The Quarterly Journal of Economics*, 138(3):1349–1402.

Beran, R. and Millar, P. W. (1987). Stochastic estimation and testing. *The Annals of Statistics*, 15(3):1131–1154.

Bergeaud, A., Guillouzouic, A., Henry, E., and Malgouyres, C. (2025). From public labs to private firms: Magnitude and channels of local r&d spillovers. *The Quarterly Journal of Economics*.

Bernstein, S., Colonnelli, E., Giroud, X., and Iverson, B. (2019). Bankruptcy spillovers. *Journal of Financial Economics*, 133(3):608–633.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, pages 647–671.

Blattman, C., Duncan, G., Lessing, B., and Tobón, S. (2024). Gang rule: Understanding and countering criminal governance. *Review of Economic Studies*, page rdae079.

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Bobkov, S. G., Götze, F., and Sambale, H. (2019). Higher order concentration of measure. *Communications in Contemporary Mathematics*, 21(03):1850043.

Borusyak, K., Dix-Carneiro, R., and Kovak, B. (2022a). Understanding migration responses to local shocks. *Available at SSRN 4086847*.

Borusyak, K. and Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6):2155–2185.

Borusyak, K., Hull, P., and Jaravel, X. (2022b). Quasi-experimental shift-share research designs. *The Review of economic studies*, 89(1):181–213.

Boucheron, S., Bousquet, O., Lugosi, G., and Massart, P. (2005). Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560.

Boucheron, S., Lugosi, G., and Massart, P. (2003). Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614.

Boucheron, S., Lugosi, G., and Massart, P. (2013). Concentration inequalities: a non asymptotic theory of independence.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and regression trees*. Routledge.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., and Pan, M. (2020). Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–2484.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *The Annals of Probability*, 1(1):19–42.

Cai, J., Janvry, A. D., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Cai, J. and Szeidl, A. (2024). Indirect effects of access to finance. *American Economic Review*, 114(8):2308–2351.

Cantoni, E. and Pons, V. (2022). Does context outweigh individual characteristics in driving voting behavior? evidence from relocations within the united states. *American Economic Review*, 112(4):1226–1272.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, 135(1):269–327.

Casey, K., Kamara, A. B., and Meriggi, N. F. (2021a). Data and code for: An experiment in candidate selection. openICPSR Project 124501, Version V1.

Casey, K., Kamara, A. B., and Meriggi, N. F. (2021b). An experiment in candidate selection. *American Economic Review*, 111(5):1575–1612.

Chakravorty, B., Arulampalam, W., Bhatiya, A. Y., Imbert, C., and Rathelot, R. (2024a). Can information about jobs improve the effectiveness of vocational training? experimental evidence from india. *Journal of Development Economics*, 169:103273.

Chakravorty, B., Arulampalam, W., Bhatiya, A. Y., Imbert, C., and Rathelot, R. (2024b). Data for: Can information about jobs improve the effectiveness of vocational training? experimental evidence from india. Author-provided dataset (personal communication).

Chatterjee, S. (2005). Concentration inequalities with exchangeable pairs (Ph. D. thesis). *arXiv preprint math/0507526*.

Chatterjee, S. (2007). Stein's method for concentration inequalities. *Probability Theory and Related Fields*, 1(138):305–321.

Chen, J. and Ritzwoller, D. M. (2023). Semiparametric estimation of long-term treatment effects. *Journal of Econometrics*, 237(2):105545.

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2011). *Normal approximation by Stein's method*, volume 2. Springer.

Chen, Q., Syrgkanis, V., and Austern, M. (2022). Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems*, 35:3096–3109.

Chen, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications.

Chen, X. and Kato, K. (2019). Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics*, 47(6):3127–3156.

Chen, Z., Liu, Z., Suárez Serrato, J. C., and Xu, D. Y. (2021). Notching r&d investment with corporate income tax cuts in china. *American Economic Review*, 111(7):2065–2100.

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: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Detailed proof of Nazarov's inequality. *arXiv preprint arXiv:1711.10696*.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2023). Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india. *arXiv preprint arxiv:1712.04802*.

Chetverikov, D. (2024). Tuning parameter selection in econometrics. *arXiv preprint arXiv:2405.03021*.

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317.

Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 36(2):457–462.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Cicala, S. (2022). Imperfect markets versus imperfect regulation in us electricity generation. *American Economic Review*, 112(2):409–441.

Conley, T. G. and Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *American economic review*, 100(1):35–69.

Cooper, Z., Scott Morton, F., and Shekita, N. (2020). Surprise! out-of-network billing for emergency care in the united states. *Journal of Political Economy*, 128(9):3626–3677.

Cornec, M. (2010). Concentration inequalities of the cross-validation estimate for stable predictors. *arXiv preprint arXiv:1011.5133*.

Covert, T. R. and Sweeney, R. L. (2023). Relinquishing riches: Auctions versus informal negotiations in texas oil and gas leasing. *American Economic Review*, 113(3):628–663.

Cox, D. R. (1958). Planning of experiments.

Dasgupta, P. and Stiglitz, J. (1980). Industrial structure and the nature of innovative activity. *The Economic Journal*, 90(358):266–293.

De la Pena, V. and Giné, E. (1999). *Decoupling: from dependence to independence*. Springer.

De la Peña, V. H. and Montgomery-Smith, S. J. (1995). Decoupling inequalities for the tail probabilities of multivariate U-statistics. *The Annals of Probability*, pages 806–816.

Delfino, A. (2024). Breaking gender barriers: Experimental evidence on men in pink-collar jobs. *American Economic Review*, 114(6):1816–1853.

Dembo, A. (2021). Probability theory: Stat310/math230.

Demirkaya, E., Fan, Y., Gao, L., Lv, J., Vossler, P., and Wang, J. (2024). Optimal nonparametric inference with two-scale distributional nearest neighbors. *Journal of the American Statistical Association*, 119(545):297–307.

Derenoncourt, E. (2022). Can you move to opportunity? evidence from the great migration. *American Economic Review*, 112(2):369–408.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., and Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219.

Diamond, R. and McQuade, T. (2019). Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. *Journal of Political Economy*, 127(3):1063–1117.

DiCiccio, C. and Romano, J. (2022). CLT for U-statistics with growing dimension. *Statistica Sinica*, 32(1).

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020). Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Donaldson, D. and Hornbeck, R. (2016). Railroads and american economic growth: A “market access” approach. *The Quarterly Journal of Economics*, 131(2):799–858.

Dubé, J.-P. and Misra, S. (2023). Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189.

Duchi, J. (2024). Statistics and information theory. *URL: https://web.stanford.edu/class/stats311/lecture-notes.pdf.*, 2:23.

Dunn, R., Ramdas, A., Balakrishnan, S., and Wasserman, L. (2023). Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022a). Data and code for: General equilibrium effects of cash transfers: Experimental evidence from kenya. Econometric Society replication files for Econometrica.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022b). General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica*, 90(6):2603–2643.

Elisseeff, A., Evgeniou, T., Pontil, M., and Kaelbing, L. P. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1).

Ellison, G. and Glaeser, E. L. (1997). Geographic concentration in us manufacturing industries: a dartboard approach. *Journal of political economy*, 105(5):889–927.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

Ellison, G. and Pathak, P. A. (2021). The efficiency of race-neutral alternatives to race-based affirmative action: Evidence from chicago's exam schools. *American Economic Review*, 111(3):943–975.

Emmenegger, C., Spohn, M.-L., Elmer, T., and Bühlmann, P. (2025). Treatment effect estimation with observational network data using machine learning. *Journal of Causal Inference*, 13(1):20230082.

Fafchamps, M. and Quinn, S. (2018). Networks and manufacturing firms in africa: Results from a randomized field experiment. *The World Bank Economic Review*, 32(3):656–675.

Faridani, S. and Niehaus, P. (2022). Linear estimation of average global effects. *arXiv preprint arXiv:2209.14181*.

Farronato, C., Fradkin, A., Larsen, B. J., and Brynjolfsson, E. (2024). Consumer protection in an online world: An analysis of occupational licensing. *American Economic Journal: Applied Economics*, 16(3):549–579.

Feyrer, J., Mansur, E. T., and Sacerdote, B. (2017). Geographic dispersion of economic shocks: Evidence from the fracking revolution. *American Economic Review*, 107(4):1313–1334.

Fix, E. and Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918.

Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908.

Franklin, S., Imbert, C., Abebe, G., and Mejia-Mantilla, C. (2024). Urban public works in spatial equilibrium: Experimental evidence from ethiopia. *American Economic Review*, 114(5):1382–1414.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2021). Package 'glmnet'. *CRAN R Repositary*, 595.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75.

Fudenberg, D. and Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*, 109(12):4112–4141.

Gao, M. (2024). Endogenous interference in randomized experiments. *arXiv preprint arXiv:2412.02183*.

Gao, M. and Ding, P. (2023). Causal inference in network experiments: regression-based analysis and design-based properties.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Giné, E., Latała, R., and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, pages 13–38. Springer.

Giroud, X., Lenzu, S., Maingi, Q., and Mueller, H. (2024). Propagation and amplification of local productivity spillovers. *Econometrica*, 92(5):1589–1619.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

Götze, F., Sambale, H., and Sinulis, A. (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26(none):1–22.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation. Technical report, National Bureau of Economic Research.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of political economy*, 118(3):536–598.

Griliches, Z. (1992). The search for r&d spillovers. *The Scandinavian Journal of Economics*, pages S29–S47.

Guo, W. and Romano, J. P. (2017). Analysis of error control in large scale two-stage multiple hypothesis testing. *arXiv preprint arXiv:1703.06336*.

Gut, A. (2009). *Stopped random walks*. Springer.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, 14(1):1–16.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083.

Hanson, G. H. (2005). Market potential, increasing returns and geographic concentration. *Journal of international economics*, 67(1):1–24.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Haushofer, J., Niehaus, P., Paramo, C., Miguel, E., and Walker, M. W. (2022). Targeting impact versus deprivation. Technical report, National Bureau of Economic Research.

Head, K. and Mayer, T. (2011). Gravity, market potential and economic development. *Journal of Economic Geography*, 11(2):281–294.

Heinonen, J. (2001). *Lectures on analysis on metric spaces*. Springer Science & Business Media.

Helm, I. (2020). National industry trade shocks, local labour markets, and agglomeration spillovers. *The Review of Economic Studies*, 87(3):1399–1431.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Statist.*, 23(4):169–192.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098.

Hoover, D. N. (1979). Relations on probability spaces and arrays of. *Preprint, Institute for Advanced Study*.

Hornbeck, R. and Moretti, E. (2024). Estimating who benefits from productivity growth: local and distant effects of city productivity growth on wages, rents, and inequality. *Review of Economics and Statistics*, 106(3):587–607.

Hu, Y., Li, S., and Wager, S. (2022). Average direct and indirect causal effects under interference. *Biometrika*, 109(4):1165–1172.

Huber, K. (2018). Disentangling the effects of a banking crisis: Evidence from german firms and counties. *American Economic Review*, 108(3):868–898.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the american statistical association*, 103(482):832–842.

Huo, Y., Fan, Y., and Han, F. (2023). On the adaptation of causal forests to manifold data. *arXiv preprint arXiv:2311.16486*.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Jackson, M. O. and Rogers, B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915.

Jaffe, A. B. (1986). Technological opportunity and spillovers of r & d: Evidence from firms' patents, profits, and market value. *The American Economic Review*, pages 984–1001.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598.

Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-validation and mean-square stability. In *ICS*, pages 487–495.

Kelly, M., Mokyr, J., and Ó Gráda, C. (2023). The mechanics of the industrial revolution. *Journal of Political Economy*, 131(1):59–94.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049.

Khosravi, K., Lewis, G., and Syrgkanis, V. (2019). Non-parametric inference adaptive to intrinsic dimension. *arXiv preprint arXiv:1901.03719*.

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. *The Quarterly journal of economics*, 129(1):275–331.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Koijen, R. S., Richmond, R. J., and Yogo, M. (2024). Which investors matter for equity valuations and expected returns? *Review of Economic Studies*, 91(4):2387–2424.

Kojevnikov, D. (2021). The bootstrap for network dependent processes. *arXiv preprint arXiv:2101.12312*.

Kolesár, M. and Plagborg-Møller, M. (2024). Dynamic causal effects in a nonlinear world: the good, the bad, and the ugly. *arXiv preprint arXiv:2411.10415*.

Kovak, B. K. (2013). Regional effects of trade reform: What is the correct measure of liberalization? *American Economic Review*, 103(5):1960–1976.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. *Advances in Neural Information Processing Systems*, 24.

Kraay, A. and McKenzie, D. (2014). Do poverty traps exist? assessing the evidence. *Journal of Economic Perspectives*, 28(3):127–148.

Krugman, P. (1987). The narrow moving band, the dutch disease, and the competitive consequences of mrs. thatcher: Notes on trade in the presence of dynamic scale economies. *Journal of development Economics*, 27(1-2):41–55.

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013). Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35. PMLR.

Lachièze-Rey, R. and Peccati, G. (2017). New berry–esseen bounds for functionals of binomial point processes. *The Annals of Applied Probability*, 27(4):1992–2031.

Landers, D. and Rogge, L. (1976). The exact approximation order in the central-limit-theorem for random summation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 36(4):269–283.

Landers, D. and Rogge, L. (1988). Sharp orders of convergence in the random central limit theorem. *Journal of Approximation Theory*, 53(1):86–111.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

Lee, A. J. (1990). *U-statistics: theory and practice*. Marcel Dekker.

Lehmann, E. L. and Romano, J. P. (2022). *Testing statistical hypotheses*, volume 4. Springer.

Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lerche, A. (2025). Direct and indirect effects of investment tax incentives. *American Economic Review*.

Leung, M. P. (2020). Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2):368–380.

Leung, M. P. (2022a). Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293.

Leung, M. P. (2022b). Rate-optimal cluster-randomized designs for spatial interference. *The Annals of Statistics*, 50(5):3064–3087.

Leung, M. P. (2023). Network cluster-robust inference. *Econometrica*, 91(2):641–667.

Leung, M. P. and Loupos, P. (2025). Graph neural networks for causal inference under network confounding. *arXiv preprint arXiv:2211.07823*.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Lewis, G. and Syrgkanis, V. (2021). Double/debiased machine learning for dynamic treatment effects. In *NeurIPS*, pages 22695–22707.

Li, S. and Wager, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358.

Liu, S.-H., Shao, Q.-M., and Shi, H. (2025). Berry-esseen bounds for degenerate u-statistics with application to distance correlation. *Science China Mathematics*, pages 1–36.

Lucking, B., Bloom, N., and Van Reenen, J. (2019). Have r&d spillovers declined in the 21st century? *Fiscal Studies*, 40(4):561–590.

Lychagin, S., Pinkse, J., Slade, M. E., and Reenen, J. V. (2016). Spillovers in space: does geography matter? *The Journal of Industrial Economics*, 64(2):295–335.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Marshall, A. (1890). *Principles of Economics*. Macmillan, London.

Mastrobuoni, G. (2020). Crime is terribly revealing: Information technology and police productivity. *The Review of Economic Studies*, 87(6):2727–2753.

Maurer, A. (2019). A Bernstein-type inequality for functions of bounded interaction.

McCormick, T. H. and Zheng, T. (2015). Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association*, 110(512):1684–1695.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(1):841–881.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143–2188.

Menzel, K. (2025). Fixed-population causal inference for models of equilibrium. *arXiv preprint arXiv:2501.19394*.

Miguel, E. and Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Minsker, S. (2023). U-statistics of growing order and sub-Gaussian mean estimators with sharp constants. *Mathematical Statistics and Learning*.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, migration, and local employment elasticities. *American Economic Review*, 108(12):3855–3890.

Montgomery-Smith, S. J. (1993). Comparison of sums of independent identically distributed random variables. *Probability and Mathematical Statistics*, 14:281–285.

Moretti, E. (2004). Workers' education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review*, 94(3):656–690.

Moscona, J. and Seck, A. A. (2024). Age set versus kin: Culture and financial ties in east africa. *American Economic Review*, 114(9):2748–2791.

Motwani, R. and Raghavan, P. (1995). *Randomized algorithms*. Cambridge university press.

Muendler, M.-A. and Becker, S. O. (2010). Margins of multinational labor substitution. *American Economic Review*, 100(5):1999–2030.

Munro, E. (2025). Causal inference under interference through designed markets. *arXiv preprint arXiv:2504.07217*.

Munro, E., Wager, S., and Xu, K. (2025). Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*, 5.

Muralidharan, K., Niehaus, P., and Sukhtankar, S. (2023). General equilibrium effects of (improving) public employment programs: Experimental evidence from india. *Econometrica*, 91(4):1261–1295.

Myers, K. R. and Lanahan, L. (2022). Estimating spillovers from publicly funded r&d: Evidence from the us department of energy. *American Economic Review*, 112(7):2393–2423.

Nadeau, C. and Bengio, Y. (1999). Inference for the generalization error. *Advances in Neural Information Processing Systems*, 12.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

Ogburn, E. L., Sofrygin, O., Diaz, I., and Van der Laan, M. J. (2024). Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611.

Okunogbe, O. and Pouliquen, V. (2022). Technology, taxation, and corruption: evidence from the introduction of electronic tax filing. *American Economic Journal: Economic Policy*, 14(1):341–372.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR.

Paulin, D., Mackey, L., and Tropp, J. A. (2013). Deriving matrix concentration inequalities from kernel couplings. *arXiv preprint arXiv:1305.0612*.

Paulin, D., Mackey, L., and Tropp, J. A. (2016). Efron-Stein inequalities for random matrices. *The Annals of Probability*, 44(5):3431 – 3473.

Peel, T., Anthoine, S., and Ralaivola, L. (2010). Empirical Bernstein inequalities for U-statistics. *Advances in Neural Information Processing Systems*, 23.

Peng, W., Coleman, T., and Mentch, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics*, 16(1):232–292.

Penrose, M. (2003). *Random geometric graphs*, volume 5. OUP Oxford.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer.

Pollmann, M. (2023). Causal inference for spatial treatments. *arXiv preprint arXiv:2011.00373*.

Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability*, pages 2053–2086.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A. and Manole, T. (2023). Randomized and exchangeable improvements of markov's, chebyshev's and chernoff's inequalities. *arXiv preprint arXiv:2304.02611*.

Redding, S. and Venables, A. J. (2004). Economic geography and international inequality. *Journal of international Economics*, 62(1):53–82.

Rényi, A. (1957). On the asymptotic distribution of the sum of a random number of independent random variables. *Acta Math*, 8:193–199.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.

Ritzwoller, D. M. and Romano, J. P. (2023). Reproducible aggregation of sample-split statistics. *arXiv preprint arXiv:2311.14204*.

Ritzwoller, D. M., Romano, J. P., and Shaikh, A. M. (2024). Randomization inference: Theory and applications. *arXiv preprint arXiv:2406.09521*.

Ritzwoller, D. M. and Syrgkanis, V. (2024). Uniform inference for subsampled moment regression. *arXiv preprint arXiv:2405.07860*.

Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys*, 8:210–293.

Rotemberg, M. (2019). Equilibrium effects of firm subsidies. *American Economic Review*, 109(10):3475–3513.

Rubin, D. B. (1980). Discussion of a paper by D. Basu. *Journal of the American Statistical Association*, 75(371):591–593.

Rüger, B. (1978). Das maximale signifikanzniveau des tests:"lehne h_o ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 25:171–178.

Sadka, J., Seira, E., and Woodruff, C. (2024). Information and bargaining through agents: Experimental evidence from mexico's labour courts. *Review of Economic Studies*, page rdae003.

Sävje, F. (2024). Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):1–15.

Sävje, F., Aronow, P., and Hudgens, M. (2021). Average treatment effects in the presence of unknown interference. *Annals of statistics*, 49(2):673.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151.

Scotchmer, S. (2004). *Innovation and incentives*. MIT press.

Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

Serfling, R. (1980). Approximation theorems of mathematical statistics. *Wiley series in probability and mathematical statistics*.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.

Shao, Q.-M. and Zhang, Z.-S. (2019). Berry–Esseen bounds of normal and nonnormal approximation for unbounded exchangeable pair. *The Annals of Probability*, 47(1):61–108.

Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, 22(1):439–459.

Shevtsova, I. (2011). On the absolute constants in the berry-esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*.

Song, Y., Chen, X., and Kato, K. (2019). Approximating high-dimensional infinite-order u-statistics: statistical and computational guarantees. *Electronic Journal of Statistics*, 13(2).

Spence, M. (1984). Cost reduction, competition, and industry performance. *Econometrica*, 52(1):101–122.

Steiger, W. (1970). Bernstein's inequality for martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 16(2):104–106.

Stein, C. (1986). Approximate computation of expectations. *IMS Lecture Notes—Monograph Series*, 7.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR.

Tse, T. and Davison, A. C. (2022). A note on universal inference. *Stat*, 11(1):e501.

van der Vaart, A. (2000). Asymptotic statistics. *Cambridge Books*.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence*. Springer.

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, 237(1):105237.

Veitch, V., Wang, Y., and Blei, D. (2019). Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, 32.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Viviano, D., Lei, L., Imbens, G., Karrer, B., Schrijvers, O., and Shi, L. (2023). Causal clustering: design of cluster experiments under network interference. *arXiv preprint arXiv:2310.14983*.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, Y., Samii, C., Chang, H., and Aronow, P. (2025). Design-based inference for spatial experiments under unknown interference. *The Annals of Applied Statistics*, 19(1):744 – 768.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2023). Evaluating treatment prioritization rules via rank-weighted average treatment effects.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2024). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, pages 1–14.

Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statistics*, 14(4):478–486.

Zacchia, P. (2020). Knowledge spillovers through networks of scientists. *The Review of economic studies*, 87(4):1989–2018.

Zhang, Z.-S. (2022). Berry–esseen bounds for generalized u-statistics. *Electronic Journal of Probability*, 27:1–36.