Three Essays on Identification and Inference in Econometric Models

by

Kenichi Nagasawa

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2019

Doctoral Committee:

Professor Matias D. Cattaneo, Chair
Assistant Professor Andreas Hagemann
Professor Lutz Kilian
Professor Rocío Titiunik

Kenichi Nagasawa

kenagasa@umich.edu

ORCID iD: 0000-0002-9633-1132

© Kenichi Nagasawa 2019

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Matias Cattaneo. Without his guidance, patience, and encouragement, I would not have made it thus far in my academic journey. Also, it has been very enjoyable to work with Matias on different projects.

I would also like to thank Lutz Kilian, Andreas Hagemann, and Rocío Titiunik for their support and encouragement. Their support was invaluable especially during the job market process. I am also grateful to Xinwei Ma, Gonzalo Vazques-Bare, Max Farrel, Sebastian Calonico, Yingjie Feng, Mel Stephens, and Michael Jansson for their help. Chapters 3 and 4 of this dissertation are partially based on the joint work with Michael and Matias.

Finally, I would like to express my appreciation to Daisaku Ikeda, whose examples have inspired me to live my life to the fullest. I will keep learning from his examples and aspire to make impact on the world.

TABLE OF CONTENTS

ACKNOWLEDGMENTS ii

LIST OF TABLES vi

LIST OF APPENDICES vii

ABSTRACT viii

CHAPTER

I Introduction 1

II Identification and Estimation of Group-Level Partial Effects 2

II.1 Introduction . . . 2

II.1.1 Related Literature . . . 5

II.2 Setup and Overview of Results . . . 7

II.2.1 Identification Problem . . . 10

II.2.2 Heuristic Discussion of Identification Result . . . 12

II.3 Identification Result . . . 15

II.3.1 Main Result . . . 18

II.4 Estimation and Inference . . . 19

II.5 Numerical Results . . . 22

II.5.1 Empirical Application . . . 22

II.5.2 Monte Carlo Experiments . . . 25

II.6 Extensions . . . 26

II.6.1 Instrumental Variables for Omitted Variable Bias . . . 26

II.6.2 Nonparametric Identification . . . 27

III Bootstrap-based Inference for Cube Root Consistent Estimators 33

III.1 Introduction . . . 33

III.2 Heuristics . . . 34
III.3 Main Result . . . 38
III.4 Examples . . . 43
III.4.1 The Maximum Score Estimator . . . 43
III.4.2 The Conditional Maximum Score Estimator . . . 49
III.5 Discussion . . . 52
III.6 Extension: Isotonic Density Estimation . . . 54
III.6.1 Bootstrap-based Inference . . . 57
III.6.2 Drift Estimation . . . 58
III.6.3 Simulation Evidence . . . 58

IV Smoothed Pairwise Difference Estimators: Distribution Theory 61
IV.1 Introduction . . . 61
IV.2 Setup . . . 62
IV.3 Main Result . . . 65
IV.4 Conclusion . . . 70

APPENDIX 71

A Proofs for Chapter II 71
A.1 Proofs . . . 71
A.1.1 Proof of Lemma 1 . . . 71
A.1.2 Proof of Theorem 1 . . . 74
A.1.3 Proof of Theorem 2 . . . 75
A.1.4 Proof of Theorem 3 . . . 80
A.1.5 Proof of Theorem 4 . . . 82
A.1.6 Proof of Theorem 5 . . . 83
A.2 Details of DGPs for Monte Carlo Studies . . . 84

B Proofs for Chapter III 87
B.1 Proofs . . . 87
B.1.1 Weak Convergence of Original Estimator . . . 87
B.1.2 Proof of Theorem 6 . . . 94
B.1.3 Proof of Lemmas 2 and 3 . . . 101
B.1.4 Proof of Corollary MS . . . 107
B.1.5 Proof of Lemma MS . . . 109

B.1.6 Proof of Corollary CMS 115
B.1.7 Proof of Lemma CMS 120
B.1.8 Proof of Theorem ID 123
B.2 Rule-of-Thumb Bandwidth Selection 127
B.2.1 Maximum Score Estimation 127
B.2.2 Isotonic Density Estimation 129

C Proofs for Chapter IV 131
C.1 Proof of Theorem 7 131
C.1.1 Consistency 132
C.1.2 Rates of Convergence 133
C.1.3 Weak Convergence 135

BIBLIOGRAPHY 143

LIST OF TABLES

TABLE

II.1 Estimates of School/Neighborhood Effects on Wage and Years of Post Secondary Education 30
II.2 Results of Monte Carlo Experiments 31
II.3 Empirical Application: Random Effects Model 32

III.1 Simulations, Maximum Score Estimator, 95% Confidence Intervals. 48
III.2 Simulations, Isotonic Density Estimator, 95% Confidence Intervals. 60

LIST OF APPENDICES

APPENDIX A. Proofs for Chapter II 71

APPENDIX B. Proofs for Chapter III 87

APPENDIX C. Proofs for Chapter IV 131

ABSTRACT

The dissertation studies identification and inference problems in econometric models. In the first chapter, I present a new identification result for partial effects of group-level variables when economic agents select into groups. The econometric model allows for selection on individual unobserved heterogeneity, which generally causes endogeneity of group-level variables. I use individual-level variables that have sufficient correlation with the unobserved heterogeneity to develop a control function method. The second chapter studies statistical inference in a class of M-estimation problems in which the standard nonparametric bootstrap fails due to non-smoothness in the objective function. I propose a way to "reshape" the objective function to restore bootstrap consistency. In the third chapter, I present a novel distributional theory for a class of pairwise smoothed estimators, which highlights non-robustness of conventional inference methods with respect to choice of tuning parameters.

CHAPTER I

Introduction

This dissertation studies two important areas in econometrics analysis, identification and inference. The first topic, identifiability of model parameters, is particularly important in Economics, where researchers often face data issues such as sample selection, unobserved key variables, and missing data. Absent of these data issues, connections between observable quantities and parameters of interest are often transparent. Yet, data issues make the connection opaque, and in some cases, parameters of interest are not even uniquely determined from the observed information.

Chapter II examines identification of partial effects of group-level variables when economic agents select into groups. In such circumstance, a complication arises because agents potentially base their group selection on group-level variables of interest, which creates a well-known endogeneity issue due to sample selection.

The other two chapters study the issue of statistical inference. They focus on issues of "robust" inference, broadly defined. Specifically, by robustness, I refer to the property that inference procedures remain valid under different smoothness in underlying data generating processes or different choices of tuning parameters.

Chapter III studies bootstrap distributional approximation of a class of cube root convergent estimators. I show that the standard nonparametric bootstrap fails for this class of estimators due to inability to capture the Hessian matrix. The chapter proposes reshaping of the objective function to restore bootstrap consistency.

Chapter IV develops novel distributional results for a class of smoothed pairwise difference estimators. I show that pairwise difference estimators have a range of limit distributions depending on tuning parameters. An important implication of this new distributional approximation is that a deviation from stringent bandwidth conditions renders conventional inference procedures invalid.

CHAPTER II

Identification and Estimation of Group-Level Partial Effects

II.1 Introduction

Policy makers often design interventions to influence individual outcomes through group-level variables. For instance, a government may relocate disadvantaged children to higher quality schools to improve their academic performance. Given their potential impact, many studies in economics have sought to evaluate group-level policy interventions (see Durlauf, 2004; Durlauf and Ioannides, 2010; Graham, 2018, and references therein). Nevertheless, estimation of group-level treatment effects is challenging. The problem is that individuals select into groups in part based on their unobserved characteristics, and this sorting causes systematic dependence among group-level variables and those individual characteristics. Therefore, comparing outcomes across groups without accounting for differences in unobserved heterogeneity is subject to selection bias.

This paper presents a novel identification result for group-level partial effects using observable variables that "shift" the distribution of unobserved individual heterogeneity. Informally, a random vector W is a "shifter" for another (unobservable) random vector Θ if the conditional distribution of Θ, given $W = w$, varies sufficiently with w. A key insight of this paper is that controlling for group-level distributions of observable shifters accounts for variation in unobserved individual heterogeneity. Thus, if there remains enough variation in group-level variables conditional on the control variables, group-level causal effects become identifiable.

For simplicity, suppose that unobserved individual heterogeneity Θ takes two values, "high" and "low," denoted by θ_H and θ_L. Then, an observable W is a shifter for Θ if $\Pr(\Theta = \theta_H | W = w)$ is a non-constant function of w, or equivalently, if there exist w_1, w_2 such that the 2×2 matrix $\mathbf{\Pi} = [\Pr(\Theta = \theta_l | W = w_k)]_{k=1,2;l=H,L}$ is non-singular. The availability of a shifter is useful because a group-level distribution of W can be written as

a linear transformation of the distribution of Θ within the same group, where the matrix of this linear transformation is $\mathbf{\Pi}$. Since $\mathbf{\Pi}$ is non-singular, there exists a one-to-one mapping between group-level distributions of the unobserved variable Θ and the observable shifter W. Therefore, we can use an observable shifter to control for across-group variation in unobserved heterogeneity distributions.

To fix ideas, consider a setting where students choose schools. Academically motivated students may prefer high quality schools, and when they sort into schools based on this preference, there will be a positive correlation between school-level teacher quality and the proportion of highly motivated students within the school. Here, student's motivation corresponds to unobserved individual heterogeneity and school-level teacher quality is the group-level variable of interest. A possible shifter for academic motivation is educational attainment of a child's mother. It seems reasonable to assume that the higher is a mother's educational attainment, the more motivated her child tends to be.

Now suppose that student's motivation takes two values, high and low. Then, mother's education serves as a shifter if a child of a college graduate mother has a larger probability of being highly motivated than a student whose mother has only a high school diploma. Provided that mother's education is a valid shifter, there exists a one-to-one mapping between a school-level fraction of college graduate mothers and the distribution of student types within the school, which enables us to control for across-school variation in student's motivation. If school-wide teacher quality has independent variation from the school-level distribution of mother's education, then ceteris paribus effects of teacher quality are identifiable. Note that mother's education is not an instrumental variable (IV): a valid IV would be independent of student's motivation and be related to teacher quality.

To provide further intuition on the above argument, I describe it with simple equations. Denote an outcome of interest by Y_{is} (e.g., test score), where i and s index agent and group, respectively, group-level covariates of interest by X_s (e.g., school-wide teacher quality), an observable shifter by W_i (e.g., mother's education), and unobserved individual heterogeneity by Θ_i (e.g., academic motivation). Like the above discussion, assume Θ_i takes two values, θ_H and θ_L (e.g., high and low motivation). Also, let J_i be individual i's group choice. Here, the interest lies in the causal effect of teacher quality on test scores. The regression equation turns out to be

$$Y_{iJ_i} = \alpha + X'_{J_i}\beta + W'_i\gamma + \delta \Pr(\Theta_i = \theta_H | X_{J_i}, J_i) + \varepsilon_{iJ_i}, \qquad \mathbb{E}[\varepsilon_{is} | X_s, W_i, J_i = s] = 0 \quad \text{(II.1)}$$

where subscript J_i is present because a researcher only observes the outcome for the group individuals selected in the data. Since $\Pr(\Theta_i = \theta_H | X_s, J_i = s)$ cannot be estimated, its

dependence on X_s hinders the identification of β. In the school example, $\Pr(\Theta_i = \theta_H | X_s, J_i = s)$ corresponds to the fraction of highly motivated students in a school with teacher quality X_s. In this framework, the inability to control for a fraction of high motivation types confounds naive estimates of teacher-quality effects.

As noted above, the observable shifter condition implies that there exists a linear map from the distribution of the observable shifter to the unobserved heterogeneity distribution, i.e., $\Pr(\Theta_i = \theta_H | X_s, J_i = s) = \sum_{l=1}^{2} d_l \Pr(W_i = w_l | X_s, J_i = s)$ for some $d_1, d_2 \in \mathbb{R}$. Then, (II.1) becomes

$$Y_{iJ_i} = \alpha + X'_{J_i}\beta + W'_i\gamma + \sum_{l=1}^{2} \delta d_l \Pr(W_i = w_l | X_{J_i}, J_i) + \varepsilon_{iJ_i}$$

where the error term ε_{iJ_i} is uncorrelated with all the other right-hand side variables. Since $[\Pr(W_i = w_l | X_s, J_i = s)]_{l=1,2}$ is identifiable from the data (e.g., school-level fractions of college graduate and high school graduate mothers), β becomes identifiable.

I generalize the argument based on the two-point supported distribution of Θ to one with a general distribution. To formalize the idea of shifter, I use the notion of statistical completeness, which has been applied in a wide range of nonparametric identification problems. I argue that the availability of a shifter is a reasonable assumption, especially when a researcher observes a rich array of individual-level characteristics.

When completeness is used for identification, estimation often faces ill-posed inverse problems, which may lead to poor finite-sample properties of estimators. The estimator in this paper circumvents this problem by applying the completeness assumption to estimation of a nuisance parameter of the model. By additive separability of the nuisance function from the parameter of interest, the estimator for group-level partial effects converges at a rate of the square root of the sample size as happens in partially linear regression models.

For implementation, a researcher first picks a set of basis functions and computes the group-means of the observable shifters transformed by those basis functions. She then runs a linear regression of the outcome on the group-level covariates, individual-level regressors, and the group-means computed in the first step. I show that this procedure leads to a consistent and asymptotically normal estimator, and I provide a consistent variance estimator. Monte Carlo experiments demonstrate that the proposed estimator has low mean squared errors and that the associated confidence intervals have high coverage accuracy in samples of moderate size.

I apply this method to the problem of studying the effects of school/neighborhood characteristics on years of post-secondary education and adulthood wages. I build on the analysis of Altonji and Mansfield (2018), who also use a control function approach. This paper shows

that the estimation equation arising from my identification result encompasses that of Altonji and Mansfield. Thus, my estimator is more robust to possible mis-specifications. I find that the two estimators produce very similar estimates, which supports the results in Altonji and Mansfield.

This paper employs assumptions distinct from those of existing identification strategies. In particular, I do not require natural/quasi experimental variation as often used in IV methods. Instead, I use the observable shifter condition, which can hold in many applications of interest. Thus, I provide an alternative approach to identifying group-level treatment effects. In addition, my identification result has wide applicability since the argument based on the shifter condition extends to non-linear and nonseparable models in a straightforward manner. From a theoretical perspective, this paper achieves new identification results using unexploited yet empirically relevant features in triangular models, such as group structures. Lastly, this paper applies statistical completeness in a novel way to develop a control function approach. To elaborate on these contributions, I now review the related literature.

II.1.1 Related Literature

This paper contributes to the large empirical literature that examines causal effects of neighborhood, school, and, more generally, group-level characteristics.[1] A non-trivial challenge in this literature is how to control for endogeneity arising from agents selecting into groups (Durlauf, 2004; Graham, 2018). Many studies exploit exogenous variation in group selection via instrumental variables (e.g., Angrist, Pathak and Walters, 2013; Kling, Liebman and Katz, 2007; Ludwig, Duncan and Hirschfield, 2001; Oreopoulos, 2003), some studies use detailed data on individual choice behavior to control for selection bias (e.g., Abdulkadiroğlu, Pathak, Schellenberg and Walters, 2017; Dale and Krueger, 2002), and other studies use aggregation to mitigate influence of selection (e.g., Card and Rothstein, 2007). In contrast, I take a control function approach that complements the existing methods by exploiting a novel, yet empirically relevant identifying condition.

This paper is most closely related to recent work by Altonji and Mansfield (2018), who also employ a control function method. They show that group-means of individual covariates play the role of control functions in their model. In contrast, I use group-means of *transformations* of individual covariates as control functions, where these transformations are basis functions (e.g., polynomials and splines). Whereas Altonji and Mansfield exploit

[1] A partial list includes Aaronson (1998); Abdulkadiroğlu, Pathak and Walters (2018); Altonji, Elder and Taber (2005); Angrist and Lang (2004); Chetty and Hendren (2018); Chetty, Hendren and Katz (2016); Dobbie and Fryer (2011); Gould, Lavy and Paserman (2004); Hanushek, Kain and Rivkin (2009); Hoxby (2000). Also, Chetverikov, Larsen and Palmer (2016) study effects of group-level endogenous variables. They do not consider endogeneity from selection and focus on other issues.

specific functional forms in their model to justify their control function method, I use restrictions on the conditional distribution of individual-level unobserved heterogeneity given observables to construct a control function. An advantage of my approach is that the identification argument extends to non-linear, nonseparable models in a relatively straightforward manner and thus my results apply to a wide class of models. In addition, the estimation equation arising from my identification result encompasses that of Altonji and Mansfield. In fact, when certain regression functions in the model are linear, the two approaches produce effectively the same identification result.

Another strand of the related literature is the one on control function methods, which is also connected to triangular models. A general triangular system takes the form

$$Y = m(X, \varepsilon)$$
$$X = h(Z, \eta)$$

where Y is the outcome of interest, X is potentially endogenous, and Z is an instrument independent of the unobserved heterogeneity (ε, η). In his seminal work, Heckman (1974, 1979) develops a control function approach where X is a binary variable, using additive separability and joint normality of (ε, η). Subsequent papers (e.g., Dahl, 2002; Das, Newey and Vella, 2003; Dubin and McFadden, 1984; Lee, 1983) extend the model in different directions. Examples include allowing for multinomial X and weakening parametric distributional assumptions. Also, Newey, Powell and Vella (1999) develop a control function method for nonparametric triangular systems and Blundell and Powell (2004) use a control function to identify average partial effects in semiparametric limited dependent variable models. Recently, Chesher (2003) and Imbens and Newey (2009) exploit strict monotonicity of functions in unobserved heterogeneity to construct a control variate to identify the ceteris paribus effect of X on Y (see also Matzkin, 2016, and references therein). I consider a version of the triangular model that exploits availability of multiple measurements of groups as well as an observable shifter to construct a control function. In contrast, the earlier literature seems to have largely focused on monotonicity restrictions and distributional assumptions on unobserved heterogeneity as identifying conditions. The use of empirically relevant features such as multiple measurements of groups offers a new approach to achieving identification in this class of models.

In addition, this paper is related to the growing literature on nonparametric identification using statistical completeness. Since the seminal work of Newey and Powell (2003), completeness has been applied to a wide range of econometric identification problems. Examples include nonparametric IV (e.g., Chernozhukov and Hansen, 2005; Darolles, Fan, Florens

and Renault, 2011; Hall and Horowitz, 2005; Newey and Powell, 2003), errors-in-variables models (Hu and Schennach, 2008), nonparametric discrete choice models with unobserved product characteristics (Berry and Haile, 2014), and nonseparable (dynamic) panel data models (e.g., Arellano, Blundell and Bonhomme, 2017; Cunha, Heckman and Schennach, 2010; Freyberger, 2018; Sasaki, 2015). See also the survey article Hu (2017) and references therein. My paper applies the completeness assumption in a novel way to construct a control function.

This paper shares with the literature on network formation the feature that agents form groups endogenously within models (e.g., Blume, Brock, Durlauf and Jayaraman, 2015; de Paula, 2017; Graham, 2017, and see references therein). However, I consider a different group selection mechanism from those in the network literature. In particular, in my analysis, the utility function for group selection does not depend on other individuals' group choices and their characteristics. Therefore, my model excludes certain selection patterns that are possible under network formation models. Nonetheless, the model still covers many empirical settings of interest.

The remainder of the paper is organized as follows. Section II.2 describes the econometric model and provides a heuristic discussion of the identification strategy. In Section II.3, I formalize the identification idea. In Section II.4, I propose a simple estimator of group-level partial effects and study its asymptotic properties. Section II.5 discusses the empirical application and Monte Carlo experiments to examine finite-sample accuracy of the proposed estimator. Section II.6 introduces two extensions of the model described in Section II.2. Details of the proofs can be found in Appendix A.1.

II.2 Setup and Overview of Results

In this section, I describe the econometric model. A researcher observes "cities," indexed by $g = 1, \ldots, G$. Within each city, there exist groups and agents, indexed by $s \in \mathcal{S} \equiv \{1, \ldots, S\}$ and $i = 1, 2, \ldots, N$, respectively. The outcome of interest, denoted by Y_{isg}, takes the form

$$Y_{isg} = \beta' X_{sg} + \gamma' W_{ig} + \chi_{sg} + \omega_{ig} + \epsilon_{isg}, \qquad \mathbb{E}[\epsilon_{isg} | X_{sg}, W_{ig}, \chi_{sg}, \omega_{ig}] = 0 \tag{II.2}$$

where $\{X_{sg} : s \in \mathcal{S}\}$ and W_{ig} are observable variables for groups and individuals, respectively, $\{\chi_{sg} : s \in \mathcal{S}\}$ and ω_{ig} denote unobservable characteristics for groups and individuals, respectively, and $\{\epsilon_{isg} : s \in \mathcal{S}\}$ represents unobserved idiosyncratic terms. The focus of this paper is on group-level causal effects, captured by β in (II.2). To identify this parameter, I require variation in group-level variables. The presence of cities provides multiple

measurements, or independent copies, of groups and agents, which enables identifying the distribution of group-level covariates and within-group distributions of individual variables. Below, I omit city index g to reduce notational burden when appropriate.

To ground the discussion on concrete terms, consider the following examples.

Example 1 (Residential Segregation and Youth Outcomes)**.** Graham (2018) surveys issues of residential segregation and its consequence on youth outcomes. In one example, the outcome Y_{is} measures adulthood wage, W_i denotes the indicator of a resident being a minority, and X_s represents the proportion of minorities in one's neighborhood. Here a neighborhood corresponds to a group. Unobservables that enter into the equation include resident's innate cognitive ability (ω_i) and distance from the city center to one's neighborhood (χ_s). The parameter of interest is β, which measures how the proportion of minorities in one's neighborhood affects an outcome of interest. □

Example 2 (Effects of School District/Neighborhood on Student Performance)**.** Altonji and Mansfield (2018) examine effects of school district/neighborhood on various student outcomes. In one of their empirical specifications, Y_{is} is years of post-secondary education, X_s includes teacher-student ratio and distance to four-year college, W_i contains student's scores on standardized tests and parents' years of education, ω_i includes how much parents value child's academic learning, and χ_s represents unobserved characteristics of school. Here, a school district/neighborhood represents a group. □

Example 3 (Effects of Hospital Ownership on Quality of Care)**.** Sloan, Picone, Taylor and Chou (2001) study whether private, for-profit hospitals have lower quality of care. Since different types of patients may select into for-profit hospitals compared to public ones, selection bias is of potential concern. Here, a hospital corresponds to a group, the outcome Y_{is} is patient's condition after hospitalization, X_s includes the indicator of whether the hospital is private and for-profit, and W_i includes measures of health conditions before hospitalization. □

In these examples, a researcher observes the outcome variable only for the group an individual selects into. To be precise, let $J_{ig} \in \mathcal{S}$ be the group membership of agent i. Then, we only observe $Y_{ig} := \sum_{s \in \mathcal{S}} Y_{isg} \mathbb{1}\{J_{ig} = s\}$. I model the group membership determination as following.

$$J_{ig} = J(\Theta_{ig}, A_{1g}, \ldots, A_{Sg}, \eta_{i1g}, \ldots, \eta_{iSg}) \tag{II.3}$$

where $J(\cdot)$ is an unknown function belonging to a nonparametric class, Θ_{ig} is individual heterogeneity affecting the selection, $\{A_{sg} : s \in \mathcal{S}\}$ are characteristics of groups affecting agents' group choice, and $\{\eta_{isg} : s \in \mathcal{S}\}$ are idiosyncratic terms. One example of the selection

equation $J(\cdot)$ that fits into this framework is the widely used random utility discrete choice model. In that setting, $J(\Theta_i, A_1, \ldots, A_S, \eta_{i1}, \ldots, \eta_{iS}) = \arg\max_{s \in \mathcal{S}} V_i(s)$ and $V_i(s) = \Theta_i' A_s + \eta_{is}$, where $V_i(s)$ represents agent i's utility choosing the group s, and the utility is specified as the group-level features A_s weighted by the taste coefficient Θ_i plus the idiosyncratic term η_{is}. Random coefficient specification allows for heterogeneous tastes for group characteristics and accommodates complicated selection patterns.

A key restriction on the selection function $J(\cdot)$ is that, loosely speaking, if a researcher were to observe all the variables, she could compare groups with different X_s holding within-group distributions of Θ_i constant. That is, there must exist separate variation in X_s from within-group distributions of individual characteristics. This condition holds, for instance, in the random utility discrete choice model satisfying the following: given $x_1, x_2 \in \text{supp}(X_s)$, there exist $\mathbf{a}^k \in \text{supp}(\{A_s : s \in \mathcal{S}\} | X_s = x_k), k = 1, 2$ satisfying $a_s^1 - a_{s'}^1 = a_s^2 - a_{s'}^2$ for all pairs of $(s, s') \in \mathcal{S}^2$. Alternative conditions are possible, and I formalize this assumption as non-singularity of some conditional variance in Section II.3. Also, I accommodate the possibility that any or all of $(\Theta_i, \{\eta_{is}, A_s : s \in \mathcal{S}\})$ in the selection equation is not observed, and instead I impose that the observable W_i has non-trivial relationship with the preference heterogeneity Θ_i.

An important feature of the selection equation (II.3) is that it does not depend on choices and characteristics of other individuals, which excludes the types of models considered in network formation literature. Yet, this does not exclude equilibrium effects. For instance, X_s can include features of the distribution of individual characteristics within group s, which can be empirically important since the coefficients on such variables may represent "peer effects," e.g., the effect of school-level minority fraction on student performance. As I will outline below, the control function method I propose uses group-level averages of (transformed) variables W_i. If X_s includes some features of the within-group distribution of a subvector of W_i (e.g., mean), the subvector needs to be excluded from the construction of the control functions.

Given the outcome and selection equations, I now state the following sampling assumptions to complete the model description.[2]

$$(\mathbf{X}_g, \boldsymbol{\chi}_g, \mathbf{A}_g) \overset{iid}{\sim} F, \qquad (W_{ig}, \omega_{ig}, \Theta_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig}) \overset{iid}{\sim} H \tag{II.4}$$

$$(W_{ig}, \omega_{ig}, \Theta_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig}) \perp\!\!\!\perp (\mathbf{X}_g, \boldsymbol{\chi}_g, \mathbf{A}_g) \tag{II.5}$$

$$(W_{ig}, \omega_{ig}, \boldsymbol{\epsilon}_{ig}) \perp\!\!\!\perp \boldsymbol{\eta}_{ig} | \Theta_{ig} \tag{II.6}$$

[2]These conditions are slightly stronger than those in Section II.3. I maintain them here for ease of exposition.

where $\mathbf{X}_g = \{X_{sg} : s \in \mathcal{S}\}$, $\boldsymbol{\chi}_g = \{\chi_{sg} : s \in \mathcal{S}\}$, $\mathbf{A}_g = \{A_{sg} : s \in \mathcal{S}\}$, $\boldsymbol{\epsilon}_{ig} = \{\epsilon_{isg} : s \in \mathcal{S}\}$, $\boldsymbol{\eta}_{ig} = \{\eta_{isg} : s \in \mathcal{S}\}$, and F and H are distribution functions. Restriction (II.4) states that the group-level features are i.i.d. copies across cities and individual-level variables are a random draw across individuals and cities.

Condition (II.5) requires independence between individual variables and group-level variables. One way to rationalize this independence assumption is that "nature" first draws group-level characteristics, then subsequently samples individual characteristics whose distributions are independent of group-level variables, and finally individuals select into groups. In Example 1, this scenario means that first neighborhoods are formed, then individuals are drawn from the distribution that does not depend on neighborhood characteristics, and after the realizations of group- and individual-level variables, agents make decisions on which neighborhood to live in. This assumption seems reasonable under static frameworks.

Condition (II.6) states that the idiosyncratic term in the choice equation is independent of other individual-level variables conditional on the taste variable Θ_i. Independence is a strong assumption, but such restriction has been commonly used in the literature of discrete choice models (e.g., Briesch, Chintagunta and Matzkin, 2010). Also, here the independence only needs to hold conditioning on the taste heterogeneity.

II.2.1 Identification Problem

In this subsection, I describe a challenge in identifying the group-level effect β. Specifically, there are two sources of endogeneity, *omitted variable bias* and *selection bias*. The former occurs from unobservability of group-level characteristics χ_s and potential correlation between X_s and χ_s. Although this issue is practically relevant, it is well understood in the literature and this paper has little new insight to offer on this problem. Instead, I focus on selection bias and assume that there is no correlation between X_s and χ_s except in Section II.6.1. In particular, I impose $\mathbb{E}[\chi_s|W_i, X_s, J_i = s] = 0$ for all s in this section.

For selection bias, the problem arises because the distribution of ω_i, the unobservable in the outcome equation (II.2), varies across groups and this variation is systematically related to X_s. In the school example, the unobservable ω_i represents student's motivation, and school characteristics affecting selection A_s contain the school-level teacher quality variable X_s. If highly motivated students prefer high quality educational programs, students with high ω_i are then more likely to be in schools with higher X_s. This sorting pattern will cause bias on the coefficient on X_s.

I now formalize selection bias in the model. Recall that we only observe the outcome for one group, i.e., $Y_i = \sum_{s \in \mathcal{S}} \mathbb{1}\{J_i = s\} Y_{is}$, and computing the regression function given

observable variables yields

$$\mathbb{E}[Y_i|J_i = s, X_s, W_i] = \mathbb{E}[Y_{is}|J_i = s, X_s, W_i] = \beta' X_s + \gamma' W_i + \mathbb{E}[\omega_i|J_i = s, X_s, W_i].$$

Identifiability of β depends on whether the conditional expectation $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$ is constant in X_s. This model implication is closely related to the characterization of selection bias in Heckman (1976). He formulates that $\mathbb{E}[Y|Z, \{\text{selection rule}\}] = Z'\beta + \mathbb{E}[\epsilon|Z, \{\text{selection rule}\}]$, where Z is a vector of covariates and ϵ is an error term. In my model $\{J_i = s\}$ corresponds to the "selection rule." Heckman points out that many econometric models have this characterization and that agents' self-selection prevents identification of β when the term $\mathbb{E}[\epsilon|Z, \{\text{selection rule}\}]$ is non-constant.

To gain some intuition on how the selection rule operates in my model, take $A_s = X_s$ and $J_i = \arg\max_{s \in \mathcal{S}} \{\Theta_i' X_s\}$ (i.e., I set $\eta_{is} = 0$ for all i and s). Then, conditioning on $\{J_i = s\}$ and $\mathbf{X}$ implies $\Theta_i \in \mathcal{R}_s(\mathbf{X})$ where $\mathcal{R}_s(\mathbf{X}) = \{\Theta : (X_s - X_{s'})'\Theta > 0 \text{ for all } s' \neq s\}$ is a region formed by intersections of half planes. See Figure II.1 for a simple illustration. Then, for the conditional expectation $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$,

$$\begin{aligned} \mathbb{E}[\omega_i|J_i = s, X_s, W_i] &= \mathbb{E}[\mathbb{E}[\omega_i|J_i = s, \mathbf{X}, W_i]|J_i = s, X_s, W_i] \\ &= \mathbb{E}[\mathbb{E}[\omega_i|\Theta_i \in \mathcal{R}_s(\mathbf{X}), W_i]|J_i = s, X_s, W_i] \end{aligned} \tag{II.7}$$

where the second equality follows from $(\omega_i, \Theta_i, W_i) \perp\!\!\!\perp \mathbf{X}$. Now, to see implications of the last display, suppose $\mathbb{E}[\omega_i|\Theta_i, W_i] = \mathbb{E}[\omega_i|W_i]$. Then (II.7) implies $\mathbb{E}[\omega_i|J_i = s, X_s, W_i] = \mathbb{E}[\omega_i|W_i]$ and therefore no selection bias occurs. This result resembles the one in Heckman (1976): if unobserved terms in outcome and selection equations (ω_i and Θ_i in this model) are independent, there is no endogeneity due to selection. On the other hand, if the conditional mean of ω_i given $\Theta_i = \theta$ and W_i non-trivially changes with θ, then the term $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$ varies with X_s since the shape of $\mathcal{R}_s(\mathbf{X})$ changes with X_s. The dependence of $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$ on X_s results in non-identification of β.

Note that the above argument continues to hold even if we deviate from the simplifying assumption $A_s = X_s$, provided that A_s is related to X_s, which is likely the case. For example, X_s measures school-level teacher quality, and students are likely to use it or its proxy as a basis of their school choice decision. Also, the conditional mean independence of ω_i given Θ_i is likely to fail in applications if ω_i contains unobserved heterogeneity that affects people's preference for group characteristics Θ_i, e.g., if student's motivation affects preference for school quality.

Example 1 (Continued)**.** Graham (2018) discusses potential sources of endogeneity that

hinder identification of causal effects of neighborhood segregation of minorities. Among the conditions he discusses, the *no sorting on (individual) unobservables* condition is relevant to my paper. Roughly speaking, this condition requires that individuals with the same observed covariates are similar across neighborhoods. It excludes, for example, racial minorities living in one neighborhood differ substantially in their cognitive skills from minorities in another neighborhood.

To formalize the "no sorting" condition, redefine γ and ω_i in (II.2) by projecting ω_i onto W_i to have $\text{Cov}(W_i, \omega_i) = 0$. This is little loss of generality since γ is a nuisance parameter. Then, the no sorting condition can be represented as

$$\mathbb{E}[\omega_i|W_i, J_i] = \mathbb{E}[\omega_i|W_i],$$

where for Graham W_i is the indicator of whether the person is minority. This equation imposes no systematic difference in the average of unobserved individual heterogeneity across neighborhoods. In my model, this condition corresponds to $\mathbb{E}[\omega_i|\Theta_i, W_i] = \mathbb{E}[\omega_i|W_i]$ (i.e., no selection bias condition in the above). To see this point, note $\mathbb{E}[\omega_i|W_i, J_i] = \int \mathbb{E}[\omega_i|W_i, J_i, \mathbf{A}, \Theta] f_{\Theta\mathbf{A}|W,J} = \int \mathbb{E}[\omega_i|W_i, \Theta] f_{\Theta\mathbf{A}|W,J}$, where I use (II.5) and (II.6). If ω_i and Θ_i are conditionally mean independent given W_i, the integral integrates to one, which implies $\mathbb{E}[\omega_i|W_i, J_i]$ is independent of J_i. Therefore, no selection bias in my model translates to the no sorting condition in Graham's paper. Whereas Graham explores scenarios under which this no sorting condition is plausible, I allow for failure of this condition and instead use observable shifter variables to construct control functions. □

II.2.2 Heuristic Discussion of Identification Result

The previous subsection indicated the source of selection bias. Here, I discuss this paper's approach for identification emphasizing main ideas rather than technical details. For simplicity, assume finite support of Θ_i, i.e., $\text{supp}(\Theta_i) = \{\theta_t : 1 \leq t \leq T\}$. We can view this discrete-valued heterogeneity Θ_i as agents' type and these types differ in preferences for group attributes A_s.

I can rewrite the outcome equation by decomposing ω_i into its mean within group J_i and the deviation.

$$Y_i = \beta' X_i + \gamma' W_i + \mathbb{E}[\omega_i|J_i, \mathbf{A}] + \varepsilon_i, \qquad \varepsilon_i = \varepsilon_{iJ_i}, \qquad \varepsilon_{is} = \omega_i - \mathbb{E}[\omega_i|J_i = s, \mathbf{A}] + \chi_s + \epsilon_{is} \quad \text{(II.8)}$$

where I write $X_i \equiv X_{J_i}$, $\varepsilon_i \equiv \varepsilon_{iJ_i}$ to avoid double subscripts, and $\mathbb{E}[\omega_i|J_i = s, \mathbf{A}]$ represents the expectation of ω_i given that the individual selected group s and the underlying group

features are $\mathbf{A}$. Conditioning on $\mathbf{A}$ is necessary because an agent observes and takes as given the group features $\mathbf{A}$ when making group choice.[3] In the above decomposition, X_i is uncorrelated with ε_i, and therefore, endogeneity is present if and only if X_s has non-zero correlation with the within-group mean $\mathbb{E}[\omega_i|J_i = s, \mathbf{A}]$. For this conditional expectation, we have

$$\begin{aligned}\mathbb{E}[\omega_i|J_i = s, \mathbf{A}] &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|J_i = s, \mathbf{A}, \Theta_i = \theta_t] \Pr(\Theta_i = \theta_t|J_i = s, \mathbf{A}) \\ &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|\mathbf{A}, \Theta_i = \theta_t] \Pr(\Theta_i = \theta_t|J_i = s, \mathbf{A}) \\ &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|\Theta_i = \theta_t] \Pr(\Theta_i = \theta_t|J_i = s, \mathbf{A}). \end{aligned} \tag{II.9}$$

where the second equality follows from $J_i = J(\Theta_i, \mathbf{A}, \boldsymbol{\eta}_i)$ and independence $\omega_i \perp\!\!\!\perp \boldsymbol{\eta}_i|\Theta_i$, and the third equality uses $\mathbf{A} \perp\!\!\!\perp (\Theta_i, \omega_i)$.

The last equation indicates that across-group variation in $\mathbb{E}[\omega_i|J_i = s, \mathbf{A}]$ comes from variation in $\Pr(\Theta_i = \theta|J_i = s, \mathbf{A})$. The key insight of this paper is that $\Pr(\Theta_i = \theta|J_i = s, \mathbf{A})$ can be expressed as a linear function of the conditional distribution of the observable shifter W_i given $(J_i, \mathbf{A})$, and therefore, accounting for across-group variation in observable shifter distributions solves the endogeneity problem of X_s.

To sketch the identification argument, I assume finite support of W_i, i.e., $\text{supp}(W_i) = \{w_l : 1 \leq l \leq L\}$. This is without loss of generality since I can always create a discrete random variable from continuous one by binning. In the case of discretely distributed Θ_i, W_i is a shifter for Θ_i if the matrix

$$\mathbf{\Pi} = \begin{bmatrix} \Pr(\Theta_i = \theta_1|W_i = w_1) & \dots & \Pr(\Theta_i = \theta_T|W_i = w_1) \\ & \ddots & \\ \Pr(\Theta_i = \theta_1|W_i = w_L) & \dots & \Pr(\Theta_i = \theta_T|W_i = w_L) \end{bmatrix}$$

is of full column rank. This condition formalizes the idea that the conditional distribution of Θ_i given $W_i = w$ exhibits enough variation in w. Also it requires that the cardinality of $\text{supp}(W_i)$ is larger than that of $\text{supp}(\Theta_i)$ since full column rank fails if $L < T$. This full rank condition may be reasonable when a researcher observes a rich array of individual-level

[3]Another way to see this point is that $\mathbf{A}$ is invariant across individuals given $J_i = s$. The sample group mean is average over i, i.e., $\sum_{i=1}^{N} \mathbb{1}\{J(\Theta_i, \mathbf{A}, \boldsymbol{\eta}_i) = s\}\omega_i / \sum_{i=1}^{N} \mathbb{1}\{J(\Theta_i, \mathbf{A}, \boldsymbol{\eta}_i) = s\}$ and $\mathbf{A}$ is independent of i. Then, the sample group means converges in probability to the expecatation conditional on $\mathbf{A}$. This calculation is akin to computing probability limits of sample means over long time horizon in panel data models. Given a variable X_{it} and fixed effects α_i, $T^{-1}\sum_{t=1}^{T} X_{it} \to_{\mathbb{P}} \mathbb{E}[X_{it}|\alpha_i]$ under suitable conditions.

variables that are related to individual taste for group-level attributes. In the neighborhood example, variables such as education level, health status, type of occupation, and ethnicity are likely to have non-trivial relationships with individual taste for various neighborhood characteristics.

Doing similar calculations to those for (II.9) and by Bayes' rule,

$$\begin{aligned}\Pr(W_i = w|J_i, \mathbf{A}) &= \sum_{t=1}^{T} \Pr(W_i = w|\Theta_i = \theta_t) \Pr(\Theta_i = \theta_t|J_i, \mathbf{A}) \\ &= \sum_{t=1}^{T} \Pr(\Theta_i = \theta_t|W_i = w) \Pr(W_i = w) \frac{\Pr(\Theta_i = \theta_t|J_i, \mathbf{A})}{\Pr(\Theta_i = \theta_t)}\end{aligned}$$

and dividing the both sides by $\Pr(W_i = w)$,

$$\frac{\Pr(W_i = w|J_i, \mathbf{A})}{\Pr(W_i = w)} = \sum_{t=1}^{T} \Pr(\Theta_i = \theta_t|W_i = w) \frac{\Pr(\Theta_i = \theta_t|J_i, \mathbf{A})}{\Pr(\Theta_i = \theta_t)}$$

We can write this equation in matrix form

$$\boldsymbol{\pi}_w = \boldsymbol{\Pi}\boldsymbol{\pi}_\theta \tag{II.10}$$

where $\boldsymbol{\pi}_w = [\Pr(W_i = w_l|J_i, \mathbf{A})/\Pr(W_i = w_l)]_{l=1}^{L}$ and $\boldsymbol{\pi}_\theta = [\Pr(\Theta_i = \theta_t|J_i, \mathbf{A})/\Pr(\Theta_i = \theta_t)]_{t=1}^{T}$. Since $\boldsymbol{\Pi}$ is of full column rank by the observable shifter assumption, we have $\boldsymbol{\pi}_\theta = \mathbf{G}\boldsymbol{\pi}_w$ where $\mathbf{G} = (\boldsymbol{\Pi}'\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}'$, and the t-th element of this linear equation looks like

$$\Pr(\Theta_i = \theta_t|J_i, \mathbf{A}) = \Pr(\Theta_i = \theta_t) \sum_{l=1}^{L} G_{tl} \frac{\Pr(W_i = w_l|J_i, \mathbf{A})}{\Pr(W_i = w_l)} \equiv \sum_{l=1}^{L} d_{tl} \Pr(W_i = w_l|J_i, \mathbf{A})$$

where G_{tl} is the (t, l)th element of $\mathbf{G}$ and d_{tl} is a constant independent of $(J_i, \mathbf{A})$. Substituting this last result into (II.9), we obtain

$$\begin{aligned}\mathbb{E}[\omega_i|J_i, \mathbf{A}] &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|\Theta_i = \theta_t] \sum_{l=1}^{L} d_{tl} \Pr(W_i = w_l|J_i, \mathbf{A}) \\ &= \sum_{l=1}^{L} \delta_l \Pr(W_i = w_l|J_i, \mathbf{A}), \qquad \delta_l = \sum_{t=1}^{T} d_{tl}\mathbb{E}[\omega_i|\Theta_i = \theta_t]\end{aligned}$$

and finally the outcome equation becomes

$$Y_i = \beta' X_i + \gamma' W_i + \sum_{l=1}^{L} \delta_l \Pr(W_i = w_l | J_i, \mathbf{A}) + \varepsilon_i$$

where $\Pr(W_i = w_l | J_i, \mathbf{A})$ is identifiable from the data and $\text{Cov}(X_i, \varepsilon_i) = 0$. Therefore, the vector $[\Pr(W_i = w_l | J_i, \mathbf{A})]_{l=1}^{L}$ plays the role of a control function to account for the endogeneity of X_s with respect to unobserved individual heterogeneity ω_i.

II.3 Identification Result

In this section, I formalize the heuristics described in the preceding section. For that purpose, I impose the following conditions on the model. Below, let $L^p(\mu)$ be the class of functions whose pth power is integrable with respect to μ. For a random object V, I write F_V for its distribution function.

Assumption 1. *Let* $\mathbf{B}_g \equiv (\mathbf{X}_g, \boldsymbol{\chi}_g, \mathbf{A}_g)$.

1. $\mathbf{B}_g$ *is a random draw across* g. *Conditional on* $\mathbf{B}_g$, $(W_{ig}, \Theta_{ig}, \omega_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig})$ *is i.i.d. across* i, *and* $(W_{ig}, \Theta_{ig}, \omega_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig})$ *is independent across* g.

2. $(W_{ig}, \omega_{ig}, \Theta_{ig}) \perp\!\!\!\perp \mathbf{B}_g$ *and* $\boldsymbol{\eta}_{ig} \perp\!\!\!\perp (W_{ig}, \omega_{ig}) | \Theta_{ig}, \mathbf{B}_g$.

Assumption 2. *The distribution of* $\mathbf{B}$ *is dominated by some measure. The random vectors* (W_i, Θ_i) *have a joint density with respect to the product measure of some* σ*-finite measures* μ *and* λ. *The densities* $f_{\Theta|W}, f_\Theta, f_W$ *are bounded.*

Assumption 3. *The following mapping* $\boldsymbol{\Psi}$ *defined on* $L^2(F_\Theta)$ *is injective;*

$$(\boldsymbol{\Psi} h)(w) = \int h(\theta) f_{\Theta|W}(\theta|w) d\lambda(\theta).$$

Also, $\{f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})/f_\Theta(\theta) : s \in \mathcal{S}, \mathbf{b} \in \text{supp}(\mathbf{B})\} \subset L^2(F_\Theta)$.

From this assumption, Theorem 15.16 in Kress (2014) implies that there exist $\{\tau_j \geq 0 : j \in \mathbb{N}\}$ and an orthonormal basis $\{\phi_j : j \in \mathbb{N}\}$ on $L^2(F_\Theta)$ such that $\tau_j^2 \phi_j = \boldsymbol{\Psi}^* \boldsymbol{\Psi} \phi_j$ where $\boldsymbol{\Psi}^*$ is the adjoint operator of $\boldsymbol{\Psi}$. Assume τ_j is ordered such that $\tau_j \geq \tau_{j+1}$ for all $j \geq 1$.

Assumption 4. *With* $\{(\tau_j, \phi_j) : j \in \mathbb{N}\}$ *defined above,*

$$f_{\Theta|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_\Theta(\cdot) \in \mathcal{F}_1 = \Big\{ f \in L^2(F_\Theta) : \sum_{j=1}^{M} \big|\langle f, \phi_j \rangle_\Theta\big| < \infty \Big\} \text{ for all } s \in \mathcal{S}, \mathbf{b} \in \text{supp}(\mathbf{B})$$

$$\mathbb{E}[\omega_i|\Theta_i = \cdot] \in \mathcal{F}_2 = \Big\{ f \in L^2(F_\Theta) : \sum_{j=1}^{M} \tau_j^{-1} \big|\langle f, \phi_j \rangle_\Theta\big| < \infty \Big\},$$

where $M = \sup\{j : \tau_j > 0\}$ *and* $\langle f, g \rangle_\Theta = \int f g dF_\Theta$. M *can be positive infinity.*

Assumption 1 describes the sampling. Part 1 formalizes the idea that the distribution of group-level variables and the within-group distribution of individual observables are identifiable from the data. Part 2 restates conditions (II.5) and (II.6) in a weaker form. Assumption 2 imposes mild restrictions on the distribution of $(\Theta_i, W_i, \mathbf{B})$.

Assumption 3 is a generalization of the full column rank condition used in the previous section. It ensures that the integral equation, which is the infinite-dimensional analogue of (II.10), is "invertible" in a suitable sense. This injectivity condition, usually referred to as L^2-completeness, is a high-level assumption but has been widely used in the recent econometric literature on nonparametric identification (see Section II.1.1). Intuitively, injectivity requires the density of $f_{\Theta|W}(\cdot|w)$ to sufficiently vary in the conditioning value w. One example of (Θ, W) satisfying completeness is $\Theta = \Gamma W + \nu$ where ν given W is distributed as Normal$(0, \Sigma)$, Σ is invertible, and Γ is of full column rank. More sufficient conditions for different types of completeness can be found in the literature (Andrews, 2017; D'Haultfoeuille, 2011; Hu, Schennach and Shiu, 2017; Hu and Shiu, 2018; Mattner, 1993). One way to justify this technical condition is to assume that in the population, there are a finite number of agent types which differ in preferences. Under this assumption, injectivity reduces to the full column rank of the conditional probability matrix as analyzed in the previous section. The strategy of modeling unobserved heterogeneity as finite number of types has been employed in empirical studies. For instance, recent papers of Abowd, McKinney and Schmutte (2018) and Bonhomme, Lamadon and Manresa (2018) study the consequence of worker-firm matching on labor market earnings and model worker and/or firm heterogeneity as discrete types. Viewing unobserved heterogeneity as finite types provides one way to rationalize Assumption 3. Nonetheless, completeness is applicable more generally and the condition can hold with continuous heterogeneity variable as well.

Also, Assumption 3 imposes restrictions on the relative tail of $f_{\Theta|J,\mathbf{B}}$ and f_Θ. In this model, selection into groups allows the within-group distribution of Θ to differ from the original distribution of Θ to the extent that this tail condition is satisfied. For instance, the model excludes the case where Θ has a Gaussian tail and the within-group distribution has a

polynomial tail. This is a high-level condition since it is not straightforward to characterize the within-group distribution from the model primitives. Also, note the conditioning on $\mathbf{B}$. In Section II.2, I impose stronger independence assumptions and only condition on $\mathbf{A}$, but here I slightly weaken the independence assumptions and thus conditioning on $\mathbf{B}$ is necessary.

Assumption 4 restricts the permissible classes for $f_{W|J,\mathbf{B}}$ and $\mathbb{E}[\omega_i|\Theta_i]$. It requires that

$$f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})/f_{\Theta}(\theta) = \sum_{j=1}^{\infty} c_j(s,\mathbf{b})\phi_j(\theta), \qquad \sum_{j=1}^{\infty} |c_j(s,\mathbf{b})| < \infty,$$
$$\mathbb{E}[\omega_i|\Theta_i = \theta] = \sum_{j=1}^{\infty} d_j\phi_j(\theta), \qquad \sum_{j=1}^{\infty} |d_j/\tau_j| < \infty,$$

for some $c_j(s,\mathbf{b}) \in \mathbb{R}$, $d_j \in \mathbb{R}$, $j \in \mathbb{N}$. A restrictive, yet easily interpretable sufficient condition is that there exist integers M and L, possibly dependent on $(s,\mathbf{b})$, such that $c_j(s,\mathbf{b}) = 0$ for all $j > L$ and $d_j = 0$ for all $j > M$. This condition implies that $f_{\Theta|J,\mathbf{B}}/f_{\Theta}$ and $\mathbb{E}[\omega_i|\Theta_i = \cdot]$ can be represented as finite linear combinations of the L^2-basis $\{\phi_j : j \in \mathbb{N}\}$.

Now, I state a key lemma for the identification of group-level partial effects.

Lemma 1. *Suppose Assumptions 1-4 hold. Then, there exists some function ψ such that*

$$\mathbb{E}[\omega_i|J_i = s, \mathbf{B}] = \int \psi(w) dF_{W|J,\mathbf{B}}(w|s,\mathbf{B})$$

where $F_{W|J,\mathbf{B}}$ is the conditional distribution function of W_i given $J_i, \mathbf{B}$ and $\int \psi^2 dF_W < \infty$.

This lemma suggests a way to control for the selection bias. To see the implication, from (II.8),

$$\begin{aligned} Y_i &= \beta' X_i + \gamma' W_i + \int \psi(w) dF_{W|J,\mathbf{B}}(w|J_i,\mathbf{B}) + \{\omega_i - \mathbb{E}[\omega_i|J_i,\mathbf{B}]\} + \chi_i + \epsilon_i \\ &= \beta' X_i + \gamma' W_i + \sum_{k=1}^{\infty} \delta_k \mathbb{E}[p_k(W_i)|J_i,\mathbf{B}] + \varepsilon_i \end{aligned} \tag{II.11}$$

where $\chi_i = \chi_{J_i}$, $\epsilon_i = \epsilon_{iJ_i}$, $\varepsilon_i = \varepsilon_{iJ_i}$, $\varepsilon_{is} = \omega_i - \mathbb{E}[\omega_i|J_i,\mathbf{B}] + \chi_s + \epsilon_{is}$, $\{p_k : k \in \mathbb{N}\}$ is a basis for $L^2(F_W)$, and $\{\delta_k : k \in \mathbb{N}\}$ satisfies $\int(\psi - \sum_{k=1}^{K} \delta_k p_k)^2 dF_W \to 0$ as $K \to \infty$. The class of approximating functions is chosen by the researcher and the conditional distribution of W_i given J_i and $\mathbf{B}$ is just a within group distribution of W_i, which is identifiable from the data. Since ε_i is uncorrelated from X_i, inclusion of $\{\mathbb{E}[p_k(W_i)|J_i,\mathbf{B}] : k \in \mathbb{N}\}$ controls for endogeneity arising from selection.

Example 2 (Continued)**.** Altonji and Mansfield (2018) also use a control function approach to address selection bias under a very similar econometric model. Despite some differences

in imposed conditions,[4] Lemma 1 generalizes their result. In particular, Proposition 1 in Altonji and Mansfield states

$$\mathbb{E}[\omega_i|J_i, \mathbf{B}] = \pi'\mathbb{E}[W_i|J_i, \mathbf{B}]$$

for some π. By taking $\psi(w) = \pi' w$, Lemma 1 in this paper and their Proposition 1 coincide.

Lemma 1 encompassing their Proposition 1 has an important implication for estimation. The estimation method considered in this paper approximates ψ using a series basis expansion. Suppose a researcher chooses polynomial or spline basis functions. Then, the estimation equation based on Lemma 1 specializes to the version used by Altonji and Mansfield if series expansion terminates after the constant and linear terms. Therefore, we can view the series-based estimation proposed in this paper as a robustified version of Altonji and Mansfield's control function method. □

II.3.1 Main Result

Now I formally state the identification of group-level partial effects. The object $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$ is a random distribution function, i.e., a function-valued random element. Thus, there exists a σ-field generated by $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$ and the conditional expectation given $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$ is well defined.

Assumption 5. *1. For $s \in \mathcal{S}$, $\mathbb{E}[\varepsilon_{ig}|J_{ig} = s, F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)] = 0$ and $\mathbb{E}[V_{ig}\varepsilon_{ig}] = 0$ where $\varepsilon_{ig} = \varepsilon_{iJ_{ig}g}$ and $V_{ig} = (X'_{ig}\, W'_{ig})'$.*

2. The matrix

$$\mathbb{E}[\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)]\}\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)]\}']$$

is invertible.

Assumption 5 1 imposes exogeneity of χ_s and excludes the case of omitted variable bias. It also assumes that W_i is uncorrelated with $\{\omega_i - \mathbb{E}[\omega_i|J_i, \mathbf{B}]\}$. Since γ is not the parameter of interest, I can rewrite the regression equation by projecting $\{\omega_i - \mathbb{E}[\omega_i|J_i, \mathbf{B}]\}$ onto W_i to make them uncorrelated. Part 2 is a version of the key identification condition for partially linear models. In the regression equation (II.11), the interest lies in β and we want to "partial out" the nuisance parameter $\int \psi dF_{W|J,\mathbf{B}}$. Here the within-group distribution of W_i, denoted by $F_{W|J,\mathbf{B}}$, is estimable from the data and conditional on the value of $F_{W|J,\mathbf{B}}$, $\int \psi dF_{W|J,\mathbf{B}}$ is a constant and thus differencing eliminates this nuisance parameter.

[4] I elaborate on different assumptions between the two papers in Section II.5.1.

Part 2 requires that X_s does not include functions of the within-group distribution of W_i, e.g., the within-group mean of W_i. This follows from $\mathbb{E}[h(W_i)|F_{W|J,\mathbf{B}}(\cdot|J_i,\mathbf{B})] = \mathbb{E}[h(W_i)|J_i,\mathbf{B}]$ for every measurable function h.[5] Sometimes it is desirable to include within-group distributional features in X_s since they represent "peer effects" as in Example 1. To accommodate such situations, it suffices to have a subvector $W_i^{\texttt{sub}}$ of W_i that satisfies Assumption 3, i.e., $(\Theta_i, W_i^{\texttt{sub}})$ is L^2-complete. Then, a researcher can include distributional features of W_i in X_s provided that the included elements are not part of $W_i^{\texttt{sub}}$.

Building on Lemma 1, the following theorem formalizes the identification of coefficients on group-level variables.

Theorem 1. *Suppose that Assumptions 1-5 hold. Then, β in the equation* (II.2) *is identified.*

II.4 Estimation and Inference

Lemma 1 suggests an estimation method by series approximation. If a researcher observed within-group distributions of W_{ig}, then the estimator based on least squares is

$$\hat{\beta}_{\text{oracle}} = \mathbf{S}(\mathbf{P}_K'\mathbf{P}_K)^{-1}\mathbf{P}_K\mathbf{Y}$$

where $\mathbf{P}_K = [P_{K11}P_{K21}\dots P_{KNG}]'$, $P_{Kig} = (X_{ig}', W_{ig}', \mathbb{E}[p_1(W_{ig})|J_{ig},\mathbf{B}_g], \dots, \mathbb{E}[p_K(W_{ig})|J_{ig},\mathbf{B}_g])'$, K represents the number of series terms, $\mathbf{Y} = (Y_{11},\dots,Y_{NG})'$, and $\mathbf{S} = [\mathrm{I}_{d_x}\mathbf{O}_{d_x\times(d_x+d_w+K)}]$ where d_x, d_w denote the dimensions of X, W, respectively, I_d is the $d\times d$ identity matrix, and $\mathbf{O}_{d_1\times d_2}$ is the $d_1\times d_2$ matrix with all elements equal to zero.

In practice, a researcher needs to estimate within-group distributions. Define the following object

$$\hat{\mathbb{E}}[p_k(W_{ig})|J_{ig},\mathbf{B}_g] = \sum_{j=1}^{N} p_k(W_{jg})\mathbb{1}\{J_{jg}=J_{ig}\}\Big/\sum_{j=1}^{N}\mathbb{1}\{J_{jg}=J_{ig}\}.$$

Then, the feasible version of the above estimator is

$$\hat{\beta} = \mathbf{S}(\hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K)^{-1}\hat{\mathbf{P}}_K\mathbf{Y}$$

where $\hat{\mathbf{P}}_K = [\hat{P}_{K11}\ \dots\ \hat{P}_{KNG}]'$ and $\hat{P}_{Kig} = (X_{ig}', W_{ig}', \hat{\mathbb{E}}[p_1(W_{ig})|J_{ig},\mathbf{B}_g], \dots, \hat{\mathbb{E}}[p_K(W_{ig})|J_{ig},\mathbf{B}_g])'$.

[5] This claim follows if $\mathbb{E}[h(W_i)|J_i,\mathbf{B}]$ is measurable with respect to $\sigma(F_{W|J,\mathbf{B}})$, the σ-field generated by $F_{W|J,\mathbf{B}}(\cdot|J_i,\mathbf{B})$. A σ-field on the space of distribution functions can be defined by the σ-field generated by maps $h\to\int h dF$ where $h\geq 0$ is a measurable function from $\mathbb{R}^k\to\mathbb{R}$ and F is a random distribution function (see e.g., Kallenberg, 2017). The measurability of $\mathbb{E}[h(W_i)|J_i,\mathbf{B}]$ follows from the representation $\mathbb{E}[h(W_i)|J_i,\mathbf{B}] = \int h(w)dF_{W|J,\mathbf{B}}(w|J_i,\mathbf{B})$.

To analyze asymptotic properties of $\hat{\beta}$, I impose additional assumptions. To state the conditions, write $\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$ for the maximum and minimum of eigenvalues of the argument. Let $\|\cdot\|$ be the Euclidean norm for vectors and the induced norm for matrices. Also, write $\varepsilon_{ig} = \varepsilon_{iJ_{ig}g}$ to avoid double subscripts.

Assumption 6. *The basis functions $\{p_k : k \in \mathbb{N}\}$ are uniformly bounded and the eigenvalues of the matrix $\mathbf{Q}_K = \mathbb{E}[P_{Kig}P'_{Kig}]$ are bounded and bounded away from zero uniformly in K. In addition, the class $\{p_k : k \in \mathbb{N}\}$ is a subset of some VC-class of functions as defined in van der Vaart and Wellner (1996). In addition, for some $m_x, m_w > 4$, $\mathbb{E}[\|W_{ig}\|^{m_w}]+\mathbb{E}[\|X_{sg}\|^{m_x}] < \infty$ for all $s \in \mathcal{S}$.*

Assumption 7. *Write $\mathcal{A}_{ig}$ for the σ-field generated by $\{W_{ig}, X_{ig}, J_{ig}, \mathbb{E}[p_k(W_{ig})|J_{ig}, \mathbf{B}_g]$ $k \in \mathbb{N}\}$. Then, $\mathbb{E}[\varepsilon_{ig}|\mathcal{A}_{ig}] = 0$, and there exist fixed constants c_1, c_2, C_1, C_2 such that, with probability one, $0 < c_1 \leq \mathbb{E}[\varepsilon_{ig}\varepsilon_{jg}|\mathcal{A}_{ig}, \mathcal{A}_{jg}] \leq C_1$, $\mathbb{E}[\varepsilon_{ig}^4|\mathcal{A}_{ig}, J_{ig}] \leq C_2$, and $\lambda_{\min}(\mathbb{E}[P_{Kig}P'_{Kjg}\varepsilon_{ig}\varepsilon_{jg}]) \geq c_2 > 0$ for all K.*

Assumption 8. *Let $r_{Kig} = \sum_{k=K+1}^{\infty} \delta_k \mathbb{E}[p_k(W_{ig})|J_{ig}, \mathbf{B}_g]$. There exists some $b > 0$ such that $\mathbb{E}[r_{Kig}^2] = O(K^{-2b})$.*

The above assumptions are relatively standard in the series estimation literature. However, the "basis functions" used in this procedure are conditional expectations of basis functions. Therefore, there are no general sufficient conditions for $\lambda_{\max}(\mathbf{Q}_K)$ to be bounded above and for $\lambda_{\min}(\mathbf{Q}_K)$ and $\lambda_{\min}(\mathbb{E}[P_{Kig}P'_{Kjg}\varepsilon_{ig}\varepsilon_{jg}])$ to be bonded away from zero. Since these high-level conditions may not be easily verifiable, one approach is to take a "flexible parametric" view, which is to assume that the function ψ in Lemma 1 can be approximated by finitely many but unknown number of basis functions. In that case, the model is essentially parametric and verifications of the conditions become straightforward.

The boundedness of basis functions in Assumption 6 holds for most of basis functions used in practice if the support of W_{ig} is compact. The unbounded support case would require use of different norms. For its second part, I assume that the second moment matrix of P_{Kig} is positive definite for all K and, in particular, that the eigenvalues are bounded above and away from zero. However, as mentioned above, there are no general primitive conditions for bounded eigenvalues as the series terms are the conditional expectations of original basis functions. In the proof, I make it explicit how the estimator depends on the minimum/maximum eigenvalues to understand how eigenvalues tending to zero/infinity may affect the asymptotic distributional properties of the estimator.

Assumption 7 imposes boundedness of the conditional variance of ε_{isg} from above and below and boundedness of the conditional fourth moment. The restriction on the conditional

variance seems to be standard in the literature and I use the finite fourth moment to verify Lindberg condition for central limit theorem. Assumption 8 controls the bias term in series approximation. This condition can be verified for specific $\{p_k : k \in \mathbb{N}\}$ if the function ψ has enough smoothness (see e.g., Belloni, Chernozhukov, Chetverikov and Kato, 2015, and references therein).

Now, I formally state the asymptotic properties of $\hat{\beta}$. Define $\xi_K = \sup_w \|(p_1(w), \ldots, p_K(w))'\|$, and $\zeta_K = \xi_K + (NG)^{1/m_w} + G^{1/m_x}$ where m_w, m_x are defined in Assumption 6. In practice, this quantity ζ_K is at least as large as $\sqrt{K}$, and below I take $\zeta_K \geq \sqrt{K}$.

Theorem 2. *Suppose β is identified in* (II.11) *and for some $c > 0$, $\Pr(N_{sg}/N_g \geq c$ for all $s \in \mathcal{S}$ and $g) \to 1$ where $N_{sg} = \sum_{i=1}^{N} \mathbb{1}\{J_{ig} = s\}$ and $N_g = \sum_{s=1}^{S} N_{sg}$. Under Assumptions 1, 6-8, if $\zeta_K^2 \log K/G \to 0$ and $G/K^{2b} = o(1)$, then*

$$\sqrt{G}(\hat{\beta}_{oracle} - \beta) = \mathbf{S}\mathbf{Q}_K^{-1} \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \psi_{Ng} + o_{\mathbb{P}}(1)$$

where

$$\psi_{Ng} = \frac{1}{N} \sum_{i=1}^{N} P_{Kig} \varepsilon_{ig}.$$

Furthermore, if $\mathrm{Var}(\chi_{sg} - \mathbb{E}[\omega_{ig}|J_{ig} = s, \mathbf{B}_g]) > 0$ for each s and $\zeta_K^4/G \to 0$, then

$$\mathbf{\Omega}_K^{-1/2} \mathbf{S}\mathbf{Q}_K^{-1} \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \psi_{Ng} \rightsquigarrow \mathrm{Normal}(\mathbf{0}, \mathrm{I}_{d_x})$$

where

$$\mathbf{\Omega}_K = \mathbf{S}\mathbf{Q}_K^{-1} \mathbb{E}[\psi_{Ng}\psi_{Ng}'] \mathbf{Q}_K^{-1} \mathbf{S}'$$

and $\|\mathbf{\Omega}_K^{-1}\|$ is uniformly bounded. Finally, if $\zeta_K \sqrt{K \log(NG)/N} \to 0$, then

$$\sqrt{G}(\hat{\beta} - \hat{\beta}_{oracle}) = o_{\mathbb{P}}(1).$$

Theorem 2 characterizes asymptotic distribution of $\hat{\beta}$ and provides a set of sufficient conditions under which $\hat{\beta}$ is asymptotically equivalent to the oracle estimator $\hat{\beta}_{\text{oracle}}$. It also states that the convergence rate of $\hat{\beta}$ is $\sqrt{G}$ rather than $\sqrt{GN}$. That N does not show up in the convergence rate comes from within-group correlations of residuals. The residual ε_{isg} contains group-level unobservable χ_{sg}, which causes within-group dependence. The within-group correlation is formalized by the requirement $\mathrm{Var}(\chi_{sg} - \mathbb{E}[\omega_{ig}|J_{ig} = s, \mathbf{B}_g]) > 0$. As seen below, this within-group correlation requires the standard error estimator clustered at

group levels.

In the first part of the theorem, I assume that the number of individuals in each group grows at a proportional rate with the city-wide number of individuals for every city. This requires that the probability of selecting into a group is uniformly bounded away from zero for all groups. This assumption guarantees that I can estimate the within-group means of $p_k(W_{ig})$ uniformly well across groups and cities.

In the theorem, I also impose restrictions on (relative) growth rates of K, N, and G. The first two are $\zeta_K^2 \log K/N = o(1)$ and $G/K^{2b} \to 0$. These conditions are standard in the literature except that I have ζ_K^2, which includes terms related to the number of finite moments of W_{ig} and X_{sg}. The non-standard part of the requirement is $\zeta_K^4/G \to 0$. This condition requires at least fourth moments of W_{ig} and X_{sg} to be finite. Particularly, it implies that $\mathbb{E}[\|X_{sg}\|^{m_x}] < \infty$ with $m_x > 4$ and $\mathbb{E}[\|W_{ig}\|^{m_w}] < \infty$ with $m_w > 4$ satisfying $N^4/G^{(m_w-4)} \to 0$. The finite moment condition allows control of the rate at which the maximum of $\|W_{ig}\|$ and $\|X_{ig}\|$ over (i, g) grows. In addition, this rate condition requires that $\xi_K^4/G \to 0$, which is stronger than what is used in the literature (e.g., Cattaneo, Farrell and Feng, 2018). Since this paper's emphasis is on the constructive identification result, I maintain this assumption and plan to improve on this aspect of the theoretical result in the future research.

For inference, we need a consistent estimator of the variance. A natural estimator of $\mathbf{\Omega}_K$ is

$$\hat{\mathbf{\Omega}}_K = \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{\Sigma}}_K\tilde{\mathbf{Q}}_K^{-1}\mathbf{S}'$$

where $\tilde{\mathbf{Q}}_K = \hat{\mathbf{P}}'_K\hat{\mathbf{P}}_K/GN$, $\hat{\mathbf{\Sigma}}_K = \frac{1}{G}\sum_{g=1}^{G}\hat{\psi}_{Ng}\hat{\psi}'_{Ng}$, $\hat{\psi}_{Ng} = \frac{1}{N}\sum_{i=1}^{N}\hat{P}_{Kig}(Y_{ig} - \hat{P}'_{Kig}\hat{\theta})$, and $\hat{\theta} = (\hat{\mathbf{P}}'_K\hat{\mathbf{P}}_K)^{-1}(\hat{\mathbf{P}}'_K\mathbf{Y})$. The following theorem formalizes that this variance estimator is consistent.

Theorem 3. *In addition to the hypothesis of Theorem 2, if* $\zeta_K^3\sqrt{K/G} + \zeta_K^3 K^{-b} \to 0$ *and for some* $m_\varepsilon > 2$, $\mathbb{E}[|\varepsilon_{isg}|^{m_\varepsilon}] < \infty$ *and* $\zeta_K\sqrt{(GN)^{1/m_\varepsilon}\log K/G} \to 0$ *hold, then*

$$\|\hat{\mathbf{\Omega}}_K^{-1/2} - \mathbf{\Omega}_K^{-1/2}\| = o_{\mathbb{P}}(1).$$

II.5 Numerical Results

II.5.1 Empirical Application

I employ the results of this paper to study the neighborhood/school-district effects on student outcomes. Particularly, I use the National Longitudinal Study of 1972 (NLS72), which was

one of the datasets analyzed by Altonji and Mansfield (2018).[6] They consider an econometric model very similar to (II.2)-(II.3). However, the selection equation in this paper is more general and I impose a set of different conditions to achieve identification. Their Proposition 1 implies that the outcome equation can be written as

$$Y_{is} = X_s'\beta + W_i'\gamma + \pi'\mathbb{E}[W_i|J_i, \mathbf{B}] + \varepsilon_{is}$$

and the term $\pi'\mathbb{E}[W_i|J_i, \mathbf{B}]$ plays the role of a control function to eliminate selection bias. Lemma 1 in this paper suggests that in general the control function may not be linear in within-group means of observable shifter W_i. Since least squares estimation based on series expansions covers the general nonlinear case and nests the linearity case, the estimation method proposed in this paper can test whether the linear specification is a reasonable approximation of the data generating process.

An empirically relevant concern is that even after including control functions, there remains "omitted variable bias" because X_s and χ_s are potentially correlated and a researcher does not observe χ_s. That is, the least squares estimate of β converges in probability to $\tilde{\beta} = \beta + \mathbb{E}[X_s X_s']^{-1}\mathbb{E}[X_s\chi_s]$, which includes the coefficient of linear projection of χ_s on X_s. Altonji and Mansfield address this concern by developing lower bounds for some measures of group-level effects on outcomes. They look at the impact of shifting from the 10th to the 90th quantile of the school/neighborhood characteristics, $\mathrm{Q}_{90}(X_s'\beta + \chi_s) - \mathrm{Q}_{10}(X_s'\beta + \chi_s)$. If $X_s'\beta + \chi_s$ is normally distributed, this difference in quantile can be expressed as $2 * 1.28 * \sqrt{\mathrm{Var}(X_s'\beta + \chi_s)}$, and Altonji and Mansfield provide conditions under which $\mathrm{Var}(X_s'\tilde{\beta}) \leq \mathrm{Var}(X_s'\beta + \chi_s)$. Since $\mathrm{Var}(X_s'\tilde{\beta})$ is estimable from the data, this approach produces feasible lower bounds for the object of interest.[7]

I follow the same lower bound approach to study the school/neighborhood contribution to early adulthood wage earnings of students and years of post-secondary education. Table II.1 displays estimates of measures of school/neighborhood effects on early adulthood log wage (with two sets of regressors) and years of post-secondary education. For each outcome variable, I estimated the regression model in two specifications. The first one uses group means of individual covariates, which is the specification of Altonji and Mansfield, and the second contains group means of interaction and squared terms as additional controls. In the table, "AM" denotes the first specification and "This Paper" refers to the second. The first and second rows are the impact of moving a student from a 10th percentile neighborhood to

[6] They analyze three other datasets in their paper, all of which are restricted-use. The dataset based on NLS72 is publicly available.

[7] Altonji and Mansfield also look at the fraction of variance attributable to school/neighborhood quality. They employ random effects modeling to estimate the variance of the school-level unobserved term and decompose the variance to estimate the (lower bound of the) variance fraction.

a 90th/50th percentile one i.e., $\{\Phi^{-1}(q) - \Phi^{-1}(0.1)\}\sqrt{\text{Var}(X_s'\beta)}$, $q \in \{0.9, 0.5\}$ where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function.

Across outcome variables and estimands, the linear and quadratic specifications produce very similar estimates. For the log wage outcome, moving a student from a 10th percentile school to a 90th percentile increases the wage by $17.58\% \approx (100 * \exp(0.162) - 100)$ with post-secondary education as a control and by 17.11% without the control. For the 10th versus the 50th percentile, the increases are 8.44% and 8.22% for the two specifications. For years of post-secondary education, the shift from a 10th to 90th/50th percentile school induces increases of 0.37 and 0.19 years, which correspond to 0.22 and 0.11 standard deviation, respectively. These estimates suggest non-trivial effects of school/neighborhood quality on how many years of education students attain after graduating from high school.

The estimates in this paper differ from those in Altonji and Mansfield to some degree. For instance, they report the point estimates of 0.121 and 0.125 for log wage increase by moving from the 10th percentile to the 90th percentile, which contrast with 0.158 and 0.162 in this paper. The discrepancy arises because Altonji and Mansfield use a random effects model to estimate the variance of group-level unobservable term whereas I use the ordinary least squares (OLS) method. When I employ the same estimation approach, the quadratic specification still produces estimates very similar to the ones based on the specification of Altonji and Mansfield (see Table II.3). I choose the OLS results as my main estimates for consistency with the theoretical results proven in this paper. However, one can expect that under appropriate conditions, methods based on random effects models will be valid using the control function method.

Differences in Assumptions

In this subsection, I discuss the main differences between the assumptions in this paper and those of Altonji and Mansfield (2018). Their key conditions include linearity of the regression functions $\mathbb{E}[W_{ig}|\Theta_{ig}], \mathbb{E}[\omega_{ig}|\Theta_{ig}]$ and what they term "spanning assumption." To define the spanning assumption, introduce the new random vector U_{ig} and write

$$\begin{aligned}\omega_{ig} &= c'U_{ig} \\ \Theta_{ig} &= \Gamma W_{ig} + \Delta U_{ig} + \nu_{ig}\end{aligned}$$

where $c \in \mathbb{R}^{d_u}$, $\Gamma \in \mathbb{R}^{d_\theta \times d_w}$, and $\Delta \in \mathbb{R}^{d_\theta + d_u}$ are fixed parameters, and $\nu_{ig} \in \mathbb{R}^{d_\theta}$ is a random vector uncorrelated with $(W_{ig}', U_{ig}')'$. Let Π_{UW} be the matrix of coefficients for linear projection of U_{ig} onto W_{ig}. Then, the spanning assumption of Altonji and Mansfield is that

there exists a matrix R such that

$$\Delta = (\Gamma + \Delta\Pi_{UW})R.$$

Intuitively, the spanning assumption requires that for each element of A_s, if U_i influences the taste coefficient for that element of A_s, then either one of them has to hold: W_i affects the taste coefficient directly by Γ having non-zero elements or indirectly by non-zero correlation with U_i.

In this paper, I use an injectivity condition (Assumption 3) as well as restrictions on permissible classes of densities and conditional expectations to construct a control function. Unlike Altonji and Mansfield, I do not impose linearity in conditional expectations and instead use restrictions on the distribution, which makes the spanning assumption and Assumption 3 quite different. In fact, there are examples where one of the two holds but the other fails. However, if $(W_{ig}', U_{ig}', \nu_{ig}')'$ are jointly normal, the two conditions coincide. Despite the differences, Lemma 1 nests Proposition 1 in Altonji and Mansfield and therefore, the estimator proposed in this paper is generally more robust than the one in their paper.

II.5.2 Monte Carlo Experiments

In this subsection, I present results of simulation studies to investigate finite-sample properties of the proposed estimator. For the Monte Carlo experiments, I first generate characteristics of individuals and groups based on distributions mimicking the empirical distributions of the NLS72 data. Then, for each individual, I compute utility functions of choosing different groups and then assign people to groups to maximize the sum of utilities under group size constraints. This design builds on Altonji and Mansfield, and the realized allocation can be viewed as an approximation to the equilibrium of a competitive market through price mediation.

I consider three data generating processes (DGPs). The first design has $\mathbb{E}[\omega_i|J_i, \mathbf{A}] = \pi'\mathbb{E}[W_i|J_i, \mathbf{A}]$ for some non-stochastic vector π, and thus the linear specification of Altonji and Mansfield is correct. For the second DGP, I take ω_i to be a linear combination of some normal random variable and binary variable $\mathbb{1}\{W_i'\tau > \nu_i\}$ where τ is a non-stochastic vector and ν_i is an independent normal random variable. This design is meant to capture a more realistic scenario in which non-linearity may be a issue. Here non-linearity arises due to discreteness of part of ω_i, though ω_i itself is continuously distributed. For the third DGP, I model $\mathbb{E}[\omega_i|J_i, \mathbf{A}] = \pi_1\mathbb{E}[W_i^2|J_i, \mathbf{A}] + \pi_2\mathbb{E}[W_i^3|J_i, \mathbf{A}]$ to see how deviation from linearity affects the estimator.

Table II.2 presents the mean squared error (MSE) and the coverage probability of the

95% confidence intervals based on the proposed estimator for three different specifications: one with perfect control on selection bias (i.e., "oracle" estimator), the linear specification of Altonji and Mansfield, and the one including additional series approximation terms. For DGP 1, linear specification is correct and including additional control variables does not alter the results much. For the second DGP, the group-level expectation of ω_i is $\Phi(W_i'\tau)$ where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and both specifications are mis-specified. Yet, the estimator with cubic splines controls outperforms the linear one. For the third design, I made the DGP particularly difficult for the linear specification to see how deviation from linearity affects performance of the estimator. The estimator based on cubic spline expansion performs much better in terms of MSE and coverage probability than the linear specification. Although this Monte Carlo study is small-scale, it indicates usefulness of including additional control variables to reliably estimate group-level partial effects.

II.6 Extensions

II.6.1 Instrumental Variables for Omitted Variable Bias

In the previous sections, I assume away omitted variable bias by imposing exogeneity of χ_{sg}. If instrument variables Z_{sg} for χ_{sg} are available, combination of the control function and IV methods identifies the partial effects of group-level covariates under the presence of both selection and omitted variable biases.

Assumption 9. *1. In addition to* $\{(Y_{ig}, X_{ig}', W_{ig}', J_{ig})'\}_{1\le i\le N, 1\le g\le G}$*, a researcher observes* $\{Z_{sg}\}_{s\in\mathcal{S}, 1\le g\le G}$ *and, after redefining* $\mathbf{B}_g$ *by including* $\mathbf{Z}_g = \{Z_{sg} : s \in \mathcal{S}\}$*, Assumption 1 holds.*

2. For all $s \in \mathcal{S}$*,* $\mathbb{E}[\varepsilon_{isg}|F_{W|J,\mathbf{B}}(\cdot|s, \mathbf{B}_g), J_{ig} = s] = 0$ *and* $\mathbb{E}[\tilde{Z}_{ig}\varepsilon_{isg}|J_{ig} = s] = 0$ *where* $\tilde{Z}_{ig} = (Z_{ig}', W_{ig}')'$*.*

3. Recall $V_{ig} = (X_{ig}', W_{ig}')'$*. The matrix* $\mathbb{E}[\tilde{Z}_{ig}\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)]\}']$ *is of full column rank.*

Theorem 4. *Under the hypothesis of Lemma 1 and Assumption 9,* β *in* (II.2) *is identified.*

This result is a straightforward extension of Theorem 1 but can have an important application. For instance, if a policy intervention creates an exogenous variation in teacher assignment to different schools and households select schools after the intervention, then

omitted variable bias (i.e., correlation between X_{sg} and χ_{sg}, where X_{sg} measures school-level teacher quality) can be resolved through IV but selection bias remains problematic for identification of coefficients on X_{sg}. Theorem 4 establishes that the combination of IV and control function methods achieves identification of group-level partial effects.

II.6.2 Nonparametric Identification

I extend the identification result to a more general outcome equation

$$Y_{isg} = m(X_{sg}, W_{ig}, \varepsilon_{isg})$$
$$J_{ig} = J(\mathbf{B}_{sg}, \Theta_{ig}, \boldsymbol{\eta}_{ig}).$$

In the linear model, I distinguish among $(\omega_{ig}, \chi_{sg}, \epsilon_{isg})$ but in this model ε_{isg} subsumes all the unobservable components due to its nonseparability. In this model, a family of parameters can be defined as

$$M(x) := \int m(x, w, e) f_{W\varepsilon}(w, e) d(w, e) \qquad x \in \mathcal{X}$$

where $f_{W\varepsilon}$ is the joint density of $(W_{ig}, \varepsilon_{isg})$ and $\mathcal{X} \subset \text{supp}(X_{sg})$ is some non-empty set. This object is called Average Structural Function (ASF) in the literature and it summarizes partial effects of a covariate X on the outcome Y. For identification, I impose the following conditions.

Assumption 10. *1.* $(W_{ig}, \boldsymbol{\varepsilon}_{isg}) \perp\!\!\!\perp \boldsymbol{\eta}_{ig} | \Theta_{ig}, \mathbf{B}_g$ *and* $(W_{ig}, \boldsymbol{\varepsilon}_{isg}, \Theta_{ig}) \perp\!\!\!\perp \mathbf{B}_g$. *Also,* $(W_{ig}, \varepsilon_{isg})$ *has identical distributions across* $s \in \mathcal{S}$.

2. $\sup_{s \in \mathcal{S}, \mathbf{b} \in \text{supp}(\mathbf{B}_g)} \mathbb{E}[|m(x, W_{ig}, \varepsilon_{isg})| | J_{ig} = s, \mathbf{B}_g = \mathbf{b}] < \infty$.

3. Given non-empty $\mathcal{X} \subset \text{supp}(X_{sg})$, $\text{supp}(f_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g))$ *is invariant conditional on* $X_{ig} = x$ *for* $x \in \mathcal{X}$.

Assumption 1 imposes independence between χ_{sg} and X_{sg}, stronger than mean exogeneity of χ_{sg} given X_{sg}. Thus, I assume away omitted variable bias and focus on the issue of selection bias. Imposition of identical distributions across $s \in \mathcal{S}$ simplifies some arguments and interpretation of the parameter. Assumption 2 is used to justify interchanging orders of certain integrations. The requirement 3, which states that the support of $f_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)$ is invariant conditional on X_{ig}, is referred to as support condition in the literature on triangular models, and it is considered to be a stringent assumption. Still, it holds for some special case. If the conditional support of $\mathbf{A}$ given $X_s = x$ does not vary with x, then the support

condition holds. This sufficient condition excludes, among other things, that X_s is part of A_s. This may not be so restrictive if individuals only observe a coarse version of X_s when they make group decision. That is, A_s is a noisy measure of X_s.

The following theorem states the identification result for the nonseparable model.

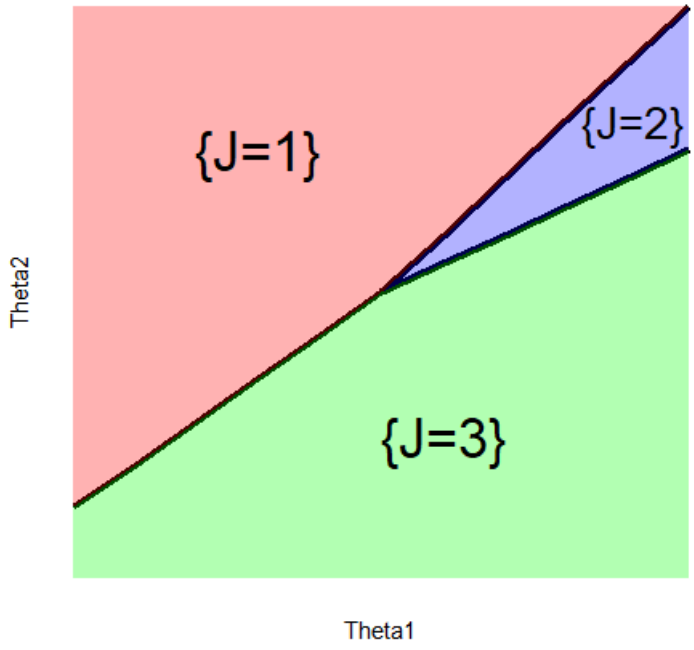
Theorem 5. *If Assumptions 1, 2, 3, and 10 hold, $M(x)$ is identified for $x \in \mathcal{X}$.*

If the support condition is not satisfied, we can still identify different versions of ASF as done in the literature. For instance, a conditional version of ASF is

$$\int m(x, w, e) f_{W\varepsilon|f_{W|J,\mathbf{B}}}(w, e|f) d(w, e) \qquad x \in \mathcal{X}$$

where I condition on the random function $f_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)$. We can see this parameter as a measure of the partial effect conditioning on a within-group distribution of individual characteristics. For example, if ε_{isg} represents an unobservable measure of student's motivation, then the conditional ASF represents the average outcome for different levels of group-level covariate X, fixing the within-school distribution of student's motivation.

Figure II.1. Support of Θ Partitioned by Group Choice



Notes. This is an example of partitioning of the support of Θ_i by $\{J_i = 1\}$, $\{J_i = 2\}$, and $\{J_i = 3\}$ ($S = 3$). The above figure is created based on $A_1 = (1, 3)$, $A_2 = (3, 2)$, and $A_3 = (4, 1)$ with $J_i = \arg\max_{s \in \{1,2,3\}} \{\Theta_i' A_s\}$. If Θ_i falls onto the red region, $J_i = 1$ will be chosen. Similarly, the blue region represents $J_i = 2$, and the green region means $J_i = 3$.

Table II.1. Estimates of School/Neighborhood Effects on Wage and Years of Post Secondary Education

	Wage w/ PS. Ed.		**Wage w/o PS. Ed.**		**Yrs. PS. Ed.**	
	AM	This Paper	AM	This Paper	AM	This Paper
Q10-90						
	0.157	0.162	0.153	0.158	0.352	0.369
	(0.017)	(0.017)	(0.017)	(0.017)	(0.079)	(0.089)
Q10-50						
	0.078	0.081	0.076	0.079	0.176	0.185
	(0.009)	(0.008)	(0.009)	(0.008)	(0.039)	(0.045)
Mean						
	2.878		2.878		1.836	

Notes. The table shows various estimates for effects of school/neighborhood quality on early adulthood wage and years of post-secondary education. The first and second row groups ("Q10-90" and "Q10-50") correspond to the impact of moving from the 10th percentile school/neighborhood to the 90th/50th percentile, respectively. The top headers indicate outcome variables in regression. "Wage w/ PS. Ed." and "Wage w/o PS. Ed." refer to log wage with/without post-secondary education as a control, respectively, and "Yrs. PS. ED." denotes years of post-secondary education. "AM" refers to the specification used in Altonji and Mansfield (2018) and "This Paper" specification adds within-school means of interaction and squared terms of W_i variables. For each row group, the number in the first line is the estimate and the number in parentheses represents the standard error estimate. The last row is the average of dependent variables in the data sample.

Table II.2. Results of Monte Carlo Experiments

	Specification		
	Oracle	AM	This Paper
DGP 1			
MSE	1.00	1.69	1.75
Coverage	0.956	0.914	0.913
DGP 2			
MSE	1.00	3.88	2.36
Coverage	0.942	0.876	0.951
DGP 3			
MSE	1.00	23.32	4.45
Coverage	0.959	0.498	0.922

Notes. The table presents the results of simulation studies, particularly the mean squared error (MSE) and coverage probability of the confidence interval for different specifications. The columns "Oracle" represents infeasible regression with unobserved heterogeneity ω_i, the column "AM" corresponds to the specification of Altonji and Mansfield, and the column "This Paper" denotes the specification with group means of individual variables using cubic polynomial splines. Each row group presents different data generating processes, the row "MSE" presents (relative) MSE of the corresponding estimator divided by the MSE of the "oracle" specification, and the row "Coverage" is the coverage probability of the 95% confidence interval constructed from the corresponding estimator. The sample size is $G = 1,000$ and for each city, $N = 300$ and $S = 3$ with each group's size constraint being no more than 115 people.

Table II.3. Empirical Application: Random Effects Model

	Yrs. PS. Ed.		**Wage w/o PS. Ed.**		**Wage w/ PS. Ed.**	
	AM	This Paper	AM	This Paper	AM	This Paper
Q10-90						
No Unobs.	0.215	0.267	0.121	0.123	0.125	0.122
w/ Unobs.	0.503	0.516	0.203	0.214	0.203	0.207
Q10-50						
No Unobs.	0.107	0.134	0.061	0.061	0.063	0.061
w/ Unobs.	0.251	0.258	0.102	0.107	0.102	0.104
Sample Means						
	1.620		2.880		2.880	

Notes. The estimates come from the random effects model assuming normality of the error terms. The column header indicates dependent variable used and "Q10-90" and "Q10-50" refer to effects of shifting from the 10th percentile neighborhood to the 90th/50th percentile neighborhood. The estimates are based on the lower bound on variance as described in Section II.5.1. "No Unobs" rows display variance lower bound estimates excluding the random effect estimate from school/neighborhood contribution. "w/ Unobs" includes the random effect estimate as part of school/neighborhood contribution.

CHAPTER III

Bootstrap-based Inference for Cube Root Consistent Estimators

III.1 Introduction

In a seminal paper, Kim and Pollard (1990) studied a class of M-estimators exhibiting cube root asymptotics. These estimators not only have a non-standard rate of convergence, but also have the property that rather than being Gaussian their limiting distributions are of Chernoff (1964) type; that is, the limiting distribution is that of the maximizer of a Gaussian process. In fact, in leading examples of cube root consistent estimators such as the maximum score estimator of Manski (1975), the covariance kernel of the Gaussian process characterizing the limiting distribution depends on an infinite-dimensional nuisance parameter. From the perspective of inference, this feature of the limiting distribution represents a nontrivial complication relative to the conventional asymptotically normal case, where the limiting distribution is known up to the value of a finite-dimensional nuisance parameter (namely, the covariance matrix of the limiting distribution). In particular, the dependence of the limiting distribution on an infinite-dimensional nuisance parameter implies that resampling-based distributional approximations seem to offer the most attractive approach to inference in estimation problems exhibiting cube root asymptotics. The purpose of this note is to propose an easy-to-implement bootstrap-based distributional approximation applicable in such cases.

As does the familiar nonparametric bootstrap, the method proposed herein employs bootstrap samples of size n from the empirical distribution function. But unlike the nonparametric bootstrap, which is inconsistent in general (e.g., Abrevaya and Huang, 2005;

Léger and MacGibbon, 2006), my method offers a consistent distributional approximation for cube root consistent estimators and therefore has the advantage that it can be used to construct asymptotically valid inference procedures. Consistency is achieved by altering the shape of the criterion function defining the estimator whose distribution I seek to approximate. Heuristically, the method is designed to ensure that the bootstrap version of a certain empirical process has a mean resembling the large sample version of its population counterpart. The latter is quadratic in the problems I study, and known up to the value of a certain matrix. As a consequence, the only ingredient needed to implement the proposed "reshapement" of the objective function is a consistent estimator of the unknown matrix entering the quadratic mean of the empirical process. Such estimators turn out to be generically available and easy to compute.

The chapter proceeds as follows. Section III.2 is heuristic in nature and serves the purpose of outlining the main idea underlying my approach in the M-estimation setting of Kim and Pollard (1990). Section III.3 then makes the heuristics of Section III.2 rigorous in a more general setting where the M-estimation problem is formed using a possibly n-varying (observation specific) objective function, as recently studied by Seo and Otsu (2018). Section III.4 discusses two examples covered by my general results, namely the maximum score estimator of Manski (1975) and the conditional maximum score estimator of Honoré and Kyriazidou (2000). Further discussion of my results is provided in Section III.5. Section III.6 describes an extension to the case of the isotonic density estimator of Grenander (1956). Finally, all derivations and proofs have been collected in Appendix B.1.

III.2 Heuristics

Suppose $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$ is an estimand admitting the characterization

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, M_0(\boldsymbol{\theta}), \qquad M_0(\boldsymbol{\theta}) = \mathbb{E}[m_0(\mathbf{z}, \boldsymbol{\theta})], \tag{III.1}$$

where m_0 is a known function, and where $\mathbf{z}$ is a random vector of which a random sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is available. Studying estimation problems of this kind for non-smooth m_0, Kim

and Pollard (1990) gave conditions under which the M-estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}), \qquad \hat{M}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_0(\mathbf{z}_i, \boldsymbol{\theta}),$$

exhibits cube root asymptotics:

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \arg\max_{\mathbf{s}\in\mathbb{R}^d} \{\mathcal{Q}_0(\mathbf{s}) + \mathcal{G}_0(\mathbf{s})\}, \tag{III.2}$$

where $\rightsquigarrow$ denotes weak convergence, $\mathcal{Q}_0(\mathbf{s})$ is a quadratic form given by

$$\mathcal{Q}_0(\mathbf{s}) = -\frac{1}{2}\mathbf{s}'\mathbf{V}_0\mathbf{s}, \qquad \mathbf{V}_0 = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} M_0(\boldsymbol{\theta}_0),$$

and $\mathcal{G}_0$ is a non-degenerate zero-mean Gaussian process with $\mathcal{G}_0(0) = 0$.

Whereas the matrix $\mathbf{V}_0$ governing the shape of $\mathcal{Q}_0$ is finite-dimensional, the covariance kernel of $\mathcal{G}_0$ in (III.2) typically involves infinite-dimensional unknown quantities. As a consequence, the limiting distribution of $\hat{\boldsymbol{\theta}}_n$ tends to be more difficult to approximate than Gaussian distributions, implying in turn that basing inference on $\hat{\boldsymbol{\theta}}_n$ is more challenging under cube root asymptotics than in the more familiar case where $\hat{\boldsymbol{\theta}}_n$ is ($\sqrt{n}$-consistent and) asymptotically normally distributed.

As a candidate method of approximating the distribution of $\hat{\boldsymbol{\theta}}_n$, consider the nonparametric bootstrap. To describe it, let $\mathbf{z}^*_{1,n}, \ldots, \mathbf{z}^*_{n,n}$ denote a random sample from the empirical distribution of $\mathbf{z}_1, \ldots \mathbf{z}_n$ and let the natural bootstrap analogue of $\hat{\boldsymbol{\theta}}_n$ be denoted by

$$\hat{\boldsymbol{\theta}}^*_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}^*_n(\boldsymbol{\theta}), \qquad \hat{M}^*_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_0(\mathbf{z}^*_{i,n}, \boldsymbol{\theta}).$$

Then the nonparametric bootstrap estimator of $\mathbb{P}[\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \leq \cdot]$ is given by $\mathbb{P}^*[\hat{\boldsymbol{\theta}}^*_n - \hat{\boldsymbol{\theta}}_n \leq \cdot]$, where $\mathbb{P}^*$ denotes a probability computed under the bootstrap distribution conditional on the data. As is well documented, however, this estimator is inconsistent under cube root asymptotics (e.g., Abrevaya and Huang, 2005; Léger and MacGibbon, 2006).

For the purpose of giving a heuristic, yet constructive, explanation of the inconsistency of the nonparametric bootstrap, it is helpful to recall that a proof of (III.2) can be based on

the representation

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max}\{Q_n(\mathbf{s}) + \hat{G}_n(\mathbf{s})\}, \tag{III.3}$$

where (for $\mathbf{s}$ such that $\boldsymbol{\theta}_0 + \mathbf{s}n^{-1/3} \in \boldsymbol{\Theta}$)

$$\hat{G}_n(\mathbf{s}) = n^{2/3}[\hat{M}_n(\boldsymbol{\theta}_0 + \mathbf{s}n^{-1/3}) - \hat{M}_n(\boldsymbol{\theta}_0) - M_0(\boldsymbol{\theta}_0 + \mathbf{s}n^{-1/3}) + M_0(\boldsymbol{\theta}_0)]$$

is a zero-mean random process, while

$$Q_n(\mathbf{s}) = n^{2/3}[M_0(\boldsymbol{\theta}_0 + \mathbf{s}n^{-1/3}) - M_0(\boldsymbol{\theta}_0)]$$

is a non-random function that is correctly centered in the sense that $\arg\max_{\mathbf{s}\in\mathbb{R}^d} Q_n(\mathbf{s}) = \mathbf{0}$. In cases where m_0 is non-smooth but M_0 is smooth, Q_n and $\hat{G}_n$ are usually asymptotically quadratic and asymptotically Gaussian, respectively, in the sense that

$$Q_n(\mathbf{s}) \to \mathcal{Q}_0(\mathbf{s}) \tag{III.4}$$

and

$$\hat{G}_n(\mathbf{s}) \rightsquigarrow \mathcal{G}_0(\mathbf{s}). \tag{III.5}$$

Under regularity conditions ensuring among other things that the convergence in (III.4) and (III.5) is suitably uniform in $\mathbf{s}$, result (III.2) then follows from an application of a continuous mapping-type theorem for the arg max functional to the representation in (III.3).

Similarly to (III.3), the bootstrap analogue of $\hat{\boldsymbol{\theta}}_n$ admits a representation of the form

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) = \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max}\{\hat{Q}_n(\mathbf{s}) + \hat{G}_n^*(\mathbf{s})\},$$

where (for $\mathbf{s}$ such that $\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3} \in \boldsymbol{\Theta}$)

$$\hat{G}_n^*(\mathbf{s}) = n^{2/3}[\hat{M}_n^*(\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3}) - \hat{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3}) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n)]$$

and

$$\hat{Q}_n(\mathbf{s}) = n^{2/3}[\hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3}) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n)].$$

Under mild conditions, $\hat{G}_n^*$ satisfies the following bootstrap counterpart of (III.5):

$$\hat{G}_n^*(\mathbf{s}) \rightsquigarrow_{\mathbb{P}} \mathcal{G}_0(\mathbf{s}), \tag{III.6}$$

where $\rightsquigarrow_{\mathbb{P}}$ denotes conditional weak convergence in probability (defined as van der Vaart and Wellner, 1996, Section 2.9). On the other hand, although $\hat{Q}_n$ is non-random under the bootstrap distribution and satisfies $\arg\max_{\mathbf{s}\in\mathbb{R}^d} \hat{Q}_n(\mathbf{s}) = \mathbf{0}$, it turns out that $\hat{Q}_n(\mathbf{s}) \nrightarrow_{\mathbb{P}} \mathcal{Q}_0(\mathbf{s})$ in general. In other words, the natural bootstrap counterpart of (III.4) typically fails and, as a partial consequence, so does the natural bootstrap counterpart of (III.2); that is, $\sqrt[3]{n}(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \not\rightsquigarrow_{\mathbb{P}} \arg\max_{\mathbf{s}\in\mathbb{R}^d}\{\mathcal{Q}_0(\mathbf{s}) + \mathcal{G}_0(\mathbf{s})\}$ in general.

To the extent that the implied inconsistency of the bootstrap can be attributed to the fact that the shape of $\hat{Q}_n$ fails to replicate that of Q_n, it seems plausible that a consistent bootstrap-based distributional approximation can be obtained by basing the approximation on

$$\tilde{\boldsymbol{\theta}}_n^* = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{M}_n^*(\boldsymbol{\theta}), \qquad \tilde{M}_n^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \tilde{m}_n(\mathbf{z}_{i,n}^*, \boldsymbol{\theta}),$$

where $\tilde{m}_n$ is a suitably "reshaped" version of m_0 satisfying two properties. First, $\tilde{Q}_n$ should be asymptotically quadratic, where $\tilde{Q}_n$ is the counterpart of $\hat{Q}_n$ associated with $\tilde{m}_n$:

$$\tilde{Q}_n(\mathbf{s}) = n^{2/3}[\tilde{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3}) - \tilde{M}_n(\hat{\boldsymbol{\theta}}_n)], \qquad \tilde{M}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \tilde{m}_n(\mathbf{z}_i, \boldsymbol{\theta}).$$

Second, $\tilde{G}_n^*$ should be asymptotically equivalent to $\hat{G}_n^*$, where

$$\tilde{G}_n^*(\mathbf{s}) = n^{2/3}[\tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3}) - \tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \tilde{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}n^{-1/3}) + \tilde{M}_n(\hat{\boldsymbol{\theta}}_n)],$$

is the counterpart of $\hat{G}_n^*$ associated with $\tilde{m}_n$.

Accordingly, let

$$\tilde{m}_n(\mathbf{z},\boldsymbol{\theta}) = m_0(\mathbf{z},\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\tilde{\mathbf{V}}_n(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n),$$

where $\tilde{\mathbf{V}}_n$ is an estimator of $\mathbf{V}_0$. Then

$$\sqrt[3]{n}(\tilde{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) = \arg\max_{\mathbf{s}\in\mathbb{R}^d}\{\tilde{Q}_n(\mathbf{s}) + \tilde{G}_n^*(\mathbf{s})\},$$

where, by construction, $\tilde{G}_n^*(\mathbf{s}) = \hat{G}_n^*(\mathbf{s})$ and $\tilde{Q}_n(\mathbf{s}) = -\mathbf{s}'\tilde{\mathbf{V}}_n\mathbf{s}/2$. Because $\tilde{G}_n^* = \hat{G}_n^*$, $\tilde{G}_n^*(\mathbf{s}) \rightsquigarrow_{\mathbb{P}} \mathcal{G}_0(\mathbf{s})$ whenever (III.6) holds. In addition, $\tilde{Q}_n(\mathbf{s}) \to_{\mathbb{P}} \mathcal{Q}_0(\mathbf{s})$ provided $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$. As a consequence, it seems plausible that $\mathbb{P}^*[\tilde{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n \leq \cdot]$ is a consistent estimator of $\mathbb{P}[\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \leq \cdot]$ if $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$.

III.3 Main Result

When making the heuristics of Section III.2 precise, it is of interest to consider the more general situation where the estimator $\hat{\boldsymbol{\theta}}_n$ is an approximate maximizer (with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$) of

$$\hat{M}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_n(\mathbf{z}_i,\boldsymbol{\theta}),$$

where m_n is a known function, and where $\mathbf{z}_1,\ldots,\mathbf{z}_n$ is a random sample of a random vector $\mathbf{z}$. This formulation of $\hat{M}_n$, which reduces to that of Section III.2 when m_n does not depend on n, is adopted in order to cover certain estimation problems where, rather than admitting a characterization of the form (III.1), the estimand $\boldsymbol{\theta}_0$ admits the characterization

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} M_0(\boldsymbol{\theta}), \qquad M_0(\boldsymbol{\theta}) = \lim_{n\to\infty} M_n(\boldsymbol{\theta}), \qquad M_n(\boldsymbol{\theta}) = \mathbb{E}[m_n(\mathbf{z},\boldsymbol{\theta})].$$

In other words, the setting considered in this section is one where $\hat{\boldsymbol{\theta}}_n$ approximately maximizes a function $\hat{M}_n$ whose population counterpart M_n can be interpreted as a regularization (in the sense of Bickel and Li, 2006) of a function M_0 whose maximizer $\boldsymbol{\theta}_0$ is the object of interest. The additional flexibility (relative to the more traditional M-estimation

setting of Section III.2) afforded by the present setting is attractive because it allows us to formulate results that cover local M-estimators such as the conditional maximum score estimator of Honoré and Kyriazidou (2000). Studying this type of setting, Seo and Otsu (2018) gave conditions under which $\hat{\boldsymbol{\theta}}_n$ converges at a rate equal to the cube root of the "effective" sample size and has a limiting distribution of Chernoff (1964) type. Analogous conclusions will be drawn below, albeit under slightly different conditions.

For any n and any $\delta > 0$, define $\mathcal{M}_n = \{m_n(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, $\bar{m}_n(\mathbf{z}) = \sup_{m \in \mathcal{M}_n} |m(\mathbf{z})|$, $\boldsymbol{\Theta}_0^\delta = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\}$, $\mathcal{D}_n^\delta = \{m_n(\cdot, \boldsymbol{\theta}) - m_n(\cdot, \boldsymbol{\theta}_0) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^\delta\}$, and $\bar{d}_n^\delta(\mathbf{z}) = \sup_{d \in \mathcal{D}_n^\delta} |d(\mathbf{z})|$.

Condition CRA (Cube Root Asymptotics) For a positive q_n with $r_n = \sqrt[3]{nq_n} \to \infty$, the following are satisfied:

(i) $\{\mathcal{M}_n : n \geq 1\}$ is uniformly manageable for the envelopes $\bar{m}_n$ and $q_n \mathbb{E}[\bar{m}_n(\mathbf{z})^2] = O(1)$.

Also, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})| = o(1)$ and, for every $\delta > 0$, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0^\delta} M_0(\boldsymbol{\theta}) < M_0(\boldsymbol{\theta}_0)$.

(ii) $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$ and, for some $\delta > 0$, M_0 and M_n are twice continuously differentiable on $\boldsymbol{\Theta}_0^\delta$ and $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^\delta} \|\partial^2 [M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})]/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'\| = o(1)$.

Also, $r_n \|\partial M_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\| = o(1)$ and $\mathbf{V}_0 = -\partial^2 M_0(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ is positive definite.

(iii) For some $\delta > 0$, $\{\mathcal{D}_n^{\delta'} : n \geq 1, 0 < \delta' \leq \delta\}$ is uniformly manageable for the envelopes $\bar{d}_n^{\delta'}$ and $q_n \sup_{0 < \delta' \leq \delta} \mathbb{E}[\bar{d}_n^{\delta'}(\mathbf{z})^2/\delta'] = O(1)$.

(iv) For every positive δ_n with $\delta_n = O(r_n^{-1})$, $q_n^2 \mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^3] + q_n^3 r_n^{-1} \mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^4] = o(1)$, and, for all $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ and for some $\mathcal{C}_0$ with $\mathcal{C}_0(\mathbf{s}, \mathbf{s}) + \mathcal{C}_0(\mathbf{t}, \mathbf{t}) - 2\mathcal{C}_0(\mathbf{s}, \mathbf{t}) > 0$ for $\mathbf{s} \neq \mathbf{t}$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta_n}} |q_n \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{s}) - m_n(\mathbf{z}, \boldsymbol{\theta})\}\{m_n(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{t}) - m_n(\mathbf{z}, \boldsymbol{\theta})\}/\delta_n] - \mathcal{C}_0(\mathbf{s}, \mathbf{t})| = o(1).$$

(v) For every positive δ_n with $\delta_n = O(r_n^{-1})$,

$$\lim_{C \to \infty} \limsup_{n \to \infty} \sup_{0 < \delta \leq \delta_n} q_n \mathbb{E}[\mathbb{1}\{q_n \bar{d}_n^\delta(\mathbf{z}) > C\} \bar{d}_n^\delta(\mathbf{z})^2/\delta] = 0$$

and $\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\boldsymbol{\Theta}_0^{\delta_n}} \mathbb{E}[|m_n(\mathbf{z},\boldsymbol{\theta}) - m_n(\mathbf{z},\boldsymbol{\theta}')|]/\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| = O(1)$.

To interpret Condition CRA, consider first the benchmark case where $m_n = m_0$ and $q_n = 1$. In this case, the condition is similar to, but slightly stronger than, assumptions (ii)-(vii) of the main theorem of Kim and Pollard (1990), to which the reader is referred for a definition of the term (uniformly) manageable. The most notable difference between Condition CRA and the assumptions employed by Kim and Pollard (1990) is probably that part (iv) contains assumptions about moments of orders three and four, that the displayed part of part (iv) is a locally uniform (with respect to $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$) version of its counterpart in Kim and Pollard (1990), and that (i) can be thought of as replacing the high level condition $\hat{\boldsymbol{\theta}}_n \to_{\mathbb{P}} \boldsymbol{\theta}_0$ of Kim and Pollard (1990) with more primitive conditions that imply it for approximate M-estimators. In all three cases, the purpose of strengthening the assumptions of Kim and Pollard (1990) is to be able to analyze the bootstrap.

More generally, Condition CRA can be interpreted as an n-varying version of a suitably (for the purpose of analyzing the bootstrap) strengthened version of the assumptions of Kim and Pollard (1990). The differences between Condition CRA and the *i.i.d.* version of the conditions in Seo and Otsu (2018) seem mostly technical in nature, but for completeness I highlight two differences here. First, to handle dependent data Seo and Otsu (2018) control the complexity of various function classes using the concept of bracketing entropy. In contrast, because I assume random sampling I can follow Kim and Pollard (1990) and obtain maximal inequalities using bounds on uniform entropy numbers implied by the concept of (uniform) manageability. Second, whereas Seo and Otsu (2018) controls the bias of $\hat{\boldsymbol{\theta}}_n$ through a locally uniform bound on $M_n - M_0$, Condition CRA controls the bias through the first and second derivatives of $M_n - M_0$.

Under Condition CRA, the effective sample size is given by nq_n. In perfect agreement with Seo and Otsu (2018), it turns out that if $\hat{\boldsymbol{\theta}}_n$ is an approximate maximizer of $\hat{M}_n$, then

$$r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \arg\max_{\mathbf{s}\in\mathbb{R}^d} \{\mathcal{Q}_0(\mathbf{s}) + \mathcal{G}_0(\mathbf{s})\}, \tag{III.7}$$

where $\mathcal{Q}_0(\mathbf{s}) = -\mathbf{s}'\mathbf{V}_0\mathbf{s}/2$, and where $\mathcal{G}_0$ is a zero-mean Gaussian process with $\mathcal{G}_0(0) = 0$ and covariance kernel $\mathcal{C}_0$. The heuristics of the previous section are rate-adaptive, so once again

it stands to reason that if $\tilde{V}_n$ is a consistent estimator of V_0, then a consistent distributional approximation can be based on an approximate maximizer $\tilde{\boldsymbol{\theta}}_n^*$ of

$$\tilde{M}_n^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{m}_n(\mathbf{z}_{i,n}^*, \boldsymbol{\theta}), \qquad \tilde{m}_n(\mathbf{z}, \boldsymbol{\theta}) = m_n(\mathbf{z}, \boldsymbol{\theta}) - \hat{M}_n(\theta) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\tilde{\mathbf{V}}_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n),$$

where $\mathbf{z}_{1,n}^*, \ldots, \mathbf{z}_{n,n}^*$ is a random sample from the empirical distribution of $\mathbf{z}_1, \ldots \mathbf{z}_n$.

Following (van der Vaart, 1998, Chapter 23), I say that the bootstrap-based estimator of the distribution of $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is consistent if

$$\sup_{\mathbf{t}\in\mathbb{R}^d} \left| \mathbb{P}^*[r_n(\tilde{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \leq \mathbf{t}] - \mathbb{P}[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \leq \mathbf{t}] \right| \to_{\mathbb{P}} 0. \tag{III.8}$$

Because the limiting distribution in (III.7) is continuous, this consistency property implies consistency of bootstrap-based confidence intervals (e.g., van der Vaart, 1998, Lemma 23.3). Moreover, continuity of the limiting distribution implies that (III.8) holds provided the estimator $\tilde{\boldsymbol{\theta}}_n^*$ satisfies the following bootstrap counterpart of (III.7):

$$r_n(\tilde{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \rightsquigarrow_{\mathbb{P}} \arg\max_{\mathbf{s}\in\mathbb{R}^d} \{\mathcal{Q}_0(\mathbf{s}) + \mathcal{G}_0(\mathbf{s})\}.$$

Theorem 6, the main result, gives sufficient conditions for this to occur.

Theorem 6. *Suppose Condition CRA holds. If* $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$ *and if*

$$\hat{M}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - o_{\mathbb{P}}(r_n^{-2}), \qquad \tilde{M}_n^*(\tilde{\boldsymbol{\theta}}_n^*) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{M}_n^*(\boldsymbol{\theta}) - o_{\mathbb{P}}(r_n^{-2}),$$

then (III.8) holds.

To implement the bootstrap-based approximation to the distribution of $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, only a consistent estimator of $\mathbf{V}_0$ is needed. A generic estimator based on numerical derivatives is $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$, the matrix whose element (k, l) is given by

$$\begin{aligned}\tilde{V}_{n,kl}^{\mathtt{ND}} = -\frac{1}{4\epsilon_n^2}\Big[&\hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{e}_k\epsilon_n + \mathbf{e}_l\epsilon_n) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{e}_k\epsilon_n - \mathbf{e}_l\epsilon_n) \\ &- \hat{M}_n(\hat{\boldsymbol{\theta}}_n - \mathbf{e}_k\epsilon_n + \mathbf{e}_l\epsilon_n) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n - \mathbf{e}_k\epsilon_n - \mathbf{e}_l\epsilon_n)\Big],\end{aligned}$$

where $\mathbf{e}_k$ is the kth unit vector in $\mathbb{R}^d$ and where ϵ_n is a positive tuning parameter. Conditions under which this estimator is consistent are given in the following lemma.

Lemma 2. *Suppose Condition CRA holds. If $\epsilon_n \to 0$ and if $r_n\epsilon_n \to \infty$, then $\tilde{\mathbf{V}}_n^{\mathtt{ND}} \to_{\mathbb{P}} \mathbf{V}_0$.*

More generally, if only consistency is assumed on the part of $\hat{\boldsymbol{\theta}}_n$, then $\tilde{\mathbf{V}}_n^{\mathtt{ND}} \to_{\mathbb{P}} \mathbf{V}_0$ holds provided $\epsilon_n \to 0$ and $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|/\epsilon_n \to_{\mathbb{P}} 0$. The proof of the lemma goes beyond consistency and develops a Nagar-type mean squared error (MSE) expansion for $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ under the additional assumption that M_0 is four times continuously differentiable near $\boldsymbol{\theta}_0$. To be specific, assume that

1. for some $\delta > 0$, M_0 and M_n are four times continuously differentiable on $\boldsymbol{\Theta}_0^\delta$, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\delta} |\partial^4[M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})]/\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}| = o(1)$ for all $j_1, j_2, j_3, j_4 \in \{1, 2, \cdots, d\}$,
2. $\mathcal{C}_0(\mathbf{s}, -\mathbf{s}) = 0$ and $\mathcal{C}_0(\mathbf{s}, \mathbf{t}) = \mathcal{C}_0(-\mathbf{s}, -\mathbf{t})$ for all $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$.

Under the additional conditions, I have the following result.

Lemma 3. *$\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ admits an approximation $\check{\mathbf{V}}_n^{\mathtt{ND}}$ satisfying*

$$\tilde{\mathbf{V}}_n^{\mathtt{ND}} - \check{\mathbf{V}}_n^{\mathtt{ND}} = o_{\mathbb{P}}(r_n^{-3/2}\epsilon_n^{-3/2}) + o(\epsilon_n^2) + O(r_n^{-1})$$

and

$$\mathbb{E}[\|\check{\mathbf{V}}_n^{\mathtt{ND}} - \mathbf{V}_n\|^2] = \epsilon_n^4\Big(\sum_{k=1}^d\sum_{l=1}^d \mathsf{B}_{kl}^2\Big) + r_n^{-3}\epsilon_n^{-3}\Big(\sum_{k=1}^d\sum_{l=1}^d \mathsf{V}_{kl}\Big) + o(\epsilon_n^2 + r_n^{-3}\epsilon_n^{-3}),$$

where $\mathbf{V}_n = -\partial^2 M_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ and

$$\begin{aligned}
\mathsf{B}_{kl} &= -\frac{1}{6}\left(\frac{\partial^4}{\partial\theta_k^3\partial\theta_l}M_0(\boldsymbol{\theta}_0) + \frac{\partial^4}{\partial\theta_k\partial\theta_l^3}M_0(\boldsymbol{\theta}_0)\right), \\
\mathsf{V}_{kl} &= \frac{1}{8}[\mathcal{C}_0(\mathbf{e}_k + \mathbf{e}_l, \mathbf{e}_k + \mathbf{e}_l) + \mathcal{C}_0(\mathbf{e}_k - \mathbf{e}_l, \mathbf{e}_k - \mathbf{e}_l) \\
&\qquad - 2\mathcal{C}_0(\mathbf{e}_k + \mathbf{e}_l, \mathbf{e}_k - \mathbf{e}_l) - 2\mathcal{C}_0(\mathbf{e}_k + \mathbf{e}_l, -\mathbf{e}_k + \mathbf{e}_l)].
\end{aligned}$$

Using this theorem, I immediately obtain the asymptotic MSE-optimal tuning parameter choice

$$\epsilon_n^{\mathtt{AMSE}} = \left(\frac{3\sum_{k=1}^d\sum_{l=1}^d \mathsf{V}_{kl}}{4\sum_{k=1}^d\sum_{l=1}^d \mathsf{B}_{kl}^2}\right)^{1/7} r_n^{-3/7}.$$

As the notation suggests, the constants B_{kl} and V_{kl} correspond to element (k,l) of $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$, so the asymptotic MSE-optimal tuning parameter choice for $\tilde{V}_{n,kl}^{\mathtt{ND}}$ is

$$\epsilon_{n,kl}^{\mathtt{AMSE}} = \left(\frac{3\mathsf{V}_{kl}}{4\mathsf{B}_{kl}^2}\right)^{1/7} r_n^{-3/7}.$$

I close this section by summarizing the algorithm for the proposed bootstrap-based distributional approximation.

Bootstrap-Based Approximation Let the notation and conditions in Theorem 6 hold.

Step 1. Compute $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\mathbf{V}}_n$ using the original sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

Step 2. Compute $\tilde{M}_n^*(\boldsymbol{\theta})$ and let $\tilde{\boldsymbol{\theta}}_n^*$ be an approximate maximizer thereof, both constructed using the nonparametric bootstrap sample $\mathbf{z}_{1,n}^*, \ldots, \mathbf{z}_{n,n}^*$. (Note that $\tilde{\mathbf{V}}_n$ is not recomputed at this step.)

Step 3. Repeat Step 2 B times, and then compute the empirical distribution of $r_n(\tilde{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n)$.

III.4 Examples

This section discusses two econometric examples covered by the main result, namely the maximum score estimator of Manski (1975) and the conditional maximum score estimator of Honoré and Kyriazidou (2000). From the perspective of this paper, the main difference between these examples is that (only) the latter corresponds to a situation where m_n depends on n.

III.4.1 The Maximum Score Estimator

To describe a version of the maximum score estimator of Manski (1975), suppose $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is a random sample of $\mathbf{z} = (y, \mathbf{x}')'$ generated by the binary response model

$$y = \mathbb{1}(\boldsymbol{\beta}_0'\mathbf{x} + u \geq 0), \qquad F_{u|\mathbf{x}}(0|\mathbf{x}) = 1/2, \tag{III.9}$$

where $\boldsymbol{\beta}_0 \in \mathcal{B} \subseteq \mathbb{R}^{d+1}$ is an unknown parameter of interest, $\mathbf{x} \in \mathbb{R}^{d+1}$ is a vector of covariates, and $F_{u|\mathbf{x}}(\cdot|\mathbf{x})$ is the conditional cumulative distribution function of the unobserved error term u given $\mathbf{x}$. Following Abrevaya and Huang (2005), I employ the parameterization $\boldsymbol{\beta}_0 = (1, \boldsymbol{\theta}_0')'$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ is unknown. In other words, I assume that the first element of $\boldsymbol{\beta}_0$ is positive and then normalize the (unidentified) scale of $\boldsymbol{\beta}_0$ by setting its first element equal to unity. Partitioning $\mathbf{x}$ conformably with $\boldsymbol{\theta}_0$ as $\mathbf{x} = (x_1, \mathbf{x}_2')'$, a maximum score estimator of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$ is any $\hat{\boldsymbol{\theta}}_n^{\texttt{MS}}$ approximately maximizing $\hat{M}_n$ for

$$m_n(\mathbf{z}, \boldsymbol{\theta}) = m^{\texttt{MS}}(\mathbf{z}, \boldsymbol{\theta}) = (2y - 1)\mathbb{1}(x_1 + \boldsymbol{\theta}'\mathbf{x}_2 \geq 0).$$

Regarded as a member of the class of M-estimators exhibiting cube root asymptotics, the maximum score estimator is representative. Under easy-to-interpret primitive conditions, the estimator is covered by the results of Section III.3. In particular, because m_n does not depend on n I can set $q_n = 1$ when formulating primitive conditions for Condition CRA. Specifically, I impose the following assumptions.

Condition MS Suppose the maximum score model is as above.

(i) Median$(u|\mathbf{x}) = 0$ and $0 < \mathbb{P}(y = 1|\mathbf{x}) < 1$ almost surely. The conditional distribution function of u given $\mathbf{x}$, denoted by $F_{u|x_1,\mathbf{x}_2}(u|x_1, \mathbf{x}_2)$, is bounded and S_F times continuously differentiable in u and x_1 with bounded derivatives for some $S_F \geq 1$.

(ii) The support of $\mathbf{x}$ is not contained in any proper linear subspace of $\mathbb{R}^{d+1}$, $\mathbb{E}[\|\mathbf{x}_2\|^2] < \infty$, and conditional on $\mathbf{x}_2$, x_1 has everywhere positive Lebesgue density.

(iii) The set $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^d : (1, \boldsymbol{\theta}')' \in \mathcal{B}\}$ is compact and $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$.

(iv) $M_0^{\texttt{MS}}(\boldsymbol{\theta}) = \mathbb{E}[(2y - 1)\mathbb{1}(x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0)]$ is S times continuously differentiable near $\boldsymbol{\theta}_0$ for some $S \geq 2$ and $\mathbf{V}_0^{\texttt{MS}} = -\partial^2 M_0^{\texttt{MS}}(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is positive definite.

(v) The conditional density of x_1 given $\mathbf{x}_2$, denoted by $f_{x_1|\mathbf{x}_2}(x_1|\mathbf{x}_2)$, is bounded and S_f times continuously differentiable in x_1 with bounded derivatives for some $S_f \geq 1$.

Under these assumptions, I can deduce the following result.

Corollary MS. *If Condition MS holds, then*

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}}_n^{\texttt{MS}} - \boldsymbol{\theta}_0) \rightsquigarrow \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max} \left\{ -\frac{1}{2}\mathbf{s}'\mathbf{V}_0^{\texttt{MS}}\mathbf{s} + \mathcal{G}_0(\mathbf{s}) \right\}$$

where $\mathcal{G}_0$ is a mean-zero Gaussian process indexed by $\mathbb{R}^d$ with covariance kernel $\mathcal{C}_0^{\mathtt{MS}}(\cdot,\cdot)$, and

$$\mathbf{V}_0^{\mathtt{MS}} = 2\mathbb{E}\left[\mathbf{x}_2\mathbf{x}_2' f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) f_{\varepsilon|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\right],$$

$$\mathcal{C}_0^{\mathtt{MS}}(\mathbf{s},\mathbf{t}) = \frac{1}{2}(\mathcal{L}_0(\mathbf{s})+\mathcal{L}_0(\mathbf{t})-\mathcal{L}_0(\mathbf{s}-\mathbf{t})), \qquad \mathcal{L}_0(\mathbf{s}) = \mathbb{E}\left[f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)|\mathbf{x}_2'\mathbf{s}|\right].$$

Furthermore, if a consistent estimator for the Hessian term is available (i.e., $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0^{\mathtt{MS}}$ for some estimator $\tilde{\mathbf{V}}_n$), then the reshaped bootstrap consistently estimates the large sample distribution of $\sqrt[3]{n}(\hat{\boldsymbol{\theta}}_n^{\mathtt{MS}} - \boldsymbol{\theta}_0)$.

In addition to the generic estimator $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ discussed above, the maximum score estimator admits an example-specific consistent estimator of the $\mathbf{V}_0$ associated with it. Let

$$\tilde{\mathbf{V}}_n^{\mathtt{MS}} = -\frac{1}{n}\sum_{i=1}^{n}(2y_i-1)\dot{K}_n(x_{1i}+\boldsymbol{\theta}'\mathbf{x}_{2i})\mathbf{x}_{2i}\mathbf{x}_{2i}'\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n^{\mathtt{MS}}},$$

where, for a kernel function K and a bandwidth h_n, $\dot{K}_n(u) = d\dot{K}_n(u)/du$ and $K_n(u) = K(u/h_n)/h_n$. As defined, $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ is simply minus the second derivative, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n^{\mathtt{MS}}$, of the criterion function associated with the smoothed maximum score estimator of Horowitz (1992). The estimator $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ is consistent under mild conditions on h_n and K. Concretely, I assume the following.

Condition K

(i) $\int_{\mathbb{R}} K(v)dv = 1$ and $\lim_{|v|\to\infty} vK(v) = 0$.

(ii) For all $v_1, v_2 \in \mathbb{R}$, $|\dot{K}(v_1)-\dot{K}(v_2)| \le B(v_1)|v_1-v_2|$ with $\int_{\mathbb{R}} B(v)dv + \int_{\mathbb{R}} B(v)^2dv < \infty$.

(iii) $\int_{\mathbb{R}} |\dot{K}(v)|^2dv + \int_{\mathbb{R}} |v\dot{K}(v)|dv < \infty$.

(iv) $\lim_{|v|\to\infty} v^2K(v) = 0$, $\int_{\mathbb{R}} vK(v)dv = 0$, and $\int_{\mathbb{R}} |v|^3|\dot{K}(v)|dv < \infty$.

With these restrictions on the kernel function K, I obtain the following result for $\tilde{\mathbf{V}}_n^{\mathtt{MS}} = \tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_n^{\mathtt{MS}})$, which improves slightly on Theorem **??** because of the additional smoothness imposed. Let $F_{u|x_1,\mathbf{x}_2}^{(s)}(u|x_1,\mathbf{x}_2) = d^sF_{u|x_1,\mathbf{x}_2}(-u|u+x_1,x_2)/du^s$ and $f_{x_1|\mathbf{x}_2}^{(s)}(x_1|\mathbf{x}_2) = d^sf_{x_1|\mathbf{x}_2}(x_1|\mathbf{x}_2)/dx_1^s$.

Lemma MS. *Suppose Conditions MS and K(i)-(iii) hold.*

If $h_n \to 0$, $nh_n^3 \to \infty$, *and* $\mathbb{E}[\|\mathbf{x}_2\|^6] < \infty$, *then* $\tilde{\mathbf{V}}_n^{\mathtt{MS}} \to_{\mathbb{P}} \mathbf{V}_0^{\mathtt{MS}}$.

Furthermore, if Condition K(iv) also holds, $S \geq 4$, $S_F \geq 3$, *and* $S_f \geq 2$, *then*

$$\tilde{\mathbf{V}}_n^{\mathtt{MS}} = \tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) + O_{\mathbb{P}}(n^{-5/6}h_n^{-5/2}) + o_{\mathbb{P}}(h_n^2) + O_{\mathbb{P}}(n^{-1/3}),$$

with the term $O_{\mathbb{P}}(n^{-1/3})$ *independent of* h_n, *and*

$$\mathbb{E}\left[\|\tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) - \mathbf{V}_0^{\mathtt{MS}}\|^2\right] = h_n^4(\sum_{k=1}^{d}\sum_{l=1}^{d}\mathsf{B}_{kl}^2) + \frac{1}{nh_n^3}(\sum_{k=1}^{d}\sum_{l=1}^{d}\mathsf{V}_{kl}) + o(h_n^4 + n^{-1}h_n^{-3})$$

where

$$\begin{aligned}\mathsf{B}_{kl} = \mathbb{E}[(\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)\{&F_{u|x_1,\mathbf{x}_2}^{(1)}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)f_{x_1|\mathbf{x}_2}^{(2)}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\\ &+ F_{u|x_1,\mathbf{x}_2}^{(2)}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)f_{x_1|\mathbf{x}_2}^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\\ &+ \frac{1}{3}F_{u|x_1,\mathbf{x}_2}^{(3)}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\}]\int v^3\dot{K}(v)dv\end{aligned}$$

and

$$\mathsf{V}_{kl} = \mathbb{E}[(\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)^2 f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)]\int|\dot{K}(u)|^2du.$$

To investigate the finite sample properties of the proposed bootstrap-based inference procedures, I conducted a Monte Carlo experiment. Following Horowitz (2002), and to allow for a comparison with his bootstrap-based inference method for the smoothed maximum score estimator, I generate data from a model with $d = 1$, where

$$\mathbf{x} = (x_1, x_2)' \sim \mathcal{N}\left(\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),$$

and where ε can take three distinct distributions. Specifically, DGP 1 sets $\varepsilon \sim \mathsf{Logistic}(0,1)/\sqrt{2\pi^2/3}$, DGP 2 sets $\varepsilon \sim \mathcal{T}_3/\sqrt{3}$, where $\mathcal{T}_k$ denotes a Student's t distribution with k degrees of freedom, and DGP 3 sets $\varepsilon \sim (1+2(x_1+x_2)^2+(x_1+x_2)^4)\mathsf{Logistic}(0,1)/\sqrt{\pi^2/48}$. The parameter is $\boldsymbol{\theta}_0 = 1$ in all cases.

The Monte Carlo experiment employs a sample size $n = 1,000$ with $B = 2,000$ bootstrap replications and $S = 2,000$ simulations. For each of the three DGPs, I implement

the standard non-parametric bootstrap, m-out-of-n bootstrap, and the proposed method using the two estimators $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ and $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ of $\mathbf{V}_0$. I report empirical coverage for nominal 95% confidence intervals and their average interval length. For the case of the proposed procedures, I investigate their performance using both (i) a grid of fixed tuning parameter values (bandwidth/derivative step) around the MSE-optimal choice and (ii) infeasible and feasible AMSE-optimal choices of the tuning parameter.

Table III.1 presents the main results, which are consistent across all three simulation designs. First, as expected, the standard nonparametric bootstrap (labeled "Standard") does not perform well, leading to confidence intervals with an average 64% empirical coverage rate. Second, the m-out-of-n bootstrap (labeled "m-out-of-n") performs somewhat better for small subsamples, but leads to very large average interval length of the resulting confidence intervals. The proposed methods, on the other hand, exhibit excellent finite sample performance in this Monte Carlo experiment. Results employing the example-specific plug-in estimator $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ are presented under the label " Plug-in" while results employing the generic numerical derivative estimator $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ are reported under the label "Num Deriv". Empirical coverage appears stable across different values of the tuning parameters for each method, with better performance in the case of $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$. I conjecture that $n = 1,000$ is too small for the numerical derivative estimator $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ to lead to as good inference performance as $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ (e.g., note that the MSE-optimal choice $\epsilon_{\mathtt{MSE}}$ is greater than 1). Nevertheless, empirical coverage of confidence intervals constructed using the proposed bootstrap-based method is close to 95% in all cases except when $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ is used with either the infeasible asymptotic choice $\epsilon_{\mathtt{AMSE}}$ or its estimated counterpart $\hat{\epsilon}_{\mathtt{AMSE}}$, and with an average interval length of at most half that of any of the m-out-of-n competing confidence intervals. In particular, confidence intervals based on $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ implemented with the feasible bandwidth $\hat{h}_{\mathtt{AMSE}}$ perform quite well across the three DGPs considered.

In sum, applying the bootstrap-based inference methods proposed in this note to the case of the Maximum Score estimator of Manski (1975) lead to confidence intervals with very good coverage and length properties in the simulation designs considered.

Table III.1. Simulations, Maximum Score Estimator, 95% Confidence Intervals.

(a) $n = 1,000$, $S = 2,000$, $B = 2,000$

	DGP 1			DGP 2			DGP 3		
	h, ϵ	Coverage	Length	h, ϵ	Coverage	Length	h, ϵ	Coverage	Length
Standard									
		0.639	0.475		0.645	0.480		0.640	0.247
m-out-of-n									
$m = \lceil n^{1/2} \rceil$		0.998	1.702		0.997	1.754		0.999	1.900
$m = \lceil n^{2/3} \rceil$		0.979	1.189		0.979	1.223		0.985	0.728
$m = \lceil n^{4/5} \rceil$		0.902	0.824		0.894	0.839		0.904	0.447
Plug-in: $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$									
$0.7 \cdot h_{\mathtt{MSE}}$	0.434	0.941	0.501	0.406	0.947	0.513	0.105	0.904	0.256
$0.8 \cdot h_{\mathtt{MSE}}$	0.496	0.946	0.503	0.464	0.952	0.516	0.120	0.917	0.260
$0.9 \cdot h_{\mathtt{MSE}}$	0.558	0.951	0.506	0.522	0.951	0.518	0.135	0.930	0.267
$1.0 \cdot h_{\mathtt{MSE}}$	0.620	0.954	0.510	0.580	0.952	0.522	0.150	0.941	0.273
$1.1 \cdot h_{\mathtt{MSE}}$	0.682	0.959	0.515	0.638	0.955	0.526	0.165	0.948	0.281
$1.2 \cdot h_{\mathtt{MSE}}$	0.744	0.961	0.522	0.696	0.959	0.532	0.180	0.958	0.288
$1.3 \cdot h_{\mathtt{MSE}}$	0.806	0.962	0.531	0.754	0.960	0.539	0.195	0.966	0.296
$h_{\mathtt{AMSE}}$	0.385	0.938	0.499	0.367	0.941	0.510	0.119	0.917	0.260
$\hat{h}_{\mathtt{AMSE}}$	0.446	0.947	0.509	0.415	0.949	0.518	0.155	0.941	0.275
Num Deriv: $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$									
$0.7 \cdot \epsilon_{\mathtt{MSE}}$	0.980	0.912	0.431	0.904	0.891	0.422	0.203	0.864	0.216
$0.8 \cdot \epsilon_{\mathtt{MSE}}$	1.120	0.922	0.442	1.033	0.897	0.432	0.232	0.888	0.228
$0.9 \cdot \epsilon_{\mathtt{MSE}}$	1.260	0.929	0.460	1.163	0.909	0.448	0.261	0.904	0.238
$1.0 \cdot \epsilon_{\mathtt{MSE}}$	1.400	0.939	0.484	1.292	0.919	0.469	0.290	0.917	0.248
$1.1 \cdot \epsilon_{\mathtt{MSE}}$	1.540	0.943	0.514	1.421	0.928	0.497	0.319	0.928	0.257
$1.2 \cdot \epsilon_{\mathtt{MSE}}$	1.680	0.948	0.549	1.550	0.932	0.531	0.348	0.939	0.265
$1.3 \cdot \epsilon_{\mathtt{MSE}}$	1.820	0.955	0.590	1.679	0.935	0.568	0.377	0.947	0.274
$\epsilon_{\mathtt{AMSE}}$	0.483	0.878	0.410	0.476	0.871	0.412	0.216	0.877	0.221
$\hat{\epsilon}_{\mathtt{AMSE}}$	0.518	0.877	0.414	0.513	0.884	0.418	0.368	0.932	0.269

Notes:
(i) Panel **Standard** refers to standard nonparametric bootstrap, Panel **m-out-of-n** refers to m-out-of-n nonparametric bootstrap with subsample m, Panel **Plug-in:** $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ refers to the proposed bootstrap-based implemented using the example-specific plug-in drift estimator, and Panel **Num Deriv:** $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ refers to the proposed bootstrap-based implemented using the generic numerical derivative drift estimator.
(ii) Column "h, ϵ" reports tuning parameter value used or average across simulations when estimated, and Columns "Coverage" and "Length" report empirical coverage and average length of bootstrap-based 95% percentile confidence intervals, respectively.
(iii) $h_{\mathtt{MSE}}$ and $\epsilon_{\mathtt{MSE}}$ correspond to the simulation MSE-optimal choices, $h_{\mathtt{AMSE}}$ and $\epsilon_{\mathtt{AMSE}}$ correspond to the AMSE-optimal choices, and $\hat{h}_{\mathtt{AMSE}}$ and $\hat{\epsilon}_{\mathtt{AMSE}}$ correspond to the ROT feasible implementation of $\hat{h}_{\mathtt{AMSE}}$ and $\hat{\epsilon}_{\mathtt{AMSE}}$ described in Appendix B.2.

III.4.2 The Conditional Maximum Score Estimator

Consider the dynamic binary response model

$$Y_t = \mathbb{1}(\boldsymbol{\beta}_0'\mathbf{X}_t + \gamma_0 Y_{t-1} + \alpha + u_t \geq 0), \qquad t = 1, 2, 3,$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^d$ and $\gamma_0 \in \mathbb{R}$ are unknown parameters of interest, α is an unobserved (time-invariant) individual-specific effect, and u_t is an unobserved error term. Honoré and Kyriazidou (2000) analyzed this model and gave conditions under which $\boldsymbol{\beta}_0$ and γ_0 are identified up to scale. Assuming these conditions hold and assuming the first element of $\boldsymbol{\beta}_0$ is positive, I can normalize that element to unity and employ the parameterization $(\boldsymbol{\beta}_0', \gamma_0)' = (1, \boldsymbol{\theta}_0')'$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ is unknown.

To describe a version of the conditional maximum score estimator of Honoré and Kyriazidou (2000), partition $\mathbf{X}_t$ after the first element as $\mathbf{X}_t = (X_{1t}, \mathbf{X}_{2t}')'$ and define $\mathbf{z} = (y, x_1, \mathbf{x}_2', \mathbf{w}')'$, where $y = Y_2 - Y_1$, $x_1 = X_{12} - X_{11}$, $\mathbf{x}_2 = ((\mathbf{X}_{22} - \mathbf{X}_{21})', Y_3 - Y_0)'$, and $\mathbf{w} = \mathbf{X}_2 - \mathbf{X}_3$. Assuming $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is a random sample of $\mathbf{z}$, a conditional maximum score estimator of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$ is any $\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}}$ approximately maximizing $\hat{M}_n$ for

$$m_n(\mathbf{z}, \boldsymbol{\theta}) = m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta}) = y\mathbb{1}(x_1 + \boldsymbol{\theta}'\mathbf{x}_2 \geq 0)L_n(\mathbf{w}),$$

where, for a kernel function L and a bandwidth b_n, $L_n(\mathbf{w}) = L(\mathbf{w}/b_n)/b_n^d$.

Through its dependence on b_n, the function $m_n^{\texttt{CMS}}$ depends on n in a non-negligible way. In particular, the effective sample size is nb_n^d (rather than n) in the current setting, one would expect primitive sufficient conditions for Condition CRA to include $q_n = b_n^d$ in this example. Apart from this predictable change, the properties of the conditional maximum score estimator $\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}}$ turn out to b qualitatively similar to those of $\hat{\boldsymbol{\theta}}_n^{\texttt{MS}}$. To be specific, under regularity conditions the conditional maximum score estimator is covered by the results of Section III.3 and an example-specific alternative (of smoothed maximum score type) to the generic numerical derivative estimator $\tilde{\mathbf{V}}_n^{\texttt{ND}}$ is available.

Now I describe one set of sufficient conditions. For convenience, I employ the following

additional notation:

$$\psi(x_1, \mathbf{x}_2, \mathbf{w}) = \{\mathbb{E}(A(Y_0, Y_3)|x_1, \mathbf{x_2}, \mathbf{w}) - \mathbb{E}(B(Y_0, Y_3)|x_1, \mathbf{x}_2, \mathbf{w})\}_{x_1}(x_1|\mathbf{x}_2, \mathbf{w}),$$

$$A(d_0, d_3) = \mathbb{1}(Y_0 = d_0, Y_1 = 0, Y_2 = 1, Y_3 = d_3),$$

$$B(d_0, d_3) = \mathbb{1}(Y_0 = d_0, Y_1 = 1, Y_2 = 0, Y_3 = d_3),$$

$$\varphi_1(\mathbf{w}; \boldsymbol{\theta}) = \int \int_{-\mathbf{x}_2'\boldsymbol{\theta}}^{\infty} \psi(x, \mathbf{x}_2, \mathbf{w}) dx dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w}),$$

$$\varphi_2(\mathbf{w}) = \int \mathbf{x}_2 \psi(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{w}) dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w}),$$

$$\mathbf{V}(\mathbf{w}; \boldsymbol{\theta}) = \int \mathbf{x}_2 \mathbf{x}_2' \left. \frac{d}{dx} \psi(x, \mathbf{x}_2, \mathbf{w}) \right|_{x=-\mathbf{x}_2'\boldsymbol{\theta}} dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w}),$$

and $\mathbf{X} = \{\mathbf{X}_t\}_{t=1}^3$, $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\| \leq \eta\}$, and $\mathcal{S} = \mathbb{R} \times \text{supp}(\mathbf{x}_2) \times \mathcal{W}$ for some $\eta > 0$.

Condition CMS Let $S \geq 1$ and $S_f \geq 1$ be some integers.

(i) The parameter space $\boldsymbol{\Theta}$ is compact and $\boldsymbol{\theta}_0$, which satisfies $(1, \boldsymbol{\theta}_0')' = (\boldsymbol{\beta}_0', \gamma_0)'$, lies in the interior of $\boldsymbol{\Theta}$.

(ii) The sample $\{\mathbf{z}_i\}_{i \geq 1}$ is i.i.d. across $i = 1, 2, \ldots, n$. The data generating process is

$$\mathbb{P}(Y_0 = 1|\mathbf{X}, \alpha) = p(\mathbf{X}, \alpha),$$

$$\mathbb{P}(Y_t = 1|\mathbf{X}, \alpha, Y_0, \ldots, Y_{t-1}) = F(\mathbf{X}_t'\boldsymbol{\beta}_0 + \gamma_0 Y_{t-1} + \alpha), \qquad t = 1, 2, 3,$$

where $F(\cdot)$ is strictly increasing on the entire real line.

(iii) $\mathbb{P}(y \neq 0|x_1, \mathbf{x}_2, \mathbf{w}) > 0$ on $\mathcal{S}$. The density $f_{\mathbf{w}}$ is continuous on $\mathcal{W}$ and $f_{\mathbf{w}}(\mathbf{0}) > 0$.

(iv) The support of $\mathbf{X}_1 - \mathbf{X}_2$ given $\mathbf{w} \equiv \mathbf{X}_2 - \mathbf{X}_3$ is not contained in any proper linear subspace of $\mathbb{R}^d$ for all $\mathbf{w} \in \mathcal{W}$ and $\mathbb{E}[\|\mathbf{x}_2\|^2|\mathbf{w}] \leq C$ for all $\mathbf{w} \in \mathcal{W}$.

(v) The conditional density $f_{x_1}(x_1|\mathbf{x}_2, \mathbf{w})$ is positive and bounded on $\mathcal{S}$. In addition, there exists a function $B_f : \mathbb{R}^d \to \mathbb{R}_+$ such that, for some $\epsilon > 0$,

$$\sup_{\mathbf{w} \in \mathcal{W}} |f_{x_1}(\tilde{v}|\mathbf{x}_2, \mathbf{w}) - f_{x_1}(v|\mathbf{x}_2, \mathbf{w})| \leq B_f(v, \mathbf{x}_2)|\tilde{v} - v|^{\epsilon},$$

with $\sup_{\mathbf{w}\in\mathcal{W}} \mathbb{E}[\|\mathbf{x}_2\|^3 B_f(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)|\mathbf{w}] < \infty$.

(vi) The function $\psi(x_1, \mathbf{x}_2, \mathbf{w})$ is bounded and S times continuously differentiable in x_1 with bounded derivatives on $\mathcal{S}$.

(vii) The functions $\varphi_1(\mathbf{w};\boldsymbol{\theta})$ and $\mathbf{V}(\mathbf{w};\boldsymbol{\theta})$ satisfy, for some $\epsilon > 0$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\varphi_1(\mathbf{w};\boldsymbol{\theta}) - \varphi_1(\mathbf{0};\boldsymbol{\theta})| + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\delta} \|\mathbf{V}(\mathbf{w};\boldsymbol{\theta}) - \mathbf{V}(\mathbf{0};\boldsymbol{\theta})\| \leq C\|\mathbf{w}\|^\epsilon, \qquad \mathbf{w}\in\mathcal{W}.$$

The function φ_2 is S_f times continuously differentiable on $\mathcal{W}$. Also, $\mathbf{V}(\mathbf{0};\boldsymbol{\theta}_0)$ is positive definite.

(viii) The kernel function $L(\cdot)$ is bounded, compactly supported, P-th order with $1 \leq P \leq S_f$, and $\int L(\mathbf{v})d\mathbf{v} = 1$.

Under these conditions, I have the following result.

Corollary CMS. *Suppose Condition CMS holds. If $nb_n^d \to \infty$ and $nb_n^{d+3P} \to 0$, then*

$$\sqrt[3]{nb_n^d}(\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}} - \boldsymbol{\theta}_0) \rightsquigarrow \operatorname*{arg\,max}_{\mathbf{s}\in\mathbb{R}^d} \left\{ -\frac{1}{2}\mathbf{s}'\mathbf{V}_0^{\texttt{CMS}}\mathbf{s} + \mathcal{G}_0(\mathbf{s}) \right\}$$

where $\mathcal{G}_0$ is a mean-zero Gaussian process with covariance kernel $\mathcal{C}_0^{\texttt{CMS}}(\cdot,\cdot)$, and

$$\mathbf{V}_0^{\texttt{CMS}} = \mathbb{E}[\mathbf{x}_2\mathbf{x}_2'\psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{0})|\mathbf{w} = \mathbf{0}]f_{\mathbf{w}}(\mathbf{0}), \qquad \psi^{(1)}(x, \mathbf{x}_2, \mathbf{w}) = \frac{d}{dx}\psi(x, \mathbf{x}_2, \mathbf{w}),$$

$$\mathcal{C}_0^{\texttt{CMS}}(\mathbf{s},\mathbf{t}) = \mathbb{E}\left[\min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} f_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2, \mathbf{0}) \mathbb{1}\{\operatorname{sgn}(\mathbf{x}_2'\mathbf{s}) = \operatorname{sgn}(\mathbf{x}_2'\mathbf{t})\} \middle| \mathbf{w} = \mathbf{0}\right] f_{\mathbf{w}}(\mathbf{0}).$$

Furthermore, if a consistent estimator for the Hessian term is available (i.e., $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0^{\texttt{CMS}}$ for some estimator $\tilde{\mathbf{V}}_n$), then the reshaped bootstrap consistently estimates the large sample distribution of $\sqrt[3]{nb_n^d}(\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}} - \boldsymbol{\theta}_0)$.

As previously mentioned, we can consider a case-specific estimator of $\mathbf{V}_0^{\texttt{AMSE}}$. The proposed estimator is

$$\tilde{\mathbf{V}}_n^{\texttt{CMS}} = \tilde{\mathbf{V}}_n^{\texttt{CMS}}(\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}}), \qquad \tilde{\mathbf{V}}_n^{\texttt{CMS}}(\boldsymbol{\theta}) = -\frac{1}{nh_n^2}\sum_{i=1}^n y_i \dot{K}\left(\frac{x_{1i} + \mathbf{x}_{2i}'\boldsymbol{\theta}}{h_n}\right) \mathbf{x}_{2i}\mathbf{x}_{2i}' L_n(\mathbf{w}_i)$$

where $h_n = o(1)$ is a bandwidth, and $\dot{K}(u) = dK(u)/du$ with K a kernel function.

I impose Condition K on the kernel function $K(\cdot)$, and obtain the following result for $\tilde{\mathbf{V}}_n^{\mathtt{CMS}}$. Let $\mathbf{V}_n^{\mathtt{CMS}} = \partial^2 M_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ where $M_n^{\mathtt{CMS}}(\boldsymbol{\theta}) = \mathbb{E}[m_n^{\mathtt{CMS}}(\mathbf{z},\boldsymbol{\theta})]$.

Lemma CMS. *Suppose Condition CMS and Assumption K hold with $S \geq 2$.*

If $h_n \to 0$ and $nb_n^d h_n^3 \to \infty$, and $\mathbb{E}[\|\mathbf{x}_2\|^6|\mathbf{w}] \leq C$ for $\mathbf{w} \in \mathcal{W}$, then $\tilde{\mathbf{V}}_n^{\mathtt{CMS}} \to_{\mathbb{P}} \mathbf{V}_0^{\mathtt{CMS}}$.

Furthermore, if $S \geq 3$, then

$$\mathbb{E}\left[\|\tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0) - \mathbf{V}_n^{\mathtt{CMS}}\|^2\right] = O(h_n^4 + (nb_n^d h_n^3)^{-1}),$$

where the rate is sharp in the sense that there exist constants B_{kl} and V_{kl} such that

$$\frac{\mathbb{E}[(\tilde{V}_{n,kl}^{\mathtt{CMS}}(\boldsymbol{\theta}_0) - V_{n,kl}^{\mathtt{CMS}})^2]}{h_n^4\mathsf{B} + (nb_n^d h_n^3)^{-1}\mathsf{V}} \to 1,$$

where $\tilde{V}_{n,kl}^{\mathtt{CMS}}(\boldsymbol{\theta}_0)$ and $V_{n,kl}^{\mathtt{CMS}}$ denote the (k,l)-th elements of $\tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0)$ and $\mathbf{V}_n^{\mathtt{CMS}}$, respectively.

III.5 Discussion

The applicability of the procedure proposed in this note extends beyond the estimators covered by Theorem 6. For instance, it is not hard to show that the bootstrap-based distributional approximation is consistent also in the more standard case where $m_n(\mathbf{z},\boldsymbol{\theta})$ is sufficiently smooth in $\boldsymbol{\theta}$ to ensure that an approximate maximizer of $\hat{M}_n$ is asymptotically normal and that the nonparametric bootstrap is consistent. In fact, $\tilde{\boldsymbol{\theta}}_n^*$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_n^*$ in that standard case, so my procedure can be interpreted as a modification of the nonparametric bootstrap that is designed to be "robust" to the types of non-smoothness that give rise to cube root asymptotics.

Moreover, and perhaps more importantly, the idea of reshaping can be used to achieve consistency of bootstrap-based approximations to the distributions of certain estimators which are not of M-estimator type, yet exhibit cube root asymptotics and have the feature that the standard bootstrap-based approximations to their distribution are known to be inconsistent. In particular, Section III.6 shows how the idea of reshaping a process can be used to achieve consistency on the part of a bootstrap-based approximation to the distribution of

the celebrated ($\sqrt[3]{n}$-consistent) isotonic density estimator of Grenander (1956).[1]

This note is not the first to propose a consistent resampling-based distributional approximation for cube root consistent estimators. For cube root asymptotic problems, the best known consistent alternative to the nonparametric bootstrap is probably subsampling (Politis and Romano, 1994), whose applicability was pointed out by Delgado *et al.* (2001). A related method is the rescaled bootstrap (Dümbgen, 1993), whose validity in cube root asymptotic M-estimation problems was established recently by Hong and Li (2019). In addition, case-specific smooth bootstrap methods have been proposed for leading examples such as maximum score estimation (Patra *et al.*, 2015) and isotonic density estimation (Kosorok, 2008; Sen *et al.*, 2010). Like ours, each of these methods can be viewed as offering a "robust" alternative to the nonparametric bootstrap but, unlike ours, they all achieve consistency by modifying the distribution used to generate the bootstrap counterpart of the estimator whose distribution is being approximated. In contrast, my method achieves consistency by means of an analytic modification to the objective function used to construct the bootstrap-based distributional approximation.

An alternative interpretation of my approach is available. Restating the result in (III.7) as

$$r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{S}_0(\mathcal{G}_0), \qquad \mathcal{S}_0(\mathcal{G}) = \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max}\{\mathcal{Q}_0(\mathbf{s}) + \mathcal{G}(\mathbf{s})\},$$

my procedure approximates the distribution of $\mathcal{S}_0(\mathcal{G}_0)$ by that of $\tilde{\mathcal{S}}_n(\hat{G}_n^*)$, where

$$\hat{G}_n^*(\mathbf{s}) = r_n^2[\hat{M}_n^*(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) - \hat{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n)]$$

is a bootstrap process whose distribution approximates that of $\mathcal{G}_0(\mathbf{s})$ and where

$$\tilde{\mathcal{S}}_n(\mathcal{G}) = \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max}\{\tilde{Q}_n(\mathbf{s}) + \mathcal{G}(\mathbf{s})\}, \qquad \tilde{Q}_n(\mathbf{s}) = -\frac{1}{2}\mathbf{s}'\tilde{\mathbf{V}}_n\mathbf{s},$$

is an estimator of $\mathcal{S}_0(\mathcal{G})$. In other words, my procedure replaces the functional $\mathcal{S}_0$ with a con-

[1] The asymptotic properties of the Grenander estimator have been studied by Prakasa Rao (1969), Groeneboom (1985a), and Kim and Pollard (1990), among others. Inconsistency of the standard bootstrap-based approximation to the distribution of the Grenander estimator has been pointed out by Kosorok (2008) and Sen *et al.* (2010), among others.

sistent estimator (namely, $\tilde{\mathcal{S}}_n$) and its random argument $\mathcal{G}_0$ with a bootstrap approximation (namely, $\hat{G}_n^*$). Similar constructions have been successfully applied in other settings, two relatively recent examples being Andrews and Soares (2010) and Fang and Santos (2018).

Finally, Seo and Otsu (2018) give conditions under which results of the form (III.7) can be obtained also when the data is weakly dependent. It seems plausible that a version of my procedure, implemented with resampling procedure suitable for dependent data, can be shown to be consistent under similar conditions, but it is beyond the scope of this note to substantiate that conjecture.

III.6 Extension: Isotonic Density Estimation

The main ideas and results can also be used to conduct valid bootstrap-based inference in the context of isotonic density estimation, another cube root consistent estimator with a Chernoff-type limiting distribution. See, e.g., van der Vaart and Wellner (1996, Example 3.2.14) for a modern textbook treatment of this example.

Suppose we observe $x_1, \ldots, x_n \overset{iid}{\sim} F$ where F is continuous distribution function supported on $[0, \infty)$ with non-increasing density f. An extensively-studied method of estimating $f(x_0)$, $x_0 \in (0, \infty)$, is nonparametric maximum likelihood estimation (NPMLE):

$$\arg\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i)$$

where $\mathcal{F}$ is the class of non-increasing densities supported on $[0, \infty)$. It is well known that the solution to this problem, denoted by $\hat{f}_n(x)$, is the left derivative at x of the least concave majorant (LCM) of the empirical distribution function $\hat{F}_n(\cdot) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(x_i \leq \cdot)$.

The asymptotic distributional result for the NPMLE in this setting is as follows.

Lemma ID. *If f is differentiable at x_0 with $f^{(1)}(x_0) < 0$, then*

$$\sqrt[3]{n}(\hat{f}_n(x_0) - f(x_0)) \rightsquigarrow |4f^{(1)}(x_0) f(x_0)|^{1/3} \arg\max_{s \in \mathbb{R}} \left\{ \mathcal{W}(s) - s^2 \right\}$$

where $\mathcal{W}$ is a standard two-sided Brownian motion with $\mathcal{W}(0) = 0$.

The NPMLE estimator $\hat{f}_n(x_0)$ is not a member of the class of M-estimators considered in this paper, but nonetheless a distributional approximation of $\sqrt[3]{n}(\hat{f}_n(x_0) - f(x_0))$ can be established using the results for cube root consistent estimators discussed herein via the switching technique of Groeneboom (1985b). Therefore, I first outline the main idea underlying this distributional approximation because it provides an insight to the proposed bootstrap-based inference approach.

The NPMLE estimator is the left-derivative of the least concave majorant (LCM) of $\hat{F}_n$ at x_0. Using this property, I have

$$\hat{S}_n(a) \leq x \iff \hat{f}_n(x) \leq a$$

where

$$\hat{S}_n(a) = \arg\max_{s \geq 0} \{\hat{F}_n(s) - as\}.$$

This is the switching technique alluded above. Then, I have

$$\mathbb{P}\left[n^{1/3}(\hat{f}_n(x_0) - f(x_0)) \leq t\right] = \mathbb{P}\left[\hat{S}_n(f(x_0) + n^{-1/3}t) \leq x_0\right].$$

Using the definition of $\hat{S}_n(a)$ and change-of-variable $x = x_0 + sn^{-1/3}$,

$$\begin{aligned}
&\hat{S}_n(f(x_0) + n^{-1/3}t) \\
&= \arg\max_{x \in [0,\infty)} \left\{\hat{F}_n(x) - (f(x_0) + n^{-1/3}t)x\right\} \\
&= x_0 + n^{-1/3} \arg\max_{s \in [-n^{1/3}x_0, \infty)} \left\{\hat{F}_n(x_0 + sn^{-1/3}) - (f(x_0) + n^{-1/3}t)(x_0 + sn^{-1/3})\right\} \\
&= x_0 + n^{-1/3} \arg\max_{s \in [-n^{1/3}x_0, \infty)} \left\{\hat{F}_n(x_0 + sn^{-1/3}) - \hat{F}_n(x_0) - (f(x_0) + n^{-1/3}t)sn^{-1/3}\right\} \\
&= x_0 + n^{-1/3} \arg\max_{s \in [-n^{1/3}x_0, \infty)} \left\{n^{2/3}(\hat{F}_n(x_0 + sn^{-1/3}) - \hat{F}_n(x_0) - f(x_0)sn^{-1/3}) - ts\right\}
\end{aligned}$$

where I use that a maximizer is invariant under shifting and scaling of the objective function.

Therefore, to study the limit law of $n^{1/3}(\hat{f}_n(x_0) - f(x_0))$, it suffices to look at the asymptotic distribution of the maximizer of $n^{2/3}(\hat{F}_n(x_0 + sn^{-1/3}) - \hat{F}_n(x_0) - f(x_0)sn^{-1/3}) - ts$. To do that, I can first establish weak convergence of the objective function and then apply

an argmax continuous mapping theorem.

To establish uniform weak convergence of the objective function, I let

$$n^{2/3}(\hat{F}_n(x_0 + sn^{-1/3}) - \hat{F}_n(x_0) - f(x_0)sn^{-1/3}) - ts = Q_n(s) + G_n(s)$$

where

$$\begin{aligned} Q_n(s) &= n^{2/3}\left(F(x_0 + sn^{-1/3}) - F(x_0) - f(x_0)sn^{-1/3}\right) - ts, \\ G_n(s) &= n^{2/3}\left(\hat{F}_n(x_0 + sn^{-1/3}) - \hat{F}_n(x_0) - (F(x_0 + sn^{-1/3}) - F(x_0))\right). \end{aligned}$$

For the first term, using differentiability of f at x_0, a Taylor expansion gives

$$Q_n(s) = \frac{f^{(1)}(x_0)}{2}s^2 - ts + o(1),$$

which is non-random and the $o(1)$ term converges uniformly with respect to s over a compact subset of $\mathbb{R}$. Note that in this M-estimation problem, the Hessian term $-V_0$ equals $f^{(1)}(x_0)$, the first derivative of f at x_0.

For the second term, Lindeberg-Feller CLT implies finite-dimensional convergence and a standard maximal inequality shows stochastic equicontinuity, so I can establish that for any compact subset $\mathcal{S} \subset \mathbb{R}$,

$$G_n(s) \rightsquigarrow \mathcal{W}(f(x_0)s), \qquad s \in \mathcal{S}.$$

Then, an argmax continuous mapping theorem implies

$$\mathbb{P}\left(n^{1/3}(\hat{f}_n(x_0) - f(x_0)) \leq t\right) \to \mathbb{P}\left(\arg\max_s \left\{\mathcal{W}(f(x_0)s) + f^{(1)}(x_0)s^2/2 - ts\right\} \leq 0\right).$$

Rewriting the object inside the arg max operator using properties of Brownian motion (c.f., van der Vaart and Wellner, 1996, Problem 3.2.5), I obtain the conclusion of Lemma ID.

III.6.1 Bootstrap-based Inference

The above discussion indicates that, if $\tilde{\Delta}_n$ is the left-derivative at t_0 of the LCM of some generic function $U_n(x)$, I have

$$\mathbb{P}[n^{1/3}\tilde{\Delta}_n \leq t] = \mathbb{P}\left(\arg\max_s \left\{n^{2/3}(U_n(x_0 + sn^{-1/3}) - U_n(x_0)) - ts\right\} \leq 0\right).$$

Therefore, for bootstrap validity, I need to study the bootstrap counterpart $n^{2/3}(U_n^*(x_0 + sn^{-1/3}) - U_n^*(x_0))$ and, in particular, establish uniform weakly convergence in probability in order to establish validity of the bootstrap distributional approximation.

However, as discussed in this chapter, in the cube root consistent case as in Kim and Pollard (1990) the bootstrap approximation of the objective function fails at the Hessian term. Therefore, I reshape the M-estimator as before prior to applying the bootstrap. In particular, let $\tilde{\Delta}_n^*$ be the left-derivative at $x = x_0$ of the LCM of $\tilde{F}_n^*(x)$, where

$$\tilde{F}_n^*(x) = \hat{F}_n^*(x) - \hat{F}_n(x) + \frac{1}{2}\tilde{V}_n(x - x_0)^2, \qquad \tilde{V}_n = \tilde{f}_n^{(1)}(x_0),$$

$\hat{F}_n^*(\cdot) = \sum_{i=1}^n \mathbb{1}(x_i^* \leq \cdot)$, and $\tilde{f}_n^{(1)}(x_0)$ is a consistent estimator of $f^{(1)}(x_0)$.

The form of the above objective function exactly mimics the intuition described in the paper: I remove the problematic drift term under the bootstrap distribution and add back a plug-in estimate thereof.

Theorem ID. *Suppose the conditions in Lemma ID hold. If $\tilde{f}_n^{(1)}(x_0) \to_{\mathbb{P}} f^{(1)}(x_0)$, then*

$$\sqrt[3]{n}\tilde{\Delta}_n^* \rightsquigarrow_{\mathbb{P}} |4f^{(1)}(x_0)f(x_0)|^{1/3} \arg\max_{s\in\mathbb{R}} \left\{\mathcal{W}(s) - s^2\right\},$$

where $\tilde{\Delta}_n^*$ be the left-derivative at $x = x_0$ of the LCM of $\tilde{F}_n^*(x) = \hat{F}_n^*(x) - \hat{F}_n(x) + \frac{1}{2}\tilde{V}_n(x - x_0)^2$.

The above bootstrap-based construction employs the reshaped objective function $\tilde{F}_n^*(x)$. Another possibility is to use $\tilde{F}_n^*(x) + \hat{f}_n(x_0)x$ with $\hat{f}_n(x_0)$ being the NPMLE from the original sample. If I let $\tilde{f}_n^*(x_0)$ be the left-derivative at x_0 of the LCM of the alternative objective function, I have

$$\tilde{f}_n^*(x_0) - \hat{f}_n(x_0) = \tilde{\Delta}_n^*.$$

This follows because given $G(x) = g(x) + cx$ for some constant c, the LCM of G equals the LCM of g plus cx. In particular, suppose that this claim is false. Then, there exists a concave function F such that $G(x) \leq F(x)$ for all x and $F(\tilde{x}) < Lg(\tilde{x}) + c\tilde{x}$ for at least one $\tilde{x}$, where Lg denotes the LCM of g. Thus, $F(\tilde{x}) - c\tilde{x} < Lg(\tilde{x})$ and this implies $F(\breve{x}) - c\breve{x} < g(\breve{x})$ for some $\breve{x}$, but this contradicts $G(x) \leq F(x)$ for all x.

III.6.2 Drift Estimation

In this example the counterpart of $-\mathbf{V}_0$ is simply $f^{(1)}(x_0)$, the first derivative of the Lebesgue density $f(x)$ at x_0. Thus, the estimation of the drift term can be done using standard nonparametric estimation techniques, already available in most software platforms. For methodological and technical details see, for example, Wand and Jones (1995).

III.6.3 Simulation Evidence

I investigate the finite sample properties of confidence intervals for $f(x_0)$ constructed using the bootstrap-based distributional approximation whose consistency was established in Theorem ID. I employ the DGPs and simulation setting previously considered in Sen *et al.* (2010). This, as in the case of the Maximum Score estimator discussed in the paper, allows for a direct comparison with other bootstrap-based inference methods and their numerical performance already studied in previous work available in the literature.

I estimate $f(x_0)$ at the evaluation point $x_0 = 1$ using a random sample of observations, where three distinct distributions are considered: DGP 1 sets $x \sim \mathsf{Exponential}(1)$, DGP 2 sets $x \sim |\mathsf{Normal}(0,1)|$, and DGP 3 sets $x \sim |\mathcal{T}_3|$. As in the case of the Maximum Score example, the Monte Carlo experiment employs a sample size $n = 1,000$ with $B = 2,000$ bootstrap replications and $S = 2,000$ simulations, and compares three types of bootstrap-based inference procedures: the standard non-parametric bootstrap, m-out-of-n bootstrap, and the proposed method using two distinct estimators of $f'(x_0)$ (plug-in and numerical derivative).

Table III.2 presents the numerical results for this example. I continue to report empirical coverage for nominal 95% confidence intervals and their average interval length. For the

case of the proposed procedures, I again investigate their performance using both (i) a grid of fixed tuning parameter value (derivative step/bandwidth) and (ii) infeasible and feasible AMSE-optimal choice of tuning parameter. Also in this case, the numerical evidence is very encouraging. The proposed bootstrap-based inference method leads to confidence intervals with excellent empirical coverage and average interval length, outperforming both the standard nonparametric bootstrap (which is inconsistent) and the m-out-of-n bootstrap (which is consistent). In particular, in this example, the plug-in method employs an off-the-shelf kernel derivative estimator, which in this case leads to confidence intervals that are very robust (i.e., insensitive) to the choice of bandwidth. Furthermore, when the corresponding feasible off-the-shelf MSE-optimal bandwidth is used, the resulting confidence intervals continue to perform excellently. Finally, the generic numerical derivative estimator also leads to very good performance of bootstrap-based infeasible and feasible confidence intervals.

In sum, this example provides a second numerical illustration of the very good finite sample performance of inference based on the proposed bootstrap-based distributional approximation for cube root consistent estimators.

Table III.2. Simulations, Isotonic Density Estimator, 95% Confidence Intervals.

(a) $n = 1,000$, $S = 2,000$, $B = 2,000$

	DGP 1			DGP 2			DGP 3		
	h, ϵ	Coverage	Length	h, ϵ	Coverage	Length	h, ϵ	Coverage	Length
Standard									
		0.828	0.146		0.808	0.172		0.821	0.155
m-out-of-n									
$m = \lceil n^{1/2} \rceil$		1.000	0.438		0.995	0.495		0.998	0.452
$m = \lceil n^{2/3} \rceil$		0.989	0.314		0.979	0.360		0.989	0.328
$m = \lceil n^{4/5} \rceil$		0.953	0.235		0.937	0.274		0.948	0.248
Plug-in: $\tilde{\mathbf{V}}_n^{\mathtt{ID}}$									
$0.7 \cdot h_{\mathtt{MSE}}$	0.264	0.955	0.157	0.202	0.947	0.183	0.209	0.957	0.165
$0.8 \cdot h_{\mathtt{MSE}}$	0.302	0.954	0.157	0.231	0.946	0.182	0.239	0.952	0.165
$0.9 \cdot h_{\mathtt{MSE}}$	0.339	0.951	0.156	0.260	0.944	0.181	0.269	0.949	0.164
$1.0 \cdot h_{\mathtt{MSE}}$	0.377	0.949	0.154	0.289	0.941	0.180	0.299	0.948	0.163
$1.1 \cdot h_{\mathtt{MSE}}$	0.415	0.940	0.151	0.318	0.938	0.178	0.329	0.944	0.161
$1.2 \cdot h_{\mathtt{MSE}}$	0.452	0.934	0.147	0.347	0.934	0.176	0.359	0.939	0.158
$1.3 \cdot h_{\mathtt{MSE}}$	0.490	0.922	0.142	0.376	0.928	0.173	0.389	0.935	0.155
$h_{\mathtt{AMSE}}$	0.380	0.949	0.154	0.300	0.940	0.180	0.333	0.943	0.161
$\hat{h}_{\mathtt{AMSE}}$	0.364	0.950	0.155	0.290	0.941	0.180	0.401	0.930	0.154
Num Deriv: $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$									
$0.7 \cdot \epsilon_{\mathtt{MSE}}$	0.726	0.954	0.158	0.527	0.947	0.183	0.554	0.952	0.165
$0.8 \cdot \epsilon_{\mathtt{MSE}}$	0.830	0.956	0.159	0.602	0.947	0.182	0.633	0.950	0.164
$0.9 \cdot \epsilon_{\mathtt{MSE}}$	0.933	0.956	0.160	0.678	0.944	0.181	0.712	0.949	0.163
$1.0 \cdot \epsilon_{\mathtt{MSE}}$	1.037	0.956	0.159	0.753	0.942	0.180	0.791	0.948	0.162
$1.1 \cdot \epsilon_{\mathtt{MSE}}$	1.141	0.955	0.159	0.828	0.940	0.179	0.870	0.946	0.161
$1.2 \cdot \epsilon_{\mathtt{MSE}}$	1.244	0.956	0.160	0.904	0.936	0.177	0.949	0.943	0.160
$1.3 \cdot \epsilon_{\mathtt{MSE}}$	1.348	0.960	0.163	0.979	0.935	0.176	1.028	0.940	0.159
$\epsilon_{\mathtt{AMSE}}$	0.927	0.956	0.160	0.731	0.943	0.180	0.812	0.948	0.162
$\hat{\epsilon}_{\mathtt{AMSE}}$	0.888	0.956	0.159	0.708	0.943	0.181	0.978	0.942	0.159

Notes:
(i) Panel **Standard** refers to standard nonparametric bootstrap, Panel **m-out-of-n** refers to m-out-of-n nonparametric bootstrap with subsample m, Panel **Plug-in:** $\tilde{\mathbf{V}}_n^{\mathtt{ID}}$ refers to the proposed bootstrap-based implemented using the example-specific plug-in drift estimator, and Panel **Num Deriv:** $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ refers to the proposed bootstrap-based implemented using the generic numerical derivative drift estimator.
(ii) Column "h, ϵ" reports tuning parameter value used or average across simulations when estimated, and Columns "Coverage" and "Length" report empirical coverage and average length of bootstrap-based 95% percentile confidence intervals, respectively.
(iii) $h_{\mathtt{MSE}}$ and $\epsilon_{\mathtt{MSE}}$ correspond to the simulation MSE-optimal choices, $h_{\mathtt{AMSE}}$ and $\epsilon_{\mathtt{AMSE}}$ correspond to the AMSE-optimal choices, and $\hat{h}_{\mathtt{AMSE}}$ and $\hat{\epsilon}_{\mathtt{AMSE}}$ correspond to the ROT feasible implementation of $\hat{h}_{\mathtt{AMSE}}$ and $\hat{\epsilon}_{\mathtt{AMSE}}$ described in Appendix B.2.

CHAPTER IV

Smoothed Pairwise Difference Estimators: Distribution Theory

IV.1 Introduction

Semiparametric econometric models provide a useful way to balance flexibility and parsimonious parametrization, which allows for researchers to mitigate concerns with misspecifications while retaining a reasonable requirement on sample sizes of data. Yet, implementation of semiparametric models can be challenging because it faces estimation and inference issues distinct from more traditional parametric models. For example, semiparametric estimators often require tuning parameters, and sample-finite properties of such estimators are sensitive to selection of tuning parameters.

In this paper, I study a statistical inference problem in semiparametric models, explicitly accounting for impact of tuning parameter selection. Specifically, I investigate a class of smoothed pairwise difference estimators, which involves bandwidth as a tuning parameter, and I highlight non-robustness of the conventional Gaussian distributional approximation with respect to bandwidth choice. The estimator exhibits a range of different limit distributions depending on the bandwidth sequence. From a practical standpoint, the results of this chapter indicate that performance of conventional inference methods is sensitive to specific bandwidth chosen. This property makes statistical inference less credible, and this chapter provides another piece of evidence corroborating the widespread concern about non-robust behavior of semiparametric estimators with respect to tuning parameters (see e.g., Ichimura and Todd, 2007).

There is an extensive literature on distribution theory and inference using semiparametric estimators (see e.g., Andrews, 1994; Newey and McFadden, 1994; Chen, 2007). More recently, several studies have investigated impact of tuning parameter choices on statistical inference. Cattaneo *et al.* (2014) examine asymptotic distributional approximation of density-weighted average derivative estimators, allowing for bandwidth sequences shrinking at a faster rate than those entertained in previous work. Cattaneo *et al.* (2013) study related distribution results for weighted average derivative estimators where the limit distribution features asymptotic bias term, and they propose a bias-correction inference procedure. Both of these papers develop distributional approximations valid under weaker restrictions on tuning parameter sequences. Chernozhukov *et al.* (2018) also establish distribution theory, allowing for a wider range of tuning parameters. Whereas the first two papers focus on what they term "small bandwidth asymptotics," the last paper examines the case where bandwidth is larger. This difference stems from whether one focuses on higher-order variance or bias. The first two look at the higher-order variance, and the latter deal with bias term. In this paper, I follow the first approach by accommodating the higher-order variance term in distributional approximation.

IV.2 Setup

Let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ be an estimand of interest, which solves the following optimization problem:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, M_0(\boldsymbol{\theta}), \qquad M_0(\boldsymbol{\theta}) = \mathbb{E}[\mathbb{E}[s(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta})|\mathbf{w}_1 = \mathbf{w}_2]] \tag{IV.1}$$

where $s : \mathbb{R}^{d_z} \times \mathbb{R}^{d_z} \times \boldsymbol{\Theta} \to \mathbb{R}$ is known, $\mathbf{z} \in \mathbb{R}^{d_z}$ is a random vector, and $\mathbf{w} \in \mathbb{R}^d$ is a subvector of $\mathbf{z}$. Note that by the conditioning $\mathbf{w}_1 = \mathbf{w}_2$, the characterization of $\boldsymbol{\theta}_0$ uses only the subpopulation in which a pair of observations has the same values of covariates $\mathbf{w}$.

In their influential study, Honoré and Powell (2005) study an estimator of the form

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, \hat{M}_n(\boldsymbol{\theta}), \qquad \hat{M}_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} s(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}) K_{h_n}(\mathbf{w}_i - \mathbf{w}_j)$$

where $\{\mathbf{z}_i\}_{i=1}^n$ is a random sample of $\mathbf{z}$, $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function, $K_h(\mathbf{u}) = h^{-d}K(\mathbf{u}/h)$, and h_n is a bandwidth sequence shrinking to zero.

The kernel function assigns weights to pairs of observations. If $\mathbf{w}_i$ and $\mathbf{w}_j$ are close to each other, the kernel gives a large weight to the pair. On the other hand, if $\mathbf{w}_i$ and $\mathbf{w}_j$ are far apart, the kernel weight is small. Then, in asymptotics where h_n shrinks to zero, the value of $\hat{M}_n$ is determined by pairs of observations whose values of $\mathbf{w}$ are identical. This behavior of kernel weighting is key to consistently estimating $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}_n$ because, to precisely characterize $\boldsymbol{\theta}_0$, one needs to use only the subpopulation in which $\mathbf{w}_i = \mathbf{w}_j$. Yet, in a finite sample, the event $\{\mathbf{w}_i = \mathbf{w}_j\}$ has zero probability if $\mathbf{w}$ has a continuous distribution. Kernel weighting enables estimation of $\boldsymbol{\theta}_0$ in the absence of pairs with $\mathbf{w}_i = \mathbf{w}_j$ by including pairs with $\mathbf{w}_i \approx \mathbf{w}_j$ in finite samples.

Due to this kernel weighting, estimation of $\boldsymbol{\theta}_0$ faces a trade-off of bias and variance. When pairs with not so close values of $\mathbf{w}$ get non-trivial weights, the estimation is based on a sample with $\mathbf{w}_i \not\approx \mathbf{w}_j$, which induces bias. On the other hand, only using pairs with very close $\mathbf{w}$'s leads to a smaller sample size, which increases variance. This trade off is governed by the bandwidth h_n. A larger h_n gives non-trivial weights to more pairs, and a smaller bandwidth results in a smaller number of observations. Then, for good finite-sample properties of the estimator, choice of a bandwidth plays an important role.

To better understand the impact of bandwidths on the estimator, think about the case where $\hat{M}_n(\boldsymbol{\theta})$ is smooth in the parameter and $\hat{\boldsymbol{\theta}}_n$ admits a closed form representation. In this situation, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \approx \mathcal{Z} + B\sqrt{n}h_n^P + \mathcal{W}/\sqrt{nh_n^d} \tag{IV.2}$$

where $\mathcal{Z}$ has a mean-zero normal distribution, the term $B\sqrt{n}h_n^P$ is the leading bias term, and $\mathcal{W}$ is a mean-zero random vector. For the bias term $B\sqrt{n}h_n^P$, the power on the bandwidth P coincides with the order of the kernel function $K(\cdot)$ under enough smoothness in the data generating process. Here, the trade-off occurs between $B\sqrt{n}h_n^P$ and $\mathcal{W}/\sqrt{nh_n^d}$. For a large h_n, the first term (bias) becomes dominant, and for a small h_n, the second term (variance) dominates.

In the literature, existing studies impose $\sqrt{n}h_n^P \to 0$ and $nh_n^d \to \infty$ and use a Gaussian approximation for the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. Although this approach may seem natural, this requirement can impose stringent conditions on the bandwidth. Specifically, h_n needs to satisfy $n^{-1/d} \ll h_n \ll n^{-1/(2P)}$, where $a_n \ll b_n$ means $a_n/b_n = o(1)$. When the covariate $\mathbf{w}$ is multi-dimensional, this range of bandwidths can be tight. For instance, if $P = 2$, then the bandwidth sequence has to lie in a tight range between $n^{-1/d}$ and $n^{-1/4}$, and d has to be less than four. Although one may be able to alleviate this problem using a higher-order kernel, higher-order kernels often have poor finite sample properties.

In finite samples, this requirement on bandwidth rates manifests as non-robustness of distributional approximation with respect to bandwidth choice. Specifically, when the theory indicates a tight bandwidth range, the limit distribution that assumes away the bias and variance term tends to give a poor approximation to the finite-sample distribution. This non-robustness issue has been documented by Cattaneo *et al.* (2014), who show by Monte Carlo experiments that 95% confidence intervals based on normal approximation have correct coverage on a very tight bandwidth range.

In this paper, I present novel distributional approximation results under weaker conditions on bandwidths. This new approximation provides a key step to a new inference procedure that is less sensitive to bandwidth choice. To describe the asymptotic distribution result, define $r_n = (\frac{1}{\sqrt{n}} + \frac{1}{n^2 h_n^d})^{-1}$, which turns out to be the convergence rate of the estimator $\hat{\boldsymbol{\theta}}_n$. Also let $\pi_0 = \lim_{n\to\infty} \sqrt{nh_n^d}$, which I assume to exist on the extended real line. Then, under regularity conditions,

$$r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \frac{\pi_0}{1+\pi_0}\mathcal{Z} + \frac{1}{1+\pi_0}\mathcal{W}, \qquad \pi_0 \in [0, +\infty]$$

where $\rightsquigarrow$ denotes convergence in distribution, $\mathcal{Z}$ and $\mathcal{W}$ are independent mean-zero Gaussian random vectors, $\pi_0/(1+\pi_0) = 1$ and $1/(1+\pi_0) = 0$ if $\pi_0 = +\infty$.

When the bandwidth sequence satisfies the conventional restriction, namely $nh_n^d \to \infty$, the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ can be approximated by the mean-zero normal random vector $\mathcal{Z}$ because $\pi_0 = +\infty$ in this case and the term $\mathcal{W}$ vanishes from the expression. This distribution result has been established by the previous studies.

However, as previously mentioned, the requirement $nh_n^d \to \infty$ may impose stringent conditions, which makes the conventional Gaussian approximation less credible. Indeed, when the bandwidth requirement is violated i.e., $nh_n^d \not\to \infty$, then the limiting distribution needs to account for the higher-order term $\mathcal{W}/\sqrt{nh_n^d}$ in (IV.2). I show that this higher-order term has a mean-zero Gaussian distribution. Then, for the case where π_0 is some positive constant (i.e., nh_n^d converges to a non-zero constant), the limit distribution of $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ remains normal but with larger variance. This result corresponds with the intuition that a smaller bandwidth leads to higher variance. For the case where $\pi_0 = 0$, the bandwidth sequence is even "smaller," or vanishes at a faster rate. In this case, the convergence rate becomes $nh_n^{d/2} = \sqrt{n}\sqrt{nh_n^d}$, which is slower than $\sqrt{n}$ because $\sqrt{nh_n^d} \to 0$, and the limit distribution features $\mathcal{W}$, which has a different variance from $\mathcal{Z}$.

This new distribution result highlights the impact of bandwidth on the limit distribution. In particular, it implies that the conventional normal approximation may not be accurate when $\sqrt{nh_n^d}$ is not large. To be more precise, whether $\sqrt{nh_n^d}$ is sufficiently large depends on the underlying DGP, and thus researchers usually have little knowledge on which limit distribution provides a good approximation. In other words, accuracy of the conventional normal approximation may be sensitive to bandwidth choice.

IV.3 Main Result

In this section, I formally present the distributional result discussed in the previous section. First I state the main assumption, and for the purpose, define the following objects.

$$m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) = s(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) K_{h_n}(\mathbf{w}_1 - \mathbf{w}_2),$$

$$\ell_n(\mathbf{z}, \boldsymbol{\theta}) = \mathbb{E}[m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) | \mathbf{z}_1 = \mathbf{z}] - \mathbb{E}[m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta})],$$

$$w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) = m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) - \mathbb{E}[m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) | \mathbf{z}_1] - \mathbb{E}[m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) | \mathbf{z}_2] + \mathbb{E}[m_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta})],$$

$$\mathcal{L}_n = \{\ell_n(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}, \qquad \bar{\ell}_n(\mathbf{z}) = \sup_{\ell \in \mathcal{L}_n} |\ell(\mathbf{z})|$$

$$\mathcal{W}_n = \{w_n(\cdot,\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}, \qquad \bar{w}_n(\mathbf{z}_1,\mathbf{z}_2) = \sup_{w \in \mathcal{W}_n} |w(\mathbf{z}_1,\mathbf{z}_2)|$$

$$\boldsymbol{\Theta}_0^\delta = \{\boldsymbol{\theta} \in \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\}$$

$$\mathcal{L}_n^\delta = \{\ell_n(\cdot,\boldsymbol{\theta}) - \ell_n(\cdot,\boldsymbol{\theta}_0) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^\delta\}, \qquad \bar{\ell}_n^\delta(\mathbf{z}) = \sup_{\ell \in \mathcal{L}_n^\delta} |\ell(\mathbf{z})|$$

$$\mathcal{W}_n^\delta = \{w_n(\cdot,\cdot,\boldsymbol{\theta}) - w_n(\cdot,\cdot,\boldsymbol{\theta}_0) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^\delta\}, \qquad \bar{w}_n^\delta(\mathbf{z}_1,\mathbf{z}_2) = \sup_{w \in \mathcal{W}_n^\delta} |w(\mathbf{z}_1,\mathbf{z}_2)|.$$

Here is the set of assumptions used to derive the distributional result.

Assumption SP Let $r_n = (n^{-1/2} + (n^2 h_n^d)^\beta)^{-1} \to \infty$ where $\beta = 1/(4-\gamma)$ and $\gamma \in \{1,2\}$.

(i) $\{\mathcal{L}_n : n \geq 1\}$ and $\{\mathcal{W}_n : n \geq 1\}$ are uniformly manageable for the envelops $\bar{\ell}_n$ and $\bar{w}_n$, respectively, $\mathbb{E}[|\bar{\ell}_n(\mathbf{z})|^2] = o(n)$, and $\mathbb{E}[|\bar{w}_n(\mathbf{z}_1,\mathbf{z}_2)|^2] = o(n^2)$. Also, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})| = o(1)$ and, for every $\delta > 0$, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}\setminus\boldsymbol{\Theta}_0^\delta} M_0(\boldsymbol{\theta}) < M_0(\boldsymbol{\theta}_0)$.

(ii) $\boldsymbol{\theta}$ is an interior point of $\boldsymbol{\Theta}$ and, for some $\delta > 0$, M_0 and M_n are twice continuously differentiable on $\boldsymbol{\Theta}_0^\delta$ and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\delta}$. Also, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\delta} \|\partial^2[M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})]/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\| = o(1)$, $r_n\|\partial M_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\| = o(1)$, and $\mathbf{V}_0 = -\partial^2 M_0(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is positive definite.

(iii) For some $\delta > 0$, $\{\mathcal{L}_n^{\delta'} : n \geq 1, \delta' \in (0,\delta]\}$ and $\{\mathcal{W}_n^{\delta'} : n \geq 1, \delta' \in (0,\delta]\}$ are uniformly manageable for the envelopes $\bar{\ell}_n^{\delta'}$ and $\bar{w}_n^{\delta'}$, respectively. Also, there exist functions $B_n : \mathbb{R}^{d_z} \to \mathbb{R}$ such that $|\ell_n(\mathbf{z},\boldsymbol{\theta}') - \ell_n(\mathbf{z},\boldsymbol{\theta})| \leq B_n(\mathbf{z})\|\boldsymbol{\theta}' - \boldsymbol{\theta}_0\|$ for $\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}_0^\delta$, $\mathbb{E}|B_n(\mathbf{z})|^2 = O(1)$, and $\sup_{\delta'\in(0,\delta]} h_n^d \mathbb{E}[|\bar{w}_n^{\delta'}|^2/\delta'^\gamma] = O(1)$.

(iv) For every positive $\delta_n = O(r_n^{-1})$ and all $\mathbf{s},\mathbf{t} \in \mathbb{R}^k$, there exist a positive definite matrix $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{k\times k}$ and a function $\mathcal{C}_0 : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ such that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^{\delta_n}} |r_n^2 \mathbb{E}[\{\ell_n(\mathbf{z},\boldsymbol{\theta} + \mathbf{s}\delta_n) - \ell_n(\mathbf{z},\boldsymbol{\theta})\}\{\ell_n(\mathbf{z},\boldsymbol{\theta} + \mathbf{t}\delta_n) - \ell_n(\mathbf{z},\boldsymbol{\theta})\}] - \mathbf{s}'\boldsymbol{\Sigma}_0\mathbf{t}| = o(1),$$

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^{\delta_n}} |r_n^\gamma h_n^d \mathbb{E}[\{w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta} + \mathbf{s}\delta_n) - w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta})\}\{w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta} + \mathbf{t}\delta_n) - w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta})\}] - \mathcal{C}_0(\mathbf{s},\mathbf{t})| = o(1),$$

and $\mathcal{C}_0(\mathbf{s},\mathbf{s}) + \mathcal{C}_0(\mathbf{t},\mathbf{t}) - 2\mathcal{C}(\mathbf{s},\mathbf{t}) > 0$.

(v) For any positive $\delta_n = O(r_n^{-1})$, $r_n^8 n^{-3}\mathbb{E}|\bar{\ell}_n^{\delta_n}(\mathbf{z})|^4 = o(1)$, $r_n^8 n^{-5}\mathbb{E}[\mathbb{E}[|\bar{w}_n^{\delta_n}(\mathbf{z}_1,\mathbf{z}_2)|^2|\mathbf{z}_1]^2] = o(1)$, $r_n^8 n^{-6}\mathbb{E}[|\bar{w}_n^{\delta_n}(\mathbf{z}_1,\mathbf{z}_2)|^4] = o(1)$, $r_n^8 n^{-4}\mathbb{E}[\mathbb{E}[|\bar{w}_n^{\delta_n}(\mathbf{z}_1,\mathbf{z}_2)\bar{w}_n^{\delta_n}(\mathbf{z}_1,\mathbf{z}_3)||\mathbf{z}_2,\mathbf{z}_3]^2] = o(1)$, and $r_n^8 n^{-3}\mathbb{E}[\mathbb{E}[\bar{\ell}_n^{\delta_n}(\mathbf{z}_1)\bar{w}_n^{\delta_n}(\mathbf{z}_1,\mathbf{z}_2)|\mathbf{z}_2]] = o(1)$.

(vi) One of the following conditions holds:

(a) [$\gamma = 2$] For some $\delta > 0$, there exist functions $B_{2n} : \mathbb{R}^{d_z} \times \mathbb{R}^{d_z} \to \mathbb{R}$ such that $|w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta}') - w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta})| \le B_{2n}(\mathbf{z}_1,\mathbf{z}_2)\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|$ for $\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}_0^{\delta}$, and $h_n^d\mathbb{E}[|B_{2n}(\mathbf{z}_1,\mathbf{z}_2)|^2] = O(1)$.

(b) [$\gamma = 1$] For any $\delta_n = O(r_n^{-1})$, $\sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}_0^{\delta_n}} \mathbb{E}[|w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta}') - w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta})|/\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|] = O(1)$, and $\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{\delta' \in (0,\delta_n]} h_n^d\mathbb{E}[\mathbb{1}\{h_n^d|\bar{w}_n^{\delta'}(\mathbf{z}_1,\mathbf{z}_2)| > C\}|\bar{w}_n^{\delta'}(\mathbf{z}_1,\mathbf{z}_2)|^2/\delta'] = 0$.

The term uniform manageability has been employed in empirical process literature. This condition is satisfied by many classes of functions encountered in practice, and it ensures applicability of uniform law of large numbers and maximal inequalities. For definition, refer to Kim and Pollard (1990).

In the above conditions, the integer $\gamma \in \{1, 2\}$ shows up in multiple occasions. This number denotes smoothness of the function $s(\cdot,\cdot,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. For instance, if $s(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$ almost surely, under some regularity conditions, γ will be 2. On the other hand, in some econometric models, $s(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta})$ can be discontinuous in $\boldsymbol{\theta}$ (e.g., indicator function), and in such situations, $\gamma = 1$. There is an important qualitative difference in the limit distribution for different values of γ. When $\gamma = 2$, $\mathcal{C}_0(\mathbf{s},\mathbf{t}) = \mathbf{s}'\boldsymbol{\Omega}_0\mathbf{t}$ for some positive definite matrix $\boldsymbol{\Omega}_0$, and the limit distribution remains Gaussian. In contrast, $\gamma = 1$ implies $\mathcal{C}_0(\alpha\mathbf{s},\alpha\mathbf{t}) = |\alpha|\mathcal{C}_0(\mathbf{s},\mathbf{t})$ and the limit distribution becomes non-Gaussian. Theorem 7 below unifies distribution theory under varying smoothness of objective functions.

The sequence r_n, which is the convergence rate of $\hat{\boldsymbol{\theta}}_n$, depends on the smoothness measure γ. From the definition, when $n^{\gamma/2}h_n^d$ is bounded away from zero, $r_n = O(\sqrt{n})$, and if $n^{\gamma/2}h_n^d \to 0$, $r_n/(n^2h_n^d)^{-1/(4-\gamma)} \to 1$. Here smoothness of the objective function can affect the convergence rate in two ways. If the function is not smooth i.e., $\gamma = 1$, then $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_n$ is retained for a narrower range of bandwidth sequences. In addition, $\gamma = 1$ implies that when $n^{1/2}h_n^d \to 0$, the convergence rate is cubic root of $n^2h_n^d$. This resembles cubic root

convergence rate of Kim and Pollard (1990) and Seo and Otsu (2018), who also study a class of M-estimation problems with non-smooth objective functions. Notable differences from these previous studies are that I study a maximizer of second-order U-statistic as opposed to a single sum and that the estimator in this problem exhibits a range of different limit distributions.

Most of the other conditions are standard in M-estimation literature. One exception is the second condition, which requires $r_n\|\partial M_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\| = o(1)$, and it ensures that "smoothing" bias is of small order. In light of the discussion on (IV.2), this paper focuses on the variance term and is not concerned about the bias term. This perspective has been termed as "small bandwidth asymptotics" in the literature (e.g., Cattaneo and Jansson, 2018). Another approach has been also considered, where a "larger" bandwidth is assumed so that the variance is of smaller order and the smoothing bias has been dealt explicitly (e.g., Chernozhukov *et al.*, 2018). This paper takes small bandwidth asymptotics as its framework.

Now I present the formal distributional result. For the theory, the estimator needs to maximize the objective function only approximately. The precise formulation is that $\hat{\boldsymbol{\theta}}_n$ satisfies

$$\hat{M}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - o_{\mathbb{P}}(r_n^{-2})$$

where r_n as defined in Assumption SP. Also, let $\mathcal{G}_n$ be a Gaussian process indexed by $\mathbf{s} \in \mathbb{R}^k$ whose mean is $-\mathbf{s}'\mathbf{V}_0\mathbf{s}/2$ and whose covariance kernel is

$$4\Big(\frac{\pi_n}{1+\pi_n}\Big)^2 \mathbf{s}'\boldsymbol{\Sigma}_0'\mathbf{t} + 2\Big(\frac{1}{1+\pi_n}\Big)^{4-\gamma}\mathcal{C}_0(\mathbf{s},\mathbf{t}), \qquad \pi_n = \sqrt{n^{\gamma/2}h_n^d}$$

Theorem 7. *Under Assumption SP, if $n^2h_n^d \to \infty$, then*

$$\lim_{n\to\infty}\sup_{\mathbf{t}\in\mathbb{R}^k}\left|\Pr\left(r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \leq \mathbf{t}\right) - \Pr\left(\arg\max_{\mathbf{s}\in\mathbb{R}^k}\{\mathcal{G}_n(\mathbf{s})\} \leq \mathbf{t}\right)\right| = 0.$$

To understand the implications of this result, first consider cases where $\pi_n = \sqrt{n^{\gamma/2}h_n^d}$ has a limit in the extended real line i.e., $\pi_0 = \lim_{n\to\infty}\pi_n \in [0,+\infty]$ and $\gamma = 2$ i.e., smooth objective function. When $\pi_0 = +\infty$, $\arg\max_{\mathbf{s}\in\mathbb{R}^k}\{\mathcal{G}_n(\mathbf{s})\}$ converges in distribution

to Normal$(\mathbf{0}, 4\mathbf{V}_0^{-1}\mathbf{\Sigma}_0\mathbf{V}_0^{-1})$. This scenario corresponds to the conventional case as discussed in Section IV.2, and inference about $\boldsymbol{\theta}_0$ is straightforward.

When $0 < \pi_0 < \infty$, then the limit distribution has two variance components: one is the normal random vector with the variance $4[\pi_0/(1+\pi_0)]^2\mathbf{\Sigma}_0$ and the other part involves $\mathcal{C}_0(\mathbf{s}, \mathbf{t})$. As previously mentioned, when $\gamma = 2$, $\mathcal{C}_0(\mathbf{s}, \mathbf{t}) = \mathbf{s}'\mathbf{\Omega}_0\mathbf{t}$ for some positive definite matrix $\mathbf{\Omega}_0$, and the limit distribution becomes a mean-zero normal vector with the covariance matrix $\mathbf{V}_0^{-1}[4(\pi_0/(1+\pi_0))^2\mathbf{\Sigma}_0 + 2(1/(1+\pi_0))^2\mathbf{\Omega}_0]\mathbf{V}_0^{-1}$. Note that the middle matrix is the covariance matrix of a convex combination of two random vectors whose covariance matrices are $4\mathbf{\Sigma}_0$ and $2\mathbf{\Omega}_0$, where the weight is given by $\pi_0/(1+\pi_0)$. For $\pi_0 = 0$, the limit distribution becomes a mean-zero normal vector with the covariance matrix $2\mathbf{V}_0^{-1}\mathbf{\Omega}_0\mathbf{V}_0^{-1}$. Therefore, as π_0 ranges over $[0, +\infty]$, the limit distribution varies "continuously," which can be seen from the weight $\pi_0/(1+\pi_0)$ going to 1 as $\pi_0 \to \infty$ and it vanishing as $\pi_0 \to 0$.

For the case $\gamma = 1$, similar results hold although there is an important difference. If $\pi_0 = +\infty$, then the limit distribution is Gaussian, and the conventional distributional approximation works. Yet, when π_0 is finite, then the limit distribution becomes

$$\arg\max_{\mathbf{s}\in\mathbb{R}^k} \Big\{ -\frac{1}{2}\mathbf{s}'\mathbf{V}_0\mathbf{s} + \frac{\pi_0}{1+\pi_0}\mathbf{s}'\mathcal{Z} + \Big(\frac{1}{1+\pi_0}\Big)^{3/2}\mathcal{W}(\mathbf{s})\Big\}$$

where $\mathcal{Z} = \text{Normal}(\mathbf{0}, 4\mathbf{\Sigma}_0)$, $\mathcal{W}$ is a mean-zero Gaussian process whose covariance kernel is $2\mathcal{C}_0(\cdot,\cdot)$, and $\mathcal{Z} \perp\!\!\!\perp \mathcal{W}$. The argmax generally does not have a closed form, and the distribution is non-normal. As in the case with $\gamma = 2$, the limit distribution changes continuously with π_0, although interpretation as a convex combination of two normal vectors does not apply.

Note that when π_n does not converge on the extended real line, arguing along subsequences establishes that the law of $\arg\max_{\mathbf{s}\in\mathbb{R}^k}\{\mathcal{G}_n(\mathbf{s})\}$ continues to well approximate the distribution of $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. Thus, we do not need to assume that the sequence π_n has a well-defined limit.

From practical point of view, Theorem 7 may provide a better distributional approximation in samples of moderate size. The key object here is π_n. If π_n is "large" given the sample size and bandwidth, then we expect that the conventional approximation works well, provided that the bias term is also small. On the other hand, if π_n is not sufficiently large,

accounting for the additional source of variance can improve the distributional approximation. There are different ways in which π_n may be small. For instance, the dimension of conditioning covariates $\mathbf{w}$ can be large (e.g., $d \geq 3$), and in such case, choosing a bandwidth that makes the smoothing bias small and π_n large simultaneously may not be feasible. Furthermore, in practice, researchers may not know whether or not given π_n is sufficiently large because it depends on the underlying data generating process.

IV.4 Conclusion

This chapter develops a novel distributional approximation to a class of smoothed pairwise difference estimators. I not only allow for weaker restrictions on bandwidth sequence but also highlight potential non-robustness of conventional Gaussian approximation. Given this robustness issue, it is desirable to construct inference procedures that automatically adapt to different bandwidth sequences. One may conjecture that resampling methods capture the additional variance term when the bandwidth sequence is small. However, one can show that the standard nonparametric bootstrap fails to approximate the correct limit distribution. In future work, I plan to address this inference problem by developing appropriately modified bootstrap procedures that are robust with respect to bandwidth choice.

APPENDIX A

Proofs for Chapter II

A.1 Proofs

Let $L^p(\mu)$ be the class of functions whose pth power is integrable with respect to μ. I write F_W, F_Θ for the distribution of random vectors W, Θ, and so on.

A.1.1 Proof of Lemma 1

Inverse of $\mathbf{\Psi}$

Recall

$$\mathbf{\Psi} h(w) = \int h(\theta) f_{\Theta|W}(\theta|w) d\lambda(\theta).$$

Let $\mathcal{H}$ be the range of $\mathbf{\Psi}$ on $L^2(F_\Theta)$. I show that $h \in \mathcal{H}$ is square-integrable with respect to F_W. For any $h \in L^2(F_\Theta)$,

$$\begin{aligned}
\int (\mathbf{\Psi} h)^2 dF_W &= \int \Big[\int h(\theta) f_{\Theta|W}(\theta|w) d\lambda(\theta) \Big]^2 dF_W(w) \\
&= \int \big(\mathbb{E}[h(\Theta_{ig})|W_{ig} = w] \big)^2 dF_W(w) \\
&\leq \int \big[\mathbb{E}[|h(\Theta_{ig})|^2 | W_{ig} = w] dF_W(w) \\
&= \mathbb{E}[|h(\Theta_{ig})|^2] < \infty.
\end{aligned}$$

By boundedness of $f_{\Theta|W}$, $\mathbf{\Psi}$ is a compact linear operator. Then, Theorem 15.16 in Kress (2014) implies there exist non-negative reals $\{\tau_j : j \in \mathbb{N}\}$ and orthonormal sequences $\{\phi_j :$

$j \in \mathbb{N}\}$, $\{\varphi_j : j \in \mathbb{N}\}$ in $L^2(F_\Theta)$ and $\mathcal{H}$, respectively, such that

$$\mathbf{\Psi} h = \sum_{j=1}^{\infty} \tau_j \langle h, \phi_j \rangle_\Theta \varphi_j$$

where $\langle h, \phi \rangle_\Theta = \int h \phi dF_\Theta$. Also let $\langle m, \varphi \rangle_W = \int m \varphi dF_W$.

Without loss of generality, assume $\tau_j > 0$ for all j. Define

$$\mathbf{\Psi}^\dagger m = \sum_{j=1}^{\infty} \tau_j^{-1} \langle m, \varphi_j \rangle_W \phi_j.$$

By definition, $\mathbf{\Psi}^\dagger \mathbf{\Psi} h = h$ on $L^2(F_\Theta)$.

Integral with respect to $f_{W|J,\mathbf{B}}$

$f_{W|J,\mathbf{B}}/f_W \in \mathcal{H}$ because

$$\begin{aligned} f_{W|J,\mathbf{B}}(w|s,\mathbf{b}) &= \int f_{W|\Theta,J,\mathbf{B}}(w|\theta,s,\mathbf{b}) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \int f_{W|\Theta,\mathbf{B}}(w|\theta,\mathbf{b}) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \int f_{W|\Theta}(w|\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \mathbf{\Psi}(f_{\Theta|J,\mathbf{B}}/f_\Theta)(w) f_W(w) \end{aligned}$$

and $f_{\Theta|J,\mathbf{B}}/f_\Theta \in L^2(F_\Theta)$, where the second equality holds by $J = J(\Theta, \mathbf{A}, \boldsymbol{\eta})$ and $\boldsymbol{\eta} \perp\!\!\!\perp (W, \Theta)|\mathbf{B}$ and the third follows from independence $W \perp\!\!\!\perp \mathbf{B}|\Theta$.

Then, $\mathbf{\Psi}^\dagger(f_{W|J,\mathbf{B}}/f_W) = f_{\Theta|J,\mathbf{B}}/f_\Theta$. Write $m(\theta) = \mathbb{E}[\omega_{ig}|\Theta_{ig} = \theta]$. If the interchange of

integral and infinite sum is permitted,

$$\begin{aligned}
\mathbb{E}[\omega_{ig}|J_{ig}=s,\mathbf{B}_g=\mathbf{b}] &= \int \mathbb{E}[\omega_{ig}|\Theta_{ig}=\theta,J_{ig}=s,\mathbf{B}_g=\mathbf{b}]f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})d\lambda(\theta) \\
&= \int \mathbb{E}[\omega_{ig}|\Theta_{ig}=\theta]f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})d\lambda(\theta) \\
&= \int m(\theta)\frac{f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})}{f_\Theta(\theta)}dF_\Theta(\theta) \\
&= \int m(\theta)\big(\mathbf{\Psi}^\dagger f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot)\big)(\theta)dF_\Theta(\theta) \\
&= \int m(\theta)\bigg[\sum_{j=1}^\infty \tau_j^{-1}\langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot),\varphi_j\rangle_W\phi_j(\theta)\bigg]dF_\Theta(\theta) \\
&= \sum_{j=1}^\infty \tau_j^{-1}\langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot),\varphi_j\rangle_W \int m(\theta)\phi_j(\theta)dF_\Theta(\theta) \qquad \text{(A.1)} \\
&= \int \bigg[\sum_{j=1}^\infty \tau_j^{-1}\int m(\theta)\phi_j(\theta)dF_\Theta(\theta)\varphi_j(w)\bigg]f_{W|J,\mathbf{B}}(w|s,\mathbf{b})d\mu(w) \qquad \text{(A.2)}
\end{aligned}$$

where I use $\omega_{ig} \perp\!\!\!\perp \mathbf{B}_g|\Theta_{ig}$ for the second equality. To justify (A.1), it suffices to show

$$\sum_{j=1}^\infty \int \tau_j^{-1}\big|\langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot),\varphi_j\rangle_W\big||m(\theta)\phi_j(\theta)|dF_\Theta(\theta) < \infty.$$

Using $f_{W|J,\mathbf{B}}/f_W = \mathbf{\Psi} f_{\Theta|J,\mathbf{B}}/f_\Theta$,

$$\begin{aligned}
\langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot),\varphi_j\rangle_W &= \langle \mathbf{\Psi}(f_{\Theta|J,\mathbf{B}}/f_\Theta),\varphi_j\rangle_W \\
&= \Big\langle \sum_{\ell=1}^\infty \lambda_\ell\langle f_{\Theta|J,\mathbf{B}}/f_\Theta,\phi_\ell\rangle_\Theta\varphi_\ell,\varphi_j\Big\rangle_W \\
&= \sum_{\ell=1}^\infty \lambda_\ell\langle f_{\Theta|J,\mathbf{B}}/f_\Theta,\phi_\ell\rangle_\Theta\langle\varphi_\ell,\varphi_j\rangle_W \\
&= \tau_j\langle f_{\Theta|J,\mathbf{B}}/f_\Theta,\phi_j\rangle_\Theta.
\end{aligned}$$

Thus, it suffices to have

$$\sum_{j=1}^\infty \big|\langle f_{\Theta|J,\mathbf{B}}/f_\Theta,\phi_j\rangle_\Theta\big|\langle|m|,|\phi_j|\rangle_\Theta \le \|m\|_\Theta \sum_{j=1}^\infty \big|\langle f_{\Theta|J,\mathbf{B}}/f_\Theta,\phi_j\rangle_\Theta\big| < \infty$$

where $\|\cdot\|_\Theta$ is the norm induced by $\langle\cdot,\cdot\rangle_\Theta$ and I use $\|\phi_j\|_\Theta = 1$ for $j \in \mathbb{N}$. For (A.2), it suffices to show

$$\begin{aligned}
&\sum_{j=1}^{\infty} \int \tau_j^{-1} |f_{W|J,\mathbf{B}}(w|s,\mathbf{b})\varphi_j(w)| \big|\langle m, \phi_j\rangle_\Theta\big| d\mu(w) \\
&\leq \sum_{j=1}^{\infty} \tau_j^{-1} \big|\langle m, \phi_j\rangle_\Theta\big| \Big[\int |f_{W|J,\mathbf{B}}(w|s,\mathbf{b})|^2 d\mu(w) \int |\varphi_j(w)|^2 d\mu(w) \Big]^{1/2} \\
&\leq C \sum_{j=1}^{\infty} \tau_j^{-1} \big|\langle m, \phi_j\rangle_\Theta\big| < \infty.
\end{aligned}$$

Thus, the interchageability follows from Assumption 4.

Since

$$\psi(w) = \sum_{j=1}^{\infty} \tau_j^{-1} \alpha_j \varphi_j(w), \qquad \alpha_j = \langle m, \phi_j\rangle_\Theta,$$

$\{\alpha_j/\tau_j : j \in \mathbb{N}\}$ is square-summable, and $\{\varphi_j : j \in \mathbb{N}\}$ is an orthonormal set with respect to F_W, ψ is square integrable with respect to F_W.

A.1.2 Proof of Theorem 1

By taking conditional expectations given $F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)$,

$$\begin{aligned}
&\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\
&= \mathbb{E}[X_{J_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]'\beta + \mathbb{E}[W_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]'\gamma + \int \psi(w) dF_{W|J,\mathbf{B}}(w|J_{ig},\mathbf{B}_g)
\end{aligned}$$

and subtracting this conditional expectation from the equation (II.11),

$$\begin{aligned}
&Y_{iJ_{ig}g} - \mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\
&= \big\{X_{J_{ig}g} - \mathbb{E}[X_{J_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]\big\}'\beta + \big\{W_{ig} - E[W_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]\big\}'\gamma + \varepsilon_{ig}.
\end{aligned}$$

Writing $V_{ig} = (X'_{J_{ig}g}\, W'_{ig})'$,

$$\mathbb{E}[\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}_g}]\}Y_{iJ_{ig}g}] = \mathbb{E}[\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}_g}]\}\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}_g}]\}'] \begin{pmatrix} \beta \\ \gamma \end{pmatrix}.$$

Then, the invertibility condition 2 guarantees identifiability of β.

A.1.3 Proof of Theorem 2

Below, I use the following notational conventions. For two sequences of real numbers a_n and b_n, $a_n \lesssim b_n$ means there exists a constant C not dependent on n such that $a_n/b_n \leq C$ for n large enough. For sequences of random variables X_n and Y_n, $X_n \lesssim_{\mathbb{P}} Y_n$ means that there exists a fixed constant C satisfying $\Pr(X_n \leq CY_n) \to 1$ as $n \to \infty$. Given a symmetric matrix $\mathbf{X}$, I write $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ for largest and smallest eigenvalues of the matrix $\mathbf{X}$. Define $\bar{\lambda}_K = \lambda_{\max}(\mathbf{Q}_K)$ and $\tilde{\lambda} = 1/\lambda_{\min}(\mathbf{Q}_K)$.

Also, define

$$\boldsymbol{r}_K = \Big(\sum_{k=K+1}^{\infty} \delta_k \mathbb{E}[\rho_k(W_{11})|J_{11}, \mathbf{B}_1], \ldots, \sum_{k=K+1}^{\infty} \delta_k \mathbb{E}[\rho_k(W_{NG})|J_{NG}, \mathbf{B}_G] \Big)'$$
$$\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{NG})'.$$

Proof. Lemma 4 below implies the first part. By $\mathrm{Var}(\chi_s - \mathbb{E}[\omega_{ig}|J_{ig} = s, \mathbf{B}_g]) > 0$, $\mathbb{E}[\boldsymbol{\varepsilon}_{isg}\boldsymbol{\varepsilon}_{jsg}] > 0$ and the minimum eigenvalue of

$$\mathbb{E}[\psi_{Ng}\psi_{Ng}'] = \frac{1}{N}\mathbb{E}[P_{Kig}P_{Kig}'\varepsilon_{ig}^2] + (1 - 1/N)\mathbb{E}[P_{Kig}P_{Kjg}'\varepsilon_{ig}\varepsilon_{jg}]$$

is bounded from below uniformly in K. In addition, $\|\mathbf{Q}_K^{-1}\| \geq 1/\lambda_{\max}(\mathbf{Q}_K)$, which is bounded away from zero uniformly. Thus, $\lambda_{\min}(\boldsymbol{\Omega}_K)$ is bounded away from zero uniformly.

For the asymptotic normality of the oracle estimator, note ψ_{Ng} is mean zero and independent across t. I verify Lindberg's condition. The variance matrix is the identity matrix by construction. For any vector $\mathbf{v}$, I want to show for all $d > 0$,

$$\mathbb{E}[|\mathbf{v}'\boldsymbol{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}|^2 I\{|\mathbf{v}'\boldsymbol{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}| \geq d\sqrt{G}\}] \to 0.$$

Letting $\varpi_{ig} = \mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}P_{Kig}$, we have $\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng} = \sum_i \varpi_{ig}\varepsilon_{ig}/N$ and $|\varpi_{ig}| \lesssim \zeta_K$.

$$
\begin{aligned}
&\mathbb{E}[|\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}|^2 I\{|\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}| \geq d\sqrt{G}\}] \\
&\leq \mathbb{E}\frac{1}{N}\sum_{i=1}^N |\varpi_{ig}|^2 \frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2 I\Big\{\Big|\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2\Big|^{1/2} \geq d\sqrt{G}/\zeta_K\Big\} \\
&\leq \mathbb{E}[|\varpi_{ig}|^2] \sup_{\mathbf{P}_{Kg}} \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2 I\Big\{\Big|\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2\Big|^{1/2} \geq d\sqrt{G}/\zeta_K\Big\}\Big|\mathbf{P}_{Kg}\Big] \\
&\leq \mathbb{E}[|\varpi_{ig}|^2]\frac{\zeta_K^2}{d^2 G} \sup_{\mathbf{P}_{Kg}} \mathbb{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2\Big)^2\Big|\mathbf{P}_{Kg}\Big] \lesssim \zeta_K^4/G \to 0.
\end{aligned}
$$

□

Now I prove the asymptotic (first-order) equivalency between the oracle and feasible estimators. Let $\theta = (\beta', \gamma', \delta_1, \ldots, \delta_K)$ and we have

$$
\begin{aligned}
\hat{\beta} - \beta &= \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}'_K(\mathbf{P}_K - \hat{\mathbf{P}}_K)\theta/NG + \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}'_K\boldsymbol{r}_K/NG + \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}'_K\boldsymbol{\varepsilon}/NG \\
&\equiv \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}
$$

For I, Lemmas 4, 5, and 7 imply that it is $O_{\mathbb{P}}(\tilde{\theta}_K \max\{\zeta_K, \sqrt{K}\}\sqrt{K\log(NG)/NG})$. For II, the calculation done in Lemma 4 yields that it is $O_{\mathbb{P}}(\sqrt{\tilde{\lambda}_K}K^{-b})$, and for III,

$$
\begin{aligned}
\mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}'_K\boldsymbol{\varepsilon}/NG &= \mathbf{S}\{\tilde{\mathbf{Q}}_K^{-1} - \hat{\mathbf{Q}}_K^{-1}\}(\hat{\mathbf{P}}_K - \mathbf{P}_K)'\boldsymbol{\varepsilon}/NG + \mathbf{S}\{\tilde{\mathbf{Q}}_K^{-1} - \hat{\mathbf{Q}}_K^{-1}\}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG \\
&\quad + \hat{\mathbf{Q}}_K^{-1}(\hat{\mathbf{P}}_K - \mathbf{P}_K)'\boldsymbol{\varepsilon}/NG + \hat{\mathbf{Q}}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG \\
&= \hat{\mathbf{Q}}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG + O_{\mathbb{P}}(\tilde{\lambda}_K^2 \max\{\zeta_K, \sqrt{K}\}\sqrt{K\log(NG)/NG} + \tilde{\lambda}_K\sqrt{K/NG}).
\end{aligned}
$$

Lemmas for Proof of Theorem 2

Lemma 4. *Assume* $\zeta_K^2\tilde{\lambda}_K^2\bar{\lambda}_K \log K/G \to 0$.

$$
\hat{\beta}_{oracle} - \beta = \mathbf{S}\mathbf{Q}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG + O_{\mathbb{P}}((\tilde{\lambda}_K)^{1/2}K^{-b}) + O_{\mathbb{P}}\Big(\frac{\tilde{\lambda}_K^2\zeta_K\bar{\lambda}_K\sqrt{\log K}}{G}\Big)
$$

where $\zeta_K = (NG)^{1/m_w} + G^{1/m_x} + \xi_K$.

Proof. By decomposition, $\hat{\beta}_{\text{oracle}} = \beta + \mathbf{S}(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K(\boldsymbol{r}_K + \boldsymbol{\varepsilon})$ and

$$\hat{\beta}_{\text{oracle}} - \beta = \mathbf{S}\mathbf{Q}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG + \mathbf{S}(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K\boldsymbol{r}_K + \mathbf{S}\big\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1} - \mathbf{Q}_K^{-1}\big\}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG.$$

First, Lemma 6.2 in Belloni *et al.* (2015) implies

$$\|(\mathbf{P}'_K\mathbf{P}_K/NG) - \mathbf{Q}_K\| = O_{\mathbb{P}}\Big(\sqrt{\frac{\zeta_K^2\bar{\lambda}_K\log K}{G}}\Big). \tag{A.3}$$

Thus, the smallest eigenvalue of $\hat{\mathbf{Q}}_K = (\mathbf{P}'_K\mathbf{P}_K/NG)$ is bounded below by $\lambda_{\min}(\mathbf{Q}_K)\{1 + o_{\mathbb{P}}(1)\}$ if $\zeta_K^2\tilde{\lambda}_K^2\bar{\lambda}_K\log K/G \to 0$. Using Frobenius norm $\|\cdot\|_F$,

$$\begin{aligned}\|(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K\boldsymbol{r}_K\|_F^2 &= \boldsymbol{r}'_K\mathbf{P}_K\hat{\mathbf{Q}}_K^{-1/2}\hat{\mathbf{Q}}_K^{-1}\hat{\mathbf{Q}}_K^{-1/2}\mathbf{P}'_K\boldsymbol{r}_K/(NG)^2\\ &\lesssim_{\mathbb{P}} \tilde{\lambda}_K\boldsymbol{r}'_K\mathbf{P}_K(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K\boldsymbol{r}_K/NG\\ &\lesssim \tilde{\lambda}_K\boldsymbol{r}'_K\boldsymbol{r}_K/NG = O_{\mathbb{P}}(\tilde{\lambda}_K K^{-2b}).\end{aligned}$$

For $\mathbf{S}\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1} - \mathbf{Q}_K^{-1}\}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG$, since ε_{ig} is conditionally mean zero given P_{Kig}, it suffices to look at its variance. For the conditional variance given $\mathbf{P}_K$,

$$\begin{aligned}&N^{-2}G^{-2}\Big\|\mathbf{S}\Big\{\Big(\frac{\mathbf{P}'_K\mathbf{P}_K}{NG}\Big)^{-1} - \mathbf{Q}_K^{-1}\Big\}\sum_{g=1}^{G}\mathbf{P}'_{Kg}\mathbb{E}[\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g|\mathbf{P}_{Kg}]\mathbf{P}_{Kg}\Big\{\Big(\frac{\mathbf{P}'_K\mathbf{P}_K}{NG}\Big)^{-1} - \mathbf{Q}_K^{-1}\Big\}'\mathbf{S}'\Big\|\\ &\le G^{-1}\Big\|\mathbf{S}\big\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1} - \mathbf{Q}_K^{-1}\big\}\frac{1}{NG}\sum_{g=1}^{G}\sum_{i=1}^{N}P_{Kig}P'_{Kig}\big\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1} - \mathbf{Q}_K^{-1}\big\}'\mathbf{S}'\Big\|\\ &= G^{-1}\Big\|\mathbf{S}\big\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1} - \mathbf{Q}_K^{-1}\big\}\Big\|^2\|\mathbf{P}'_K\mathbf{P}_K/NG\| = O_{\mathbb{P}}\Big(G^{-1}\tilde{\lambda}_K^4\frac{\zeta_K^2\bar{\lambda}_K^2\log K}{G}\Big)\end{aligned}$$

where I use $A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$. □

Lemma 5. *Let* $\tilde{\mathbf{Q}}_K = (\hat{\mathbf{P}}'_K\hat{\mathbf{P}}_K/NG)^{-1}$. *Under* $\max\{\zeta_K, \sqrt{K}\}\tilde{\lambda}_K\sqrt{K\log(NG)/NG} \to 0$,

$$\|\tilde{\mathbf{Q}}_K^{-1}\| \lesssim_{\mathbb{P}} \|\hat{\mathbf{Q}}_K^{-1}\|$$

Proof. Using Lemma 7

$$\begin{aligned}\|\hat{\mathbf{P}}'_K\hat{\mathbf{P}}_K/NG - \mathbf{P}'_K\mathbf{P}_K/NG\| &\leq \|\hat{\mathbf{P}}_K - \mathbf{P}_K\|^2/NG + 2\|(\hat{\mathbf{P}}_K - \mathbf{P}_K\|/NG \\ &= O_{\mathbb{P}}\big(\max\{\zeta_K, \sqrt{K}\}\sqrt{K\log(NG)/NG}\big).\end{aligned}$$

□

Lemma 6. *If Assumption 6 holds, then*

$$\max_{\substack{1\leq g\leq G\\ 1\leq i\leq N}} \|\hat{P}_{Kig} - P_{Kig}\| = O_{\mathbb{P}}(\sqrt{K/N}).$$

Proof. By definition,

$$\|\hat{P}_{Kig} - P_{Kig}\|^2 = \sum_{\ell=1}^{K}\Big(\frac{1}{N_{J_{ig}g}}\sum_{j=1}^{N}\rho_\ell(W_{jg})I\{J_{jg} = J_{ig}\} - \mathbb{E}[\rho_\ell(W_{jg})|J_{jg} = J_{ig}, \mathbf{B}_g]\Big)^2$$

and using $N_{sg}/N \geq c > 0$ for all $s \in \mathcal{S}$ and t with probability approaching one,

$$\begin{aligned}&\frac{1}{N_{J_{ig}g}}\sum_{j=1}^{N}\rho_\ell(W_{jg})I\{J_{jg} = J_{ig}\} - \mathbb{E}[\rho_\ell(W_{jg})|J_{jg} = J_{ig}, \mathbf{B}_g] \\ &= O_{\mathbb{P}}(1)\frac{1}{N}\sum_{j\neq i}\big(\rho_\ell(W_{jg})I\{J_{jg} = J_{ig}\} - \mathbb{E}[\rho_\ell(W_{jg})|J_{jg} = J_{ig}, \mathbf{B}_g]\big) + \frac{\rho(W_{ig})}{N}O_{\mathbb{P}}(1).\end{aligned}$$

Conditional on J_{ig} and $\mathbf{B}_g$, the summand is independent across j and Corollary 4.3 in Pollard (1989) implies that

$$\sup_{\ell\in\mathbb{N}}\Big|\frac{1}{\sqrt{N}}\sum_{j\neq i}\big(\rho_\ell(W_{jg})I\{J_{jg} = J_{ig}\} - \mathbb{E}[\rho_\ell(W_{jg})|J_{jg} = J_{ig}, \mathbf{B}_g]\big)\Big| = O_{\mathbb{P}}(1).$$

The uniformity of the bound with respect to the conditioning variables $(J_{ig}, \mathbf{B}_g)$ indicates that the above bound hold uniformly in $(i, g) \in \{1, \ldots, N\} \times \{1, \ldots, G\}$. □

Lemma 7. *Under the hypothesis of Lemma 6,*

$$\|\mathbf{P}'_K(\hat{\mathbf{P}}_K - \mathbf{P}_K)\| = O_{\mathbb{P}}(\sqrt{GNK\log(NG)}\max\{\zeta_K, \sqrt{K}\}).$$

Proof.

$$\mathbf{P}_K'(\hat{\mathbf{P}}_K - \mathbf{P}_K) = \sum_{g=1}^{G}\sum_{i=1}^{N} P_{Kig}(\hat{P}_{Kig} - P_{Kig})' \equiv \sum_{g=1}^{G} \Xi_g.$$

I use Theorem 1.6 in Tropp (2012). Using Lemma 6, $\|\hat{P}_{Kig} - P_{Kig}\| \leq C\sqrt{K/N}$ with probability arbitrary close to one with some large constant C and

$$\|\Xi_g\| \leq C\sum_{i=1}^{N}\sum_{j\neq i} \|P_{Kig}\|\sqrt{K/N} \leq C\zeta_K\sqrt{NK}$$

For the variance terms,

$$\begin{aligned}
\|\mathbb{E}[\Xi_g\Xi_g']\| &= \left\| \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[P_{Kig}P_{Kjg}'(\hat{P}_{Kig} - P_{Kig})'(\hat{P}_{Kjg} - P_{Kjg})] \right\| \\
&\lesssim \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[\|P_{Kig}P_{Kjg}'\|]K/N \leq N^2\zeta_K^2(K/N) = \zeta_K^2 NK \\
\|\mathbb{E}[\Xi_g'\Xi_g]\| &= \left\| \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[(\hat{P}_{Kjg} - P_{Kjg})(\hat{P}_{Kig} - P_{Kig})'P_{Kig}'P_{Kjg}] \right\| \\
&\leq \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[\|\hat{P}_{Kjg} - P_{Kjg}\|\|\hat{P}_{Kig} - P_{Kig}\||P_{Kig}'P_{Kjg}|] \lesssim N^2(K/N)K = K^2N.
\end{aligned}$$

Then, Theorem 1.6 in Tropp (2012) implies

$$\Pr\left(\left\|\sum_{g=1}^{G}\Xi_g\right\| \geq M\right) \leq 2NG\exp\left(\frac{-M^2/2}{NGK\max\{\zeta_K^2, K\} + \zeta_K\sqrt{NK}M/3}\right)$$

and taking $M = \sqrt{TNK\log(NG)}\max\{\zeta_K, \sqrt{K}\}$ shows the desired result. □

Lemma 8. *If Assumptions 6 and 8 hold,*

$$\|(\hat{\mathbf{P}}_K - \mathbf{P}_K)'\boldsymbol{\varepsilon}/NG\| = O_{\mathbb{P}}(\sqrt{K/NG}).$$

Proof.

$$\begin{aligned}\mathbb{E}\|(\hat{\mathbf{P}}_K - \mathbf{P}_K)'\boldsymbol{\varepsilon}/NG\|^2 &= \mathbb{E}\Big\|\frac{1}{NG}\sum_{g=1}^{G}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})\varepsilon_{ig}\Big\|^2 \\ &= \frac{1}{N^2G^2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[(\hat{P}_{Kig} - P_{Kig})(\hat{P}_{Kjg} - P_{Kjg})\varepsilon_{ig}\varepsilon_{jg}] \\ &\lesssim G^{-1}\frac{K}{N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[|\varepsilon_{ig}\varepsilon_{jg}|].\end{aligned}$$

□

A.1.4 Proof of Theorem 3

Recall

$$\hat{\mathbf{\Omega}}_K = \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{\Sigma}}_K\tilde{\mathbf{Q}}_K^{-1}\mathbf{S}'$$

where $\tilde{\mathbf{Q}}_K = \hat{\mathbf{P}}'_K\hat{\mathbf{P}}_K/NG$,

$$\hat{\mathbf{\Sigma}}_K = \frac{1}{G}\sum_{g=1}^{G}\hat{\psi}_{Ng}\hat{\psi}'_{Ng} \qquad \hat{\psi}_{Ng} = \frac{1}{N}\sum_{i=1}^{N}\hat{P}_{Kig}(Y_{ig} - \hat{P}'_{Kig}\hat{\theta})$$

and $\hat{\theta} = (\hat{\mathbf{P}}'_K\hat{\mathbf{P}}_K)^{-1}(\hat{\mathbf{P}}'_K\mathbf{Y})$. Using the lemmas below,

$$\|\hat{\mathbf{\Omega}}_K^{-1/2} - \mathbf{\Omega}_K^{-1/2}\| \leq \|\hat{\mathbf{\Omega}}_K^{-1/2}\|\|\hat{\mathbf{\Omega}}_K^{1/2} - \mathbf{\Omega}_K^{1/2}\|\|\mathbf{\Omega}_K^{-1/2}\| \leq \tilde{\lambda}_K\{2\bar{\lambda}_K^{-1/2}\}^{-1}\|\hat{\mathbf{\Omega}}_K - \mathbf{\Omega}_K\|$$

and

$$\|\hat{\mathbf{\Omega}}_K - \mathbf{\Omega}_K\| = O_{\mathbb{P}}(\|\hat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}\|\|\hat{\mathbf{\Sigma}}_K\|\|\hat{\mathbf{Q}}_K^{-1}\| + \|\hat{\mathbf{Q}}_K^{-1}\|^2\|\hat{\mathbf{\Sigma}}_K - \mathbf{\Sigma}_K\|),$$

which shows the result.

Lemma 9. *Under the hypothesis of Theorem 3,*

$$\Big\|\hat{\mathbf{\Sigma}}_K - \mathbf{\Sigma}\Big\| = O_{\mathbb{P}}(\zeta_K^2(\zeta_K\sqrt{K/G} + \sqrt{K/N} + K^{-\alpha}) + \zeta_K\sqrt{(NG)^{1/m_\varepsilon}\log K/G}).$$

Proof.

$$\|\hat{\mathbf{\Omega}}_K - \mathbf{\Omega}_K\| \leq \|\hat{\mathbf{\Omega}}_K - \frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\psi_{Ng}'\| + \|\frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\psi_{Ng}' - \mathbf{\Omega}_K\|.$$

For the second term, Lemma 6.2 in Belloni *et al.* (2015) implies that it is $O_{\mathbb{P}}(\zeta_K\sqrt{(NG)^{1/m_\varepsilon}\log K/G})$. For the first term,

$$\hat{\mathbf{\Sigma}}_K = \frac{1}{G}\sum_{g=1}^{G}\big(\hat{\psi}_{Ng} - \psi_{Ng}\big)\hat{\psi}_{Ng}' + \frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\big(\hat{\psi}_{Ng} - \psi_{Ng}\big)' + \frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\psi_{Ng}'$$

and

$$\begin{aligned}
\hat{\psi}_{Ng} - \psi_{Ng} &= \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})(Y_{ig} - \hat{P}_{Kig}'\hat{\theta}) - \frac{1}{N}\sum_{i=1}^{N}P_{Kig}(\hat{P}_{Kig}'\hat{\theta} - P_{Kig}'\theta) + \frac{1}{N}\sum_{i=1}^{N}P_{Kig}r_{ig} \\
&= \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})(\varepsilon_{ig} + r_{ig}) + \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})(P_{Kig}'\theta - \hat{P}_{Kig}'\hat{\theta}) \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}P_{Kig}(\hat{P}_{Kig}'\hat{\theta} - P_{Kig}'\theta) + \frac{1}{N}\sum_{i=1}^{N}P_{Kig}r_{ig}.
\end{aligned}$$

Note $\|\hat{P}_{Kig} - P_{Kig}\| = O_{\mathbb{P}}(\sqrt{K/N})$ and $\|\hat{\theta} - \theta\| = O_{\mathbb{P}}(\sqrt{K/G})$. The latter follows from

$$\|\hat{\theta} - \theta\| = \|\mathbf{Q}_K^{-1}\mathbf{P}_K'\boldsymbol{\varepsilon}/NG\| + o_{\mathbb{P}}(G^{-1/2}) = O_{\mathbb{P}}(\sqrt{K/G}).$$

Then,

$$\begin{aligned}
\|P_{Kig}'\theta - \hat{P}_{Kig}'\hat{\theta}\| &\leq \|P_{Kig}\|\|\hat{\theta} - \theta\| + \|\hat{P}_{Kig} - P_{Kig}\|\hat{\theta}\| \\
&= O_{\mathbb{P}}(\zeta_K\sqrt{K/G} + \sqrt{K/N})
\end{aligned}$$

and

$$\left\|\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig}-P_{Kig})(\varepsilon_{ig}+r_{ig})\right\| = O_{\mathbb{P}}(\sqrt{K/N})$$
$$\left\|\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig}-P_{Kig})(P'_{Kig}\theta-\hat{P}'_{Kig}\hat{\theta})\right\| = O_{\mathbb{P}}(K/N+\zeta_K K/\sqrt{NG})$$
$$\left\|\frac{1}{N}\sum_{i=1}^{N}P_{Kig}(\hat{P}'_{Kig}\hat{\theta}-P'_{Kig}\theta)\right\| = O_{\mathbb{P}}(\zeta_K^2\sqrt{K/G}+\zeta_K\sqrt{K/N})$$
$$\left\|\frac{1}{N}\sum_{i=1}^{N}P_{Kig}r_{ig}\right\| = O_{\mathbb{P}}(\zeta_K K^{-\alpha}).$$

Thus, $\|\hat{\psi}_{Ng}-\psi_{Ng}\| = O_{\mathbb{P}}(\zeta_K(\zeta_K\sqrt{K/G}+\sqrt{K/N}+K^{-\alpha}))$ and the conclusion follows from $\|\psi_{Ng}\|+\|\hat{\psi}_{Ng}\| = O_{\mathbb{P}}(\zeta_K)$. □

Lemma 10 (Lemma 2.2 in Schmitt (1992))**.** *Let A and B two symmetric matrices satisfying $A \succ \mu_a^2\mathrm{I}$, $B \succ \mu_b^2\mathrm{I}$, where $A \succ B$ denotes $A-B$ is positive definite. Then,*

$$\|A^{1/2}-B^{1/2}\| \leq \{\mu_a+\mu_b\}^{-1}\|A-B\|.$$

A.1.5 Proof of Theorem 4

By taking conditional expectation give $F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)$,

$$\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}] = \mathbb{E}[X_{J_{ig}g}|F_{W|J,\mathbf{B}}]'\beta + \mathbb{E}[W_{ig}|F_{W|J,\mathbf{B}}]'\gamma + \int \psi(w)dF_{W|J,\mathbf{B}}(w|J_{ig},\mathbf{B}_g).$$

Then, subtracting it from the original equation,

$$Y_{iJ_{ig}g}-\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}] = \{V_{ig}-\mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}]\}'\begin{pmatrix}\beta\\ \gamma\end{pmatrix}+\varepsilon_{ig}$$

where $V_{ig} = (X'_{J_{ig}g}, W'_{ig})'$. Then, letting $\tilde{Z}_{ig} = (Z'_{J_{ig}g}, W'_{ig})'$,

$$\mathbb{E}[\tilde{Z}_{ig}\{Y_{iJ_{ig}g}-\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}]\}] = \mathbb{E}[\tilde{Z}_{ig}\{V_{ig}-\mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}]\}']\begin{pmatrix}\beta\\ \gamma\end{pmatrix}$$

and full column rank of the matrix implies identifiability of β.

A.1.6 Proof of Theorem 5

By the independence assumption

$$\begin{aligned} f_{W,\varepsilon|J,\mathbf{B}}(w,e|s,\mathbf{b}) &= \int f_{W,\varepsilon|J,\mathbf{B},\Theta}(w,e|s,\mathbf{b},\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\theta \\ &= \int f_{W,\varepsilon|\Theta}(w,e|\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\theta \end{aligned}$$

and similarly,

$$f_{W|J,\mathbf{B}}(w|s,\mathbf{b}) = \int f_{W|\Theta}(w|\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\theta.$$

By injectivity of $\mathbf{\Psi}$, we have

$$\mathbf{\Psi}^{\dagger}\big(f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})\big)(\theta) = f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}).$$

Then, conditioning on $F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)$, the function $f_{W,\varepsilon|J,\mathbf{B}}$ is non-stochastic. Therefore,

$$\begin{aligned} &\mathbb{E}[Y_{iJ_{ig}g}|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\ &= \mathbb{E}[m(x,W_{ig},\varepsilon_{iJ_{ig}})|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\ &= \mathbb{E}[\mathbb{E}[m(x,W_{ig},\varepsilon_{iJ_{ig}})|X_{J_{ig}g}=x, J_{ig},\mathbf{B}_g]|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\ &= \mathbb{E}\Big[\int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|X_{J_{ig}g}=x,\mathbf{B}_g]\Big|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)\Big] \\ &= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]d(w,e) \\ &= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]d(w,e) \end{aligned}$$

where the third equality use the display below, the second-to-last equality uses the Fubini theorem to interchange the order of integration, and the last equality uses $f_{W\varepsilon|J,\mathbf{B}}$ is non-

stochastic conditional on $F_{W|J,\mathbf{B}}$. For the third equality in the above display,

$$\begin{aligned}\mathbb{E}[m(x, W_{ig}, \varepsilon_{iJ_{ig}g})|J_{ig}, \mathbf{B}_g] &= \sum_{s=1}^{S} \mathbb{E}[m(x, W_{ig}, \varepsilon_{isg})|J_{ig} = s, \mathbf{B}_g] \Pr(J_{ig} = s|\mathbf{B}_g) \\ &= \sum_{s=1}^{S} \int m(x, w, e) f_{W,\varepsilon|J,\mathbf{B}}(w, e|s, \mathbf{B}_g) d(w, e) \Pr(J_{ig} = s|\mathbf{B}_g) \\ &= \int m(x, w, e) \mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w, e|J_{ig}, \mathbf{B}_g)|\mathbf{B}_g] d(w, e).\end{aligned}$$

Finally, letting ν be the measure corresponding to the distribution of $F_{W|J,\mathbf{B}}$, which is identifiable from the data,

$$\begin{aligned}&\int \mathbb{E}[Y_{iJ_{ig}g}|X_{J_{ig}g} = x, F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g) = F] d\nu(F) \\ &= \int m(x, w, e) \mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w, e|J_{ig}, \mathbf{B}_g)|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g) = F] d(w, e) d\nu(F) \\ &= \int m(x, w, e) \mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w, e|J_{ig}, \mathbf{B}_g)] d(w, e) \\ &= \int m(x, w, e) f_{W,\varepsilon}(w, e) d(w, e).\end{aligned}$$

A.2 Details of DGPs for Monte Carlo Studies

The details of the DGPs used in the simulation studies are following. I use $G = 1,000$ and for each city, $N = 300$ and $S = 3$ with the constraint that each group can have at most 115 people. The final sample from the NLS72 dataset contains 917 schools, which I treated as independent draws. This assumption is valid if schools are geographically isolated from each other.

The econometric model generating the data is

$$\begin{aligned}Y_{is} &= X_s\beta + W_i'\gamma + \omega_i + \chi_s + \epsilon_{is} \\ J_i &= \underset{s\in\{1,2,3\}}{\arg\max}\{\Theta_i' A_s + \eta_{is}\}\end{aligned}$$

where $\beta = -0.15$ and $\gamma = (-0.0003, 0.06)'$ are taken from an estimate in the NLS72. Also,

W_i has a bivariate normal distribution with mean $(12.37, 10.92)$ and covariance matrix

$$\begin{bmatrix} 4.25 & 0.49 \\ 0.49 & 0.42 \end{bmatrix}.$$

The means and covariance matrix are based on the NLS72 data. The preference coefficient Θ_i is generated by

$$\Theta_i = W_i + \omega_i \mathbf{1} + \nu_i$$

where $\mathbf{1}$ is 2×1 vector whose elements are unity and ν_i is a mean-zero bivariate normal with covariance matrix $\begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$. For χ_s and ϵ_{is}, they have a mean-zero normal distribution with standard deviation 0.446, which imitates the standard error of a regression from the data. The idiosyncratic term η_{is} is drawn from mean-zero normal distribution with standard deviation 2. For X_s, I use the empirical distribution of a school-level variable observed in the NLS72 dataset; A_s consists of X_s and a binary variable that takes value 1 with probability 0.29 and has correlation of 0.25 with X_s.

For ω_i, each DGP generates this variable in different ways. For DGP1, ω_i has a joint normal distribution with W_i, its mean is 2.16, variance is 0.51, and the covariance with W_i is 0.051, 0.024. The joint normality leads to the linear specification of the control function. For DGP2, I model

$$\omega_i = 0.1 * u_i + 2 * I\{W_{1i} + W_{2i} - \mathbb{E}[W_{1i} + W_{2i}] > v_i\}$$

where u_i has the same distribution as ω_i in DGP1 (jointly normal with W_i) and $v_i =_d$ Normal$(0, 0.2)$. For the third DGP,

$$\omega_i = 0.5 * W_{1i}^2 + 0.5 * W_{2i}^3$$

where W_{1i} and W_{2i} represent the first and second component of W_i.

In terms of implementation, I generate the random variables and compute utilities of choosing different groups for each individual. Then, I use linear programming to find an

allocation that maximizes the sum of utilities under the group size constraint. Altonji and Mansfield use this step in their simulation and I follow their procedure.

APPENDIX B

Proofs for Chapter III

B.1 Proofs

B.1.1 Weak Convergence of Original Estimator

Under the assumptions of the theorem,

$$r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max}\{Q_n(\mathbf{s}) + \hat{G}_n(\mathbf{s})\} + o_{\mathbb{P}}(1),$$

where

$$Q_n(\mathbf{s}) = r_n^2[M_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - M_n(\boldsymbol{\theta}_0)]\mathbb{1}(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta})$$

and

$$\hat{G}_n(\mathbf{s}) = r_n^2[\hat{M}_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - \hat{M}_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) + M_n(\boldsymbol{\theta}_0)]\mathbb{1}(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta}).$$

Moreover, $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}}(1)$ by Lemmas 11 and 13 and $Q_n + \hat{G}_n \rightsquigarrow \mathcal{Q}_0 + \mathcal{G}_0$ by Lemmas 12, 14, and 15. The result now follows from the argmax continuous mapping theorem.

Consistency

Lemma 11. *Suppose Condition CRA(i) holds and suppose $r_n \to \infty$. If*

$$\hat{M}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - o_{\mathbb{P}}(1),$$

then $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_{\mathbb{P}}(1)$.

Proof of Lemma 11. It suffices to show that every $\delta > 0$ admits a positive constant c_δ such that

$$\mathbb{P}\left[\hat{M}_n(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^\delta} \hat{M}_n(\boldsymbol{\theta}) > c_\delta\right] \to 1. \tag{B.1}$$

By assumption, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})| = o(1)$. Also, it follows from (Pollard, 1989, Theorem 4.2) that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\hat{M}_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta})| = O_{\mathbb{P}}(\sqrt{n^{-1}\mathbb{E}[\bar{m}_n(\mathbf{z})^2]}) = O_{\mathbb{P}}(1/\sqrt{nq_n}) = O_{\mathbb{P}}(r_n^{-3/2}) = o_{\mathbb{P}}(1).$$

As a consequence, for any $\delta > 0$,

$$\hat{M}_n(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^\delta} \hat{M}_n(\boldsymbol{\theta}) = M_0(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^\delta} M_0(\boldsymbol{\theta}) + o_{\mathbb{P}}(1),$$

so (B.1) is satisfied with $c_\delta = [M_0(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^\delta} M_0(\boldsymbol{\theta})]/2$. □

Local Behavior of M_n

If M_n is twice continuously differentiable on a neighborhood $\boldsymbol{\Theta}_n$ of $\boldsymbol{\theta}_0$, then it follows from Taylor's theorem that

$$\left|M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{V}_n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right| \leq \dot{M}_n\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \frac{1}{2}\ddot{M}_n\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2, \tag{B.2}$$

for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_n$, where

$$\mathbf{V}_n = -\frac{1}{2}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}M_n(\boldsymbol{\theta}_0), \qquad \dot{M}_n = \left\|\frac{\partial}{\partial\boldsymbol{\theta}}M_n(\boldsymbol{\theta}_0)\right\|, \qquad \ddot{M}_n = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}[M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0)]\right\|.$$

As an immediate consequence of (B.2), we have the following result about Q_n.

Lemma 12. *Suppose Condition CRA(ii) holds and suppose* $r_n \to \infty$. *Then, for any finite* $K > 0$,

$$\sup_{\|\mathbf{s}\|\leq K} |Q_n(\mathbf{s}) - \mathcal{Q}_0(\mathbf{s})| \to 0.$$

Proof of Lemma 12. Let $K > 0$ be given and suppose n is large enough that $Kr_n^{-1} \leq \delta$, where $\delta > 0$ is as in Condition CRA(ii). Using (B.2) with $\boldsymbol{\Theta}_n = \boldsymbol{\Theta}_0^{Kr_n^{-1}}$, we have, uniformly in $\mathbf{s}$ with $\|\mathbf{s}\| \leq K$,

$$\begin{aligned}|Q_n(\mathbf{s}) - \mathcal{Q}_0(\mathbf{s})| &= \left| r_n^2[M_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - M_n(\boldsymbol{\theta}_0)] + \frac{1}{2}\mathbf{s}'\mathbf{V}_0\mathbf{s} \right| \\ &\leq \frac{1}{2}\left|\mathbf{s}'(\mathbf{V}_n - \mathbf{V}_0)\mathbf{s}\right| + r_n\dot{M}_n\|\mathbf{s}\| + \frac{1}{2}\ddot{M}_n\|\mathbf{s}\|^2 = K^2 o(1),\end{aligned}$$

where the last equality uses

$$\mathbf{V}_n - \mathbf{V}_0 = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}[M_n(\boldsymbol{\theta}_0) - M_0(\boldsymbol{\theta}_0)] \to 0, \qquad r_n\dot{M}_n = r_n\left\|\frac{\partial}{\partial\boldsymbol{\theta}}M_n(\boldsymbol{\theta}_0)\right\| \to 0,$$

and

$$\begin{aligned}\ddot{M}_n &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^{Kr_n^{-1}}}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}[M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0)]\right\| \\ &\leq 2\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^{Kr_n^{-1}}}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}[M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})]\right\| + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^{Kr_n^{-1}}}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}[M_0(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta}_0)]\right\| \to 0,\end{aligned}$$

which completes the proof. □

Rate of Convergence

Lemma 13. *Suppose Conditions CRA(ii) and CRA(iii) hold and suppose* $r_n \to \infty$. *If* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_{\mathbb{P}}(1)$ *and if*

$$\hat{M}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\hat{M}_n(\boldsymbol{\theta}) - o_{\mathbb{P}}(r_n^{-2}),$$

then $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_{\mathbb{P}}(r_n^{-1})$.

Proof of Lemma 13. For any $\delta > 0$ and any positive integer C, $\mathbb{P}[r_n\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > 2^C]$

is no greater than

$$\mathbb{P}[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n) \geq \delta r_n^{-2}] + \mathbb{P}[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \delta/2]$$
$$+ \sum_{j\geq C, 2^j \leq \delta r_n} \mathbb{P}\left[\sup_{2^{j-1}<r_n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq 2^j} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) \geq -\delta r_n^{-2}\right].$$

By assumption, the probabilities on the first line go to zero for any $\delta > 0$. As a consequence, it suffices to show that the sum on the last line can be made arbitrarily small (for large n) by making $\delta > 0$ small and C large.

To do so, let $\delta > 0$ be small enough so that Conditions CRA(ii) and CRA(iii) are satisfied and

$$\liminf_{n\to\infty}[\lambda_{\min}(\mathbf{V}_n) - \ddot{M}_n^\delta] > 0,$$

where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of the argument and

$$\mathbf{V}_n = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} M_n(\boldsymbol{\theta}_0), \qquad \ddot{M}_n^\delta = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\delta} \left\| \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}[M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0)] \right\|.$$

Then, for all n large enough and for any integer pair (j, C) with $j \geq C$, we have:

$$M_n(\boldsymbol{\theta}_0) - \sup_{2^{j-1}<r_n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq 2^j} M_n(\boldsymbol{\theta}) - \delta r_n^{-2} \geq 2^{2j} c_n(\delta, C) r_n^{-2}$$

where

$$c_n(\delta, C) = \frac{1}{8}[\lambda_{\min}(\mathbf{V}_n) - \ddot{M}_n^\delta] - 2^{-C} r_n \left\| \frac{\partial}{\partial\boldsymbol{\theta}} M_n(\boldsymbol{\theta}_0) \right\| - 2^{-2C}\delta,$$

and where the inequality uses the following implication of (B.2): If $\lambda_{\min}(\mathbf{V}_n) - \ddot{M}_n^\delta \geq 0$, then

$$M_n(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n'} M_n(\boldsymbol{\theta}) \geq \frac{1}{2}[\lambda_{\min}(\mathbf{V}_n) - \ddot{M}_n^\delta] \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n'} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 - \dot{M}_n \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n'} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$$

for any subset $\boldsymbol{\Theta}_n'$ of $\boldsymbol{\Theta}_n$.

Choosing C large enough that $\liminf_{n\to\infty} c_n(\delta, C) > 2c > 0$ for some $c > 0$, we may

assume that $c_n(\delta, C) > c$ for every n, in which case

$$\begin{aligned}
&\sum_{j \geq C, 2^j \leq \delta r_n} \mathbb{P}\left[\sup_{2^{j-1} < r_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^j} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) \geq -\delta r_n^{-2}\right] \\
&\leq \sum_{j \geq C, 2^j \leq \delta r_n} \mathbb{P}\left[\sup_{2^{j-1} < r_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^j} \{\hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}) + M_n(\boldsymbol{\theta}_0)\} \geq 2^{2j} c_n(\delta, C) r_n^{-2}\right] \\
&\leq \sum_{j \geq C, 2^j \leq \delta r_n} \mathbb{P}\left[\sup_{r_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^j} \|\hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}) + M_n(\boldsymbol{\theta}_0)\| \geq 2^{2j} c r_n^{-2}\right] \\
&\leq \sum_{j \geq C, 2^j \leq \delta r_n} \frac{r_n^2 \mathbb{E}\left[\sup_{r_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^j} \|\hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}) + M_n(\boldsymbol{\theta}_0)\|\right]}{2^{2j} c},
\end{aligned}$$

where the last inequality uses the Markov inequality.

Under Condition CRA(iii), $q_n \sup_{0 \leq \delta' \leq \delta} \mathbb{E}[\bar{d}_n^{\delta'}(\mathbf{z})^2/\delta'] = O(1)$ and it follows from (Pollard, 1989, Theorem 4.2) that the sum on the last line is bounded by a constant multiple of

$$\sum_{j \geq C, 2^j \leq \delta r_n} \frac{r_n^2 \sqrt{\mathbb{E}[\bar{d}_n^{2^j/r_n}(\mathbf{z})^2]}}{2^{2j}} \leq \sqrt{q_n \sup_{0 \leq \delta' \leq \delta} \mathbb{E}[\bar{d}_n^{\delta'}(\mathbf{z})^2/\delta']} \sum_{j \geq C} \frac{1}{2^{3j/2}},$$

which can be made arbitrarily small by making C large. □

Weak Convergence

To show that $\hat{G}_n \rightsquigarrow \mathcal{G}_0$, it suffices to show finite-dimensional convergence and stochastic equicontinuity.

Lemma 14. *Suppose Conditions CRA(iii) and CRA(iv) hold, $r_n \to \infty$, and suppose $Q_n(\mathbf{s}) = o(\sqrt{n})$ for every $\mathbf{s} \in \mathbb{R}^d$. Then $\hat{G}_n$ converges to $\mathcal{G}_0$ in the sense of weak convergence of finite-dimensional projections.*

Proof of Lemma 14. Because $\hat{G}_n(\mathbf{s}) = n^{-1/2} \sum_{i=1}^n \psi_n(\mathbf{z}_i; \mathbf{s})$, where

$$\psi_n(\mathbf{z}; \mathbf{s}) = \sqrt{r_n q_n}[m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) + M_n(\boldsymbol{\theta}_0)] \mathbb{1}(\boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1} \in \boldsymbol{\Theta})$$

the result follows from the Cramér-Wold device and the Berry-Esseen inequality if it can be

shown that, for any $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$,

$$\mathbb{E}[\psi_n(\mathbf{z}; \mathbf{s})\psi_n(\mathbf{z}; \mathbf{t})] \to \mathcal{C}_0(\mathbf{s}, \mathbf{t})$$

and

$$n^{-1/2}\mathbb{E}[|\psi_n(\mathbf{z}; \mathbf{s})|^3] \to 0.$$

Let $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ be given and suppose without loss of generality that $\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}, \boldsymbol{\theta}_0 + \mathbf{t}r_n^{-1} \in \boldsymbol{\Theta}$. Then, using $Q_n(\mathbf{s}) = o(\sqrt{n})$ and the representation

$$\psi_n(\mathbf{z}; \mathbf{s}) = \sqrt{r_n q_n}[m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)] - n^{-1/2}Q_n(\mathbf{s}),$$

I have:

$$\begin{aligned}
&\mathbb{E}[\psi_n(\mathbf{z}; \mathbf{s})\psi_n(\mathbf{z}; \mathbf{t})] \\
&= r_n q_n \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \mathbf{t}r_n^{-1}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}] - n^{-1}Q_n(\mathbf{s})Q_n(\mathbf{t}) \\
&\to \mathcal{C}_0(\mathbf{s}, \mathbf{t})
\end{aligned}$$

and, using $\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^3] = o(q_n^{-2})$ (for $\delta_n = O(r_n^{-1})$),

$$\begin{aligned}
n^{-1/2}\mathbb{E}[|\psi(\mathbf{z}; \mathbf{s})|^3] &\leq 8n^{-1/2}r_n^{3/2}q_n^{3/2}\mathbb{E}[|m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)|^3] + 8n^{-2}|Q_n(\mathbf{s})|^3 \\
&= o(n^{-1/2}r_n^{3/2}q_n^{3/2}q_n^{-2}) + o(n^{-1/2}) = o(1),
\end{aligned}$$

as was to be shown. □

Lemma 15. *Suppose Conditions CRA(iii) and CRA(v) hold and suppose $r_n \to \infty$. Then, for any finite $K > 0$ and for any positive Δ_n with $\Delta_n = o(1)$,*

$$\sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \leq \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \leq K}} |\hat{G}_n(\mathbf{s}) - \hat{G}_n(\mathbf{t})| \to_{\mathbb{P}} 0.$$

Proof of Lemma 15. Let $K > 0$ be given. Proceeding as in the proof of (Kim and

Pollard, 1990, Lemma 4.6) and using the fact that $q_n\delta_n^{-1}\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^2] = O(1)$ (for $\delta_n = O(r_n^{-1})$), it suffices to show that

$$q_n r_n \sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\Delta_n \\ \|\mathbf{s}\|,\|\mathbf{t}\|\le K}} \frac{1}{n}\sum_{i=1}^n d_n(\mathbf{z}_i;\mathbf{s},\mathbf{t})^2 \to_{\mathbb{P}} 0,$$

where $d_n(\mathbf{z};\mathbf{s},\mathbf{t}) = |m_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}r_n^{-1}) - m_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{t}r_n^{-1})|/2$.

For any $C>0$ and any $\mathbf{s},\mathbf{t}\in\mathbb{R}^d$ with $\|\mathbf{s}\|,\|\mathbf{t}\|\le K$,

$$\begin{aligned} q_n\frac{1}{n}\sum_{i=1}^n d_n(\mathbf{z}_i;\mathbf{s},\mathbf{t})^2 \le\ & q_n\frac{1}{n}\sum_{i=1}^n \bar{d}_n^{Kr_n^{-1}}(\mathbf{z}_i)^2\mathbb{1}\{q_n\bar{d}_n^{Kr_n^{-1}}(\mathbf{z}_i)>C\} \\ & + C\mathbb{E}[d_n(\mathbf{z};\mathbf{s},\mathbf{t})] \\ & + C\frac{1}{n}\sum_{i=1}^n\{d_n(\mathbf{z}_i;\mathbf{s},\mathbf{t}) - \mathbb{E}[d_n(\mathbf{z};\mathbf{s},\mathbf{t})]\}, \end{aligned}$$

and therefore

$$\begin{aligned} q_n r_n\mathbb{E}\left[\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\Delta_n \\ \|\mathbf{s}\|,\|\mathbf{t}\|\le K}} \frac{1}{n}\sum_{i=1}^n d_n(\mathbf{z};\mathbf{s},\mathbf{t})^2\right] \le\ & q_n r_n\mathbb{E}\left[\bar{d}_n^{Kr_n^{-1}}(\mathbf{z})^2\mathbb{1}\{q_n\bar{d}_n^{Kr_n^{-1}}(\mathbf{z})>C\}\right] \\ & + Cr_n\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\Delta_n \\ \|\mathbf{s}\|,\|\mathbf{t}\|\le K}}\mathbb{E}[d_n(\mathbf{z};\mathbf{s},\mathbf{t})] \\ & + Cr_n\mathbb{E}\left[\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\Delta_n \\ \|\mathbf{s}\|,\|\mathbf{t}\|\le K}} |\frac{1}{n}\sum_{i=1}^n\{d_n(\mathbf{z}_i;\mathbf{s},\mathbf{t}) - \mathbb{E}[d_n(\mathbf{z};\mathbf{s},\mathbf{t})]\}|\right]. \end{aligned}$$

For large n, the first term on the majorant side can be made arbitrarily small by making C large. Also, for any fixed C, the second term tends to zero because $\Delta_n\to 0$. Finally, (Pollard, 1989, Theorem 4.2) can be used to show that for fixed C and for large n, the last term is bounded by a constant multiple of

$$r_n n^{1/2}\sqrt{\mathbb{E}[\bar{d}_n^{Kr_n^{-1}}(\mathbf{z})^2]} = \sqrt{K}r_n^{-1}\sqrt{q_n\mathbb{E}[\bar{d}_n^{Kr_n^{-1}}(\mathbf{z})^2/(Kr_n^{-1})]} = O(r_n^{-1}) = o(1),$$

which gives the result. □

B.1.2 Proof of Theorem 6

Under the assumptions of the theorem,

$$r_n(\tilde{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) = \arg\max_{\mathbf{s}\in\mathbb{R}^d}\{\tilde{Q}_n(\mathbf{s}) + \tilde{G}_n^*(\mathbf{s})\} + o_{\mathbb{P}}(1),$$

where

$$\tilde{Q}_n(\mathbf{s}) = r_n^2[\tilde{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) - \tilde{M}_n(\hat{\boldsymbol{\theta}}_n)]\mathbb{1}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta}) = -\frac{1}{2}\mathbf{s}'\tilde{\mathbf{V}}_n\mathbf{s}\mathbb{1}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta})$$

and

$$\tilde{G}_n^*(\mathbf{s}) = r_n^2[\tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) - \tilde{M}_n^*(\hat{\theta}_n) - \tilde{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) + \tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n)]\mathbb{1}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta}).$$

Moreover, $r_n(\tilde{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) = O_{\mathbb{P}}(1)$ by Lemmas 16 and 18 and $\tilde{Q}_n + \tilde{G}_n^* \rightsquigarrow_{\mathbb{P}} \mathcal{Q}_0 + \mathcal{G}_0$ by Lemmas 17, 19, and 20. The result now follows from the argmax continuous mapping theorem.

Consistency

Lemma 16. *Suppose Condition CRA(i) holds, $r_n \to \infty$, and suppose $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$, where $\mathbf{V}_0$ is positive definite. If*

$$\tilde{M}_n^*(\tilde{\theta}_n^*) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{M}_n^*(\boldsymbol{\theta}) - o_{\mathbb{P}}(1),$$

then $\tilde{\theta}_n^* - \hat{\boldsymbol{\theta}}_n = o_{\mathbb{P}}(1)$.

Proof of Lemma 16. It suffices to show that every $\delta > 0$ admits a positive constant c_δ^* such that

$$\mathbb{P}\left[\tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\hat{\boldsymbol{\Theta}}_n^\delta} \tilde{M}_n^*(\boldsymbol{\theta}) > c_\delta^*\right] \to 1, \tag{B.3}$$

where $\hat{\boldsymbol{\Theta}}_n^\delta = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| \leq \delta\}$. The process $\tilde{M}_n^*$ satisfies

$$\tilde{M}_n^*(\boldsymbol{\theta}) = \hat{M}_n^*(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\theta}_n)'\tilde{\mathbf{V}}_n(\boldsymbol{\theta} - \hat{\theta}_n), \qquad \hat{M}_n^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m_n(\mathbf{z}_{i,n}^*, \boldsymbol{\theta}).$$

It follows from (Pollard, 1989, Theorem 4.2) that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\hat{M}_n^*(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta})| = O_{\mathbb{P}}(\sqrt{n^{-1}\mathbb{E}[\bar{m}_n(\mathbf{z})^2]}) = O_{\mathbb{P}}(r_n^{-3/2}) = o_{\mathbb{P}}(1).$$

As a consequence, for any $\delta > 0$,

$$\tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\hat{\boldsymbol{\Theta}}_n^\delta} \tilde{M}_n^*(\boldsymbol{\theta}) = \frac{1}{2}\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\hat{\boldsymbol{\Theta}}_n^\delta} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\tilde{\mathbf{V}}_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + o_{\mathbb{P}}(1),$$

so (B.1) is satisfied with $c_\delta^* = \delta^2\lambda_{\min}(\mathbf{V}_0 + \mathbf{V}_0')/8$. □

Local Behavior of $\tilde{M}_n$

Because

$$\tilde{M}_n(\boldsymbol{\theta}) = \mathbb{E}^*[\tilde{M}_n^*(\boldsymbol{\theta})] = \frac{1}{n}\sum_{i=1}^n \tilde{m}_n(\mathbf{z}_i, \boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\tilde{\mathbf{V}}_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n),$$

we have the following result about $\tilde{Q}_n$.

Lemma 17. *Suppose $r_n \to \infty$, $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$, and suppose $\hat{\boldsymbol{\theta}}_n \to_{\mathbb{P}} \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$. Then, for any finite $K > 0$,*

$$\sup_{\|\mathbf{s}\|\leq K} |\tilde{Q}_n(\mathbf{s}) - \left(-\frac{1}{2}\mathbf{s}'\mathbf{V}_0\mathbf{s}\right)| \to_{\mathbb{P}} 0.$$

Proof of Lemma 17. Uniformly in $\mathbf{s}$ with $\|\mathbf{s}\| \leq K$, we have:

$$|\tilde{Q}_n(\mathbf{s}) - \left(-\frac{1}{2}\mathbf{s}'\mathbf{V}_0\mathbf{s}\right)| \leq \frac{1}{2}\left|\mathbf{s}'(\tilde{\mathbf{V}}_n - \mathbf{V}_0)\mathbf{s}\right| + \frac{1}{2}\left|\mathbf{s}'\mathbf{V}_0\mathbf{s}\right| \mathbb{1}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \notin \boldsymbol{\Theta}) \leq K^2 o_{\mathbb{P}}(1),$$

where the last equality uses $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$ and $\mathbb{P}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \notin \boldsymbol{\Theta}) \to 0$. □

Rate of Convergence

Lemma 18. *Suppose Condition CRA(iii) holds,* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_{\mathbb{P}}(r_n^{-1}) = o_{\mathbb{P}}(1)$*, and suppose* $\tilde{\mathbf{V}}_n \to_{\mathbb{P}} \mathbf{V}_0$*, where* $\mathbf{V}_0$ *is positive definite. If* $\tilde{\theta}_n^* - \hat{\theta}_n = o_{\mathbb{P}}(1)$ *and if*

$$\tilde{M}_n^*(\tilde{\theta}_n^*) \geq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \tilde{M}_n^*(\boldsymbol{\theta}) - o_{\mathbb{P}}(r_n^{-2}),$$

then $\tilde{\theta}_n^* - \hat{\boldsymbol{\theta}}_n = O_{\mathbb{P}}(r_n^{-1})$.

Proof of Lemma 18. For any $\delta > 0$ and any positive integer C, $\mathbb{P}[r_n \|\tilde{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n\| > 2^{C+1}]$ is no greater than

$$\begin{aligned}
&\mathbb{P}[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \tilde{M}_n^*(\boldsymbol{\theta}) - \tilde{M}_n^*(\tilde{\theta}_n^*) \geq \delta r_n^{-2}] + \mathbb{P}[\|\tilde{\mathbf{V}}_n - \mathbf{V}_0\| > \delta] + \mathbb{P}[\|\tilde{\theta}_n^* - \hat{\boldsymbol{\theta}}_n\| > \delta/4] \\
&\qquad + \mathbb{P}[r_n \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > 2^C] \\
&+ \sum_{j \geq C, 2^{j+1} \leq \delta r_n} \mathbb{P}\left[\sup_{2^{j-1} < r_n \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| \leq 2^j, r_n \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq 2^C, \|\tilde{\mathbf{V}}_n - \mathbf{V}_0\| \leq \delta} \tilde{M}_n^*(\boldsymbol{\theta}) - \tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n) \geq -\delta r_n^{-2} \right].
\end{aligned}$$

By assumption, the probabilities on the first line go to zero for any $\delta > 0$ and the probability on the second line can be made arbitrarily small by making C large. As a consequence, it suffices to show that the sum on the last line can be made arbitrarily small (for large n) by making $\delta > 0$ small and C large.

To do so, let $\delta > 0$ be small enough so that Condition CRA(iii) holds and

$$\inf_{\|\mathbf{V} - \mathbf{V}_0\| \leq \delta} \lambda_{\min}(\mathbf{V} + \mathbf{V}') > \lambda_{\min}(\mathbf{V}_0 + \mathbf{V}_0')/2.$$

Then, if $\|\tilde{\mathbf{V}}_n - \mathbf{V}_0\| \leq \delta$, we have, for any integer pair (j, C) with $j \geq C$:

$$\tilde{M}_n(\hat{\boldsymbol{\theta}}_n) - \sup_{2^{j-1} < r_n \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| \leq 2^j} \tilde{M}_n(\boldsymbol{\theta}) - \delta r_n^{-2} \geq 2^{2j} c^*(\delta, C) r_n^{-2}$$

where

$$c^*(\delta, C) = \frac{1}{32} \lambda_{\min}(\mathbf{V}_0 + \mathbf{V}_0') - 2^{-2C} \delta.$$

Choosing C large enough that $c^*(\delta, C) > c^* > 0$ for some $c^* > 0$ and using the fact that

$$\tilde{M}_n^*(\boldsymbol{\theta}) - \tilde{M}_n^*(\hat{\theta}_n) - \tilde{M}_n(\boldsymbol{\theta}) + \tilde{M}_n(\hat{\theta}_n) = \hat{M}_n^*(\boldsymbol{\theta}) - \hat{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \hat{M}_n(\boldsymbol{\theta}) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n),$$

we therefore have:

$$\begin{aligned}
&\sum_{j \geq C, 2^{j+1} \leq \delta r_n} \mathbb{P}\left[\sup_{2^{j-1} < r_n \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| \leq 2^j, r_n \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq 2^C, \|\tilde{\mathbf{V}}_n - \mathbf{V}_0\| \leq \delta} \tilde{M}_n^*(\boldsymbol{\theta}) - \tilde{M}_n^*(\hat{\boldsymbol{\theta}}_n) \geq -\delta r_n^{-2}\right] \\
&\leq \sum_{j \geq C, 2^{j+1} \leq \delta r_n} \mathbb{P}\left[\sup_{2^{j-1} < r_n \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| \leq 2^j, r_n \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq 2^C} \{\hat{M}_n^*(\boldsymbol{\theta}) - \hat{M}_n^*(\hat{\theta}_n) - \hat{M}_n(\boldsymbol{\theta}) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n)\} \geq 2^{2j} c^* r_n^{-2}\right] \\
&\leq \sum_{j \geq C, 2^{j+1} \leq \delta r_n} \mathbb{P}\left[\sup_{r_n \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| \leq 2^j, r_n \|\hat{\theta}_n - \boldsymbol{\theta}_0\| \leq 2^C} \|\hat{M}_n^*(\boldsymbol{\theta}) - \hat{M}_n^*(\hat{\boldsymbol{\theta}}_n) - \hat{M}_n(\boldsymbol{\theta}) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n)\| \geq 2^{2j} c^* r_n^{-2}\right] \\
&\leq \sum_{j \geq C, 2^{j+1} \leq \delta r_n} \frac{r_n^2 \mathbb{E}\left[\sup_{r_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^{j+1}, r_n \|\boldsymbol{\theta}' - \boldsymbol{\theta}_0\| \leq 2^C} \|\hat{M}_n^*(\boldsymbol{\theta}) - \hat{M}_n^*(\boldsymbol{\theta}') - \hat{M}_n(\boldsymbol{\theta}) + \hat{M}_n(\boldsymbol{\theta}')\|\right]}{2^{2j} c^*},
\end{aligned}$$

where the last inequality uses the Markov inequality.

Under Condition CRA(iii), $q_n \sup_{0 \leq \delta' \leq \delta} \mathbb{E}[\bar{d}_n^{\delta'}(\mathbf{z})^2/\delta'] = O(1)$ and (Pollard, 1989, Theorem 4.2) can be used to show that the sum on the last line is bounded by a constant multiple of

$$\sum_{j \geq C, 2^{j+1} \leq \delta r_n} \frac{r_n^2 \sqrt{\mathbb{E}[\bar{d}_n^{2^{j+1}/r_n}(\mathbf{z})^2]}}{2^{2j}} \leq \sqrt{2 q_n \sup_{0 \leq \delta' \leq \delta} \mathbb{E}[\bar{d}_n^{\delta'}(\mathbf{z})^2/\delta']} \sum_{j \geq C} \frac{1}{2^{3j/2}},$$

which can be made arbitrarily small by making C large. □

Weak Convergence

To show that $\tilde{G}_n^* \rightsquigarrow_{\mathbb{P}} \mathcal{G}_0$, it suffices to show finite-dimensional conditional weak convergence in probability and stochastic equicontinuity.

Lemma 19. *Suppose Conditions CRA(iii) and CRA(iv) hold,* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_{\mathbb{P}}(r_n^{-1}) = o_{\mathbb{P}}(1)$, *and suppose* $\sup_{\|\mathbf{s}\| \leq K} |\hat{G}_n(\mathbf{s}) + Q_n(\mathbf{s})| = o_{\mathbb{P}}(\sqrt{n})$ *for every finite* $K > 0$. *Then* $\tilde{G}_n^*$ *converges to* $\mathcal{G}_0$ *in the sense of conditional weak convergence in probability of finite-dimensional projections.*

Proof of Lemma 19. Because $\tilde{G}_n^*(\mathbf{s}) = n^{-1/2}\sum_{i=1}^n \hat{\psi}_n(\mathbf{z}_{i,n}^*;\mathbf{s})$, where

$$\hat{\psi}_n(\mathbf{z};\mathbf{s}) = \sqrt{r_n q_n}[m_n(\mathbf{z},\hat{\theta}_n + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z},\hat{\boldsymbol{\theta}}_n) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) + \hat{M}_n(\hat{\boldsymbol{\theta}}_n)]\mathbb{1}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta}),$$

the result follows from the Cramér-Wold device and the Berry-Esseen inequality if it can be shown that, for any $\mathbf{s},\mathbf{t} \in \mathbb{R}^d$,

$$\mathbb{E}^*[\hat{\psi}_n(\mathbf{z}^*;\mathbf{s})\hat{\psi}_n(\mathbf{z}^*;\mathbf{t})] = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_n(\mathbf{z}_i;\mathbf{s})\hat{\psi}_n(\mathbf{z}_i;\mathbf{t}) \to_{\mathbb{P}} \mathcal{C}_0(\mathbf{s},\mathbf{t})$$

and

$$n^{-1/2}\mathbb{E}^*[|\hat{\psi}_n(\mathbf{z}^*;\mathbf{s})|^3] = \frac{1}{n^{3/2}}\sum_{i=1}^n |\hat{\psi}_n(\mathbf{z}_i;\mathbf{s})|^3 \to_{\mathbb{P}} 0.$$

Let $\mathbf{s},\mathbf{t} \in \mathbb{R}^d$ be given and suppose without loss of generality that $\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}, \hat{\boldsymbol{\theta}}_n + \mathbf{t}r_n^{-1} \in \boldsymbol{\Theta}$. Because $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}}(1)$, we have:

$$\begin{aligned}
\hat{Q}_n(\mathbf{s}) &= r_n^2[\hat{M}_n(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n)]\mathbb{1}(\hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1} \in \boldsymbol{\Theta}) \\
&= \{\hat{G}_n[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \mathbf{s}] + Q_n[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \mathbf{s}]\} - \{\hat{G}_n[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] + Q_n[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]\} \\
&= o_{\mathbb{P}}(\sqrt{n})
\end{aligned}$$

and, using $\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^4] = o(q_n^{-3}r_n)$ (for $\delta_n = O(r_n^{-1})$) and (Pollard, 1989, Theorem 4.2),

$$\begin{aligned}
&r_n q_n \mathbb{E}^*[\{m_n(\mathbf{z}^*,\hat{\theta}_n + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}^*,\hat{\boldsymbol{\theta}}_n)\}\{m_n(\mathbf{z}^*,\hat{\boldsymbol{\theta}}_n + \mathbf{t}r_n^{-1}) - m_n(\mathbf{z}^*,\hat{\boldsymbol{\theta}}_n)\}] - \hat{\mathcal{C}}_n(\mathbf{s},\mathbf{t}) \\
&= r_n q_n \frac{1}{n}\sum_{i=1}^n \{m_n(\mathbf{z}_i,\hat{\theta}_n + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}_i,\hat{\boldsymbol{\theta}}_n)\}\{m_n(\mathbf{z}_i,\hat{\boldsymbol{\theta}}_n + \mathbf{t}r_n^{-1}) - m_n(\mathbf{z}_i,\hat{\boldsymbol{\theta}}_n)\} - \hat{\mathcal{C}}_n(\mathbf{s},\mathbf{t}) \\
&= o_{\mathbb{P}}(r_n q_n n^{-1/2}\sqrt{q_n^{-3}r_n}) = o_{\mathbb{P}}(1),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mathcal{C}}_n(\mathbf{s},\mathbf{t}) &= r_n q_n\, \mathbb{E}[\{m_n(\mathbf{z},\boldsymbol{\theta} + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z},\boldsymbol{\theta})\}\{m_n(\mathbf{z},\boldsymbol{\theta} + \mathbf{t}r_n^{-1}) - m_n(\mathbf{z},\boldsymbol{\theta})\}]\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \\
&= \mathcal{C}_0(\mathbf{s},\mathbf{t}) + o_{\mathbb{P}}(1).
\end{aligned}$$

Using these facts and the representation

$$\hat{\psi}_n(\mathbf{z};\mathbf{s}) = \sqrt{r_n q_n}[m_n(\mathbf{z}, \hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}, \hat{\boldsymbol{\theta}}_n)] - n^{-1/2}\hat{Q}_n(\mathbf{s}),$$

we have:

$$\begin{aligned}
&\mathbb{E}^*[\hat{\psi}_n(\mathbf{z}^*;\mathbf{s})\hat{\psi}_n(\mathbf{z}^*;\mathbf{t})] \\
&= r_n q_n \mathbb{E}^*[\{m_n(\mathbf{z}^*, \hat{\theta}_n + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}^*, \hat{\theta}_n)\}\{m_n(\mathbf{z}^*, \hat{\boldsymbol{\theta}}_n + \mathbf{t}r_n^{-1}) - m_n(\mathbf{z}^*, \hat{\boldsymbol{\theta}}_n)\}] - n^{-1}\hat{Q}_n(\mathbf{s})\hat{Q}_n(\mathbf{t}) \\
&= \mathcal{C}_0(\mathbf{s},\mathbf{t}) + o_{\mathbb{P}}(1).
\end{aligned}$$

and, using $\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^3] = o(q_n^{-2})$ (for $\delta_n = O(r_n^{-1})$),

$$\begin{aligned}
n^{-1/2}\mathbb{E}^*[|\hat{\psi}_n(\mathbf{z}^*;\mathbf{s})|^3] &= \frac{1}{n^{3/2}}\sum_{i=1}^n |\hat{\psi}_n(\mathbf{z}_i;\mathbf{s})|^3 \\
&\leq 8n^{-1/2}r_n^{3/2}q_n^{3/2}\frac{1}{n}\sum_{i=1}^n |m_n(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_n + \mathbf{s}r_n^{-1}) - m_n(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_n)|^3 + 8n^{-2}|\hat{Q}_n(\mathbf{s})|^3 \\
&= o_{\mathbb{P}}(n^{-1/2}r_n^{3/2}q_n^{3/2}q_n^{-2}) + o_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(1),
\end{aligned}$$

as was to be shown. □

Lemma 20. *Suppose Conditions CRA(iii) and CRA(v) hold and suppose* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_{\mathbb{P}}(r_n^{-1}) = o_{\mathbb{P}}(1)$. *Then, for any finite* $K > 0$ *and for any positive* Δ_n *with* $\Delta_n = o(1)$,

$$\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\leq\Delta_n \\ \|\mathbf{s}\|,\|\mathbf{t}\|\leq K}} |\tilde{G}_n^*(\mathbf{s}) - \tilde{G}_n^*(\mathbf{t})| \to_{\mathbb{P}} 0.$$

Proof of Lemma 20. Let $K > 0$ be given. Proceeding as in the proof of (Kim and Pollard, 1990, Lemma 4.6) and using $q_n\delta_n^{-1}\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^2] = O(1)$ (for $\delta_n = O(r_n^{-1})$) along with

the fact that $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_{\mathbb{P}}(r_n^{-1})$, it suffices to show that, for every finite $k > 0$,

$$q_n r_n \mathbb{1}\{r_n \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \le k\} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \le K}} \frac{1}{n} \sum_{i=1}^{n} \hat{d}_n(\mathbf{z}_{i,n}^*; \mathbf{s}, \mathbf{t})^2$$
$$\le q_n r_n \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \le K+k}} \frac{1}{n} \sum_{i=1}^{n} d_n(\mathbf{z}_{i,n}^*; \mathbf{s}, \mathbf{t})^2 \to_{\mathbb{P}} 0,$$

where

$$\hat{d}_n(\mathbf{z}; \mathbf{s}, \mathbf{t}) = |m_n(\mathbf{z}, \hat{\theta}_n + \mathbf{s} r_n^{-1}) - m_n(\mathbf{z}, \hat{\boldsymbol{\theta}}_n + \mathbf{t} r_n^{-1})|/2 = d_n(\mathbf{z}; r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \mathbf{s}, r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \mathbf{t}).$$

Let $k > 0$ be given. For any $C > 0$ and any $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ with $\|\mathbf{s}\|, \|\mathbf{t}\| \le K + k$,

$$q_n r_n \frac{1}{n} \sum_{i=1}^{n} d_n(\mathbf{z}_{i,n}^*; \mathbf{s}, \mathbf{t})^2 \le q_n r_n \frac{1}{n} \sum_{i=1}^{n} \bar{d}_n^{(K+k) r_n^{-1}}(\mathbf{z}_{i,n}^*)^2 \mathbb{1}\{q_n \bar{d}_n^{(K+k) r_n^{-1}}(\mathbf{z}_{i,n}^*) > C\}$$
$$+ C r_n \mathbb{E}[d_n(\mathbf{z}; \mathbf{s}, \mathbf{t})]$$
$$+ C r_n \frac{1}{n} \sum_{i=1}^{n} \{d_n(\mathbf{z}_{i,n}; \mathbf{s}, \mathbf{t}) - \mathbb{E}[d_n(\mathbf{z}; \mathbf{s}, \mathbf{t})]\}$$
$$+ C r_n \frac{1}{n} \sum_{i=1}^{n} \{d_n(\mathbf{z}_{i,n}^*; \mathbf{s}, \mathbf{t}) - \mathbb{E}^*[d_n(\mathbf{z}^*; \mathbf{s}, \mathbf{t})]\},$$

and therefore

$$q_n r_n \mathbb{E}\left[\sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \le K+k}} \frac{1}{n} \sum_{i=1}^{n} d_n(\mathbf{z}_{i,n}^*; \mathbf{s}, \mathbf{t})^2\right] \le q_n r_n \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \bar{d}_n^{(K+k) r_n^{-1}}(\mathbf{z}_{i,n}^*)^2 \mathbb{1}\{q_n \bar{d}_n^{(K+k) r_n^{-1}}(\mathbf{z}_{i,n}^*) > C\}\right]$$
$$+ C r_n \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \le K+k}} \mathbb{E}[d_n(\mathbf{z}; \mathbf{s}, \mathbf{t})]$$
$$+ C r_n \mathbb{E}\left[\sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \le K+k}} \left|\frac{1}{n} \sum_{i=1}^{n} \{d_n(\mathbf{z}_{i,n}; \mathbf{s}, \mathbf{t}) - \mathbb{E}[d_n(\mathbf{z}; \mathbf{s}, \mathbf{t})]\}\right|\right]$$
$$+ C r_n \mathbb{E}\left[\sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \Delta_n \\ \|\mathbf{s}\|, \|\mathbf{t}\| \le K+k}} \left|\frac{1}{n} \sum_{i=1}^{n} \{d_n(\mathbf{z}_{i,n}^*; \mathbf{s}, \mathbf{t}) - \mathbb{E}^*[d_n(\mathbf{z}^*; \mathbf{s}, \mathbf{t})]\}\right|\right].$$

For large n, the first term on the majorant side can be made arbitrarily small by making C large. Also, for any fixed C, the second term tends to zero because $\Delta_n \to 0$. Finally, (Pollard, 1989, Theorem 4.2) can be used to show that for fixed C and for large n, each of the last two terms is bounded by a constant multiple of

$$r_n n^{1/2}\sqrt{\mathbb{E}[\bar{d}_n^{(K+k)r_n^{-1}}(\mathbf{z})^2]} = \sqrt{K+k}\, r_n^{-1}\sqrt{q_n \mathbb{E}[\bar{d}_n^{(K+k)r_n^{-1}}(\mathbf{z})^2/(\{K+k\}r_n^{-1})]} = O(r_n^{-1}) = o(1),$$

which gives the result. □

B.1.3 Proof of Lemmas 2 and 3

Consistency

Without loss of generality, assume $\mathbb{P}[r_n|\hat{\theta} - \boldsymbol{\theta}_0| \leq C] = 1$ for some fixed constant C. Write $\bar{V}_{n,kl}(\boldsymbol{\theta}) = \mathbb{E}[\tilde{V}_{n,kl}(\boldsymbol{\theta})]$. Consider the following decomposition

$$\tilde{V}_{n,kl}^{\texttt{ND}}(\hat{\theta}_n) = \bar{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0) + \left[\tilde{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0) - \bar{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0)\right] + R_{1,n} + R_{2,n},$$

where

$$\begin{aligned} R_{1,n} &= \bar{V}_{n,kl}^{\texttt{ND}}(\hat{\theta}_n) - \bar{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0) \\ R_{2,n} &= \tilde{V}_{n,kl}^{\texttt{ND}}(\hat{\theta}_n) - \tilde{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0) - \bar{V}_{n,kl}^{\texttt{ND}}(\hat{\theta}_n) + \bar{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0). \end{aligned}$$

By definition,

$$\begin{aligned} \bar{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}) = -\frac{1}{4\epsilon_n^2}\big[&M_n(\boldsymbol{\theta} + \epsilon_n\mathbf{e}_k + \epsilon_n\mathbf{e}_l) - M_n(\boldsymbol{\theta} - \epsilon_n\mathbf{e}_k + \epsilon_n\mathbf{e}_l) \\ &- M_n(\boldsymbol{\theta} + \epsilon_n\mathbf{e}_k - \epsilon_n\mathbf{e}_l) + M_n(\boldsymbol{\theta} - \epsilon_n\mathbf{e}_k - \epsilon_n\mathbf{e}_l)\big]. \end{aligned}$$

Using (B.2),

$$\left|M_n(\boldsymbol{\theta}_0 + \epsilon_n\mathbf{e}_k + \epsilon_n\mathbf{e}_l) - M_n(\boldsymbol{\theta}_0) + \frac{\epsilon_n^2}{2}(\mathbf{e}_k + \mathbf{e}_l)'\mathbf{V}_n(\mathbf{e}_k + \mathbf{e}_l)\right| \leq o(r_n^{-1}\epsilon_n + \epsilon_n^2)$$

where I use $\dot{M}_n = o(r_n^{-1})$ and $\ddot{M}_n = o(1)$. Then,

$$\begin{aligned}\bar{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0) &= \frac{1}{4}\left[(\mathbf{e}_k+\mathbf{e}_l)'\mathbf{V}_n(\mathbf{e}_k+\mathbf{e}_l) - (\mathbf{e}_k-\mathbf{e}_l)'\mathbf{V}_n(\mathbf{e}_k-\mathbf{e}_l)\right] + o(r_n^{-1}\epsilon_n^{-1}+1)\\ &= \mathbf{e}_k'\mathbf{V}_n\mathbf{e}_l + o(r_n^{-1}\epsilon_n^{-1}+1).\end{aligned}$$

By condition CRA(ii), we have $\mathbf{V}_n \to \mathbf{V}_0$. Therefore, $\bar{\mathbf{V}}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0)$ converges to $\mathbf{e}_k'\mathbf{V}_0\mathbf{e}_l = V_{0,kl}$.

It remains to show $\tilde{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0) - \bar{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0) = o_{\mathbb{P}}(1)$, $R_{1,n} = o_{\mathbb{P}}(1)$, and $R_{2,n} = o_{\mathbb{P}}(1)$. First, note that $\tilde{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0) - \bar{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0)$ is mean zero and its variance is bounded by a constant multiple of $\epsilon_n^{-4}n^{-1}\mathbb{E}[\bar{d}_n^{2\epsilon_n}(\mathbf{z})^2] = O(n^{-1}q_n^{-1}\epsilon_n^{-3})$. Then, $\tilde{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0) - \bar{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0) = O_{\mathbb{P}}((\epsilon_n r_n)^{-3/2}) = o_{\mathbb{P}}(1)$.

Next, using (B.2),

$$\begin{aligned}&M_n(\hat{\theta}_n + \epsilon_n[\mathbf{e}_k+\mathbf{e}_l]) - M_n(\boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k+\mathbf{e}_l])\\ &= -\frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k+\mathbf{e}_l])'\mathbf{V}_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k+\mathbf{e}_l]) + \frac{\epsilon_n^2}{2}(\mathbf{e}_k+\mathbf{e}_l)'\mathbf{V}_n(\mathbf{e}_k+\mathbf{e}_l) + o(r_n^{-1}\epsilon_n+\epsilon_n^2)\\ &\leq C(\|\hat{\theta}-\boldsymbol{\theta}_0\|^2 + \epsilon_n\|\hat{\theta}-\boldsymbol{\theta}_0\|) + o(r_n^{-1}\epsilon_n+\epsilon_n^2),\end{aligned}$$

and therefore $R_{1,n} = O_{\mathbb{P}}(r_n^{-2}\epsilon_n^{-2} + \epsilon_n^{-1}r_n^{-1}) + o(r_n^{-1}\epsilon_n^{-1}+1) = o_{\mathbb{P}}(1)$.

Finally, by adding and subtracting $\hat{M}_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0)$ to the left-hand side of $R_{2,n}$, it suffices to analyze

$$\epsilon_n^{-2}\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq Cr_n^{-1}+2\epsilon_n}\left|\hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) - \{M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0)\}\right|.$$

Letting $t_n = Cr_n^{-1} + 2\epsilon_n$, and applying the maximal inequality in (Pollard, 1989, Theorem 4.2),

$$\begin{aligned}&\epsilon_n^{-2}\mathbb{E}\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq t_n}\left|\hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) - \{M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0)\}\right|\\ &\leq C\epsilon_n^{-2}n^{-1/2}(\mathbb{E}|\bar{d}_n^{t_n}(\mathbf{z})|^2)^{1/2}\\ &\leq C\epsilon_n^{-2}n^{-1/2}q_n^{-1/2}(r_n^{-1}+\epsilon_n)^{1/2} = O((\epsilon_n r_n)^{-3/2}).\end{aligned}$$

This implies that $R_{2,n} = o_{\mathbb{P}}(1)$.

Approximate Mean Squared Error

Define $\check{\mathbf{V}}_n^{\mathtt{ND}} = \tilde{\mathbf{V}}_n^{\mathtt{ND}}(\boldsymbol{\theta}_0)$. Since $\tilde{V}_{n,kl}^{\mathtt{ND}} = \check{V}_{n,kl}^{\mathtt{ND}} + R_{1,n} + R_{2,n}$, the first goal is to show that $R_{1,n} = o_{\mathbb{P}}(\epsilon_n^2) + O_{\mathbb{P}}(r_n^{-1})$ and $R_{2,n} = o_{\mathbb{P}}(r_n^{-3/2}\epsilon_n^{-3/2})$. For $\boldsymbol{\theta}, \vartheta \in \boldsymbol{\Theta}_0^{\delta}$, using a Taylor approximation,

$$
\begin{aligned}
&M_n(\boldsymbol{\theta}) - M_n(\vartheta) \\
&= (\boldsymbol{\theta} - \vartheta)' \frac{\partial}{\partial \boldsymbol{\theta}} M_n(\vartheta) - \frac{1}{2}(\boldsymbol{\theta} - \vartheta)' \mathbf{V}_n (\boldsymbol{\theta} - \boldsymbol{\vartheta}) \\
&\quad + \frac{1}{6} \sum_{j_1,j_2,j_3} \frac{\partial^3 M_n(\boldsymbol{\vartheta})}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} (\theta_{j_1} - \vartheta_{j_1})(\theta_{j_2} - \vartheta_{j_2})(\theta_{j_3} - \vartheta_{j_3}) \\
&\quad + \frac{1}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\vartheta + \xi)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3} \partial \theta_{j_4}} (\theta_{j_1} - \vartheta_{j_1})(\theta_{j_2} - \vartheta_{j_2})(\theta_{j_3} - \vartheta_{j_3})(\theta_{j_4} - \vartheta_{j_4})
\end{aligned} \tag{B.4}
$$

where ξ is a mean-value between $\mathbf{0}$ and $\boldsymbol{\theta} - \boldsymbol{\vartheta}$. Then,

$$
\begin{aligned}
&M_n(\hat{\boldsymbol{\theta}}_n + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) - M_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) + M_n(\boldsymbol{\theta}_0) \\
&= (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial}{\partial \boldsymbol{\theta}} M_n(\boldsymbol{\theta}_0) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \mathbf{V}_n (\hat{\theta}_n - \boldsymbol{\theta}_0) - \epsilon_n (\mathbf{e}_k + \mathbf{e}_l)' \mathbf{V}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{1}{6} \sum_{j_1,j_2,j_3} \frac{\partial^3 M_n(\boldsymbol{\theta}_0)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} \mathbf{e}_{j_1}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \mathbf{e}_{j_2}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \mathbf{e}_{j_3}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n}{2} \sum_{j_1,j_2,j_3} \frac{\partial^3 M_n(\boldsymbol{\theta}_0)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} \mathbf{e}_{j_1}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_2}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \mathbf{e}_{j_3}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n^2}{2} \sum_{j_1,j_2,j_3} \frac{\partial^3 M_n(\boldsymbol{\theta}_0)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} \mathbf{e}_{j_1}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_2}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_3}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n^3}{6} \sum_{j_1,j_2,j_3} \frac{\partial^3 M_n(\boldsymbol{\theta}_0)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} \mathbf{e}_{j_1}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_2}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_3}'(\mathbf{e}_k + \mathbf{e}_l) \\
&\quad + \frac{1}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0 + \xi_1)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3} \partial \theta_{j_4}} \mathbf{e}_{j_1}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \mathbf{e}_{j_2}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \\
&\qquad\qquad \times \mathbf{e}_{j_3}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \mathbf{e}_{j_4}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \\
&\quad - \frac{\epsilon_n^4}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0 + \xi_2)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3} \partial \theta_{j_4}} \mathbf{e}_{j_1}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_2}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_3}'(\mathbf{e}_k + \mathbf{e}_l) \mathbf{e}_{j_4}'(\mathbf{e}_k + \mathbf{e}_l)
\end{aligned}
$$

where ξ_1 and ξ_2 are also mid-points as ξ. Using uniform convergence of fourth derivatives, the last two lines equal

$$
\begin{aligned}
&= \frac{1}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l])\mathbf{e}'_{j_2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \\
&\qquad \times \mathbf{e}'_{j_3}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l])\mathbf{e}'_{j_4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \\
&- \frac{\epsilon_n^4}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_3}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_4}(\mathbf{e}_k + \mathbf{e}_l) + o_{\mathbb{P}}(\epsilon_n^4)
\end{aligned}
$$

where we use $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_{\mathbb{P}}(\epsilon_n)$. By expanding the terms,

$$
\begin{aligned}
&\frac{1}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l])\mathbf{e}'_{j_2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \\
&\qquad \times \mathbf{e}'_{j_3}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l])\mathbf{e}'_{j_4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) \\
&= \frac{1}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\hat{\theta}_n - \boldsymbol{\theta}_0)\mathbf{e}'_{j_2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\mathbf{e}'_{j_3}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\mathbf{e}'_{j_4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n}{6} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\mathbf{e}'_{j_3}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\mathbf{e}'_{j_4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n^2}{2} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_3}(\hat{\theta}_n - \boldsymbol{\theta}_0)\mathbf{e}'_{j_4}(\hat{\theta}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n^3}{4} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_3}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + \frac{\epsilon_n^4}{24} \sum_{j_1,j_2,j_3,j_4} \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}} \mathbf{e}'_{j_1}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_3}(\mathbf{e}_k + \mathbf{e}_l)\mathbf{e}'_{j_4}(\mathbf{e}_k + \mathbf{e}_l).
\end{aligned}
$$

Then, because

$$
\begin{aligned}
&- 4\epsilon_n^2 \left[\bar{V}^{\mathtt{ND}}_{n,kl}(\hat{\theta}_n) - \bar{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0)\right] \\
&= M_n(\hat{\boldsymbol{\theta}}_n + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) - M_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) + M_n(\boldsymbol{\theta}_0) \\
&\quad - M_n(\hat{\boldsymbol{\theta}}_n + \epsilon_n[\mathbf{e}_k - \mathbf{e}_l]) + M_n(\boldsymbol{\theta}_0) + M_n(\boldsymbol{\theta}_0 + \epsilon_n[\mathbf{e}_k - \mathbf{e}_l]) - M_n(\boldsymbol{\theta}_0) \\
&\quad - M_n(\hat{\boldsymbol{\theta}}_n - \epsilon_n[\mathbf{e}_k - \mathbf{e}_l]) + M_n(\boldsymbol{\theta}_0) + M_n(\boldsymbol{\theta}_0 - \epsilon_n[\mathbf{e}_k - \mathbf{e}_l]) - M_n(\boldsymbol{\theta}_0) \\
&\quad + M_n(\hat{\boldsymbol{\theta}}_n - \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) - M_n(\boldsymbol{\theta}_0) - M_n(\boldsymbol{\theta}_0 - \epsilon_n[\mathbf{e}_k + \mathbf{e}_l]) + M_n(\boldsymbol{\theta}_0),
\end{aligned}
$$

we have

$$
\begin{aligned}
&-4\epsilon_n^2\left[\bar{V}^{\mathtt{ND}}_{n,kl}(\hat{\theta}_n)-\bar{V}^{\mathtt{ND}}_{n,kl}(\boldsymbol{\theta}_0)\right]\\
&=\epsilon_n^2\sum_{j_1,j_2,j_3}\frac{\partial^3 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}\mathbf{e}'_{j_1}(\mathbf{e}_k+\mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k+\mathbf{e}_l)\mathbf{e}'_{j_3}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\\
&\quad-\epsilon_n^2\sum_{j_1,j_2,j_3}\frac{\partial^3 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}\mathbf{e}'_{j_1}(\mathbf{e}_k-\mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k-\mathbf{e}_l)\mathbf{e}'_{j_3}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\\
&\quad+\epsilon_n^2\sum_{j_1,j_2,j_3,j_4}\frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}}\mathbf{e}'_{j_1}(\mathbf{e}_k+\mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k+\mathbf{e}_l)\mathbf{e}'_{j_3}(\hat{\theta}_n-\boldsymbol{\theta}_0)\mathbf{e}'_{j_4}(\hat{\theta}_n-\boldsymbol{\theta}_0)\\
&\quad-\epsilon_n^2\sum_{j_1,j_2,j_3,j_4}\frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}\partial\theta_{j_4}}\mathbf{e}'_{j_1}(\mathbf{e}_k-\mathbf{e}_l)\mathbf{e}'_{j_2}(\mathbf{e}_k-\mathbf{e}_l)\mathbf{e}'_{j_3}(\hat{\theta}_n-\boldsymbol{\theta}_0)\mathbf{e}'_{j_4}(\hat{\theta}_n-\boldsymbol{\theta}_0)+o_{\mathbb{P}}(\epsilon_n^4)\\
&=O_{\mathbb{P}}(\epsilon_n^2 r_n^{-1})+o_{\mathbb{P}}(\epsilon_n^4)
\end{aligned}
$$

where we use $r_n^{-1}=o(\epsilon_n)$. Note that the term $O_{\mathbb{P}}(\epsilon_n^2 r_n^{-1})$ is independent of ϵ_n after divided by ϵ_n^2. Therefore, $R_{1,n}=o_{\mathbb{P}}(\epsilon_n^2)+O_{\mathbb{P}}(r_n^{-1})$.

For term $R_{2,n}$, consider

$$
\epsilon_n^{-2}\left[\hat{M}_n(\hat{\boldsymbol{\theta}}_n+\epsilon_n\mathbf{e})-\hat{M}_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{e})-M_n(\hat{\boldsymbol{\theta}}_n+\epsilon_n\mathbf{e})+M_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{e})\right]
$$

for some d-dimensional vector $\mathbf{e}$. It suffices to show

$$
\epsilon_n^{-2}\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\Delta_n\\ \|\mathbf{s}\|,\|\mathbf{t}\|\le C}}\left|\hat{M}_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{s})-\hat{M}_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{t})-M_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{s})+M_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{t})\right|=o_{\mathbb{P}}(r_n^{-3/2}\epsilon_n^{-3/2})
$$

for $\Delta_n=o(1)$. Note that this is almost identical to the statement in Lemma 15, except that r_n in Lemma 15 is replaced by ϵ_n^{-1}. Doing the same calculation,

$$
\begin{aligned}
&\epsilon_n^{-2}\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\Delta_n\\ \|\mathbf{s}\|,\|\mathbf{t}\|\le C}}\left|\hat{M}_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{s})-\hat{M}_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{t})-M_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{s})+M_n(\boldsymbol{\theta}_0+\epsilon_n\mathbf{t})\right|\\
&\le\epsilon_n^{-2}r_n^{-2}(r_n\epsilon_n)^{1/2}o_{\mathbb{P}}(1)=o_{\mathbb{P}}(r_n^{-3/2}\epsilon_n^{-3/2}).
\end{aligned}
$$

Now, we calculate the constants for leading bias and variance. Using (B.4),

$$\begin{aligned}\mathbb{E}[\check{V}_{n,kl}^{\mathtt{ND}}] &= \bar{V}_{n,kl}^{\mathtt{ND}}(\boldsymbol{\theta}_0) \\ &= \frac{1}{2}\mathbf{e}_k'\mathbf{V}_n\mathbf{e}_l - \frac{\epsilon_n^2}{6}\left(\frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_k^3\partial\theta_l} + \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_k\partial\theta_l^3}\right).\end{aligned}$$

Thus, the constant for the leading bias is

$$\mathsf{B}_{kl} = -\frac{1}{6}\lim_{n\to\infty}\left(\frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_k^3\partial\theta_l} + \frac{\partial^4 M_n(\boldsymbol{\theta}_0)}{\partial\theta_k\partial\theta_l^3}\right) = -\frac{1}{6}\left(\frac{\partial^4 M_0(\boldsymbol{\theta}_0)}{\partial\theta_k^3\partial\theta_l} + \frac{\partial^4 M_0(\boldsymbol{\theta}_0)}{\partial\theta_k\partial\theta_l^3}\right).$$

Next, consider the variance of $\check{V}_{n,kl}^{\mathtt{ND}}$. To save space, write

$$m_{ni}(\epsilon_1,\epsilon_2) = m_n(\mathbf{z}_i, \boldsymbol{\theta}_0 + \epsilon_1\mathbf{e}_k + \epsilon_2\mathbf{e}_l)$$

Recall that

$$\check{V}_{n,kl}^{\mathtt{ND}} = -\frac{1}{4\epsilon_n^2 n}\sum_{i=1}^{n}\left[m_{ni}(\epsilon_n,\epsilon_n) - m_{ni}(\epsilon_n,-\epsilon_n) - m_{ni}(-\epsilon_n,\epsilon_n) + m_{ni}(-\epsilon_n,-\epsilon_n)\right]$$

and

$$\begin{aligned}\mathbb{V}[\tilde{V}_{n,kl}^{\mathtt{ND}}(\boldsymbol{\theta}_0)] &= \frac{1}{16\epsilon_n^4 n}\mathbb{V}\left(m_{ni}(\epsilon_n,\epsilon_n) - m_{ni}(\epsilon_n,-\epsilon_n) - m_{ni}(-\epsilon_n,\epsilon_n) + m_{ni}(-\epsilon_n,-\epsilon_n)\right) \\ &= \frac{1}{16\epsilon_n^4 n}\mathbb{E}\left[\left\{m_{ni}(\epsilon_n,\epsilon_n) - m_{ni}(\epsilon_n,-\epsilon_n) - m_{ni}(-\epsilon_n,\epsilon_n) + m_{ni}(-\epsilon_n,-\epsilon_n)\right\}^2\right] \\ &\quad - \frac{1}{n}[\bar{\mathbf{V}}_{n,kl}^{\mathtt{ND}}(\boldsymbol{\theta}_0)]^2\end{aligned}$$

From the above calculation, the second term in the last equation is $O(n^{-1})$. By condition CRA(iv)

$$\epsilon_n^{-1}q_n\mathbb{E}[\{m_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}\epsilon_n) - m_n(\mathbf{z},\boldsymbol{\theta}_0)\}\{m_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{t}\epsilon_n) - m_n(\mathbf{z},\boldsymbol{\theta}_0)\}] \to \mathcal{C}_0(\mathbf{s},\mathbf{t}).$$

We have

$$\lim_{n\to\infty} \epsilon_n^{-1} q_n \mathbb{E}\left[\{m_{ni}(\epsilon_n,\epsilon_n) - m_{ni}(\epsilon_n,-\epsilon_n) - m_{ni}(-\epsilon_n,\epsilon_n) + m_{ni}(-\epsilon_n,-\epsilon_n)\}^2\right]$$
$$= 2\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,\mathbf{e}_k+\mathbf{e}_l) + 2\mathcal{C}_0(\mathbf{e}_k-\mathbf{e}_l,\mathbf{e}_k-\mathbf{e}_l) - 4\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,\mathbf{e}_k-\mathbf{e}_l) - 4\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,-\mathbf{e}_k+\mathbf{e}_l)$$

where we use $\mathcal{C}_0(\mathbf{s},-\mathbf{s}) = 0$ and $\mathcal{C}_0(\mathbf{s},\mathbf{t}) = \mathcal{C}_0(-\mathbf{s},-\mathbf{t})$. Then,

$$\mathbb{V}[\tilde{V}_{n,kl}^{\texttt{ND}}(\boldsymbol{\theta}_0)] = \frac{1}{r_n^3\epsilon_n^3}[\mathsf{V}_{kl} + o(1)] + O(n^{-1})$$

where $\mathsf{V}_{kl} = \frac{1}{8}[\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,\mathbf{e}_k+\mathbf{e}_l)+\mathcal{C}_0(\mathbf{e}_k-\mathbf{e}_l,\mathbf{e}_k-\mathbf{e}_l)-2\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,\mathbf{e}_k-\mathbf{e}_l)-2\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,-\mathbf{e}_k+\mathbf{e}_l)]$.

B.1.4 Proof of Corollary MS

It suffices to verify that Condition MS implies Condition CRA_0.

Condition CRA_0(i). The manageability assumption can be verified using the same argument as in Kim and Pollard (1990). Note that the function $|m^{\texttt{MS}}(\mathbf{z},\boldsymbol{\theta})|$ is bounded by unity in this example, and thus finite second moment condition holds. Arguing as in Manski (1985), we can show that $\boldsymbol{\theta}_0$ uniquely maximizes $M_0(\boldsymbol{\theta})$ over the parameter set. Well-separatedness follows from unique maximum, compactness of the parameter space, and continuity of the function $M_0(\boldsymbol{\theta})$.

Condition CRA_0(ii). Conditions MS(iii) and MS(iv) imply this condition.

Condition CRA_0(iii). Uniform manageability can be verified using the same argument as in Kim and Pollard (1990). Note $d_0^\delta(\mathbf{z}) = \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta} |\mathbb{1}(x_1+\mathbf{x}_2'\boldsymbol{\theta}\ge 0) - \mathbb{1}(x_1+\mathbf{x}_2'\boldsymbol{\theta}_0\ge 0)|$. The condition $\sup_{0<\delta'\le\delta}\mathbb{E}[\bar{d}_0^{\delta'}(\mathbf{z})]/\delta' < \infty$ is verified in Abrevaya and Huang (2005).

Condition CRA_0(iv). Since $d_0^\delta(\mathbf{z})^2 = d_0^\delta(\mathbf{z})$, $\mathbb{E}[d_0^{\delta_n}(\mathbf{z})^3] = O(\delta_n)$ and $\mathbb{E}[d_0^{\delta_n}(\mathbf{z})^4] = O(\delta_n)$, which implies the first condition. As noted in Abrevaya and Huang (2005),

$$\mathcal{C}_0^{\texttt{MS}}(\mathbf{s},\mathbf{t}) = \mathbb{E}[\min\{|\mathbf{x}_2'\mathbf{s}|,|\mathbf{x}_2'\mathbf{t}|\}\mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = \text{sgn}(\mathbf{x}_2'\mathbf{t})\} f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)].$$

From this representation, it follows that $\mathcal{C}_0^{\texttt{MS}}(\mathbf{s},\mathbf{s}) + \mathcal{C}_0^{\texttt{MS}}(\mathbf{t},\mathbf{t}) - 2\mathcal{C}_0^{\texttt{MS}}(\mathbf{s},\mathbf{t}) > 0$ for $\mathbf{s}\neq\mathbf{t}$. Using

$2xy = x^2 + y^2 - (x-y)^2$, it suffices to show for $\delta_n = O(r_n^{-1})$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta_n}} \left| r_n \mathbb{E}|m(\mathbf{z}_1, \boldsymbol{\theta} + \mathbf{s}/r_n) - m(\mathbf{z}_1, \boldsymbol{\theta} + \mathbf{t}/r_n)|^2 - \mathcal{L}_0(\mathbf{s} - \mathbf{t}) \right| = o(1)$$

for $\mathcal{L}_0$ defined in Corollary MS. We have

$$\begin{aligned}
&\mathbb{E}|m^{\texttt{MS}}(\mathbf{z}, \boldsymbol{\theta}_1) - m^{\texttt{MS}}(\mathbf{z}, \boldsymbol{\theta}_2)|^2 \\
&= \mathbb{E}\mathbb{1}(x_1 + \mathbf{x}_2'\boldsymbol{\theta}_1 \geq 0 > x_1 + \mathbf{x}_2'\boldsymbol{\theta}_2) + \mathbb{E}\mathbb{1}\left(x_1 + \mathbf{x}_2'\boldsymbol{\theta}_2 \geq 0 > x_1 + \mathbf{x}_2'\boldsymbol{\theta}_1\right) \\
&= \mathbb{E}\left[\{F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_2|\mathbf{x}_2) - F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_1|\mathbf{x}_2)\}\mathbb{1}\left(\mathbf{x}_2'\boldsymbol{\theta}_2 < \mathbf{x}_2'\boldsymbol{\theta}_1\right)\right] \\
&\quad + \mathbb{E}\left[\{F_{x_1|\mathbf{x}_2}(-\boldsymbol{\theta}_1'\mathbf{x}_2|\mathbf{x}_2) - F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_2|\mathbf{x}_2)\}\mathbb{1}\left(\mathbf{x}_2'\boldsymbol{\theta}_1 < \mathbf{x}_2'\boldsymbol{\theta}_2\right)\right] \\
&= \mathbb{E}[f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\mathbb{1}\left(\mathbf{x}_2'\boldsymbol{\theta}_2 < \mathbf{x}_2'\boldsymbol{\theta}_1\right)] \\
&\quad + \mathbb{E}[f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\mathbb{1}\left(\mathbf{x}_2'\boldsymbol{\theta}_1 < \mathbf{x}_2'\boldsymbol{\theta}_2\right)] + R
\end{aligned}$$

where $\boldsymbol{\theta}_1 = \boldsymbol{\theta} + \mathbf{s}/r_n$, $\boldsymbol{\theta}_2 = \boldsymbol{\theta} + \mathbf{t}/r_n$, and

$$\begin{aligned}
R = &\mathbb{E}\left[\{F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_2|\mathbf{x}_2) - F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) - f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(-\boldsymbol{\theta}_2 + \boldsymbol{\theta}_0)\}\mathbb{1}(\mathbf{x}_2'\boldsymbol{\theta}_2 < \mathbf{x}_2'\boldsymbol{\theta}_1)\right] \\
&- \mathbb{E}\left[\{F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_1|\mathbf{x}_2) - F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) - f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(-\boldsymbol{\theta}_1 + \boldsymbol{\theta}_0)\}\mathbb{1}(\mathbf{x}_2'\boldsymbol{\theta}_2 < \mathbf{x}_2'\boldsymbol{\theta}_1)\right] \\
&+ \mathbb{E}\left[\{F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_1|\mathbf{x}_2) - F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) - f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(-\boldsymbol{\theta}_1 + \boldsymbol{\theta}_0)\}\mathbb{1}(\mathbf{x}_2'\boldsymbol{\theta}_1 < \mathbf{x}_2'\boldsymbol{\theta}_2)\right] \\
&- \mathbb{E}\left[\{F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_2|\mathbf{x}_2) - F_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) - f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(-\boldsymbol{\theta}_2 + \boldsymbol{\theta}_0)\}\mathbb{1}(\mathbf{x}_2'\boldsymbol{\theta}_1 < \mathbf{x}_2'\boldsymbol{\theta}_2)\right].
\end{aligned}$$

By continuous differentiability of $f_{x_1|\mathbf{x}_2}$ with bounded derivative,

$$|R| \leq 4 \sup_{x_1, \mathbf{x}_2} \left| f_{x_1|\mathbf{x}_2}^{(1)}(x_1|\mathbf{x}_2) \right| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta_n}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0 + (\mathbf{s} + \mathbf{t})r_n^{-1}\|^2 \mathbb{E}\|\mathbf{x}_2\|^2 = o(r_n^{-1}).$$

Then,

$$\begin{aligned}
&r_n\mathbb{E}|m^{\texttt{MS}}(\mathbf{z}, \boldsymbol{\theta} + \mathbf{s}/r_n) - m^{\texttt{MS}}(\mathbf{z}, \boldsymbol{\theta} + \mathbf{t}/r_n)|^2 \\
&= \mathbb{E}[f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(-\mathbf{t} + \mathbf{s})\mathbb{1}(\mathbf{x}_2'\mathbf{t} < \mathbf{x}_2'\mathbf{s})] + \mathbb{E}[f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2'(-\mathbf{s} + \mathbf{t})\mathbb{1}(\mathbf{x}_2'\mathbf{s} < \mathbf{x}_2'\mathbf{t})] + o(1) \\
&= \mathbb{E}[f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)|\mathbf{x}_2'\mathbf{s} - \mathbf{x}_2'\mathbf{t}|] + o(1).
\end{aligned}$$

Since the remainder term is uniformly small in $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta_n}$, the verification of Condition CRA_0(iv) is complete.

Condition CRA$_0$(v). The first part easily follows from $\bar{d}_0^{\delta}(\mathbf{z}) \leq 1$, while the second part follows from the above calculation for $\mathcal{L}_0$.

B.1.5 Proof of Lemma MS

Consistency

For consistency, we use Conditions MS and K(i)-(iii), $\mathbb{E}\|\mathbf{x}_2\|^6 < \infty$, $h_n \to 0$, and $nh_n^3 \to \infty$. Recall that $\tilde{\mathbf{V}}_n^{\mathtt{MS}} = \tilde{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\boldsymbol{\theta}}_n)$, set $\bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}) = \mathbb{E}[\tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta})]$, and note that

$$\tilde{\mathbf{V}}_n^{\mathtt{MS}} = \tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) + [\bar{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\boldsymbol{\theta}}_n^{\mathtt{MS}}) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)] + [\tilde{\mathbf{V}}_n^{\mathtt{MS}} - \tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\boldsymbol{\theta}}_n^{\mathtt{MS}}) + \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)].$$

Consistency of $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ then follows from the following four results:

$$\bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) = \mathbf{V}_0^{\mathtt{MS}} + o(1), \tag{B.5}$$

$$\tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) = o_{\mathbb{P}}(1), \tag{B.6}$$

$$\bar{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\boldsymbol{\theta}}_n^{\mathtt{MS}}) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) = o_{\mathbb{P}}(1), \tag{B.7}$$

$$\tilde{\mathbf{V}}_n^{\mathtt{MS}} - \tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\theta}_n^{\mathtt{MS}}) + \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) = O_{\mathbb{P}}(n^{-5/6}h_n^{-5/2}). \tag{B.8}$$

Equation (B.5) holds because

$$\begin{aligned}
&- \mathbb{E}\left[\tilde{V}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)\right] \\
&= h_n^{-2}\mathbb{E}\left[\mathbf{x}_2\mathbf{x}_2'(2y-1)\dot{K}\left(\mathbf{x}'\boldsymbol{\beta}_0/h_n\right)\right] \\
&= h_n^{-1}\mathbb{E}\left[\mathbf{x}_2\mathbf{x}_2'\int\left[1-2F_{u|x_1,\mathbf{x}_2}(-vh_n|vh_n-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\right]\dot{K}(v)f_{x_1|\mathbf{x}_2}(vh_n-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dv\right] \\
&= -2\mathbb{E}\left[\mathbf{x}_2\mathbf{x}_2'\int F_{u|x_1,\mathbf{x}_2}^{(1)}(-\xi|\xi-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)v\dot{K}(v)f_{x_1|\mathbf{x}_2}(vh_n-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dv\right]
\end{aligned}$$

where ξ is a mean value between vh_n and 0. By boundedness of $f_{x_1|\mathbf{x}_2}(v|\mathbf{x}_2)$ and $F_{u|x_1,\mathbf{x}_2}^{(1)}(-v|v-$

$\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)$ and continuity with respect to v, the dominated convergence theorem implies

$$\begin{aligned}
&\int \mathbf{x}_2\mathbf{x}_2' \int 2F^{(1)}_{u|x_1,\mathbf{x}_2}(-\xi|\xi - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) v\dot{K}(v) f_{x_1|\mathbf{x}_2}(vh_n - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\to 2\mathbb{E}\left[\mathbf{x}_2\mathbf{x}_2' F^{(1)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right] \int v\dot{K}(v)du = \mathbf{V}_0^{\mathtt{MS}}
\end{aligned}$$

where we use $\int v\dot{K}(v)dv = -1$.

Equation (B.6) follows directly because the variance of $\mathbb{V}[\tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)] = O(n^{-1}h_n^{-3})$, as shown below.

For Equation (B.7),

$$\begin{aligned}
&\bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}) \\
&= h_n^{-2} \int \mathbf{x}_2\mathbf{x}_2' \int \{1 - 2F_{u|x_1,\mathbf{x}_2}(-\mathbf{x}'\boldsymbol{\beta}_0|x_1, \mathbf{x}_2)\} \dot{K}(\{x_1 + \mathbf{x}_2'\boldsymbol{\theta}\}/h_n) f_{x_1|\mathbf{x}_2}(x_1|\mathbf{x}_2) dx_1 dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&= h_n^{-1} \int \mathbf{x}_2\mathbf{x}_2' \int \{1 - 2F_{u|x_1,\mathbf{x}_2}(-vh_n|vh_n - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)\} \dot{K}(v + \mathbf{x}_2'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)h_n^{-1}) \\
&\qquad\qquad \times f_{x_1|\mathbf{x}_2}(vh_n - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&= -2 \int \mathbf{x}_2\mathbf{x}_2' \int v F^{(1)}_{u|x_1,\mathbf{x}_2}(-\zeta|\zeta - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) \dot{K}(v + \mathbf{x}_2'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)h_n^{-1}) \\
&\qquad\qquad \times f_{x_1|\mathbf{x}_2}(vh_n - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2).
\end{aligned}$$

where ζ lies between vh_n and 0, and does not depend on $\boldsymbol{\theta}$. Then,

$$\left\|\bar{\mathbf{V}}_{n,kl}^{\mathtt{MS}}(\boldsymbol{\theta}) - \bar{\mathbf{V}}_{n,kl}^{\mathtt{MS}}(\boldsymbol{\theta}_0)\right\| \leq C\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| h_n^{-1} \left(\int B(v)dv\right) \mathbb{E}\|\mathbf{x}_2\|^3$$

where we use boundedness of $F^{(1)}_{u|\mathbf{x}}$, $f_{x_1|\mathbf{x}_2}$, and Assumption K(ii). Then, $\|\hat{\boldsymbol{\theta}}_n^{\mathtt{MS}} - \boldsymbol{\theta}_0\| = O_{\mathbb{P}}(n^{-1/3}) = o_{\mathbb{P}}(h_n)$, which establishes the result.

Finally, we verify Equation (B.8). Using (Pollard, 1989, Theorem 4.2), it suffices to show

$$\mathbb{E}\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq Cn^{-1/3}} \left|\dot{K}\left(\frac{x_1 + \mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right) - \dot{K}\left(\frac{x_1 + \mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\right|^2 \|\mathbf{x}_2\mathbf{x}_2'\|^2\right] = O(n^{-2/3}h_n^{-1}).$$

By the assumed Lipschitz condition on $\dot{K}$,

$$\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq Cn^{-1/3}} \left| \dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right) - \dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\right| \leq CB(\mathbf{x}'\boldsymbol{\beta}_0/h_n)\|\mathbf{x}_2\|h_n^{-1}n^{-1/3},$$

and therefore

$$\begin{aligned}
&\mathbb{E}\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq Cn^{-1/3}} \left| \dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right) - \dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\right|^2 \|\mathbf{x}_2\mathbf{x}_2'\|^2\right] \\
&\leq Cn^{-2/3}h_n^{-2}\mathbb{E}\left[B(\mathbf{x}'\boldsymbol{\beta}_0/h_n)^2\|\mathbf{x}_2\|^6\right] \\
&\leq Cn^{-2/3}h_n^{-1}\sup_{x_1,\mathbf{x}_2} f_{x_1|\mathbf{x}_2}(x_1|\mathbf{x}_2)\left(\int B(v)^2dv\right)\mathbb{E}\|\mathbf{x}_2\|^6,
\end{aligned}$$

which verifies the final result.

Approximate Mean Squared Error

First we show that $\|\bar{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\theta}_n^{\mathtt{MS}}) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)\| = o_{\mathbb{P}}(h_n^2) + O_{\mathbb{P}}(n^{-1/3})$ for Equation (B.7) above. Changing variables and using a Taylor expansion, we have

$$
\begin{aligned}
\bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}) &= -h_n^{-2}\mathbb{E}\left[\mathbf{x}_2\mathbf{x}_2' \int \{1 - 2F_{u|x_1,\mathbf{x}_2}(-\mathbf{x}'\boldsymbol{\beta}_0|x_1,\mathbf{x}_2)\}\dot{K}\left(\frac{x_1 + \mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right) f_{x_1|\mathbf{x}_2}(x_1|\mathbf{x}_2)dx_1\right] \\
&= 2h_n^{-1}\int \mathbf{x}_2\mathbf{x}_2' \int (vh_n - \mathbf{x}_2'(\boldsymbol{\theta}-\boldsymbol{\theta}_0))F^{(1)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\dot{K}(v) \\
&\qquad\qquad \times f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + 2h_n^{-1}\int \mathbf{x}_2\mathbf{x}_2' \int |vh_n - \mathbf{x}_2'(\boldsymbol{\theta}-\boldsymbol{\theta}_0)|^2 F^{(1)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\dot{K}(v) \\
&\qquad\qquad \times f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + h_n^{-1}\int \mathbf{x}_2\mathbf{x}_2' \int \{vh_n - \mathbf{x}_2'(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}^3 F^{(1)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\dot{K}(v) \\
&\qquad\qquad \times f^{(2)}_{x_1|\mathbf{x}_2}(\xi_{1,n} - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + h_n^{-1}\int \mathbf{x}_2\mathbf{x}_2' \int |vh_n - \mathbf{x}_2'(\boldsymbol{\theta}-\boldsymbol{\theta}_0)|^2 F^{(2)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\dot{K}(v) \\
&\qquad\qquad \times f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + h_n^{-1}\int \mathbf{x}_2\mathbf{x}_2' \int \{vh_n - \mathbf{x}_2'(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}^3 F^{(2)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\dot{K}(v) \\
&\qquad\qquad \times f^{(1)}_{x_1|\mathbf{x}_2}(\xi_{2,n} - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + \frac{h_n^{-1}}{3}\int \mathbf{x}_2\mathbf{x}_2' \int \{vh_n - \mathbf{x}_2'(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}^3 F^{(3)}_{u|x_1,\mathbf{x}_2}(-\xi_{3,n}|\xi_{3,n} - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\dot{K}(v) \\
&\qquad\qquad \times f_{x_1|\mathbf{x}_2}(vh_n - \mathbf{x}_2'\boldsymbol{\theta}|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2)
\end{aligned}
$$

where ξ_1, ξ_2 and ξ_3 lie between $vh_n - \mathbf{x}_2'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and 0.

Using $\int v\dot{K}(v)dv = 1$, $\int \dot{K}(v)dv = 0$, and $\int v^2\dot{K}(v)dv = 0$,

$$\begin{aligned}
\bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}) = 2 &\int F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&- 4(\boldsymbol{\theta}-\boldsymbol{\theta}_0)' \int \mathbf{x}_2 F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&+ h_n^2 \int\int v^3\dot{K}(v) F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(\xi_1-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&- 2(\boldsymbol{\theta}-\boldsymbol{\theta}_0)' \int \mathbf{x}_2 F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&+ h_n^2 \int \mathbf{x}_2\mathbf{x}_2' \int v^3\dot{K}(v) F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(\xi_2-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&+ \frac{h_n^2}{3} \int \mathbf{x}_2\mathbf{x}_2' \int v^3\dot{K}(v) F^{(3)}_{u|x_1,\mathbf{x}_2}(-\xi_3|\xi_3-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(vh_n-\mathbf{x}_2'\boldsymbol{\theta}|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&+ O(h_n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + h_n^{-1}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^3)
\end{aligned}$$

where we use boundedness and continuity of derivatives of $F_{u|x_1,\mathbf{x}_2}$ and $f_{x_1|\mathbf{x}_2}$ and $\mathbb{E}\|\mathbf{x}_2\|^6 < \infty$.

For $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$ as $n \to \infty$, we have

$$\begin{aligned}
&\lim_{n\to\infty} \int \mathbf{x}_2\mathbf{x}_2' \int v^3\dot{K}(v) F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(\xi_1-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&= \int v^3\dot{K}(v)dv \int F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dF_{\mathbf{x}_2}(\mathbf{x}_2),
\end{aligned}$$

$$\begin{aligned}
&\lim_{n\to\infty} \int \mathbf{x}_2\mathbf{x}_2' \int v^3\dot{K}(v) F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(\xi_2-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&= \int v^3\dot{K}(v)dv \int F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dF_{\mathbf{x}_2}(\mathbf{x}_2),
\end{aligned}$$

$$\begin{aligned}
&\lim_{n\to\infty} \int \mathbf{x}_2\mathbf{x}_2' \int v^3\dot{K}(v) F^{(3)}_{u|x_1,\mathbf{x}_2}(-\xi_3|\xi_3-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(vh_n-\mathbf{x}_2'\boldsymbol{\theta}|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&= \int v^3\dot{K}(v)dv \int F^{(3)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' dF_{\mathbf{x}_2}(\mathbf{x}_2).
\end{aligned}$$

Because $\mathbb{P}[|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| \le Cn^{-1/3}] \to 1$, we have $\|\bar{\mathbf{V}}_n^{\mathtt{MS}}(\hat{\boldsymbol{\theta}}_n) - \bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)\| = o_{\mathbb{P}}(h_n^2) + O_{\mathbb{P}}(n^{-1/3})$. Note that the term $O_{\mathbb{P}}(n^{-1/3})$ does not depend on h_n.

Next, we derive the leading bias and variance of $\tilde{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0)$. Specifically, we show that

$$\bar{V}_{n,kl}^{\mathtt{MS}}(\boldsymbol{\theta}_0) = V_{0,kl}^{\mathtt{MS}} + h_n^2 \left(\mathsf{B}_{kl} + o(1)\right)$$

and

$$\mathbb{V}[\tilde{V}_{n,kl}^{\mathtt{MS}}(\boldsymbol{\theta}_0)] = \frac{1}{nh_n^3}\left(\mathsf{V}_{kl} + o(1)\right)$$

where B_{kl} and V_{kl}, $1 \leq l, k \leq d$, are defined in the statement of Lemma MS.

For the bias, a Taylor expansion gives

$$\begin{aligned}
&\bar{\mathbf{V}}_n^{\mathtt{MS}}(\boldsymbol{\theta}_0) - \mathbf{V}_0^{\mathtt{MS}} \\
&= 2h_n \int F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dF_{\mathbf{x}_2}(\mathbf{x}_2)\int v^2\dot{K}(v)dv \\
&\quad + h_n^2 \int \mathbf{x}_2\mathbf{x}_2' F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\int v^3\dot{K}(v) f^{(2)}_{x_1|\mathbf{x}_2}(\xi_1 - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + h_n \int F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\mathbf{x}_2\mathbf{x}_2' f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dF_{\mathbf{x}_2}(\mathbf{x}_2)\int v^2\dot{K}(v)dv \\
&\quad + h_n^2 \int \mathbf{x}_2\mathbf{x}_2' F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\int f^{(1)}_{x_1|\mathbf{x}_2}(\xi_2 - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)v^3\dot{K}(v)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\quad + \frac{h_n^2}{3}\int \mathbf{x}_2\mathbf{x}_2'\int F^{(3)}_{u|x_1,\mathbf{x}_2}(-\xi_3|\xi_3 - \mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(vh_n|\mathbf{x}_2)v^3\dot{K}(v)dvdF_{\mathbf{x}_2}(\mathbf{x}_2)
\end{aligned}$$

where ξ_1, ξ_2 and ξ_3 denote mean-values between 0 and vh_n. Boundedness of $f^{(s)}_{x_1|\mathbf{x}_2}(v|\mathbf{x}_2)$ and $F^{(s)}_{u|x_1,\mathbf{x}_2}(-v|v+x_1,\mathbf{x}_2)$, $k = 1,2,3$, and continuity of these functions with respect to v imply that

$$\begin{aligned}
&\int \mathbf{x}_2\mathbf{x}_2' F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\int v^3\dot{K}(v) f^{(2)}_{x_1|\mathbf{x}_2}(\xi_1 - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\to \int \mathbf{x}_2\mathbf{x}_2' F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dF_{\mathbf{x}_2}(\mathbf{x}_2)\int v^3\dot{K}(v)dv,
\end{aligned}$$

$$\begin{aligned}
&\int \mathbf{x}_2\mathbf{x}_2' F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)\int f^{(1)}_{x_1|\mathbf{x}_2}(\xi_2 - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)v^3\dot{K}(v)dvdF_{\mathbf{x}_2}(\mathbf{x}_2) \\
&\to \int \mathbf{x}_2\mathbf{x}_2' F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)dF_{\mathbf{x}_2}(\mathbf{x}_2)\int v^3\dot{K}(v)dv,
\end{aligned}$$

$$\int \mathbf{x}_2\mathbf{x}_2' \int F^{(3)}_{u|x_1,\mathbf{x}_2}(-\xi_3|\xi_3 - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) f_{x_1|\mathbf{x}_2}(vh_n|\mathbf{x}_2) v^3 \dot{K}(v) dv dF_{\mathbf{x}_2}(\mathbf{x}_2)$$
$$\to \int \mathbf{x}_2\mathbf{x}_2' F^{(3)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dF_{\mathbf{x}_2}(\mathbf{x}_2) \int v^3 \dot{K}(v) dv.$$

Therefore, because $\int_{-\infty}^{\infty} v^2 \dot{K}(v) dv = 0$,

$$\bar{V}^{\mathtt{MS}}_{n,kl}(\boldsymbol{\theta}_0) = V^{\mathtt{MS}}_{0,kl} + h_n^2 \left(\mathsf{B}_{kl} + o(1)\right)$$

with B_{kl} the (k,l)-th element of

$$\begin{aligned}\mathsf{B} = \int \mathbf{x}_2\mathbf{x}_2' \Big[& F^{(1)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) \\ & + F^{(2)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) \\ & + \frac{1}{3} F^{(3)}_{u|x_1,\mathbf{x}_2}(0| - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) \Big] dF_{\mathbf{x}_2}(\mathbf{x}_2) \int v^3 \dot{K}(v) dv.\end{aligned}$$

Finally, the approximate variance is obtained by noting that

$$\mathbb{E}[\tilde{V}^{\mathtt{MS}}_{n,kl}(\boldsymbol{\theta}_0)^2] = \frac{1}{nh_n^4} \mathbb{E}\left[\dot{K}\left(\frac{\mathbf{x}'\boldsymbol{\beta}_0}{h_n}\right) \mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l\right]^2$$

and

$$\begin{aligned}\frac{1}{h_n}\mathbb{E}\left[\dot{K}\left(\frac{\mathbf{x}'\boldsymbol{\beta}_0}{h_n}\right) \mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l\right]^2 &= \int (\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)^2 \int \dot{K}(v)^2 f_{x_1|\mathbf{x}_2}(vh_n - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dv dF_{\mathbf{x}_2}(\mathbf{x}_2) \\ &\to \int (\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)^2 f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dF_{\mathbf{x}_2}(\mathbf{x}_2) \int \dot{K}(v)^2 dv\end{aligned}$$

where the last line follows by dominated convergence theorem. Since $\mathbb{E}[\tilde{V}^{\mathtt{MS}}_{n,kl}(\boldsymbol{\theta}_0)] = O(1)$,

$$\mathbb{V}[\tilde{V}^{\mathtt{MS}}_{n,kl}(\boldsymbol{\theta}_0)] = \frac{1}{nh_n^3}\left[\int (\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)^2 f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) dF_{\mathbf{x}_2}(\mathbf{x}_2) \int \dot{K}(v)^2 dv + o(1)\right].$$

B.1.6 Proof of Corollary CMS

We verify that Condition CMS implies Condition CRA. Set $q_n = b_n^d$.

Condition CRA(i). For uniform manageability, as shown for maximum score estimator,

the class $\{y\mathbb{1}\{x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0\} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ has a polynomial bound on the covering number uniform in discrete probability measures and $L_n(\cdot)$ does not depend on $\boldsymbol{\theta}$. See, e.g., van der Vaart and Wellner (1996) for details. We can take $\bar{m}_n(\mathbf{z}) = |L_n(\mathbf{w})|$ and $\mathbb{E}[L_n^2(\mathbf{w})] \leq Cb_n^{-d}$, which implies $q_n\mathbb{E}[\bar{m}_n(\mathbf{z})^2] = O(1)$.

For uniform convergence of $M_n(\boldsymbol{\theta})$ to $M_0(\boldsymbol{\theta})$,

$$\begin{aligned} M_n(\boldsymbol{\theta}) &= \int \mathbb{E}[\psi(x_1, \mathbf{x}_2, \mathbf{w})\{f_{x_1}(x_1|\mathbf{x}_2, \mathbf{w})\}^{-1}\mathbb{1}\{x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0\}|\mathbf{w} = \mathbf{v}b_n]L(\mathbf{v})f_{\mathbf{w}}(\mathbf{v}b_n)d\mathbf{v} \\ &= \int\int\int_{-\mathbf{x}_2'\boldsymbol{\theta}}^{\infty} \psi(x, \mathbf{x}_2, \mathbf{v}b_n)dxdF_{\mathbf{x}_2}(\mathbf{x}_2|\mathbf{v}b_n)f_{\mathbf{w}}(\mathbf{v}b_n)L(\mathbf{v})d\mathbf{v} \\ &= \int \varphi_1(\mathbf{v}b_n; \boldsymbol{\theta})L(\mathbf{v})d\mathbf{v}. \end{aligned}$$

Using $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\varphi_1(\mathbf{w}; \boldsymbol{\theta}) - \varphi_1(\mathbf{0}; \boldsymbol{\theta})| \leq C\|\mathbf{w}\|^{\epsilon}$ for $\mathbf{w}$ close to $\mathbf{0}$,

$$\begin{aligned} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})| &\leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int |\varphi_1(\mathbf{v}b_n; \boldsymbol{\theta}) - \varphi_1(\mathbf{0}; \boldsymbol{\theta})\|L(\mathbf{v})|d\mathbf{v} \\ &\leq Cb_n^{\epsilon} \int \|\mathbf{v}\|^{\epsilon}|L(\mathbf{v})|d\mathbf{v} = o(1). \end{aligned}$$

Well-separatedness follows from uniqueness of maximizer, compactness of the parameter space, and continuity of $M_0(\boldsymbol{\theta}) = \mathbb{E}[y\mathbb{1}\{x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0\}|\mathbf{w} = \mathbf{0}]$. Uniqueness of maximizer follows from (Honoré and Kyriazidou, 2000, Lemmas 6 and 7), compactness of $\boldsymbol{\Theta}$ is assumed, and continuity follows from differentiability of $M_0(\cdot)$ shown below.

Condition CRA(ii). By definition,

$$M_0(\boldsymbol{\theta}) = \mathbb{E}[\psi(x_1, \mathbf{x}_2, \mathbf{w})\{f_{x_1}(x_1|\mathbf{x}_2, \mathbf{w})\}^{-1}\mathbb{1}\{x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0\}|\mathbf{w} = \mathbf{0}]$$

and

$$M_n(\boldsymbol{\theta}) = \int \mathbb{E}[\psi(x_1, \mathbf{x}_2, \mathbf{w})\{f_{x_1}(x_1|\mathbf{x}_2, \mathbf{w})\}^{-1}\mathbb{1}\{x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0\}|\mathbf{w} = \mathbf{v}b_n]L(\mathbf{v})f_{\mathbf{w}}(\mathbf{v}b_n)d\mathbf{v}.$$

Since $\mathbb{E}[\psi(x_1, \mathbf{x}_2, \mathbf{w})\{f_{x_1}(x_1|\mathbf{x}_2, \mathbf{w})\}^{-1}\mathbb{1}\{x_1 + \mathbf{x}_2'\boldsymbol{\theta} \geq 0\}|\mathbf{w}] = \int\int_{-\mathbf{x}_2'\boldsymbol{\theta}}^{\infty} \psi(x, \mathbf{x}_2, \mathbf{w})dxdF_{\mathbf{x}_2}(\mathbf{x}_2|\mathbf{w})$,

we have

$$\frac{\partial}{\partial \boldsymbol{\theta}} M_n(\boldsymbol{\theta}) = \int\int \mathbf{x}_2 \psi(-\mathbf{x}_2'\boldsymbol{\theta}, \mathbf{x}_2, \mathbf{v}b_n) dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}b_n) f_{\mathbf{w}}(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v},$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} M_0(\boldsymbol{\theta}) = \int \mathbf{x}_2 \psi(-\mathbf{x}_2'\boldsymbol{\theta}, \mathbf{x}_2, \mathbf{0}) dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{0}) f_{\mathbf{w}}(\mathbf{0}),$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} M_n(\boldsymbol{\theta}) = -\int\int \mathbf{x}_2 \mathbf{x}_2' \psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}, \mathbf{x}_2, \mathbf{v}b_n) dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}b_n) f_{\mathbf{w}}(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v},$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} M_0(\boldsymbol{\theta}) = -\int \mathbf{x}_2 \mathbf{x}_2' \psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}, \mathbf{x}_2, \mathbf{0}) dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{0}) f_{\mathbf{w}}(\mathbf{0}),$$

where $\psi^{(s)}(x_1, \mathbf{x}_2, \mathbf{w}) = \frac{d^s}{dx_1^s}\psi(x_1, \mathbf{x}_2, \mathbf{w})$, $s \in \mathbb{Z}$, and we use $|\psi(x_1, \mathbf{x}_2, \mathbf{w}) + \psi^{(1)}(x_1, \mathbf{x}_2, \mathbf{w})| \leq C$ for all $(x_1, \mathbf{x}_2, \mathbf{w}) \in \mathcal{S}$, $\int \|\mathbf{x}_2\|^2 dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}) \leq C$ for all $\mathbf{v}$ with $\|\mathbf{v}\| \leq \eta$, $f_{\mathbf{w}}(\cdot)$ is bounded around $\mathbf{0}$, and $\int |L(\mathbf{v})| d\mathbf{v} < \infty$, to invoke dominated convergence theorem.

Because the function $\mathbf{V}(\mathbf{w}; \boldsymbol{\theta}) = \int \mathbf{x}_2 \mathbf{x}_2' \psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}, \mathbf{x}_2, \mathbf{w}) dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w})$ satisfies $\|\mathbf{V}(\mathbf{w}; \boldsymbol{\theta}) - \varphi(\mathbf{0}; \boldsymbol{\theta})\| \leq C\|\mathbf{w}\|^{\epsilon}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta}$,

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta}} \left| \frac{\partial [M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right| &\leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta}} \int \|\mathbf{V}(\mathbf{v}b_n; \boldsymbol{\theta}) - \mathbf{V}(\mathbf{0}; \boldsymbol{\theta})\| |L(\mathbf{v})| d\mathbf{v} \\ &\leq C b_n^{\epsilon} \int \|\mathbf{v}\|^{\epsilon} |L(\mathbf{v})| d\mathbf{v} = o(1).\end{aligned}$$

For $\partial M_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$,

$$\frac{\partial}{\partial \boldsymbol{\theta}} M_n(\boldsymbol{\theta}_0) = \int \varphi_2(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v} = \varphi_2(\mathbf{0}) + b_n^P \sum_{|\boldsymbol{\ell}|=P} \int \mathbf{v}^{\boldsymbol{\ell}} \partial^{\boldsymbol{\ell}} \varphi_2(\tilde{\mathbf{v}}) L(\mathbf{v}) d\mathbf{v}$$

where we use multi-index notation, $\int \mathbf{v}^{\boldsymbol{\ell}} L(\mathbf{v}) d\mathbf{v} = 0$ for $0 < |\boldsymbol{\ell}| < P$, and $\tilde{\mathbf{v}}$ lies between $\mathbf{v}b_n$ and $\mathbf{0}$. Honoré and Kyriazidou (2000) show that, for any $(d_0, d_3) \in \{0, 1\}^2$, $\mathbb{E}[A(d_0, d_3)|\mathbf{X}, \alpha, \mathbf{w} = \mathbf{0}] > \mathbb{E}[(B(d_0, d_3)|\mathbf{X}, \alpha, \mathbf{w} = \mathbf{0}]$ if and only if $x_1 + \mathbf{x}_2'\boldsymbol{\theta}_0 > 0$, which im-

plies $\mathbb{E}[A(d_0,d_3)|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{X},\alpha,\mathbf{w}=\mathbf{0}]=\mathbb{E}[B(d_0,d_3)|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{X},\alpha,\mathbf{w}=\mathbf{0}]$. Then,

$$
\begin{aligned}
&\psi(-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{0})\\
&=\mathbb{E}[A(Y_0,Y_3)-B(Y_0,Y_3)|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w}=\mathbf{0}]f_{x_1}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2,\mathbf{0})\\
&=\sum_{d_0,d_3\in\{0,1\}}\Pr(Y_0=d_0,Y_3=d_3|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w}=\mathbf{0})f_{x_1}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2,\mathbf{0})\\
&\qquad\times\{\mathbb{E}[A(d_0,d_3)|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w}=\mathbf{0}]-\mathbb{E}[B(d_0,d_3)|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w}=\mathbf{0}]\}\\
&=\sum_{d_0,d_3\in\{0,1\}}\Pr(Y_0=d_0,Y_3=d_3|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w}=\mathbf{0})f_{x_1}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2,\mathbf{0})\\
&\qquad\times\mathbb{E}[\mathbb{E}[A(d_0,d_3)-B(d_0,d_3)|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{X},\mathbf{w}=\mathbf{0},\alpha]|x_1=-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w}=\mathbf{0}]\\
&=0
\end{aligned}
$$

which implies $\varphi_2(\mathbf{0})=\mathbf{0}$. Finally, $b_n^P=o((nb_n^d)^{-1/3})$ implies that $\sqrt[3]{nq_n}\|\partial M_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\|=o(1)$ holds.

Positive definiteness of $\mathbf{V}_0^{\mathtt{CMS}}$ is directly imposed in Condition CMS.

Condition CRA(iii). Uniform manageability follows from the class $\{y\mathbb{1}\{x_1+\mathbf{x}_2'\boldsymbol{\theta}\geq 0\}:\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ having a polynomial bound on the uniform covering number, since $L_n(\cdot)$ does not depend on $\boldsymbol{\theta}$. See, e.g., van der Vaart and Wellner (1996) for details. For this example,

$$
\bar{d}_n^{\delta}(\mathbf{z})=\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta}\mathbb{1}\{\mathbf{x}_2'\boldsymbol{\theta}_0<-x_1\leq\mathbf{x}_2'\boldsymbol{\theta}\}|L_n(\mathbf{w})|+\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta}\mathbb{1}\{\mathbf{x}_2'\boldsymbol{\theta}<-x_1\leq\mathbf{x}_2'\boldsymbol{\theta}_0\}|L_n(\mathbf{w})|.
$$

Conditional on $(\mathbf{x}_2,\mathbf{w})$, x_1 has bounded Lebesgue density and for b_n sufficiently small,

$$
\mathbb{E}[\bar{d}_n^{\delta}(\mathbf{z})^2]\leq C\mathbb{E}[\|\mathbf{x}_2\|L_n(\mathbf{w})^2]\delta\leq C\delta b_n^{-d}
$$

by change of variables. The constant C does not depend on δ, and therefore $q_n\sup_{0<\delta'\leq\delta}\mathbb{E}[\bar{d}_n^{\delta'}(\mathbf{z})^2/\delta']=O(1)$ holds.

Condition CRA(iv). Note that $\bar{d}_n^{\delta}(\mathbf{z})^3=d_n^{\delta}(\mathbf{z})|L_n(\mathbf{w})|^2$ and $\bar{d}_n^{\delta}(\mathbf{z})^3=d_n^{\delta}(\mathbf{z})|L_n(\mathbf{w})|^3$. Doing a similar calculation as above, we have $q_n^2\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^3]=O(\delta_n)=o(1)$ and $q_n^3r_n^{-1}\mathbb{E}[\bar{d}_n^{\delta_n}(\mathbf{z})^4]=O(r_n^{-1}\delta_n)=o(1)$.

As in the Maximum Score example, $\mathcal{C}_0^{\texttt{CMS}}(\mathbf{s},\mathbf{s}) + \mathcal{C}_0^{\texttt{CMS}}(\mathbf{t},\mathbf{t}) - 2\mathcal{C}_0^{\texttt{CMS}}(\mathbf{s},\mathbf{t}) > 0$ follows from its representation

$$\mathcal{C}_0^{\texttt{CMS}}(\mathbf{s},\mathbf{t}) = \mathbb{E}[\min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} f_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2, \mathbf{0}) \mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = \text{sgn}(\mathbf{x}_2'\mathbf{t})\} | \mathbf{w} = \mathbf{0}] f_{\mathbf{w}}(\mathbf{0}).$$

Now we verify the last condition.

$$\begin{aligned}
&\{m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{s}) - m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta})\}\{m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{t}) - m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta})\} \\
&= \{\mathbb{1}(\mathbf{x}_2'(\boldsymbol{\theta} + \delta_n \mathbf{s}) \geq -x_1) - \mathbb{1}(\mathbf{x}_2'\boldsymbol{\theta} \geq -x_1)\}\{\mathbb{1}(\mathbf{x}_2'(\boldsymbol{\theta} + \delta_n \mathbf{t}) \geq -x_1) - \mathbb{1}(\mathbf{x}_2'\boldsymbol{\theta} \geq -x_1)\}|L_n(\mathbf{w})|^2 \\
&= \mathbb{1}(\delta_n \min\{\mathbf{x}_2'\mathbf{s}, \mathbf{x}_2'\mathbf{t}\} \geq -x_1 - \mathbf{x}_2'\boldsymbol{\theta} > 0)\mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = 1 = \text{sgn}(\mathbf{x}_2'\mathbf{t})\}|L_n(\mathbf{w})|^2 \\
&\quad + \mathbb{1}(\delta_n \max\{\mathbf{x}_2'\mathbf{s}, \mathbf{x}_2'\mathbf{t}\} < -x_1 - \mathbf{x}_2'\boldsymbol{\theta} \leq 0)\mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = -1 = \text{sgn}(\mathbf{x}_2'\mathbf{t})\}|L_n(\mathbf{w})|^2 \\
&= \mathbb{1}(\delta_n \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} \geq -x_1 - \mathbf{x}_2'\boldsymbol{\theta} > 0)\mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = \text{sgn}(\mathbf{x}_2'\mathbf{t})\}|L_n(\mathbf{w})|^2
\end{aligned}$$

and

$$\begin{aligned}
&q_n\delta_n^{-1}\mathbb{E}[\{m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{s}) - m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta})\}\{m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{t}) - m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta})\}] \\
&= q_n^{-1}\delta_n^{-1} \int\int \{F_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}|\mathbf{x}_2, \mathbf{w}) - F_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta} - \delta_n \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\}|\mathbf{x}_2, \mathbf{w})\} \\
&\qquad\qquad \times \mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = \text{sgn}(\mathbf{x}_2'\mathbf{t})\} dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w}) L(\mathbf{w}/b_n) d\mathbf{w} \\
&= \delta_n^{-1} \int\int \{F_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}|\mathbf{x}_2, \mathbf{v}b_n) - F_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta} - \delta_n \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\}|\mathbf{x}_2, \mathbf{v}b_n)\} \\
&\qquad\qquad \times \mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = \text{sgn}(\mathbf{x}_2'\mathbf{t})\} dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}b_n) f_{\mathbf{w}}(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v} \\
&= \int\int \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} f_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta} + \tilde{\delta}_n|\mathbf{x}_2, \mathbf{v}b_n) \\
&\qquad\qquad \times \mathbb{1}\{\text{sgn}(\mathbf{x}_2'\mathbf{s}) = \text{sgn}(\mathbf{x}_2'\mathbf{t})\} dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}b_n) f_{\mathbf{w}}(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v}
\end{aligned}$$

where $\tilde{\delta}_n$ lies between 0 and $-\delta_n \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\}$. Then,

$$\begin{aligned}
&\Big| q_n \delta_n^{-1} \mathbb{E}[\{m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{s}) - m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta})\}\{m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{t}) m_n^{\texttt{CMS}}(\mathbf{z}, \boldsymbol{\theta})\}] \\
&\quad - \int\int \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} f_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}_0 | \mathbf{x}_2, \mathbf{v} b_n) \\
&\qquad\qquad \times \mathbb{1}\{\operatorname{sgn}(\mathbf{x}_2'\mathbf{s}) = \operatorname{sgn}(\mathbf{x}_2'\mathbf{t})\} dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}b_n) f_{\mathbf{w}}(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v} \Big| \\
&\leq C\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^{\epsilon} \sup_{\mathbf{w} \in \mathcal{W}} \mathbb{E}[\min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\}^2 \|\mathbf{x}_2\| B_f(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) | \mathbf{w}] \int |L(\mathbf{v})| d\mathbf{v} = O(\delta_n^{\epsilon})
\end{aligned}$$

as $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta_n}$. By dominated convergence theorem,

$$\begin{aligned}
&\lim_{n\to\infty} \int\int \min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} f_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}_0 | \mathbf{x}_2, \mathbf{v} b_n) \\
&\qquad\qquad \times \mathbb{1}\{\operatorname{sgn}(\mathbf{x}_2'\mathbf{s}) = \operatorname{sgn}(\mathbf{x}_2'\mathbf{t})\} dF_{\mathbf{x}_2|\mathbf{w}}(\mathbf{x}_2|\mathbf{v}b_n) f_{\mathbf{w}}(\mathbf{v}b_n) L(\mathbf{v}) d\mathbf{v} \Big| \\
&= \mathbb{E}[\min\{|\mathbf{x}_2'\mathbf{s}|, |\mathbf{x}_2'\mathbf{t}|\} f_{x_1|\mathbf{x}_2,\mathbf{w}}(-\mathbf{x}_2'\boldsymbol{\theta}_0 | \mathbf{x}_2, \mathbf{0}) \mathbb{1}\{\operatorname{sgn}(\mathbf{x}_2'\mathbf{s}) = \operatorname{sgn}(\mathbf{x}_2'\mathbf{t})\} | \mathbf{w} = \mathbf{0}] f_{\mathbf{w}}(\mathbf{0}) \\
&= \mathcal{C}_0^{\texttt{CMS}}(\mathbf{s}, \mathbf{t})
\end{aligned}$$

and this convergence occurs uniformly on $\boldsymbol{\Theta}_0^{\delta_n}$ because it is independent of $\boldsymbol{\theta}$. Finally,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta_n}} |q_n \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{s}) - m_n(\mathbf{z}, \boldsymbol{\theta})\}\{m_n(\mathbf{z}, \boldsymbol{\theta} + \delta_n \mathbf{t}) - m_n(\mathbf{z}, \boldsymbol{\theta})\}/\delta_n] - \mathcal{C}_0^{\texttt{CMS}}(\mathbf{s}, \mathbf{t})| = o(1),$$

as desired.

Condition CRA(v). The first condition follows from $q_n \bar{d}_n^{\delta}(\mathbf{z}) \leq \sup_{\mathbf{v}} |L(\mathbf{v})|$.

The second condition follows from the calculation similar to the covariance kernel calculation.

B.1.7 Proof of Lemma CMS

Consistency

Define $\bar{\mathbf{V}}_n^{\texttt{CMS}}(\boldsymbol{\theta}) = \mathbb{E}[\tilde{\mathbf{V}}_n^{\texttt{CMS}}(\boldsymbol{\theta})]$. It suffices to show $\|\bar{\mathbf{V}}_n^{\texttt{CMS}}(\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}}) - \mathbf{V}_0^{\texttt{CMS}}\| = o_{\mathbb{P}}(1)$ and $\|\tilde{\mathbf{V}}_n^{\texttt{CMS}} - \bar{\mathbf{V}}_n^{\texttt{CMS}}(\hat{\boldsymbol{\theta}}_n^{\texttt{CMS}})\| = o_{\mathbb{P}}(1)$.

For the first requirement, changing variables, using a Taylor expansion, and using the

fact that $\int \dot{K}(v)dv = 0$, we have

$$\begin{aligned}
-\bar{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}) &= h_n^{-2}\mathbb{E}\left[y\dot{K}\left(\frac{x_1 + \mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right)\mathbf{x}_2\mathbf{x}_2' L_n(\mathbf{w})\right] \\
&= h_n^{-1}\int \mathbb{E}\left[\int \psi(vh_n - \mathbf{x}_2'\boldsymbol{\theta}, \mathbf{x}_2, \mathbf{w})\dot{K}(v)dv\mathbf{x}_2\mathbf{x}_2'\middle|\mathbf{w}\right] f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w} \\
&= \int \mathbb{E}\left[\int \psi^{(1)}(\vartheta - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{w})v\dot{K}(v)dv\mathbf{x}_2\mathbf{x}_2'\middle|\mathbf{w}\right] f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w} \\
&\quad - h_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\int \mathbb{E}\left[\int \psi^{(1)}(\vartheta - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{w})\dot{K}(v)dv\mathbf{x}_2\mathbf{x}_2\mathbf{x}_2'\middle|\mathbf{w}\right] f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w},
\end{aligned}$$

where ϑ lies in between 0 and vh_n. By boundedness of $\psi^{(1)}$, $\mathbb{E}[\|\mathbf{x}_2\|^3|\mathbf{w}] \leq C$ for $\mathbf{w} \in \mathcal{W}$, and $\int |\dot{K}(v)|dv < \infty$, the last term is $o(1)$ for $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + o(h_n)$. By continuous differentiability of $\psi^{(1)}$ with respect to its first argument and boundedness of the derivative,

$$\begin{aligned}
&\left|\int \mathbb{E}\left[\int \{\psi^{(1)}(\vartheta - \mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{w}) - \psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{w})\}v\dot{K}(v)dv\mathbf{x}_2\mathbf{x}_2'\middle|\mathbf{w}\right] f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w}\right| \\
&\leq C\int \mathbb{E}\left[\left(h_n\int v^2|\dot{K}(v)|dv + \|\mathbf{x}_2\|\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right)\|\mathbf{x}_2\|^2\middle|\mathbf{w}\right] f_{\mathbf{w}}(\mathbf{w})|L_n(\mathbf{w})|d\mathbf{w} = o(1).
\end{aligned}$$

Thus, for $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + o(h_n)$,

$$\begin{aligned}
\bar{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}) &= -\int v\dot{K}(v)dv\int \mathbb{E}[\psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2, \mathbf{w})\mathbf{x}_2\mathbf{x}_2'|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w} + o(1) \\
&= \mathbf{V}_0^{\mathtt{CMS}} + o(1).
\end{aligned}$$

Using $\|\hat{\boldsymbol{\theta}}_n^{\mathtt{CMS}} - \boldsymbol{\theta}_0\| = O_{\mathbb{P}}(\sqrt[3]{nb_n^d}) = o_{\mathbb{P}}(h_n)$, it follows that $\bar{\mathbf{V}}_n^{\mathtt{CMS}}(\hat{\boldsymbol{\theta}}_n^{\mathtt{CMS}}) - \mathbf{V}_0^{\mathtt{CMS}} = o_{\mathbb{P}}(1)$.

To verify $\tilde{\mathbf{V}}_n^{\mathtt{CMS}} - \bar{\mathbf{V}}_n^{\mathtt{CMS}}(\hat{\boldsymbol{\theta}}_n^{\mathtt{CMS}}) = o_{\mathbb{P}}(1)$, it suffices to show that, for any $\delta_n = O(r_n^{-1})$,

$$\sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \leq \delta_n}\left\|\tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}) - \tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0) - \bar{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}) + \bar{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0)\right\| + \left\|\tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0) - \bar{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0)\right\| = o_{\mathbb{P}}(1).$$

It can be shown that $\|\tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0) - \bar{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0)\| = o_{\mathbb{P}}(1)$ easily because below we calculate the convergence rate of the variance, that is, $\mathbb{V}[\tilde{\mathbf{V}}_n^{\mathtt{CMS}}(\boldsymbol{\theta}_0)] = O(n^{-1}h_n^{-3}b_n^{-d})$. For the rest, (Pollard,

1989, Theorem 4.2) implies that it is enough to verify

$$n^{-1}h_n^{-4}\mathbb{E}\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\delta_n}\left|\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right)-\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\right|^2\|\mathbf{x}_2\|^4|L_n(\mathbf{w})|^2\right]=o(1).$$

By the Lipschitz condition on $\dot{K}$,

$$\left|\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right)-\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\right|\leq h_n^{-1}\|\mathbf{x}_2\|\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|B\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right).$$

Then,

$$\begin{aligned}
&n^{-1}h_n^{-4}\mathbb{E}\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\delta_n}\left|\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}}{h_n}\right)-\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\right|^2\|\mathbf{x}_2\|^2|L_n(\mathbf{w})|^2\right]\\
&\leq n^{-1}h_n^{-3}(h_n^{-1}\delta_n)^2\mathbb{E}\left[\int B(v)^2 f_{x_1}(vh_n|\mathbf{x}_2,\mathbf{w})dv\|\mathbf{x}_2\|^6|L_n(\mathbf{w})|^2\right]\\
&\leq o(1)n^{-1}h_n^{-3}b_n^{-d}\int B(v)^2dv\int\mathbb{E}[\|\mathbf{x}_2\|^6|\mathbf{w}=\mathbf{v}b_n]L(\mathbf{v})d\mathbf{v}=o(1).
\end{aligned}$$

Approximate Mean Squared Error

First we calculate the order of bias. Changing variables and using a Taylor approximation, we obtain:

$$\begin{aligned}
\bar{\mathbf{V}}_n^{\texttt{CMS}}(\boldsymbol{\theta}_0)&=-h_n^{-2}\int\mathbb{E}\left[\psi(x_1,\mathbf{x}_2,\mathbf{w})\{f_{x_1}(x_1|\mathbf{x}_2,\mathbf{w})\}^{-1}\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)\mathbf{x}_2\mathbf{x}_2'\middle|\mathbf{w}\right]f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w}\\
&=-\int v\dot{K}(v)dv\int\mathbb{E}[\psi^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w})\mathbf{x}_2\mathbf{x}_2'|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w}\\
&\quad-h_n\int v^2\dot{K}(v)dv\int\mathbb{E}[\psi^{(2)}(-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w})\mathbf{x}_2\mathbf{x}_2'|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w}\\
&\quad-h_n^2\int\mathbb{E}[\int\psi^{(3)}(\vartheta-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2,\mathbf{w})\dot{K}(v)dv\mathbf{x}_2\mathbf{x}_2'|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})L_n(\mathbf{w})d\mathbf{w}
\end{aligned}$$

where ϑ lies between 0 and vh_n. Using $\int v\dot{K}(v)dv=-1$, $\int v^2\dot{K}(v)=0$, and boundedness of $\psi^{(3)}$,

$$\bar{\mathbf{V}}_n^{\texttt{CMS}}(\boldsymbol{\theta}_0)=\int\mathbf{V}(\mathbf{w};\boldsymbol{\theta}_0)L_n(\mathbf{w})d\mathbf{w}+O(h_n^2)=\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}M_n^{\texttt{CMS}}(\boldsymbol{\theta}_0)+O(h_n^2).$$

Finally, for the variance, changing variables and using a Taylor approximation, we ob-

tain:

$$\begin{aligned}
\mathbb{V}[\tilde{V}^{\texttt{CMS}}_{n,kl}(\boldsymbol{\theta}_0)] &= \frac{1}{nh_n^4}\mathbb{E}\left[\dot{K}\left(\frac{x_1+\mathbf{x}_2'\boldsymbol{\theta}_0}{h_n}\right)^2(\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)^2|L_n(\mathbf{w})|^2\right] - \frac{1}{n}[\bar{V}^{\texttt{CMS}}_{n,kl}(\boldsymbol{\theta}_0)]^2 \\
&= \frac{1}{nh_n^3}\mathbb{E}\left[\int \dot{K}^2(v)f_{x_1}(vh_n - \mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2,\mathbf{w})dv(\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l)^2|L_n(\mathbf{w})|^2\right] + O(n^{-1}) \\
&\leq \frac{C}{nh_n^3}\int \dot{K}^2(v)dv \int \mathbb{E}[\|\mathbf{x}_2\|^4|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})|L_n(\mathbf{w})|^2 d\mathbf{w} + O(n^{-1}) \\
&= O(n^{-1}h_n^{-3}b_n^{-d}),
\end{aligned}$$

for all (k,l) elements of $\tilde{\mathbf{V}}^{\texttt{CMS}}_n(\boldsymbol{\theta}_0)$.

B.1.8 Proof of Theorem ID

Let

$$\mathcal{Z}_n^*(x) = \hat{F}_n^*(x) - \hat{F}_n(x) + \hat{f}_n(x_0)(x-x_0) + \tilde{f}_n^{(1)}(x_0)(x-x_0)^2/2$$

and define

$$\begin{aligned}
\hat{s}_n^*(a) &= \sup_{s\geq 0}\left\{s : \mathcal{Z}_n^*(s) - as = \sup_{x\geq 0}\left[\mathcal{Z}_n^*(x) - ax\right]\right\}, \\
b &= \sup_{x\geq 0}\left[\mathcal{Z}_n^*(x) - ax\right].
\end{aligned}$$

Denote the LCM of a function f restricted on I by $L_I f$. Since linear functions are concave,

$$\begin{aligned}
\hat{s}_n^*(a) < x &\Rightarrow \mathcal{Z}_n^*(x) - ax - b < 0 \\
&\Rightarrow L_{[0,\infty)}\mathcal{Z}_n^*(x) \leq ax + b.
\end{aligned}$$

Now, for contradiction, suppose that the left-derivative of $L_{[0,\infty)}\mathcal{Z}_n^*(x)$ is strictly greater than a. Since $L_{[0,\infty)}\mathcal{Z}_n^*(x)$ is concave, its derivative, denoted by $DL_{[0,\infty)}\mathcal{Z}_n^*(\cdot)$, satisfies $DL_{[0,\infty)}\mathcal{Z}_n^*(\tilde{x}) > a$ for $\tilde{x} \leq x$. By definition, $L_{[0,\infty)}\mathcal{Z}_n^*(\hat{s}_n^*(a)) = a\hat{s}_n^*(a) + b$ and the derivative condition implies $L_{[0,\infty)}\mathcal{Z}_n^*(x) > ax + b$, which is a contradiction and thus proving $DL_{[0,\infty)}\mathcal{Z}_n^*(x) = \tilde{f}_n^*(x) \leq a$.

Now,

$$\begin{aligned}\tilde{f}_n^*(x) < a \Rightarrow L_{[0,\infty)}\mathcal{Z}_n^*(s) &< as + b \quad \text{for } s \geq x \\ \Rightarrow \hat{s}_n^*(a) &< x.\end{aligned}$$

For the first implication, by definition,

$$\lim_{\varepsilon \downarrow 0} \frac{L_{[0,\infty)}\mathcal{Z}_n^*(s-\varepsilon) - L_{[0,\infty)}\mathcal{Z}_n^*(s)}{-\varepsilon} = \tilde{f}_n^*(s) < a \quad \text{for } s \geq x.$$

Then, there exist $\eta > 0$ such that $a - \eta > \tilde{f}_n^*(s)$ and $\varepsilon_0 > 0$ such that $-[L_{[0,\infty)}\mathcal{Z}_n^*(s-\varepsilon) - L_{[0,\infty)}\mathcal{Z}_n^*(s)]/\varepsilon < a - \eta$ for $0 < \varepsilon < \varepsilon_0$. Then,

$$\begin{aligned}L_{[0,\infty)}\mathcal{Z}_n^*(s) &< \varepsilon(a-\eta) + L_{[0,\infty)}\mathcal{Z}_n^*(s-\varepsilon) \\ &\leq \varepsilon(a-\eta) + a(s-\varepsilon) + b \\ &= as - \varepsilon\eta + b \\ &< as + b.\end{aligned}$$

Therefore, we have

$$\hat{s}_n^*(a) < x \iff \tilde{f}_n^*(x) < a.$$

Then,

$$\mathbb{P}^*\left[n^{1/3}[\tilde{f}_n^*(x_0) - \hat{f}_n(x_0)] < t\right] = \mathbb{P}^*\left[\hat{s}_n^*\left(\hat{f}_n(x_0) + n^{-1/3}t\right) < x_0\right].$$

Recall that $\hat{s}_n^*(a)$ is the maximizer of $\{\mathcal{Z}_n^*(s) - as : s \in [0,\infty)\}$.

$$\mathbb{P}^*\left[\hat{s}_n^*\left(\hat{f}_n(x_0) + n^{-1/3}t\right) < x_0\right] = \mathbb{P}^*\left[\underset{x\in[0,\infty)}{\arg\max}\left\{\mathcal{Z}_n^*(x) - [\hat{f}_n(x_0) + n^{-1/3}t]x\right\} < x_0\right]$$

where the argmax functional picks the largest value if the set of the maximizers consists of more than one element.

Letting $x = x_0 + n^{-1/3}s$ where s is the free variable,

$$\begin{aligned}
&\arg\max_{x\in[0,\infty)} \left\{ \mathcal{Z}_n^*(x) - [\hat{f}_n(x_0) + n^{-1/3}t]x \right\} \\
&= \arg\max_{x_0+n^{-1/3}s} \left\{ \mathcal{Z}_n^*(x_0 + n^{-1/3}s) - [\hat{f}_n(x_0) + n^{-1/3}t][x_0 + n^{-1/3}s] \right\} \\
&= x_0 + n^{-1/3} \arg\max_{s\in[-n^{1/3}x_0,\infty)} \left\{ \mathcal{Z}_n^*(x_0 + n^{-1/3}s) - [\hat{f}_n(x_0) + n^{-1/3}t][x_0 + n^{-1/3}s] \right\}
\end{aligned}$$

Shifting and scaling does not change the location of maximizer so

$$\begin{aligned}
&\arg\max_{s} \left\{ \mathcal{Z}_n^*(x_0 + n^{-1/3}s) - [\hat{f}_n(x_0) + n^{-1/3}t][x_0 + n^{-1/3}s] \right\} \\
&= \arg\max_{s} \left\{ n^{2/3} \left[\mathcal{Z}_n^*(x_0 + n^{-1/3}s) - \mathcal{Z}_n^*(x_0) - \hat{f}_n(x_0)n^{-1/3}s \right] - ts \right\}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{P}^* \left[\arg\max_{x\in[0,\infty)} \left\{ \mathcal{Z}_n^*(x) - [\hat{f}_n(x_0) + n^{-1/3}t]x \right\} < x_0 \right] \\
&= \mathbb{P}^* \left[\arg\max_{s\in[-n^{1/3}x_0,\infty)} \left\{ n^{2/3} \left[\mathcal{Z}_n^*(x_0 + n^{-1/3}s) - \mathcal{Z}_n^*(x_0) - \hat{f}_n(x_0)n^{-1/3}s \right] - ts \right\} < 0 \right].
\end{aligned}$$

Now we want to establish weak convergence in probability of

$$\left\{ n^{2/3} \left[\mathcal{Z}_n^*(x_0 + n^{-1/3}s) - \mathcal{Z}_n^*(x_0) - \hat{f}_n(x_0)n^{-1/3}s \right] - ts : s \in \mathcal{S} \right\} \tag{B.9}$$

with any compact $\mathcal{S} \subset \mathbb{R}$.

Note that

$$\begin{aligned}
&n^{2/3} \left[\mathcal{Z}_n^*(x_0 + n^{-1/3}s) - \mathcal{Z}_n^*(x_0) - \hat{f}_n(x_0)n^{-1/3}s \right] \\
&\quad = n^{2/3} \left[\left(\hat{F}_n^* - \hat{F}_n \right)(x_0 + n^{-1/3}s) - \left(\hat{F}_n^* - \hat{F}_n \right)(x_0) \right] + \frac{1}{2}\tilde{f}_n^{(1)}(x_0)s^2.
\end{aligned}$$

Theorem 3.2 (i) in Sen *et al.* (2010) implies that

$$n^{2/3} \left[\left(\hat{F}_n^* - \hat{F}_n \right)(x_0 + n^{-1/3}s) - \left(\hat{F}_n^* - \hat{F}_n \right)(x_0) \right] \rightsquigarrow_{\mathbb{P}} \mathcal{W}(f(x_0)s)$$

as stochastic process indexed by $s \in \mathcal{S}$.

Then, the sequence of stochastic processes in (B.9) weakly converges in probability to

$$\mathcal{W}(f(x_0)s) + \frac{1}{2}f^{(1)}(x_0)s^2 - ts.$$

As noted in Problem 3.2.5 of van der Vaart and Wellner (1996),

$$\begin{aligned}\arg\max_s &\left\{\mathcal{W}(f(x_0)s) + \frac{1}{2}f^{(1)}(x_0)s^2 - ts\right\} \\ &=_d |4f(x_0)\{f^{(1)}(x_0)\}^{-2}|^{1/3} \arg\max_s \left\{\mathcal{W}(s) - s^2\right\} - t/|f^{(1)}(x_0)|.\end{aligned}$$

Therefore, provided that

$$\hat{s}_n^* = \arg\max_s \left\{n^{2/3}\left[\mathcal{Z}_n^*(x_0 + n^{-1/3}s) - \mathcal{Z}_n^*(x_0) - \hat{f}_n(x_0)n^{-1/3}s\right] - ts : s \in \mathcal{S}\right\}$$

is $O_{\mathbb{P}^*}(1)$, the argmax continuous mapping theorem implies

$$\mathbb{P}^*\left[n^{1/3}[\tilde{f}_n^*(x_0) - \hat{f}_n(x_0)] < t\right] \to_{\mathbb{P}} \mathbb{P}\left[|4f(x_0)f^{(1)}(x_0)|^{1/3} \arg\max_s \left\{\mathcal{W}(s) - s^2\right\} < t\right].$$

Now to show $\hat{s}_n^* = O_{\mathbb{P}^*}(1)$, define

$$\begin{aligned}\hat{g}_n^* &= \arg\max_s \left\{\mathcal{Z}_n^*(s + x_0) - \mathcal{Z}_n^*(x_0) - \hat{f}_n(x_0)s - tsn^{-1/3}\right\} \\ &= \arg\max_s \left\{\hat{F}_n^*(s + x_0) - \hat{F}_n(s + x_0) + \tilde{f}_n^{(1)}(x_0)s^2/2 - tsn^{-1/3}\right\}.\end{aligned}$$

By definition, $\hat{g}_n^* = n^{-1/3}\hat{s}_n^*$, so we want to show $n^{1/3}\hat{g}_n^* = O_{\mathbb{P}^*}(1)$.

Modifying the proofs of Lemmas 16–18 with $\hat{M}_n^*(\boldsymbol{\theta}) = \hat{F}_n^*(\boldsymbol{\theta} + x_0)$, $\hat{M}_n(\boldsymbol{\theta}) = \hat{F}_n(\boldsymbol{\theta} + x_0)$, and $\tilde{M}_n^*(\boldsymbol{\theta}) = \hat{F}_n^*(\boldsymbol{\theta}+x_0) - \hat{F}_n(\boldsymbol{\theta}+x_0) + \tilde{f}_n^{(1)}(x_0)\boldsymbol{\theta}^2/2 - t\boldsymbol{\theta}n^{-1/3}$ (no need to center the objective function at $\hat{M}_n(\boldsymbol{\theta}_0)$), it can be shown that $n^{1/3}\hat{g}_n^* = O_{\mathbb{P}^*}(1)$ with probability approaching one.

B.2 Rule-of-Thumb Bandwidth Selection

We provide details on the rule-of-thumb (ROT) bandwidth selection rules used in the simulations reported for the Maximum Score Estimation and Isotonic Density estimation examples.

B.2.1 Maximum Score Estimation

To construct a ROT bandwidth in this example, we choose a reference model involving finite dimensional parameters and calculate/approximate the corresponding leading constants entering the approximate MSE of $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$ (example-specific plug-in estimate) for the case of Maximum Score estimation, and of $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$ (generic numerical derivative estimate).

Specifically, we assume $u|\mathbf{x} \sim \mathsf{Normal}(0, \sigma_u^2(\mathbf{x}))$ and $x_1|\mathbf{x}_2 \sim \mathsf{Normal}(\mu_1, \sigma_1^2)$, where we will specify some parametric specification on $\sigma_u^2(\mathbf{x}) = \sigma_u^2(x_1, \mathbf{x}_2)$. Then, in this reference model,

$$F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) = -\frac{1}{\sigma_u(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)}\phi(0),$$

$$F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) = \frac{1}{\sigma_u^2(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)}\phi^{(1)}(0) = 0,$$

$$\begin{aligned} F^{(3)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) = &-\frac{1}{\sigma_u^3(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)}\phi^{(2)}(0) \\ &- \frac{\sigma_u^{(2)}(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)\sigma_u(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2) - 2\{\sigma_u^{(1)}(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)\}^2}{\sigma_u^3(-\mathbf{x}_2'\boldsymbol{\theta}_0, \mathbf{x}_2)}\phi(0), \end{aligned}$$

where $\sigma_u^{(s)}(x_1, \cdot) = d^s\sigma_u(x_1, \cdot)/dx_1^s$, and

$$f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) = \frac{1}{\sigma_1}\phi\left(\frac{-\mathbf{x}_2'\boldsymbol{\theta}_0 - \mu_1}{\sigma_1}\right),$$

$$f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) = \frac{1}{\sigma_1^2}\phi^{(1)}\left(\frac{-\mathbf{x}_2'\boldsymbol{\theta}_0 - \mu_1}{\sigma_1}\right) = \frac{\mathbf{x}_2'\boldsymbol{\theta}_0 + \mu_1}{\sigma_1^3}\phi\left(\frac{-\mathbf{x}_2'\boldsymbol{\theta}_0 - \mu_1}{\sigma_1}\right),$$

$$f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2) = \frac{1}{\sigma_1^3}\phi^{(2)}\left(\frac{-\mathbf{x}_2'\boldsymbol{\theta}_0 - \mu_1}{\sigma_1}\right) = \left(\left(\frac{\mathbf{x}_2'\boldsymbol{\theta}_0 + \mu_1}{\sigma_1}\right)^2 - 1\right)\frac{1}{\sigma_1^3}\phi\left(\frac{-\mathbf{x}_2'\boldsymbol{\theta}_0 - \mu_1}{\sigma_1}\right).$$

Plug-in Estimator $\tilde{\mathsf{V}}_n^{\mathtt{MS}}$

Given the reference model, I need to estimate

$$\mathsf{B}_{kl} = \mathbb{E}\left[\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2'\mathbf{e}_l\left(F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right.\right.$$
$$\left.\left. + \frac{1}{3}F^{(3)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2)f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right)\right]\mu_3(\dot{K})$$

where $\mu_p(K) = \int v^p K(u)dv$. Define

$$\hat{\mathsf{B}}_{kl} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_k'\mathbf{x}_{2i}\mathbf{x}_{2i}'\mathbf{e}_l\left\{\hat{F}^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n,\mathbf{x}_{2i})\hat{f}^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n|\mathbf{x}_{2i})\right.$$
$$\left. + \frac{1}{3}\hat{F}^{(3)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n,\mathbf{x}_{2i})\hat{f}_{x_1|\mathbf{x}_2}(-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n|\mathbf{x}_{2i})\right\}\mu_3(\dot{K})$$

where the preliminary estimators $\hat{\boldsymbol{\theta}}_n$, $\hat{F}^{(1)}_{u|x_1,\mathbf{x}_2}(\cdot)$, $\hat{F}^{(2)}_{u|x_1,\mathbf{x}_2}(\cdot)$, $\hat{f}_{x_1|\mathbf{x}_2}(\cdot)$ and $\hat{f}^{(2)}_{x_1|\mathbf{x}_2}(\cdot)$ are constructed using maximum likelihood for the parametric reference model (i.e., heteroskedastic Probit), together with a flexible parametric specification for $\sigma_u^2(\mathbf{x}) = \gamma'\mathbf{p}(\mathbf{x})$ with $\mathbf{p}(\mathbf{x})$ denoting a polynomial expansion.

Similarly, the higher order variance is

$$\mathsf{V}_{kl} = \mathbb{E}\left[(\mathbf{e}_k'\mathbf{x}_2\mathbf{x}_2\mathbf{e}_l)^2 f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right]\int \dot{K}(v)^2dv,$$

and, given the reference model, a ROT estimate is

$$\hat{\mathsf{V}}_{kl} = \int \dot{K}(v)^2dv\frac{1}{n}\sum_{i=1}^{n}(\mathbf{e}_k'\mathbf{x}_{2i}\mathbf{x}_{2i}\mathbf{e}_l)^2\hat{f}_{x_1|\mathbf{x}_2}(-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n|\mathbf{x}_{2i}).$$

Numerical Differentiation Estimator $\tilde{\mathsf{V}}_n^{\mathtt{ND}}$

The bias constant is

$$\mathsf{B}_{kl} = -\frac{1}{6}\left(\frac{\partial^4 M_0(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_k^3\partial\boldsymbol{\theta}_l} + \frac{\partial^4 M_0(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_l^3}\right)$$

Given the reference model,

$$\begin{aligned}\frac{\partial^4 M_0(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_k^3 \partial \boldsymbol{\theta}_l} = 2\mathbb{E}\left[(\mathbf{e}_k'\mathbf{x}_2)^3 \mathbf{e}_l'\mathbf{x}_2 F^{(3)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right] \\ + 6\mathbb{E}\left[(\mathbf{e}_k'\mathbf{x}_2)^3 \mathbf{e}_l'\mathbf{x}_2 F^{(2)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(1)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right] \\ + 6\mathbb{E}\left[(\mathbf{e}_k'\mathbf{x}_2)^3 \mathbf{e}_l'\mathbf{x}_2 F^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_2'\boldsymbol{\theta}_0,\mathbf{x}_2) f^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right],\end{aligned}$$

and similarly for $\partial^4 M_0(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_l^3$. Then, we can estimate the bias constant by

$$\begin{aligned}\hat{\mathsf{B}}_{kl} = -\frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{e}_k'\mathbf{x}_{2i})^3 \mathbf{e}_l'\mathbf{x}_{2i} + (\mathbf{e}_l'\mathbf{x}_{2i})^3 \mathbf{e}_k'\mathbf{x}_{2i}\right] \\ \times\left[\frac{1}{3}\hat{F}^{(3)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n,\mathbf{x}_{2i})\hat{f}_{x_1|\mathbf{x}_2}(-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n|\mathbf{x}_{2i}) + \hat{F}^{(1)}_{u|x_1,\mathbf{x}_2}(0|-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n,\mathbf{x}_{2i})\hat{f}^{(2)}_{x_1|\mathbf{x}_2}(-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n|\mathbf{x}_{2i})\right]\end{aligned}$$

where $\hat{F}^{(1)}_{\varepsilon|z,\mathbf{x}_2}$, $\hat{F}^{(3)}_{\varepsilon|z,\mathbf{x}_2}$, $\hat{f}_{z|\mathbf{x}_2}$, $\hat{f}^{(2)}_{z|\mathbf{x}_2}$, etc., are as discussed above.

Finally, the variance constant is

$$\mathsf{V}_{kl} = \frac{1}{8}\left[\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,\mathbf{e}_k+\mathbf{e}_l) + \mathcal{C}_0(\mathbf{e}_k-\mathbf{e}_l,\mathbf{e}_k-\mathbf{e}_l) - 2\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,\mathbf{e}_k-\mathbf{e}_l) - 2\mathcal{C}_0(\mathbf{e}_k+\mathbf{e}_l,-\mathbf{e}_k+\mathbf{e}_l)\right]$$

$\mathcal{C}_0$ is defined in Corollary **??**. Recall that $\mathcal{C}_0(\mathbf{s},\mathbf{t}) = [\mathcal{L}_0(\mathbf{s}) + \mathcal{L}_0(\mathbf{t}) - \mathcal{L}_0(\mathbf{s}-\mathbf{t})]/2$. Then,

$$\mathsf{V}_{kl} = \frac{1}{16}\left[2\mathcal{L}_0(\mathbf{e}_k) + 2\mathcal{L}_0(\mathbf{e}_l) - \mathcal{L}_0(\mathbf{e}_k+\mathbf{e}_l) - \mathcal{L}_0(-\mathbf{e}_k+\mathbf{e}_l)\right]$$

with $\mathcal{L}_0(\mathbf{s}) = \mathbb{E}\left[|\mathbf{x}_2'\mathbf{s}| f_{x_1|\mathbf{x}_2}(-\mathbf{x}_2'\boldsymbol{\theta}_0|\mathbf{x}_2)\right]$. Then, we can estimate the variance constant by

$$\hat{\mathsf{V}}_{kl} = \frac{1}{16}\frac{1}{n}\sum_{i=1}^{n}\left[2\left|\mathbf{x}_{2i}'\mathbf{e}_k\right| + 2\left|\mathbf{x}_{2i}'\mathbf{e}_l\right| - \left|\mathbf{x}_{2i}'(\mathbf{e}_k+\mathbf{e}_l)\right| - \left|\mathbf{x}_{2i}'(-\mathbf{e}_k+\mathbf{e}_l)\right|\right]\hat{f}_{x_1|\mathbf{x}_2}(-\mathbf{x}_{2i}'\hat{\boldsymbol{\theta}}_n|\mathbf{x}_{2i}).$$

B.2.2 Isotonic Density Estimation

Here we briefly outline the calculations leading to a ROT bandwidth selector for derivative of density estimators, which are used to construct an estimate of the drift in this example. The reference model is $z \sim \mathsf{Exponential}(\lambda)$ where λ will be estimated from the data.

Plug-in Estimator $\tilde{\mathbf{V}}_n^{\mathtt{MS}}$

The plug-in estimator is the usual kernel density estimator. Assuming we use the second-order kernel, the leading bias term is $h^2 f^{(3)}(x_0)/2 \int u^2 K(u) du$ and the leading variance term is $n^{-1}h^{-3} f(x_0) \int \dot{K}^2(u) du$. For the ROT bandwidth, we take $f(x) = \exp(-x/\lambda)/\lambda$ and $K(u) = \phi(u)$ and we use Integrated MSE (IMSE) as the criterion. That is, we minimize

$$h^4 \int \left\{ f^{(3)}(x) \right\}^2 dx \left(\int u^2 K(u) du \right)^2 /4 + \int \dot{K}^2(u) du / nh^3$$

with respect to h. Then, $\int u^2 K(u) du = 1$, $\int \dot{K}^2(u) du = \frac{1}{4\sqrt{\pi}}$, $\int \left\{ f^{(3)}(x) \right\}^2 dx = (2\lambda^7)^{-1}$. Thus, the IMSE equals

$$\frac{h^4}{8\lambda^7} + \frac{1}{4\sqrt{\pi} n h^3}$$

and the minimizer is

$$\hat{h}_{\mathtt{AMSE}} = \lambda \left(\frac{3}{2\sqrt{\pi}} \right)^{1/7} n^{-1/7}.$$

Numerical Differentiation Estimator $\tilde{\mathbf{V}}_n^{\mathtt{ND}}$

The numerical differentiation estimator is

$$\frac{M_0(x_0 + h) + M_0(x_0 - h) - 2M_0(x_0)}{h^2},$$

and we take $M_0(x) = F(x)$. Then, the bias constant is

$$\frac{1}{12} \frac{d^4}{dx^4} F(x)|_{x=x_0} = \frac{f^{(3)}(x_0)}{12}.$$

The variance constant is $2f(x_0)$. As in the plug-in estimator case, we take the IMSE, which equals

$$\frac{h^4}{12^2} \int \left\{ f^{(3)}(x) \right\}^2 dx + \frac{2}{nh^3} = \frac{h^4}{288\lambda^7} + \frac{2}{nh^3}$$

and the ROT bandwidth is

$$\hat{h}_{\mathtt{AMSE}} = \lambda \, (432)^{1/7} \, n^{-1/7}.$$

APPENDIX C

Proofs for Chapter IV

C.1 Proof of Theorem 7

First suppose that $\pi_0 = \lim_{n\to\infty} n^{\gamma/2} q_n$ exists on the extended real line. By change of variable,

$$r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \arg\max_{\mathbf{s}\in\mathbb{R}^k} \{r_n^2[\hat{M}_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - \hat{M}_n(\boldsymbol{\theta}_0)]\}$$

and by argmax continuous mapping theorem (e.g., van der Vaart and Wellner, 1996), it suffices to show $r_n^2\{\hat{M}_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - \hat{M}_n(\boldsymbol{\theta}_0)\}$ weakly converges to $\mathcal{G}_0$, Gaussian process with mean $-\mathbf{s}'\mathbf{V}_0\mathbf{s}$ and the covariance kernel $4\mathbf{s}'\boldsymbol{\Sigma}_0\mathbf{t}(\pi_0/(1+\pi_0))^2 + 2\mathcal{C}_0(\mathbf{s},\mathbf{t})(1/(1+\pi_0))^{4-\gamma}$.

Lemmas 21 and 23 establish $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}}(1)$. Finite-dimensional convergence of the stochastic processes $\{\hat{G}_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) : s \in \mathbb{R}^k\}$ where $\hat{G}_n(\boldsymbol{\theta}) = r_n^2[L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0) + W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)]$ follows from Lemmas 24-27. Lemma 28 verifies the sthocastic equicotinutiy of $\{\hat{G}_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) : s \in \mathcal{S}\}$ for any compact $\mathcal{S} \subset \mathbb{R}^k$. Lemma 29 shows $\lim_{n\to\infty} r_n^2[M_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - M_n(\boldsymbol{\theta}_0)] = \mathbf{s}'\mathbf{V}_0\mathbf{s}$ uniformly in $\mathbf{s}$ over a bounded set. These results together imply $r_n^2\{\hat{M}_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - \hat{M}_n(\boldsymbol{\theta}_0)\}$ weakly converges to $\mathcal{G}_0$.

For the case $\pi_n = n^{\gamma/2} q_n$ does not have a limit, consider convergent subsequences. Since weak convergence holds for any convergent subsequence, the original sequence must also converge.

C.1.1 Consistency

Lemma 21. *Suppose Assumption PS (i) holds. If* $\hat{M}_n(\hat{\boldsymbol{\theta}}) \geq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - o_{\mathbb{P}}(1)$*, then* $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_{\mathbb{P}}(1)$.

Proof. For any $\delta > 0$, letting $\alpha_n = o_{\mathbb{P}}(1)$,

$$\Pr(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > \delta) \leq \Pr\Big(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^{\delta}} \hat{M}_n(\boldsymbol{\theta}) \geq \hat{M}_n(\boldsymbol{\theta}_0) - \alpha_n\Big), \tag{C.1}$$

and it suffices to show the right-hand side probability converges to zero. By Hoeffding decomposition,

$$\hat{M}_n(\boldsymbol{\theta}) = M_n(\boldsymbol{\theta}) + L_n(\boldsymbol{\theta}) + W_n(\boldsymbol{\theta})$$

where

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} 2\ell_n(\mathbf{z}_i, \boldsymbol{\theta}) \\ W_n(\boldsymbol{\theta}) &= \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}). \end{aligned}$$

By Theorem 4.2 in Pollard (1989) and Theorem 6 of Nolan and Pollard (1987),

$$\begin{aligned} \mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |L_n(\boldsymbol{\theta})| &\lesssim n^{-1/2}(\mathbb{E}|\bar{\ell}_n(\mathbf{z})|^2)^{1/2} \\ \mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |W_n(\boldsymbol{\theta})| &\lesssim n^{-1}(\mathbb{E}|\bar{w}_n(\mathbf{z}_1, \mathbf{z}_2)|^2)^{1/2}. \end{aligned}$$

Both terms are $o(1)$ by the hypothesis. Thus,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\hat{M}_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta})| = o_{\mathbb{P}}(1),$$

which, combined with the assumption $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})| = o(1)$, implies

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^{\delta}} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^{\delta}} M_0(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta}_0) + o_{\mathbb{P}}(1).$$

Since $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\boldsymbol{\Theta}_0^{\delta}} M_0(\boldsymbol{\theta}) < M_0(\boldsymbol{\theta}_0)$ by the hypothesis, the desired conclusion follows. □

C.1.2 Rates of Convergence

The following lemma is used to establish the convergence rate of $\hat{\boldsymbol{\theta}}_n$.

Lemma 22. *Suppose Assumption PS (ii) holds. For $0 \leq a < b$, there exists some $c > 0$ such that, with n sufficiently large,*

$$\sup_{a \leq \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq b} |M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0)| \leq b \Big\| \frac{\partial M_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \Big\| - a^2 c.$$

Proof. By Taylor expansion,

$$\begin{aligned} M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0) &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \frac{\partial M_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \frac{\partial^2 M_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \frac{\partial^2 [M_n(\tilde{\boldsymbol{\theta}}_n) - M_n(\boldsymbol{\theta}_0)]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \Big\| \frac{\partial M_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \Big\| + \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \Big(\Big\| \frac{\partial^2 [M_n(\tilde{\boldsymbol{\theta}}_n) - M_n(\boldsymbol{\theta}_0)]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big\| - \lambda_{\min}(\mathbf{V}_n) \Big) \end{aligned}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue and $\mathbf{V}_n = -\partial^2 M_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$. For some $\delta > 0$, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^\delta} \|\partial^2 [M_n(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta})]/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'\| = o(1)$ and $\partial^2 M_0(\boldsymbol{\theta})]/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ is continuous on $\boldsymbol{\Theta}_0^\delta$. Then, positive definiteness of $\mathbf{V}_0$ implies that for some $c > 0$,

$$\limsup_{n \to \infty} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^\delta} \Big\| \frac{\partial^2 [M_n(\tilde{\boldsymbol{\theta}}_n) - M_n(\boldsymbol{\theta}_0)]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big\| - \lambda_{\min}(\mathbf{V}_n) < -2c.$$

□

Lemma 23. *Suppose $\hat{\boldsymbol{\theta}}_n \to_{\mathbb{P}} \boldsymbol{\theta}_0$ and*

$$\hat{M}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - o_{\mathbb{P}}(r_n^{-2}).$$

If Assumption PS (ii) and (iii) hold, then $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}}(1)$.

Proof. Let $\mathcal{S}_j = \{\boldsymbol{\theta} : 2^{j-1} < r_n\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^j\}$. For any $\delta > 0$,

$$\begin{aligned}\Pr(r_n\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n| > 2^C) \leq \Pr\Big(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\hat{\boldsymbol{\theta}}_n) > \delta r_n^{-2}\Big) \\ + \sum_{\substack{j\geq C \\ 2^j \leq r_n\delta}} \Pr\Big(\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) \geq -\delta r_n^{-2}\Big) + \Pr(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \delta/2)\end{aligned}$$

where the first and third terms go to zero as $n \to \infty$. For each of the summand, by (C.1),

$$\begin{aligned}&\Pr\Big(\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) \geq -\delta r_n^{-2}\Big) \\ &\leq \Pr\Big(\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0) + \sup_{\boldsymbol{\theta}\in\mathcal{S}_j} W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0) + \sup_{\boldsymbol{\theta}\in\mathcal{S}_j} M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0) \geq -\delta r_n^{-2}\Big) \\ &\leq \Pr\Big(\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} |L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0)| + \sup_{\boldsymbol{\theta}\in\mathcal{S}_j} |W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)| \geq 2^{2j}\varepsilon r_n^{-2}\Big),\end{aligned}$$

for j large enough, where we use Lemma 22. Note $\varepsilon > 0$ depends on c in Lemma 22 and δ but not on n and j as long as they are sufficiently large.

Now, Theorem 4.2 in Pollard (1989) and Theorem 6 of Nolan and Pollard (1987) imply that

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} |L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0)| &\lesssim n^{-1/2}\sqrt{\mathbb{E}|\bar{\ell}_n^{2^j r_n^{-1}}(\mathbf{z})|^2} \lesssim 2^j n^{-1/2} r_n^{-1} \\ \mathbb{E}\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} |W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)| &\lesssim n^{-1}\sqrt{\mathbb{E}|\bar{w}_n^{2^j r_n^{-1}}(\mathbf{z})|^2} \lesssim 2^{\gamma j/2} n^{-1} q_n^{-1/2} r_n^{-\gamma/2}.\end{aligned}$$

Then,

$$\sum_{\substack{j\geq C \\ 2^j \leq r_n\delta}} \Pr\Big(\sup_{\boldsymbol{\theta}\in\mathcal{S}_j} \hat{M}_n(\boldsymbol{\theta}) - \hat{M}_n(\boldsymbol{\theta}_0) \geq -\delta r_n^{-2}\Big) \lesssim \sqrt{\frac{r_n^2}{n}} \sum_{j\geq C} 2^{-j} + \sqrt{\frac{r_n^{4-\gamma}}{n^2 q_n}} \sum_{j\geq C} 2^{\frac{(\gamma-4)j}{2}}$$

and the last term can be made arbitrarily small by choosing C large. □

C.1.3 Weak Convergence

We want to show $r_n^2\{L_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - L_n(\boldsymbol{\theta}_0)\} + r_n^2\{W_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - W_n(\boldsymbol{\theta}_0)\} \rightsquigarrow \mathcal{G}_0$. For that purpose, we will establish finite-dimensional convergence and stochastic equicontinuity.

Finite-Dimensional Convergence

For finite-dimensional convergence, first note that

$$\hat{G}_n(\boldsymbol{\theta}) = r_n^2\{L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0)\} + r_n^2\{W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)\} = \sum_{i=1}^{n} g_{in}(\boldsymbol{\theta})$$

$$g_{in}(\boldsymbol{\theta}) = r_n^2\left[2n^{-1}\{\ell_n(\mathbf{z}_i, \boldsymbol{\theta}) - \ell_n(\mathbf{z}_i, \boldsymbol{\theta})\} + \binom{n}{2}^{-1} \sum_{j=1}^{i-1}\{w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}) - w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0)\}\right]$$

and with the sigma field $\mathcal{F}_{in}$ generated by $g_{1n}(\boldsymbol{\theta}), \ldots, g_{in}(\boldsymbol{\theta})$,

$$\mathbb{E}[g_{in}|\mathcal{F}_{(i-1)n}] = \mathbb{E}[\mathbb{E}[g_{in}(\boldsymbol{\theta})|\mathbf{z}_1, \ldots, \mathbf{z}_n]|\mathcal{F}_{(i-1)n}],$$

$\mathbb{E}[\ell_n(\mathbf{z}_i, \boldsymbol{\theta})|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}] = 0$, $\mathbb{E}[w_n(\mathbf{z}_i, \mathbf{z}_j)|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}] = 0$ for $j \in \{1, \ldots, i-1\}$, so $\{(g_{in}(\boldsymbol{\theta}), \mathcal{F}_{in})\}_{i\geq 1}$ forms a martingale difference sequence. Below, we write $\mathcal{F}_i$ for notational convenience.

Using this martingale structure, we apply the result in Hyde and Brown (1970).

Lemma 24. *Let $\{X_n, \mathcal{F}_n\}_{n\geq 0}$ be a martingale with X_0 a.s., $X_n = \sum_{i=1}^n Y_i$ $n \geq 1$, and $\mathcal{F}_n$ be the sigma field generated by $X_0, X_1, \ldots, X_n$. Define $\sigma_n^2 = \mathbb{E}[Y_n^2|\mathcal{F}_{n-1}]$ and $s_n^2 = \sum_{i=1}^n \mathbb{E}\sigma_n^2$. Suppose for some $\delta \in (0, 1]$, $\mathbb{E}|Y_n|^{2+2\delta} < \infty$ for all n. Then, there exists a finite constant K depending only on δ such that*

$$\sup_t |\Pr(X_n \leq s_n t) - \Phi(t)| \leq K\Big\{s_n^{-2-2\delta}\Big(\sum_{i=1}^n \mathbb{E}|Y_i|^{2+2\delta} + \mathbb{E}\Big|\sum_{i=1}^n \sigma_i^2 - s_n^2\Big|^{1+\delta}\Big)\Big\}^{\frac{1}{3+2\delta}}.$$

By the Cramer-Wold device, it suffices to consider $g_{in} := \sum_{l=1}^L \lambda_l g_{in}(\boldsymbol{\theta}_0 + \mathbf{s}_l r_n^{-1})$ for some $(\mathbf{s}_l, \lambda_l) \in \mathbb{R}^{k+1}$, $l = 1, \ldots, L$.

The above lemma implies that if, for $\sigma_{in}^2 = \mathbb{E}[g_{in}^2|\mathcal{F}_{i-1}]$ and $s_n^2 = \sum_{i=1}^n \mathbb{E}\sigma_{in}^2$,

$$\frac{1}{s_n^4}\sum_{i=1}^n \mathbb{E}|g_{in}|^4 \to 0 \tag{C.2}$$

$$\mathbb{E}\Big|\frac{1}{s_n^2}\sum_{i=1}^n \sigma_{in}^2 - 1\Big|^2 \to 0, \tag{C.3}$$

then $\hat{G}_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) \rightsquigarrow \mathcal{G}_0$ in finite-dimensional sense.

Lemma 25. *Suppose Assumption PS (iv) holds and $\pi_0 = \lim_{n\to\infty}(n^{\gamma/2}q_n)^{1/(4-\gamma)}$ exists in the extended real line. Then,*

$$\lim_{n\to\infty} s_n^2 = \Big(\frac{1}{1+1/\pi_0}\Big)^2 \sum_{l=1}^L\sum_{l'=1}^L \lambda_l\lambda_{l'} 4\mathbf{s}_l'\boldsymbol{\Sigma}_0\mathbf{s}_{l'} + \Big(\frac{1}{1+\pi_0}\Big)^{4-\gamma}\sum_{l=1}^L\sum_{l'=1}^L \lambda_l\lambda_{l'} 2\mathcal{C}_0(\mathbf{s}_l,\mathbf{s}_{l'}).$$

where we interpret $1+\infty=\infty$, $1/\infty = 0$, and $1/0=\infty$

Proof.

$$\begin{aligned}
s_n^2 &= \sum_{i=1}^n \mathrm{Var}(g_{in}) \\
&= r_n^4 n^{-1} 4\mathrm{Var}\bigg(\sum_{l=1}^L \lambda_l\{\ell_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}_l r_n^{-1}) - \ell_n(\mathbf{z},\boldsymbol{\theta}_0)\}\bigg) \\
&\quad + r_n^4\binom{n}{2}^{-1}\mathrm{Var}\bigg(\sum_{l=1}^L \lambda_l\{w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta}_0+\mathbf{s}_l r_n^{-1}) - w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta}_0)\}\bigg) \\
&= r_n^2 n^{-1}\sum_{l=1}^L\sum_{l'=1}^L \lambda_l\lambda_{l'} 4r_n^2\mathbb{E}\{\ell_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}_l r_n^{-1}) - \ell_n(\mathbf{z},\boldsymbol{\theta}_0)\}\{\ell_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}_{l'} r_n^{-1}) - \ell_n(\mathbf{z},\boldsymbol{\theta}_0)\} \\
&\quad + \frac{r_n^{4-\gamma}}{n(n-1)q_n}\sum_{l=1}^L\sum_{l'=1}^L \lambda_l\lambda_{l'} 2r_n^\gamma q_n\mathbb{E}\{w_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}_l r_n^{-1}) - w_n(\mathbf{z},\boldsymbol{\theta}_0)\}\{w_n(\mathbf{z},\boldsymbol{\theta}_0+\mathbf{s}_{l'} r_n^{-1}) - w_n(\mathbf{z},\boldsymbol{\theta}_0)\} \\
&\to \Big(\frac{1}{1+1/\pi_0}\Big)^2 \sum_{l=1}^L\sum_{l'=1}^L \lambda_l\lambda_{l'} 4\mathbf{s}_l'\boldsymbol{\Sigma}_0\mathbf{s}_{l'} + \Big(\frac{1}{1+\pi_0}\Big)^{4-\gamma}\sum_{l=1}^L\sum_{l'=1}^L \lambda_l\lambda_{l'} 2\mathcal{C}_0(\mathbf{s}_l,\mathbf{s}_{l'}).
\end{aligned}$$

□

Lemma 26. *Suppose Assumption PS (v) holds. Then,* (C.2) *holds.*

Proof. Write $\tilde{\ell}_n(\mathbf{z}) = \sum_{l=1}^L \lambda_l\{\ell_n(\mathbf{z}, \boldsymbol{\theta}_0 + \mathbf{s}_l r_n^{-1}) - \ell_n(\mathbf{z}, \boldsymbol{\theta}_0)\}$ and $\tilde{w}_n(\mathbf{z}_1, \mathbf{z}_2) = \sum_{l=1}^L \lambda_l\{w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0 + \mathbf{s}_l r_n^{-1}) - w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0)\}$. By $(a+b)^s \leq 2^{s-1}(a^s + b^s)$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}|g_{in}|^4 &\lesssim r_n^8 n^{-4} \sum_{i=1}^n \mathbb{E}|\tilde{\ell}_n(\mathbf{z}_i)|^4 + r_n^8 n^{-8} \sum_{i=1}^n \mathbb{E}\Big|\sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)\Big|^4 \\ &\lesssim r_n^8 n^{-3} \mathbb{E}|\tilde{\ell}_n(\mathbf{z})|^4 + r_n^8 n^{-5} \mathbb{E}[|\tilde{w}_n(\mathbf{z}_1, \mathbf{z}_2)|^2 |\tilde{w}_n(\mathbf{z}_1, \mathbf{z}_3)|^2] + r_n^8 n^{-6} \mathbb{E}|\tilde{w}_n(\mathbf{z}_1, \mathbf{z}_2)|^4.\end{aligned}$$

□

Lemma 27. *Suppose Assumption PS (v) holds. Then,* (C.3) *holds.*

Proof.

$$\begin{aligned}\sigma_{in}^2 = 4r_n^4 n^{-2} \mathbb{E}|\tilde{\ell}_n(\mathbf{z}_i)|^2 + r_n^4 \binom{n}{2}^{-2} \mathbb{E}\Big[\Big|\sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)\Big|^2 \Big| \mathcal{F}_{i-1}\Big] \\ + 4r_n^4 n^{-1} \binom{n}{2}^{-1} \mathbb{E}\Big[\sum_{j=1}^{i-1} \tilde{\ell}_n(\mathbf{z}_i) \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j) \Big| \mathcal{F}_{i-1}\Big].\end{aligned}$$

Add and subtract $\mathbb{E}|r_n^2 2n^{-1}(n-1)^{-1} \sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)|^2$,

$$\begin{aligned}\sum_{i=1}^n \sigma_{in}^2 = s_n^2 + \sum_{i=1}^n \Bigg\{ r_n^4 \binom{n}{2}^{-2} \mathbb{E}\Big[\Big(\sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)\Big)^2 \Big| \mathcal{F}_{i-1}\Big] - \mathbb{E}\Big(r_n^2 \binom{n}{2}^{-1} \sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)\Big)^2 \\ + 4r_n^4 n^{-1} \binom{n}{2}^{-1} \mathbb{E}\Big[\sum_{j=1}^{i-1} \tilde{\ell}_n(\mathbf{z}_i) \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j) \Big| \mathcal{F}_{i-1}\Big] \Bigg\}.\end{aligned}$$

Then, since s_n^2 is bounded away from zero, to show (C.3), it suffices to show

$$r_n^8 n^{-8} \mathbb{E}\Bigg| \sum_{i=1}^n \Bigg\{ \mathbb{E}\Big[\Big(\sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)\Big)^2 \Big| \mathcal{F}_{i-1}\Big] - \mathbb{E}\Big(\sum_{j=1}^{i-1} \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j)\Big)^2 \Bigg\} \Bigg|^2 = o(1) \tag{C.4}$$

$$r_n^8 n^{-6} \mathbb{E}\Bigg| \sum_{i=1}^n \mathbb{E}\Big[\sum_{j=1}^{i-1} \tilde{\ell}_n(\mathbf{z}_i) \tilde{w}_n(\mathbf{z}_i, \mathbf{z}_j) \Big| \mathcal{F}_{i-1}\Big] \Bigg|^2 = o(1). \tag{C.5}$$

For (C.4),

$$
\begin{aligned}
&\mathbb{E}\Big|\sum_{i=1}^{n}\Big\{\mathbb{E}\Big[\Big(\sum_{j=1}^{i-1}\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)\Big)^2\Big|\mathcal{F}_{i-1}\Big]-\mathbb{E}\Big(\sum_{j=1}^{i-1}\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)\Big)^2\Big\}\Big|^2\\
&\lesssim \mathbb{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{i-1}\{\mathbb{E}[\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)^2|\mathcal{F}_{i-1}]-\mathbb{E}[\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)^2]\}\Big)^2\\
&\quad+\mathbb{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\mathbb{E}[\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_k)|\mathcal{F}_{i-1}]\Big)^2\\
&=\mathbb{E}\Big(\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}\{\mathbb{E}[\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)^2|\mathcal{F}_{i-1}]-\mathbb{E}[\tilde{w}_n(\mathbf{z}_1,\mathbf{z}_2)^2]\}\Big)^2\\
&\quad+\mathbb{E}\Big(\sum_{k=1}^{n-2}\sum_{j=k+1}^{n-1}\sum_{i=j+1}^{n}\mathbb{E}[\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_k)|\mathcal{F}_{i-1}]\Big)^2\\
&\le n^2\mathbb{E}\sum_{j=1}^{n-1}\sum_{k=1}^{n-1}\{\mathbb{E}[\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_j)^2|\mathcal{F}_{n-1}]-\mathbb{E}[\tilde{w}_n(\mathbf{z}_1,\mathbf{z}_2)^2]\}\{\mathbb{E}[\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_k)^2|\mathcal{F}_{n-1}]-\mathbb{E}[\tilde{w}_n(\mathbf{z}_1,\mathbf{z}_2)^2]\}\\
&\quad+n^2\mathbb{E}\sum_{k_1=1}^{n-2}\sum_{j_1=k_1+1}^{n-1}\sum_{k_2=1}^{n-2}\sum_{j_2=k_2+1}^{n-1}\mathbb{E}[\mathbb{E}[\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_{j_1})\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_{k_1})|\mathcal{F}_{n-1}]\mathbb{E}[\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_{j_2})\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_{k_2})|\mathcal{F}_{n-1}]]\\
&\le n^3\mathbb{E}(\mathbb{E}[\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_1)^2|\mathcal{F}_{n-1}]-\mathbb{E}[\tilde{w}_n(\mathbf{z}_1,\mathbf{z}_2)^2])^2+n^4\mathbb{E}(\mathbb{E}[\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_1)\tilde{w}_n(\mathbf{z}_n,\mathbf{z}_2)|\mathcal{F}_{n-1}])^2.
\end{aligned}
$$

For (C.5),

$$
\begin{aligned}
&\mathbb{E}\Big|\sum_{i=1}^{n}\mathbb{E}\Big[\sum_{j=1}^{i-1}\tilde{\ell}_n(\mathbf{z}_i)\tilde{w}_n(\mathbf{z}_i,\mathbf{z}_j)\Big|\mathcal{F}_{i-1}\Big]\Big|^2\\
&=\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j=1}^{\min\{i_1,i_2\}-1}\mathbb{E}[\mathbb{E}[\tilde{\ell}_n(\mathbf{z}_{i_1})\tilde{w}_n(\mathbf{z}_{i_1},\mathbf{z}_j)|\mathcal{F}_j]\mathbb{E}[\tilde{\ell}_n(\mathbf{z}_{i_2})\tilde{w}_n(\mathbf{z}_{i_2},\mathbf{z}_j)|\mathcal{F}_j]]\\
&\le n^3\mathbb{E}[\mathbb{E}[\tilde{\ell}_n(\mathbf{z}_2)\tilde{w}_n(\mathbf{z}_2,\mathbf{z}_1)|\mathcal{F}_1]^2].
\end{aligned}
$$

□

Stochastic Equicontinuity

Lemma 28. *If Assumption PS (iv)-(v) and one of Condition γ hold, then for every $\varepsilon>0$*

and $C > 0$, we can find $\delta > 0$ such that

$$\limsup_{n\to\infty} \Pr\Big(\sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\delta \\ \|\mathbf{s}\|\vee\|\mathbf{t}\|\le C}} |\hat{G}_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - \hat{G}_n(\boldsymbol{\theta}_0 + \mathbf{t}r_n^{-1})| > \varepsilon \Big) < \varepsilon.$$

Proof. It suffices to establish stochastic equicontinuity of $r_n^2\{L_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - L_n(\boldsymbol{\theta}_0)\}$ and $r_n^2\{W_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - W_n(\boldsymbol{\theta}_0)\}$, individually. By Theorem 4.2 in Pollard (1989), for some $\eta > 0$,

$$r_n^2\mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\delta \\ \|\mathbf{s}\|\vee\|\mathbf{t}\|\le C}} |L_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - L_n(\boldsymbol{\theta}_0 + \mathbf{t}r_n^{-1})| \lesssim r_n^2 n^{-1/2} \Big(\mathbb{E} \sup_{\substack{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\le\delta \\ \boldsymbol{\theta},\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\eta}} |\ell_n(\mathbf{z},\boldsymbol{\theta}) - \ell_n(\mathbf{z},\boldsymbol{\theta}')|^2 \Big)^{1/2}$$

where using $|\ell_n(\mathbf{z},\boldsymbol{\theta}) - \ell_n(\mathbf{z},\boldsymbol{\theta}')| \le B_1(\mathbf{z})\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$, the last term can be made small by taking δ small.

Now it remains to establish stochastic equicontinuity for $r_n^2\{W_n(\boldsymbol{\theta}_0 + \cdot\, r_n^{-1}) - W_n(\boldsymbol{\theta}_0)\}$. For the smooth case i.e., $\gamma = 2$, we have

$$r_n^2\mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\|\le\delta \\ \|\mathbf{s}\|\vee\|\mathbf{t}\|\le C}} |W_n(\boldsymbol{\theta}_0+\mathbf{s}r_n^{-1}) - W_n(\boldsymbol{\theta}_0+\mathbf{t}r_n^{-1})| \lesssim r_n^2 n^{-1} \Big(\mathbb{E} \sup_{\substack{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\le\delta \\ \boldsymbol{\theta},\boldsymbol{\theta}\in\boldsymbol{\Theta}_0^\eta}} |w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta}) - w_n(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\theta}')|^2 \Big)^{1/2}$$

and by Condition: Smooth Case, the last term is of order $r_n^2 n^{-1} q_n^{-1/2} r_n^{-1}\delta = \delta\sqrt{r_n^{4-\gamma}/n^2 q_n}$,

which can be made arbitrarily small by taking δ small. For the non-smooth case ($\gamma = 1$),

$$
\begin{aligned}
& r_n^2 \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \le C}} |W_n(\boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - W_n(\boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})| \\
& \lesssim r_n^2 n^{-1} \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \le C}} \left| \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - \omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})|^2 \right|^{1/2} \\
& \quad + r_n^2 n^{-1} \mathbb{E} \Bigg[\left| \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\bar{\omega}_n^{C r_n^{-1}}(\mathbf{z}_i, \mathbf{z}_j)|^2 \right|^{1/2} \\
& \qquad \times J \Bigg(\sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \le C}} \left| \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - \omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})|^2 \right|^{1/2} \\
& \qquad\qquad \Big/ \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\bar{\omega}_n^{C r_n^{-1}}(\mathbf{z}_i, \mathbf{z}_j)|^2 \Bigg) \Bigg] \\
& \lesssim r_n^2 n^{-1} \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \le C}} \left| \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - \omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})|^2 \right|^{1/2} \\
& \quad + \eta^{1/2} + r_n^2 n^{-1} \Big(\mathbb{E} \Big| \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\bar{\omega}_n^{C r_n^{-1}}(\mathbf{z}_i, \mathbf{z}_j)|^2 \Big)^{1/2} \\
& \qquad \times \left[\mathbb{E} J^2 \left(\eta^{-1} r_n^4 n^{-2} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \le C}} \left| \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\omega_n(\mathbf{z}_i,\ \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - \omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})|^2 \right) \right]^{1/2}
\end{aligned}
$$

where

$$
\begin{aligned}
& \omega_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}) = w_n(\mathbf{z}_{2i}, \mathbf{z}_{2j}, \boldsymbol{\theta}) + w_n(\mathbf{z}_{2i}, \mathbf{z}_{2j-1}, \boldsymbol{\theta}) + w_n(\mathbf{z}_{2i-1}, \mathbf{z}_{2j}, \boldsymbol{\theta}) + w_n(\mathbf{z}_{2i-1}, \mathbf{z}_{2j-1}, \boldsymbol{\theta}) \\
& \bar{\omega}_n^{\delta}(\mathbf{z}_1, \mathbf{z}_2) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^{\delta}} |\omega_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}) - \omega_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0)|
\end{aligned}
$$

and we split the expectation calculation into the event $\{\binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |\bar{\omega}_n^{C r_n^{-1}}(\mathbf{z}_i, \mathbf{z}_j)|^2 \le \eta r_n^{-4} n\}$ and its complement. Then, using $(a+b)^2 \le 2(a^2 + b^2)$, it suffices to show

$$
r_n^4 n^{-2} \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \le \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \le C}} \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \left| w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1}) \right|^2 = o(1).
$$

Splitting into the event $\{q_n 2\bar{w}_n(\mathbf{z}_i, \mathbf{z}_j) \leq L\}$ and its complement,

$$
\begin{aligned}
& r_n^4 n^{-2} \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \leq \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \leq C}} \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \left| w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1}) \right|^2 \\
& \leq r_n^4 n^{-2} q_n^{-1} \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \leq \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \leq C}} \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} |w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})| L \\
& \quad + r_n^{4-\gamma} n^{-2} q_n^{-1} 2 r_n^{\gamma} q_n \mathbb{E} |\bar{w}_n^{C r_n^{-1}}(\mathbf{z}_1, \mathbf{z}_2)|^2 I\{q_n 2 \bar{w}_n^{C r_n^{-1}}(\mathbf{z}_1, \mathbf{z}_2) > L\}
\end{aligned}
$$

and the second term can be made arbitrarily small by choosing L large enough. For the first term,

$$
\begin{aligned}
& \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \leq \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \leq C}} \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} |w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})| \\
& \leq \mathbb{E} \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \leq \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \leq C}} \left| \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} |w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})| \right. \\
& \qquad \left. - \mathbb{E}|w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})| \right| \\
& \quad + \sup_{\substack{\|\mathbf{s}-\mathbf{t}\| \leq \delta \\ \|\mathbf{s}\| \vee \|\mathbf{t}\| \leq C}} \mathbb{E}|w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - w_n(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\theta}_0 + \mathbf{t} r_n^{-1})|.
\end{aligned}
$$

The first term can be bounded by Theorem 6 of Nolan and Pollard (1987) and the second term can be bounded by the bound in Assumption PS (iv). □

Now, for the weak convergence of the process $\{r_n^2[\hat{M}_n(\boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - \hat{M}_n(\boldsymbol{\theta}_0)] : \mathbf{s} \in \mathbb{R}^k\}$, it remains to show convergence of the expectation term.

Lemma 29. *Suppose Assumption PS (ii) holds. Then, for any finite $L > 0$,*

$$
\sup_{\|\mathbf{s}\| \leq L} \left| r_n^2[M_n(\boldsymbol{\theta}_0 + \mathbf{s} r_n^{-1}) - M_n(\boldsymbol{\theta}_0)] + \mathbf{s}' \mathbf{V}_0 \mathbf{s} \right| = o(1).
$$

Proof. By Taylor expansion,

$$
\begin{aligned}
&2\left|r_n^2[M_n(\boldsymbol{\theta}_0 + \mathbf{s}r_n^{-1}) - M_n(\boldsymbol{\theta}_0)] + \frac{1}{2}\mathbf{s}'\mathbf{V}_0\mathbf{s}\right| \\
&\leq 2\|\mathbf{s}\|r_n\left\|\frac{\partial M_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right\| + \|\mathbf{s}\|^2\left\|\frac{\partial^2 M_n(\boldsymbol{\theta}_0 + \tau\mathbf{s}r_n^{-1})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} - \frac{\partial^2 M_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right\| + \|\mathbf{s}\|^2\|\mathbf{V}_n - \mathbf{V}_0\|
\end{aligned}
$$

where $\tau \in (0,1)$ and $\mathbf{V}_n = \partial^2 M_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. The first and third terms is $o(1)$ by the hypothesis. For the middle term, add and subtract $\partial^2 M_0(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ evaluated at $\boldsymbol{\theta}_0 + \tau s r_n^{-1}$ and $\boldsymbol{\theta}_0$. Using uniform convergence and continuity of $\partial^2 M_0(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, the desired result follows. □

BIBLIOGRAPHY

Aaronson, Daniel, (1998). "Using sibling data to estimate the impact of neighborhoods on children's educational outcomes," *Journal of Human Resources*, *33*(4), pp. 915–946.

Abdulkadiroğlu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters, (2017). "Do parents value school effectiveness?", Working Paper.

Abdulkadiroğlu, Atila, Parag A. Pathak, and Christopher R. Walters, (2018). "Free to choose: Can school choice reduce student achievement?" *American Economic Journal: Applied Economics*, *10*(1), pp. 175–206.

Abowd, John M., Kevin L. McKinney, and Ian M. Schmutte, (2018). "Modeling endogenous mobility in earnings determination," Forthcoming in Journal of Business and Economic Statistics.

Abrevaya, Jason and Jian Huang, (2005). "On the bootstrap of the maximum score estimator," *Econometrica*, *73*(4), pp. 1175–1204.

Altonji, Joseph G., Todd E. Elder, and Chistopher R. Taber, (2005). "Selection on observed and unobserved variables: Assesing the effectiveness of catholic schools," *Journal of Political Economy*, *113*(1), pp. 151–184.

Altonji, Joseph G. and Richard K. Mansfield, (2018). "Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects," *American Economic Review*, *108*(10), pp. 2902–2946.

Andrews, Donald W. K., (1994). "Empirical process methods in econometrics," In James J. Heckman and Edward E. Leamer (eds.) *Handbook of Econometrics*, *4*: Elsevier, pp. 2247 – 2294.

Andrews, Donald W. K., (2017). "Examples of l^2-complete and boundedly-complete distributions," *Journal of Econometrics*, *199*, pp. 213–220.

Andrews, Donald W. K. and Gustavo Soares, (2010). "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, *78*(1), pp. 119–157.

Angrist, Joshua D. and Kevin Lang, (2004). "Does school integration generate peer effects? evidence from boston's metco program," *American Economic Review*, *94*(5), pp. 1613–1634.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters, (2013). "Explaining charter school effectiveness," *American Economic Journal: Applied Economics*, *5*(4), pp. 1–27.

Arellano, Manuel, Richard Blundell, and Stéphane Bonhomme, (2017). "Earnings and consumption dynamics: A nonlinear panel data framework," *Econometrica*, *85*(3), pp. 693–734.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato, (2015). "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, *186*, pp. 345–366.

Berry, Steven T. and Philip A. Haile, (2014). "Identification in differentiated products markets using market level data," *Econometrica*, *82*(5), pp. 1749–1797.

Bickel, Peter J. and Bo Li, (2006). "Regularization in statistics," *Test*, *15*(2), pp. 271–344.

Blume, Lawrence E., William A. Brock, Steven N. Durlauf, and Rajshri Jayaraman, (2015). "Linear social interaction models," *Journal of Political Economy*, *123*(2).

Blundell, Richard W. and James L. Powell, (2004). "Endogeneity in semiparametric binary response models," *Reviews of Economic Studies*, *71*, pp. 655–679.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa, (2018). "A distributional framework for matched employer employee data," Working paper.

Briesch, Richard, Pradeep K. Chintagunta, and Rosa L. Matzkin, (2010). "Nonparametric discrete choice models with unobserved heterogeneity," *Journal of Business and Economic Statistics*, *28*(2), pp. 291–307.

Card, David and Jesse Rothstein, (2007). "Racial segregation and the black-white test score gap," *Journal of Public Economics*, *91*, pp. 2158–2184.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson, (2013). "Generalized jackknife estimators of weighted average derivatives," *Journal of the American Statistical Association*, *108*, pp. 1243–2268.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson, (2014). "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory*, *30*, pp. 176–200.

Cattaneo, Matias D., Max H. Farrell, and Yingjie Feng, (2018). "Large sample properties of partitioning-based series estimators," Working Paper.

Cattaneo, Matias D. and Michael Jansson, (2018). "Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency," *Econometrica*, *86*(3), pp. 955–995.

Chen, Xiaohong, (2007). "Large sample sieve estimation of semi-nonparametric models," In James J. Heckman and Edward E. Leamer (eds.) *Handbook of Econometrics*, *6*: Elsevier, pp. 5549 – 5632.

Chernoff, Hermann, (1964). "Estimation of the mode," *Annals of the Institute of Statistical Mathematics*, *16*(1), pp. 31–41.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney Newey, and James M. Robins, (2018). "Locally robust semiparametric estimation," working paper.

Chernozhukov, Victor and Christian Hansen, (2005). "An iv model of quantile treatment effects," *Econometrica*, *73*(1), pp. 245–261.

Chesher, Andrew, (2003). "Identification in nonseparable models," *Econometrica*, *71*(5), pp. 1405–1441.

Chetty, Raj and Nathaniel Hendren, (2018). "The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects," *Quarterly Journal of Economics*, *133*(3), pp. 1107–1162.

Chetty, Raj, Nathaniel Hendren, and Lawrence Katz, (2016). "The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment," *American Economic Review*, *106*(4), pp. 855–902.

Chetverikov, Denis, Bradley Larsen, and Christopher Palmer, (2016). "Iv quantile regression for group-level treatments, with an application to the distributional effects of trade," *Econometrica*, *84*(2), pp. 809–833.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach, (2010). "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, *78*(3), pp. 883–931.

Dahl, Gordon B., (2002). "Mobility and the return to education: Testing a roy model with multiple markets," *Econometrica*, *70*(6), pp. 2367–2420.

Dale, Stacy Berg and Alan B. Krueger, (2002). "Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables," *Quarterly Journal of Economics*, *117*(4), pp. 1491–1527.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault, (2011). "Nonparametric instrumental regression," *Econometrica*, *79*(5), pp. 1541–1565.

Das, Mitali, Whitney K. Newey, and Francis Vella, (2003). "Nonparametric estimation of sample selection models," *Review of Economic Studies*, *70*(1), pp. 33–58.

Delgado, Miguel A, Juan M Rodriguez-Poo, and Michael Wolf, (2001). "Subsampling inference in cube root asymptotics with an application to manski's maximum score estimator," *Economics Letters*, *73*(2), pp. 241–250.

D'Haultfoeuille, Xavier, (2011). "On the completeness condition in nonparametric instrumental problems," *Econometric Theory*, *27*, pp. 460–471.

Dobbie, Will and Roland G. Fryer, Jr., (2011). "Are high-quality schools enough to increase achievement among the poor? evidence from the harlem children's zone," *American Economic Journal:Applied Economics*, *3*, pp. 158–187.

Dubin, Jeffrey A. and Daniel L. McFadden, (1984). "An econometric analysis of residential electric appliance holdings and consumption," *Econometrica*, *52*(2), pp. 345–362.

Dümbgen, Lutz, (1993). "On nondifferentiable functions and the bootstrap," *Probability Theory and Related Fields*, *95*, pp. 125–140.

Durlauf, Steven N., (2004). "Neighborhood effects," In J. Vernon Henderson and Jacques-Fançois Thisse (eds.) *Handbook of Regional and Urban Economics*, *4*: Elsevier, pp. 2173–2242.

Durlauf, Steven N. and Yannis M. Ioannides, (2010). "Social interactions," *Annual Review of Economics*, *2*, pp. 451–478.

Fang, Zheng and Andres Santos, (2018). "Inference on directionally differentiable functions," Forthcoming in Review of Economic Studies.

Freyberger, Joachim, (2018). "Nonparametric panel data models with interactive fixed effects," *Review of Economic Studies*, *85*(3), pp. 1824–1851.

Gould, Eric D., Victor Lavy, and M. Daniele Paserman, (2004). "Immigrating to opportunity: Estimating the effect of school quality using a natural experiment on ethiopians in israel," *Quarterly Journal of Economics*, *119*(2), pp. 489–526.

Graham, Bryan S., (2017). "An econometric model of network formation with degree heterogeneity," *Econometrica*, *85*(4), pp. 1033–1063.

Graham, Bryan S., (2018). "Identifying and estimating neighborhood effects," *Journal of Economic Literature*, *56*(2), pp. 450–500.

Grenander, Ulf, (1956). "On the theory of mortality measurement: Part ii," *Scandinavian Actuarial Journal*, *39*(2), pp. 125–153.

Groeneboom, Piet, (1985a). "Estimating a monotone density," In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*: Institute of Mathematical Statistics, pp. 539–555.

Groeneboom, Piet, (1985b). "Estimating a monotone density," *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer (L. M. Le Cam and R. A. Olshen, eds.)*(2), pp. 539–555.

Hall, Peter and J L. Horowitz, (2005). "Nonparametric methods for inference in the presence of instrumental variables," *Annals of Statistics*, *33*(6), pp. 1–27.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin, (2009). "New evidence about brown v. board of education: The complex effects of school racial composition on achievement," *Journal of Labor Economics*, *27*(3), pp. 349–383.

Heckman, James J., (1974). "Shadow prices, market wages, and labor supply," *Econometrica*, *42*(4), pp. 679–693.

Heckman, James J., (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables," *Annals of Economic and Social Measurement*, *5*, pp. 475–492.

Heckman, James J., (1979). "Selection bias as a specification error," *Econometrica*, *47*(1), pp. 153–161.

Hong, Han and Jessie Li, (2019). "The numerical bootstrap," Forthcoming in Annals of Statistics.

Honoré, Bo E. and Ekaterini Kyriazidou, (2000). "Panel data discrete choice models with lagged dependent variables," *Econometrica*, *68*(4), pp. 839–874.

Honoré, Bo E and James L Powell, (2005). "Pairwise difference estimation of nonlinear models," In D W K Andrews and J H Stock (eds.) *Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg*: Cambridge University Press, pp. 520–553.

Horowitz, Joel L, (1992). "A smoothed maximum score estimator for the binary response model," *Econometrica*, *60*(3), pp. 505–531.

Horowitz, Joel L, (2002). "Bootstrap critical values for tests based on the smoothed maximum score estimator," *Journal of Econometrics*, *111*(2), pp. 141–167.

Hoxby, Caroline M., (2000). "The effects of class size on student achievement: New evidence from population variation," *Quarterly Journal of Economics*, *115*(4), pp. 1239–1285.

Hu, Yingyao, (2017). "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics," *Journal of Econometrics*, *200*(2), pp. 154–168.

Hu, Yingyao and Susanne M. Schennach, (2008). "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, *76*(1), pp. 195–216.

Hu, Yingyao, Susanne M. Schennach, and Ji-Liang Shiu, (2017). "Injectivity of a class of integral operators with compactly supported kernels," *Journal of Econometrics*, *200*, pp. 48–58.

Hu, Yingyao and Ji-Liang Shiu, (2018). "Nonparametric identification using instrumental variables: Sufficient conditions for completeness," *Econometric Theory*, *34*(3), pp. 659–693.

Hyde, C. C. and B. M. Brown, (1970). "On the departure from normality of a certain class of martingales," *Annals of Mathematical Statistics*, *41*(6), pp. 2161–2165.

Ichimura, Hidehiko and Petra E. Todd, (2007). "Chapter 74 implementing nonparametric and semiparametric estimators," In James J. Heckman and Edward E. Leamer (eds.) *Handbook of Econometrics*, *6* of Handbook of Econometrics: Elsevier, pp. 5369 – 5468, URL: `http://www.sciencedirect.com/science/`

article/pii/S1573441207060746, DOI: http://dx.doi.org/https://doi.org/10.1016/S1573-4412(07)06074-6.

Imbens, Guido W. and Whitney K. Newey, (2009). "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, *77*(5), pp. 1481–1512.

Kallenberg, Olav, (2017). *Random Measures, Theory and Applications*, Cham, Switzerland: Springer.

Kim, Jeankyung and David Pollard, (1990). "Cube root asymptotics," *Annals of Statistics*, *18*(1), pp. 191–219.

Kling, Jeffry R., Jeffry B. Liebman, and Lawrence F. Katz, (2007). "Experimental analysis of neighborhood effects," *Econometrica*, *75*(1), pp. 83–119.

Kosorok, Michael R, (2008). "Bootstrapping the grenander estimator," In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*: Institute of Mathematical Statistics, pp. 282–292.

Kress, Rainer, (2014). *Linear Integral Equations*, New York, NY: Springer.

Lee, Lung-Fei, (1983). "Generalized econometric models with selectivity," *Econometrica*, *51*(2), pp. 507–512.

Léger, Christian and Brenda MacGibbon, (2006). "On the bootstrap in cube root asymptotics," *Canadian Journal of Statistics*, *34*(1), pp. 29–44.

Ludwig, Jens, Greg J. Duncan, and Paul Hirschfield, (2001). "Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment," *Quarterly Journal of Economics*, *116*(2), pp. 655–679.

Manski, Charles F, (1975). "Maximum score estimation of the stochastic utility model of choice," *Journal of econometrics*, *3*(3), pp. 205–228.

Manski, Charles F, (1985). "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of econometrics*, *27*(3), pp. 313–333.

Mattner, L, (1993). "Some incomplete but boundedly complete location families," *Annals of Statistics*, *21*(4), pp. 2158–2162.

Matzkin, Rosa L., (2016). "On independence conditions in nonseparable models: Observable and unobservable instruments," *Journal of Econometrics*, *191*, pp. 302–311.

Newey, Whitney K. and Daniel McFadden, (1994). "Large sample estimation and hypothesis testing," In James J. Heckman and Edward E. Leamer (eds.) *Handbook of Econometrics*, *4*: Elsevier, pp. 2111 – 2245.

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

Newey, Whitney K., James L. Powell, and Francis Vella, (1999). "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, *67*(3), pp. 565–603.

Nolan, Deborah and David Pollard, (1987). "U-processes: Rates of convergence," *Annals of Statistics*, *15*(2), pp. 780–799.

Oreopoulos, Philip, (2003). "The long-run consequences of living in a poor neighborhood," *Quarterly Journal of Economics*, *118*(4), pp. 1533–1575.

Patra, Rohit Kumar, Emilio Seijo, and Bodhisattva Sen, (2015). "A consistent bootstrap procedure for the maximum score estimator," arXiv:1105.1976.

de Paula, Auero, (2017). "Econometrics of network models," In Bo Honore, Ariel Pakes, M Piazzesi, and L Samuelson (eds.) *Advances in Economics and Econometrics:Theory and Applications, Eleventh World Congress*: Cambridge University Press, pp. 268–323.

Politis, Dimitri N. and Joseph P. Romano, (1994). "Large sample confidence regions based on subsamples under minimal assumptions," *Annals of Statistics*, *22*(4), pp. 2031–2050.

Pollard, David, (1989). "Asymptotics via empirical processes," *Statistical Science*, *4*(4), pp. 341–366.

Prakasa Rao, BLS, (1969). "Estimation of a unimodal density," *Sankhya: The Indian Journal of Statistics, Series A*, *31*(1), pp. 23–36.

Sasaki, Yuya, (2015). "Heterogeneity and selection in dynamic panel data," *Journal of Econometrics*, *188*, pp. 236–249.

Schmitt, Bernhard A., (1992). "Pertubation bounds for matrix square roots and pythagorean sums," *Linear Algebra and Its Applications*, *174*, pp. 215–227.

Sen, Bodhisattva, Moulinath Banerjee, and Michael Woodroofe, (2010). "Inconsistency of bootstrap: The grenander estimator," *Annals of Statistics*, *38*(4), pp. 1953–1977.

Seo, Myung Hwan and Taisuke Otsu, (2018). "Local m-estimation with discontinuous criterion for dependent and limited observations," *Annals of Statistics*, *46*(1), pp. 344–369.

Sloan, Frank A., Gabriel A. Picone, Donald H. Taylor, Jr., and Sinh-Yi Chou, (2001). "Hospital ownership and cost and quality of care: Is there a dime's worth of difference?" *Journal of Health Economics*, *20*, pp. 1–21.

Tropp, Joel A., (2012). "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, *12*, pp. 389–434.

van der Vaart, Aad W, (1998). *Asymptotic Statistics*: Cambridge University Press.

van der Vaart, Aad W and Jon A Wellner, (1996). *Weak Convergence and Empirical Processes*, New York, NY: Springer.

Wand, M.P. and M.C. Jones, (1995). *Kernel Smoothing*: Chapman & Hall/CRC.